



UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

Diagnóstico automático de tuberculosis: una decisión ante incertidumbre

Autor:

Ricardo Santiago Mozos

Directores:

Dr. Antonio Artés Rodríguez

Dr. Fernando Pérez Cruz

DEPARTAMENTO DE TEORÍA DE LA
SEÑAL Y COMUNICACIONES

Leganés, Septiembre de 2009

TESIS DOCTORAL

DIAGNÓSTICO AUTOMÁTICO DE TUBERCULOSIS: UNA DECISIÓN ANTE INCERTIDUMBRE

Autor: RICARDO SANTIAGO MOZOS

Directores: DR. ANTONIO ARTÉS RODRÍGUEZ

DR. FERNANDO PÉREZ CRUZ

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Vocal:

Vocal:

Secretario:

Calificación:

Leganés, de Septiembre de 2009.

Resumen

En un contraste de hipótesis no siempre es posible definir las hipótesis con precisión, o bien, no es posible relacionarlas directamente con los datos disponibles. Esta tesis considera el desarrollo de herramientas estadísticas para el contraste de hipótesis definidas con incertidumbre y su aplicación al diagnóstico automático de tuberculosis. Se proponen métodos para medir las prestaciones alcanzables a partir de los datos de entrenamiento y también se presentan test que consideran la incertidumbre y moderan las probabilidades con las que se decide. Por otro lado, se considera el equivalente en aprendizaje estadístico de los cocientes de verosimilitud en aquellas situaciones en las que no se conoce la distribución de los datos y una sola muestra de test no es suficiente para tomar una decisión con las prestaciones requeridas.

La tesis arranca con una introducción a la tuberculosis, por qué es un problema de salud y cuáles son las técnicas para su diagnóstico, centrándose en las propuestas automáticas basadas en el análisis de imágenes del esputo y sus principales inconvenientes. A continuación realiza una breve revisión de la teoría estadística de la decisión, que incluye las metodologías paramétrica y no paramétrica del contraste de hipótesis y los contrastes secuenciales; la determinación de regiones de confianza y la máquina de vectores soporte.

Luego, introduce el contraste de hipótesis inciertas y propone métodos para dicho contraste desde el punto de vista frecuentista y bayesiano. Asimismo, formula cotas superiores de las probabilidades de error desde el punto de vista frecuentista y cotas superiores de la probabilidad *a posteriori* de cada hipótesis desde el punto de vista bayesiano.

A continuación, considera el problema de clasificar un conjunto de muestras de la misma clase desde el punto de vista del aprendizaje estadístico. Propone un nuevo método que “extiende” los datos de entrenamiento de forma que el clasificador entrenado mediante dichos datos “extendidos” proporciona una salida única al conjunto de muestras. La bondad del método se comprueba desde un punto de vista empírico mediante varias bases de datos públicas y su complejidad es examinada cuando se emplea la máquina de vectores soporte como clasificador.

Finalmente, propone un sistema automático para el diagnóstico de pacientes de tuberculosis capaz de procesar imágenes al ritmo que se capturan del microscopio. Este sistema examina imágenes de esputo vistas al microscopio en busca del bacilo de Koch. Sin embargo, no es sencillo definir qué es un paciente sano porque es muy difícil construir un clasificador cuya probabilidad de declarar un bacilo erróneamente sea cero. Es aquí dónde los métodos descritos arriba proporcionan una decisión acerca del estado del paciente que considera la incertidumbre en la definición de paciente

sano y obtienen cotas de las prestaciones alcanzables a partir de los ejemplos de ambos tipos de pacientes.

Abstract

In hypothesis testing, it is not always possible to define the hypotheses precisely, sometimes they are not directly related with the available data. This thesis considers new statistical tools for testing uncertain hypotheses and their application to automatic tuberculosis diagnosis. Methods to measure the achievable performance using the training data are developed and test which consider the uncertainty in the hypotheses and modify the decision probabilities accordingly are proposed. Another addressed problem is the machine learning equivalent to the likelihood ratio for those situations where the data distributions are unknown and one test sample does not provide the desired performance.

The thesis starts with an introduction to tuberculosis, why is a health problem and which are the diagnosis techniques. We focus on automatic diagnosis based on sputum images analysis and their principal issues. Later, it shortly reviews decision theory, which includes parametric and non-parametric hypothesis testing methodologies and sequential testing; confidence region estimation and support vector machines.

Uncertain hypotheses testing follows and methods from frequentist and Bayesian points of view are proposed. Upper bounds of the error probabilities for frequentist view and upper bounds for the *a posteriori* hypotheses probability are presented.

The problem of classifying a set of samples of the same class is considered from a machine learning point of view. A new method to extend the training samples in such a way than a classifier trained with these “extended” training samples gives a single output for the test set of samples. This algorithm is evaluated and proved worthy in some public datasets and its complexity is analysed for the support vector machine classifier.

Finally, an automatic diagnosis system for tuberculosis patients is proposed. This system is capable to process images at the same rate as the microscope captures them. The system looks for Koch bacilli in the sputum. However, it is not clear how to define a healthy patient as it is difficult to build a classifier with zero false bacillus detection probability. The methods described above give a decision for the patient that correctly considers the uncertainty in the healthy patient definition. In addition, those methods bound the achievable performance from the available training data.

Índice general

Abreviaciones	XI
Notación	XIII
1. Motivación	1
1.1. La tuberculosis	1
1.1.1. Diagnóstico de la enfermedad	2
1.1.1.1. Baciloscopia de esputo utilizando auraminas	2
1.1.1.2. Pacientes considerados como no contagiosos o no bacilíferos	3
1.1.1.3. Efectos de un diagnóstico erróneo	3
1.1.2. Propuestas de diagnóstico automático	4
1.1.3. Nuestra propuesta	5
2. Introducción	7
2.1. Teoría estadística de la decisión	7
2.1.1. Coste bayesiano esperado	8
2.1.2. Riesgo frecuentista	8
2.1.2.1. Riesgo de Bayes	9
2.1.2.2. Minimax	10
2.1.3. Estadístico suficiente	11
2.1.4. Clases de reglas de decisión	11
2.2. Contraste de hipótesis	12
2.2.1. Enfoque frecuentista	12
2.2.1.1. Contraste de hipótesis simples	14
2.2.1.2. Contraste de hipótesis compuestas	16
2.2.2. Enfoque bayesiano	19
2.2.2.1. Elección de la distribución <i>a priori</i>	20
2.2.3. Reglas no paramétricas	21
2.2.3.1. Contraste de Wald	22

2.2.3.2.	Contrastes basados en las distribuciones binomial y χ^2	22
2.2.3.3.	Contrastes para escala ordinal	24
2.2.3.4.	Contrastes de permutación	26
2.2.3.5.	Contrastes basados en la función de distribución . . .	26
2.2.3.6.	Otras reglas no paramétricas	27
2.2.4.	Test secuenciales	27
2.2.4.1.	Hipótesis compuestas	32
2.2.4.2.	Test secuenciales truncados	34
2.2.5.	Resumen	36
2.3.	Intervalos de confianza	37
2.3.1.	Enfoque bayesiano	39
2.3.2.	Proporción de una distribución binomial	40
2.3.3.	Regiones de confianza para la proporción de una distribución multinomial	43
2.4.	Máquina de vectores soporte	44
2.4.1.	Formulación	44
2.4.2.	Reducción de complejidad	46
2.4.2.1.	Preimágenes para <i>kernel</i> polinómico	50
2.4.2.2.	Clasificadores en cascada	51
2.5.	Objetivos	52
3.	Incertidumbre en las hipótesis	53
3.1.	Variables aleatorias discretas binarias y no binarias	55
3.1.1.	Enfoque de máxima verosimilitud	55
3.1.2.	Enfoque bayesiano	56
3.1.2.1.	Distribuciones no binarias.	57
3.1.2.2.	Experimentos	58
3.1.2.3.	Desarrollo secuencial	62
3.1.2.4.	Prestaciones asintóticas	66
3.1.3.	Enfoque frecuentista	67
3.1.3.1.	Prestaciones asintóticas	68
3.1.3.2.	Variables discretas no binarias	69
3.1.3.3.	Experimentos	69
3.2.	Variables aleatorias continuas	72
3.3.	Resumen	75
4.	Extensión del espacio de entrada	77
4.1.	Motivación y método	78

4.2.	Ilustración	85
4.3.	Experimentos	89
4.4.	Conclusiones	95
5.	Diagnóstico automático de la tuberculosis	97
5.1.	El problema	97
5.2.	Sistema propuesto	98
5.3.	Clasificador de parches	99
5.3.1.	Datos	99
5.3.2.	Entrenamiento	99
5.3.2.1.	División del conjunto de entrenamiento	100
5.4.	Clasificador de pacientes	102
5.4.1.	Test secuencial clásico para variables binarias con incertidumbre	102
5.4.1.1.	Estima del número de muestras necesarias	105
5.4.2.	Test secuencial bayesiano para variables binarias con incertidumbre	106
5.5.	Experimentos	107
5.5.1.	Extracción de características	107
5.5.2.	Clasificador de bacilos	108
5.5.2.1.	Implementación en tiempo real	112
5.5.3.	Clasificador de pacientes	114
5.5.3.1.	Aproximación frecuentista	116
5.5.3.2.	Aproximación bayesiana	121
5.6.	Resumen	128
6.	Conclusiones y Líneas Futuras	129
6.1.	Conclusiones	129
6.2.	Líneas futuras	130
	Bibliografía	133

Abreviaciones

- AUC** área bajo la curva ROC (*area under the ROC curve*)
- fdp** función de densidad de probabilidad
- GLRT** test de cociente de verosimilitud generalizado (*generalized likelihood ratio test*)
- GPC** proceso gaussiano para clasificación (*gaussian process for classification*)
- HPD** con densidad *a posteriori* más alta (*highest posterior density*)
- ICA** análisis de componentes independientes (*independent component analysis*)
- iid** independientes e idénticamente distribuidas
- KPCA** análisis de componentes principales basado en núcleos (*kernel principal component analysis*)
- LDA** análisis discriminante lineal (*linear discriminant analysis*)
- MMD** máxima discrepancia en media (*maximum mean discrepancy*)
- MMI** maximización de la información mutua (*maximization of mutual information*)
- OHDR** hiperplano óptimo de decisión (*optimal hyperplane decision rule*)
- OMS** Organización Mundial de la Salud
- PCA** análisis de componentes principales (*principal component analysis*)
- ROC** característica operativa del receptor (*receiver operating characteristic*)
- SPRT** test secuencial de cociente de verosimilitud (*sequential probability ratio test*)
- SVM** máquina de vectores soporte (*support vector machine*)

UMP uniformemente más potente

VIH virus de la inmunodeficiencia humana

Notación

En la memoria se ha empleado, en general, la siguiente notación:

- H_i Hipótesis i .
- $P(H_i)$ Probabilidad *a priori* de H_i .
- x minúsculas para los escalares.
- \mathbf{x} minúsculas negrita para los vectores.
- \boldsymbol{x} minúsculas negrita y cursiva para un conjunto de muestras: $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$.
- X mayúsculas para una variable aleatoria escalar.
- \mathbf{X} mayúsculas y negrita para una matriz o una variable aleatoria vectorial.

Capítulo 1

Motivación

La algoritmia desarrollada en esta tesis puede aplicarse al diagnóstico automático de pacientes de tuberculosis. La muestra a examinar para tomar dicha decisión es un conjunto de imágenes microscópicas de esputo, que típicamente es suficientemente grande. Empezamos describiendo la tuberculosis, por qué es un problema de salud en nuestros días y cuáles son los métodos para su diagnóstico.

1.1. La tuberculosis

La tuberculosis es una enfermedad infecto-contagiosa causada por el *mycobacterium tuberculosis*, conocido como el bacilo de Koch. La tuberculosis es, posiblemente, la enfermedad infecciosa más prevalente en el mundo. En 1999 la Organización Mundial de la Salud (OMS) cifró el número de nuevos casos de tuberculosis en el mundo en 3.689.833 para un total de 8.500.000 casos totales con una tasa global de 141/100.000 habitantes. En el informe OMS de 2003, se estima en 8 millones (140/100.000) el número de nuevos casos de tuberculosis, de los cuales 3,9 millones (62/100.000) son bacilíferos, esto es, contagiosos o con tuberculosis activa y 674.000 (11/100.000) están coinfectados con el virus de la inmunodeficiencia humana (VIH). La tuberculosis mantiene una prevalencia de 245/100.000 habitantes, y una tasa de mortalidad de 28/100.000. La tendencia epidemiológica de la incidencia de tuberculosis sigue aumentando en el mundo, pero la tasa de mortalidad y prevalencia están disminuyendo (Organización Mundial de la Salud, 2003, revisada en marzo de 2006). En España hubo, en 1990, 21.644 casos de tuberculosis, 2.265 de los cuales fueron mortales; en el año 2007 hubo 13.103 casos de tuberculosis de los cuales 1.375 fueron mortales, de éstos últimos, 122 padecían también VIH (Bauquerez *et al.*, 2009).

La tuberculosis¹ se trasmite a través de partículas expelidas por un paciente

¹Robert Koch anunció el descubrimiento del bacilo de la tuberculosis el 24 de marzo de 1882.

bacilífero con la tos, estornudo, hablando, etc. Las gotas infecciosas son de un diámetro de entre 0,5 y 5 μm , pudiendo ser producidas alrededor de 400.000 con un sólo estornudo. Un paciente con tuberculosis activa sin tratamiento puede infectar entre 10-15 personas por año que en su mayoría tendrán contactos frecuentes, prolongados, o intensos con el paciente (como por ejemplo, el transporte público a horas punta). El riesgo de contagio aumenta en áreas donde la tuberculosis es frecuente o con pacientes inmunodeprimidos, como los pacientes de VIH con la enfermedad activa.

La enfermedad ataca preferentemente los pulmones, pero puede también infectar otros órganos como los riñones, el hígado, la piel, meninges, etc. Es más grave en niños y ancianos, que pueden llegar a morir de ella. Iniciando el tratamiento con los medicamentos normalizados, el enfermo deja de contagiar a partir de los quince o veinte días.

1.1.1. Diagnóstico de la enfermedad

Para el diagnóstico de la enfermedad se emplean los siguientes procedimientos:

- Radiografía de tórax: esencial en el diagnóstico de la enfermedad. Las lesiones típicas radiológicas son apicales, en segmentos posteriores y generalmente formando cavidades (Kumar y Cotran, 2003).
- Cultivo de la muestra biológica (Kumar y Cotran, 2003).
- Prueba de la Tuberculina o Test de Mantoux: test cutáneo (intradermorreacción) para detectar infección tuberculosa (Chaturvedi y Cockcroft, 1992).
- Test basados en la amplificación de ácidos nucleicos (Dinnes *et al.*, 2007).
- Baciloscopia de esputo: visión directa mediante microscopio del bacilo de tuberculosis en esputo, empleando técnicas de tinción para bacilos ácido-alcohol resistentes (Ziehl-Neelsen) o auramina (Steingart *et al.*, 2006).

Esta última técnica es la empleada para el diagnóstico en este trabajo por ello, la extenderemos brevemente en el siguiente apartado.

1.1.1.1. Baciloscopia de esputo utilizando auraminas

Las micobacterias de la tuberculosis tienen una pared celular de estructura lipídica con ácidos micólicos (>60%). Para facilitar su identificación en el microscopio se

En su memoria la OMS declara el 24 de marzo como el día mundial de la tuberculosis.

utiliza una tinción fluorescente de auramina, que incluye fenol en su composición, que se intercala en los ácidos de la pared celular de los bacilos (en los carbonos) además de en otros cuerpos presentes en el esputo. La tinción afecta tanto al color como al tamaño de los bacilos que se observan. En pacientes bajo tratamiento resultan mucho más difíciles de ver aunque los bacilos sigan estando ahí dado que las paredes de los bacilos son más finas.

El protocolo médico indica que es necesario revisar tres líneas de la muestra (recorrer el portaobjetos tres veces de izquierda a derecha). Los médicos suelen utilizar un aumento 25x, aunque es posible utilizar un zoom mayor, lo que representa 75–100 campos (o imágenes no solapadas de la muestra) en las tres líneas. La Figura 1.1 muestra un campo de un paciente enfermo. El diagnóstico del médico es positivo si encuentra 3 o más bacilos en las tres líneas. De otra forma, se considera que el paciente da negativo (no se encuentra ninguno) o se le considera no bacilífero (no contagioso, si se encuentran uno o dos bacilos solamente). En cualquier caso, se realiza un cultivo de la muestra biológica para asegurar el diagnóstico. El resultado del cultivo se asume como la situación real del paciente y a la vista de ello las prestaciones de la baciloscopia realizada por los médicos rondan el 59 % de sensibilidad, definida como la probabilidad de clasificar correctamente un paciente enfermo, y el 99 % de especificidad, definida como la probabilidad de clasificar correctamente un paciente sano.

1.1.1.2. Pacientes considerados como no contagiosos o no bacilíferos

La infección latente de tuberculosis significa que el germen de la tuberculosis se encuentra en el cuerpo (generalmente en los pulmones), pero sin que se hayan presentado aún síntomas evidentes. En el caso de la tuberculosis latente, el paciente presenta una reacción importante a la prueba cutánea de Mantoux, sin que haya síntomas de tuberculosis ni organismos de la tuberculosis en el esputo. Para contagiar los gérmenes de la tuberculosis, la persona debe tener la enfermedad activa. La tuberculosis puede permanecer como infección latente toda la vida.

1.1.1.3. Efectos de un diagnóstico erróneo

Dada la alarma social que genera un diagnóstico positivo de tuberculosis, es imprescindible lograr una especificidad lo más cercana posible al 100 %. Es decir, únicamente los casos positivos claros son declarados, y para el resto se espera a realizar un cultivo de la muestra biológica para asegurar la existencia o no de bacilos. En esos casos, se considera que el paciente no es contagioso por lo menos durante el tiempo

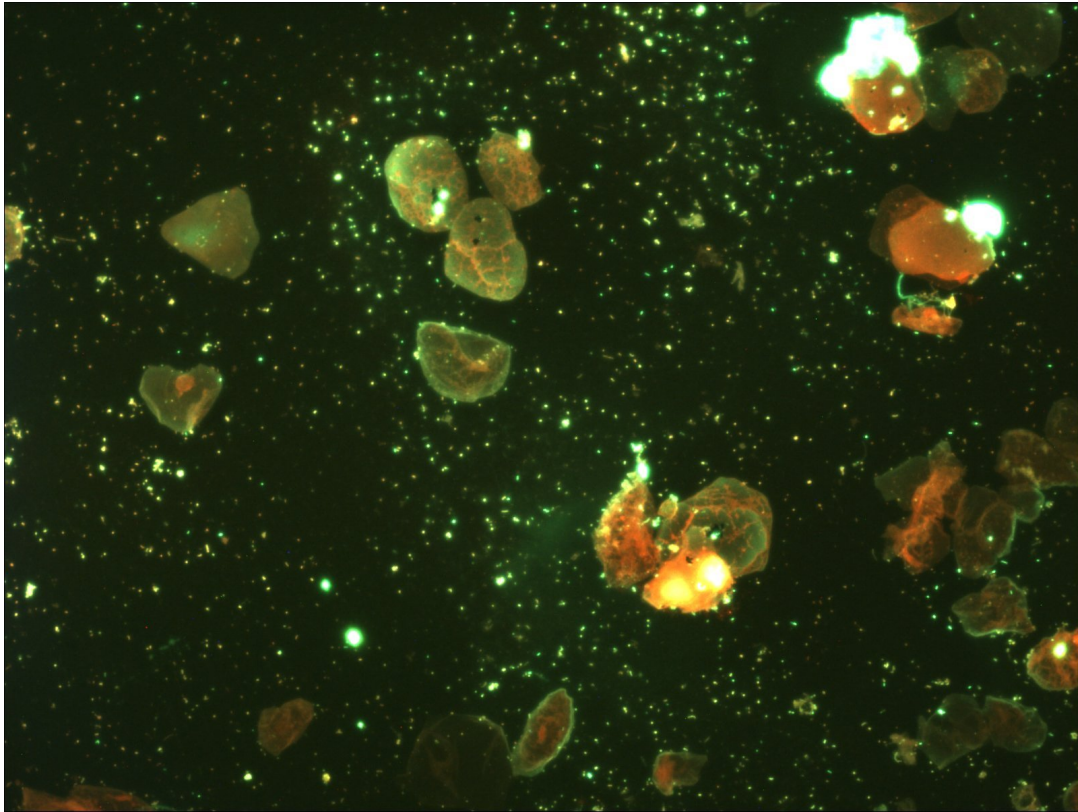


Figura 1.1: Imagen microscópica del esputo (campo) tintado con auramina.

que se tarda en realizar el cultivo, hasta cuatro semanas (Robledo *et al.*, 2006), ya que se considera no bacilífero.

1.1.2. Propuestas de diagnóstico automático

El problema de identificar automáticamente el bacilo de Koch en imágenes microscópicas del esputo ha sido atacado desde distintos puntos de vista en la literatura. Muchos de los métodos descansan en la segmentación de las imágenes, que consiste en identificar, centrar y preprocesar los posibles bacilos para luego proceder a su clasificación partiendo de distintos descriptores de los mismos (Veropoulos *et al.*, 1999; Forero *et al.*, 2006, 2003, 2004; Yu *et al.*, 1997). Los inconvenientes de estas propuestas son:

1. Descansan en la segmentación de las imágenes, que tiene una alta carga computacional de base. Sea cual fuere el método de clasificación que quiera aplicarse después, el tiempo de computación del algoritmo de detección queda lastrado por la segmentación.
2. Se limitan a la identificación o clasificación de microorganismos y no abordan

la decisión acerca del diagnóstico del paciente. Tal vez sea esta la mayor limitación de estos algoritmos. No es posible conseguir un clasificador de bacilos que tenga una probabilidad de falsa alarma nula. Es por ello que cualquier solución al problema de identificación de pacientes con tuberculosis ha de pasar por asumir que el clasificador de bacilos no es perfecto y añadir un segundo nivel centrado en el diagnóstico del paciente.

Emplear diferentes muestras de esputo de un mismo paciente para mejorar el diagnóstico ha sido analizado en (Nelson *et al.*, 1998; Cascina *et al.*, 2000), donde se muestra que no hay ventajas evidentes en la sensibilidad del diagnóstico cuando se analizan más de dos muestras y no compensa en términos económicos.

1.1.3. Nuestra propuesta

Para diagnosticar a un paciente se dispone de imágenes de pacientes sanos, imágenes de pacientes enfermos y parches (pequeñas ventanas en las imágenes) etiquetados² como bacilos. En primer lugar se implementa un clasificador de bacilos entrenado mediante los pacientes sanos y los parches etiquetados como bacilo. Sin embargo, resulta, como hemos dicho, muy difícil implementar un clasificador cuya probabilidad de detectar erróneamente un bacilo sea cero. Por tanto, no es sencillo caracterizar a un paciente sano por el número de detecciones de bacilos, dado que un paciente sano producirá falsas detecciones de bacilos.

Atacamos este problema fusionando la información del clasificador de bacilos en un test secuencial cuyas hipótesis han sido definidas con incertidumbre. Este test hace que el sistema siga adquiriendo imágenes hasta que se hayan alcanzado las prestaciones deseadas dentro del margen de prestaciones alcanzables que se puede calcular a partir de los pacientes de entrenamiento.

Por otro lado, el sistema desarrollado en este trabajo tiene en cuenta las limitaciones del procesado en tiempo real y la solución propuesta sólo considera algoritmia implementable en tiempo real cuyos requisitos computacionales son cómodamente adaptables.

²El etiquetado de bacilos es un procedimiento costoso realizado manualmente por un experto.

Capítulo 2

Introducción

Este capítulo revisa las técnicas de contraste de hipótesis, que son una aplicación de la teoría estadística de la decisión, englobada a su vez en la teoría de la decisión (North, 1968; French, 1986; Howard, 2000; Edwards *et al.*, 2007). El capítulo comienza introduciendo el problema de la decisión y posteriormente describe el contraste de hipótesis para modelos paramétricos siguiendo los enfoques frecuentista y bayesiano, y los test empleados para modelos no paramétricos. A continuación, se centra en los test secuenciales para hipótesis binarias, que son especialmente útiles para la aplicación que motiva este trabajo porque permiten especificar la especificidad y sensibilidad deseadas.

Este capítulo también considera las regiones de confianza de parámetros que, como veremos, están estrechamente relacionadas con el contraste de hipótesis y juegan un papel relevante en el capítulo siguiente. La metodología de aprendizaje estadístico a partir de muestras analiza, entre otros, el problema de clasificación que predice una variable aleatoria discreta a partir de otra variable aleatoria. Este capítulo revisa la máquina de vectores soporte: su formulación y algunas técnicas para reducir su tiempo de ejecución. Finalmente, este capítulo presenta los objetivos de este trabajo.

2.1. Teoría estadística de la decisión

La teoría estadística de la decisión se ocupa de la toma de decisiones cuando las incertidumbres presentes en el proceso de decisión están modeladas de modo estadístico (Berger, 1985). Los elementos de la teoría estadística de la decisión son: el estado $\theta \in \Theta$ que suponemos desconocido, donde Θ representa el conjunto de estados posibles; las decisiones (o acciones) $a \in \mathcal{A}$, donde \mathcal{A} representa el conjunto de decisiones posibles; y una función de coste (o utilidad) $L(\theta, a)$ que representa el

coste (o recompensa) que supone decidir a cuando el estado es θ (véase en Berger 1985, cap. 2, North 1968 y French 1986, cap. 5).

Para la toma de la decisión se dispone de una serie de observaciones que suponemos independientes e idénticamente distribuidas (iid) $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \in \mathcal{X}$, donde \mathcal{X} es el espacio de las observaciones. Se asume que \mathbf{x} es una realización de la variable aleatoria \mathbf{X} . La función de densidad de probabilidad (fdp) de \mathbf{X} depende del estado y la denotamos como $f(\mathbf{x}|\theta)$ ¹.

2.1.1. Coste bayesiano esperado

La estadística bayesiana asume que θ es una variable aleatoria y formaliza el conocimiento *a priori* sobre el estado mediante una fdp $\pi(\theta)$.

Definición 2.1 (Coste bayesiano esperado). Si $\pi^*(\theta)$ es la distribución *a posteriori* de θ , esto es, la distribución de θ en el momento de la decisión, el coste bayesiano esperado de tomar la decisión a con una función de coste L es (Berger, 1985):

$$\rho(\pi^*, a) = \int_{\Theta} L(\theta, a) \pi^*(\theta) d\theta .$$

La distribución *a posteriori* $\pi^*(\theta)$ se obtiene como la actualización de la distribución *a priori* $\pi(\theta)$ con la verosimilitud $f(\mathbf{x}|\theta)$.

$$\pi^*(\theta) = \pi(\theta | \mathbf{x}) = \frac{f(\mathbf{x} | \theta)\pi(\theta)}{f(\mathbf{x})} .$$

La cantidad $f(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}|\theta)\pi(\theta) d\theta$ se denomina evidencia. La forma habitual de hacer inferencia acerca de θ en el análisis bayesiano describe el conocimiento inicial mediante la distribución *a priori* y emplea las observaciones para derivar, vía el Teorema de Bayes, la distribución *a posteriori*. Si no hay observaciones, la distribución *a posteriori* coincide con la distribución *a priori*. La mejor regla (o criterio) de decisión bayesiana es la que minimiza el coste esperado bayesiano.

2.1.2. Riesgo frecuentista

En el enfoque frecuentista el estado no es una variable aleatoria sino un parámetro fijo, pero desconocido. Desde este punto de vista no tiene sentido asignar probabilidades *a priori*.

¹Seguiremos este tipo de notación que es la habitual en la metodología bayesiana: lo hacemos para no cambiar de notación cada vez que describimos uno y otro enfoque. En ningún momento estamos tomando partido por ninguna de las dos filosofías. La notación frecuentista sería $f(\mathbf{x}; \theta)$.

Definición 2.2 (Regla de decisión). Una regla de decisión $\delta(\mathbf{x})$ es una función de \mathcal{X} en \mathcal{A} . Si \mathbf{x} es una realización de \mathbf{X} , $\delta(\mathbf{x})$ es la decisión (acción) asociada a esta observación. Esta decisión puede ser determinista o aleatoria.

Definición 2.3 (Riesgo de $\delta(\mathbf{x})$). La función de riesgo para una regla de decisión $\delta(\mathbf{x})$ se define como:

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(\mathbf{x}))f(\mathbf{x} | \theta) d\mathbf{x} . \quad (2.1)$$

El riesgo frecuentista evalúa para cada posible estado θ la esperanza con respecto a las observaciones \mathbf{x} del coste que supone emplear la regla de decisión $\delta(\mathbf{x})$.

Definición 2.4. Una regla de decisión δ_1 es R -mejor que una regla de decisión δ_2 si $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ con desigualdad estricta para algún θ . De igual modo, δ_1 es R -equivalente a δ_2 si $R(\theta, \delta_1) = R(\theta, \delta_2)$ para todo θ .

Definición 2.5 (Admisibilidad). Una regla de decisión δ es admisible si no existe ninguna otra regla R -mejor.

Definición 2.6 (Invariancia). Si un problema de decisión es invariante bajo un grupo de transformaciones \mathcal{G} , entonces la regla $\delta(\mathbf{x})$ es invariante bajo \mathcal{G} si para todo $\mathbf{x} \in \mathcal{X}$ y $g \in \mathcal{G}$

$$\delta(g(\mathbf{x})) = \tilde{g}(\delta(\mathbf{x})) ,$$

donde la existencia de $\tilde{g}(\cdot)$ muestra que las decisiones que se toman en los problemas \mathbf{x} y $g(\mathbf{x})$ se corresponden.

2.1.2.1. Riesgo de Bayes

Hemos visto que el riesgo frecuentista es una esperanza con respecto a \mathbf{X} . Si hacemos la esperanza también con respecto a θ tenemos otra medida de riesgo, el riesgo de Bayes.

Definición 2.7 (Riesgo de Bayes). El riesgo de Bayes para una regla de decisión δ , con respecto a una distribución *a priori* π sobre Θ , se define como

$$r(\pi, \delta) = \int_{\Theta} R(\theta, \delta)\pi(\theta) d\theta ,$$

donde $R(\theta, \delta)$ se ha definido en (2.1).

La regla δ que minimiza el riesgo de Bayes se denomina regla bayesiana. Bajo condiciones generales, si $\theta \in \mathbb{R}$ y $R(\theta, \delta)$ es continuo en θ para toda regla δ , entonces

las reglas bayesianas son admisibles (Wasserman, 2005, Teor. 12.19). Desde el punto de vista frecuentista θ no es una variable aleatoria. Sin embargo, $\pi(\theta)$ se puede interpretar como la importancia que se confiere a cada valor de θ (Lehmann, 1997).

Teorema 2.1. *Si la función de coste $L(\theta, \delta(\mathbf{x})) = (\theta - \delta(\mathbf{x}))^2$ entonces la regla bayesiana es*

$$\delta(\mathbf{x}) = \int_{\Theta} \theta \pi(\theta | \mathbf{x}) d\theta ,$$

esto es, la esperanza de θ condicionada a la observación \mathbf{x} .

Si $L(\theta, \delta(\mathbf{x})) = |\theta - \delta(\mathbf{x})|$, la regla bayesiana es la mediana de $\pi(\theta|\mathbf{x})$.

Si $L(\theta, \delta(\mathbf{x}))$ es el coste “cero-uno”, que no es convexo y se define por $L(\theta, \delta(\mathbf{x})) = 0$ si $\theta = \delta(\mathbf{x})$ y $L(\theta, \delta(\mathbf{x})) = 1$ en otro caso, la regla bayesiana es la moda de $\pi(\theta|\mathbf{x})$ (Van Trees, 2001), (Wasserman, 2005, Teor. 12.8).

2.1.2.2. Minimax

Si la distribución a priori es desconocida, podemos desear minimizar el máximo riesgo, $\sup_{\theta \in \Theta} R(\theta, \delta)$. Esto es lo que propone la regla minimax.

Definición 2.8 (Regla minimax). La regla δ^* es la regla minimax si

$$\sup_{\theta \in \Theta} R(\theta, \delta^*) = \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta) ,$$

donde \mathcal{D} es el conjunto de todas las reglas de decisión.

Si una regla tiene riesgo constante y es admisible entonces es minimax (Wasserman, 2005, Teor. 12.21). Las reglas minimax son reglas bayesianas con respecto a la distribución *a priori* menos favorable, esto es, la distribución que maximiza el riesgo bayesiano. En caso que tal distribución no exista, la regla minimax es el límite de la secuencia de distribuciones menos favorables (Lehmann, 1997).

En ocasiones se dispone de cierta información *a priori* pero no se confía completamente en ella para minimizar directamente el riesgo de Bayes. En lugar de ello, se prefiere la regla que minimice el riesgo de Bayes de entre las que cumplan la siguiente condición

$$R(\theta, \delta) \leq C \quad \forall \theta ,$$

donde C es siempre mayor que el riesgo minimax y más grande cuánto más certidumbre tenga sobre su información *a priori*. Esta solución se denomina regla de Bayes restringida (Lehmann, 1997).

2.1.3. Estadístico suficiente

Las reglas de decisión basadas en estadísticos suficientes son admisibles y por tanto es suficiente con estudiar únicamente estas reglas de decisión.

Definición 2.9 (Estadístico suficiente). Sea \mathbf{X} una variable aleatoria cuya distribución depende de un parámetro desconocido θ . Una función $T(\mathbf{X})$ es un estadístico suficiente para θ si \mathbf{X} es independiente de θ dado $T(\mathbf{X})$. En otras palabras, $T(\mathbf{X})$ es suficiente para θ si contiene toda la información que \mathbf{X} proporciona sobre θ (Cover y Thomas, 2006).

Teorema 2.2 (Fisher-Neyman). Sea \mathbf{x} una muestra de \mathbf{X} , el estadístico $T(\mathbf{x})$ es suficiente para θ si y solo si

$$f(\mathbf{x} | \theta) = g(\mathbf{x})h(T(\mathbf{x}) | \theta) ,$$

donde g no depende de θ (Scharf y Demeure, 1991).

Teorema 2.3. Si T es un estadístico suficiente para θ y $\delta_0^*(\mathbf{x}, \cdot)$ es una regla de decisión, existe una regla de decisión $\delta_1^*(t, \cdot)$ R -equivalente a δ_0^* que depende únicamente de $t = T(\mathbf{x})$.

El Teorema 2.3 establece que solo hace falta considerar reglas basadas en estadísticos suficientes (Berger, 1985).

2.1.4. Clases de reglas de decisión

Definición 2.10 (Clase esencialmente completa). Una clase \mathcal{C} de reglas de decisión es esencialmente completa si para cualquier regla de decisión δ que no está en \mathcal{C} existe una regla de decisión $\delta' \in \mathcal{C}$ que es R -mejor que o R -equivalente a δ .

Definición 2.11 (Clase completa). Una clase \mathcal{C} de reglas de decisión es completa si para cualquier regla de decisión δ que no está en \mathcal{C} existe una regla de decisión $\delta' \in \mathcal{C}$ que es R -mejor que δ .

Definición 2.12 (Clase mínimamente completa). Una clase \mathcal{C} de reglas de decisión es mínimamente completa si es completa y no lo es ningún subconjunto de la clase.

El Teorema 2.3 establece que la clase de reglas no deterministas basadas en un estadístico suficiente forman una clase esencialmente completa. La clase de reglas deterministas es completa si la función de coste es convexa (Berger, 1985).

Teorema 2.4. *Si Θ es finito, \mathcal{A} es finito y $L(\theta, a) \geq K > -\infty \forall \theta, a$; entonces el conjunto de las reglas de decisión de Bayes es una clase completa y el conjunto de reglas de decisión de Bayes admisibles es una clase mínimamente completa (Berger, 1985).*

2.2. Contraste de hipótesis

El contraste de hipótesis particulariza el problema general de decisión de la siguiente manera: las decisión $a_i \in \mathcal{A}$ elige como cierta la hipótesis H_i , que consiste en que el estado θ pertenece a la región $\Theta_i \subset \Theta$. Nos centraremos en el contraste entre dos hipótesis.

2.2.1. Enfoque frecuentista

En el enfoque frecuentista el contraste de hipótesis consiste en inferir a partir de las observaciones si una determina afirmación (o hipótesis) es cierta. La hipótesis que se desea contrastar, que suele asumirse como cierta hasta que se demuestre lo contrario, se denomina hipótesis nula, y la denotaremos como $H_0 : \theta \in \Theta_0$. La otra hipótesis $H_1 : \theta \in \Theta_1$ se denomina hipótesis alternativa (que es complementaria de la hipótesis nula). La regla de decisión (contraste o test) toma una de estas dos decisiones: a_1 rechazar la hipótesis nula porque las observaciones proporcionan evidencia en favor de la hipótesis alternativa; o, a_0 no rechazar la hipótesis nula debido a la falta de evidencia en su contra (Conover, 1998).

Definición 2.13 (Región crítica). Una regla de decisión asigna a cada posible realización \mathbf{x} de \mathbf{X} una de esas dos decisiones dividiendo el espacio muestral en dos regiones complementarias: S_0 y S_1 . S_0 es la región donde la hipótesis nula no se rechaza. El conjunto S_1 se denomina región crítica o región de rechazo y el conjunto S_0 región de aceptación.

Cuando se toma una decisión pueden cometerse dos tipos de errores: rechazar la hipótesis nula cuando es cierta (error de Tipo I, o falsa alarma) o, aceptarla cuando es falsa (error de Tipo II o no detección). El enfoque frecuentista evita la utilización directa de una función de coste (aunque lo hace implícitamente (Berger, 1985; Lehmann, 1997)) y desarrolla el contraste de hipótesis por medio de las probabilidades de error limitando el error de Tipo I y minimizando el error de Tipo II.

Definición 2.14 (Nivel de significación). El nivel de significación, denotado habitualmente como α , es la probabilidad máxima de rechazar erróneamente la hipótesis nula. El valor $1 - \alpha$ se denomina nivel de confianza.

Esto es, el enfoque frecuentista limita el error de Tipo I:

$$\int_{S_1} f(\mathbf{x} | \theta) d\mathbf{x} < \alpha \quad \forall \theta \in \Theta_0 ,$$

y maximiza la probabilidad de detección:

$$\int_{S_1} f(\mathbf{x} | \theta) d\mathbf{x} \quad \forall \theta \in \Theta_1 . \quad (2.2)$$

La expresión (2.2) representa la probabilidad de rechazar correctamente la hipótesis nula.

Definición 2.15 (Tamaño de un test). Se denomina tamaño de un test o tamaño de la región crítica a

$$\sup_{\theta \in \Theta_0} \int_{S_1} f(\mathbf{x} | \theta) d\mathbf{x} .$$

Definición 2.16 (Potencia del contraste). La potencia del contraste para rechazar la hipótesis nula o potencia del contraste es la probabilidad de rechazar correctamente la hipótesis nula.

Definición 2.17 (Hipótesis simple/compuesta). Una hipótesis $H_i : \theta \in \Theta_i$ es simple si Θ_i consiste en un único punto. En otro caso, se denomina compuesta.

Si la hipótesis H_1 es simple la potencia es un número, si por el contrario es compuesta la potencia es una función de θ con $\theta \in \Theta_1$. Habitualmente disminuir el nivel de significación (probabilidad de falsa alarma) también disminuye la potencia (probabilidad de detección) del test.

Definición 2.18 (Nivel crítico del test). El nivel crítico del test (p-valor del inglés *p-value*) es el menor nivel de significación con el que se rechaza la hipótesis nula para la observación dada.

El nivel crítico del test es una medida de la evidencia en contra de H_0 , y cuanto menor sea éste, mayor evidencia contra la hipótesis nula. Sin embargo, no es una probabilidad y menos la probabilidad de que la hipótesis nula sea cierta. Un nivel crítico alto puede mostrar que H_0 es cierta, o que es falsa y el test tiene una potencia baja (Wasserman, 2005).

Definición 2.19 (Test insesgado). Un test insesgado es aquel en el que la probabilidad de rechazar correctamente H_0 es siempre mayor o igual que la probabilidad de rechazar incorrectamente H_0 .

En otras palabras, un test insesgado es aquel en el que la potencia es siempre mayor o igual que el nivel de significación.

2.2.1.1. Contraste de hipótesis simples

En el caso de hipótesis simples, tenemos perfectamente caracterizadas las fdps de las hipótesis.

Teorema 2.5 (Lema de Neyman-Pearson). *Para el contraste de dos hipótesis $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$, se dispone de una observación \mathbf{x} que es una realización de \mathbf{X} . Para $k \geq 0$ definimos la región*

$$A(k) = \left\{ \mathbf{x} : \frac{f(\mathbf{x} | \theta_0)}{f(\mathbf{x} | \theta_1)} > k \right\} .$$

Sean $A^c(k)$ la región complementaria de $A(k)$, α^ la probabilidad de decidir H_0 en $A^c(k)$ y β^* la probabilidad de decidir H_1 en $A(k)$. Sea B cualquier otra partición del espacio con probabilidades de error α y β .*

Si $\alpha \leq \alpha^$ entonces $\beta \geq \beta^*$ (Neyman y Pearson, 1933; Cover y Thomas, 2006).*

El Lema de Neyman y Pearson (1933) es el primer teorema de clase de reglas de decisión completa y demuestra que el test óptimo para contrastar hipótesis simples consiste en comparar el cociente de verosimilitudes con un umbral.

Teorema 2.6 (Test de Neyman-Pearson). *Para el contraste de dos hipótesis $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$, se dispone de una observación \mathbf{x} que es una realización de \mathbf{X} . Si existe la constante k tal que*

$$\int_{\left\{ \mathbf{x} : \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} < k \right\}} f(\mathbf{x} | \theta_0) d\mathbf{x} = \alpha ;$$

la regla de decisión, donde $0 \leq \gamma \leq 1$,

$$\delta(\mathbf{x}) = \begin{cases} H_0 : \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} > k \\ H_0 \text{ con probabilidad } \gamma : \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} = k \\ H_1 : \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} < k \end{cases}$$

es el test de tamaño α más potente. Para la demostración ver por ejemplo Kay (1998); Berger (1985); Lehmann (1997).

Este test también es una regla bayesiana, y se puede llegar a él con ciertas probabilidades *a priori* de θ_0 y θ_1 y una función de coste de tipo $0-K_i$, que asigna K_i a escoger incorrectamente la hipótesis H_i y cero a las elecciones correctas (Berger, 1985).

Definición 2.20 (Divergencia de Kullback-Leibler). La divergencia de Kullback-

Leibler entre dos fdps p y q se define como

$$D(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx ,$$

donde asumimos siguiendo argumentos de continuidad que $0 \log \frac{0}{q} = 0$ y $p \log \frac{p}{0} = \infty$ (Cover y Thomas, 2006).

Las probabilidades de error de ambos tipos descienden exponencialmente con el número de muestras de la observación (Cover y Thomas, 2006). Si hacemos que el error de Tipo I (falsa alarma) descienda arbitrariamente lento, las prestaciones asintóticas del test de Neyman-Pearson para el error de Tipo II (o no detección) vienen descritas por el Lema de Stein.

Lema 2.1 (Stein). *Si α_n y β_n son respectivamente las probabilidades de error de Tipo I y II para n muestras y β_n^ϵ es el error de Tipo II sujeto a que $\alpha_n < \epsilon$ entonces:*

$$\lim_{\epsilon \rightarrow 0} \lim_{n \rightarrow \infty} \frac{1}{n} \beta_n^\epsilon = -D(f(\mathbf{x}|\theta_0)\|f(\mathbf{x}|\theta_1)) .$$

El mayor exponente con el que decrece el error de Tipo II es la divergencia de Kullback-Leibler entre las fdps de las hipótesis $D(f(\mathbf{x}|\theta_0)\|f(\mathbf{x}|\theta_1))$ (Cover y Thomas, 2006).

La metodología bayesiana, por otro lado, puede asignar probabilidades *a priori* π_0 y π_1 a las hipótesis H_0 y H_1 , y emplear una función de coste $0 - K_i$. En ese caso, el test óptimo comparara un cociente de verosimilitudes con un umbral y la probabilidad de error global para n muestras es:

$$P_e^n = \pi_0 \alpha_n + \pi_1 \beta_n .$$

Sea

$$D^* = \lim_{n \rightarrow \infty} \min_{A_n} -\frac{1}{n} \log P_e^n ,$$

donde A_n es la región de aceptación (no rechazo) de la hipótesis H_0 para n muestras. El mejor exponente de error en este caso se denomina exponente de Chernoff y viene dado por el siguiente teorema.

Teorema 2.7 (Chernoff). *El mejor exponente de error D^* alcanzable es*

$$D^* = D(f_{\lambda^*}\|f(\mathbf{x}|\theta_0)) = D(f_{\lambda^*}\|f(\mathbf{x}|\theta_1)) ,$$

donde

$$f_\lambda = \frac{f(\mathbf{x}|\theta_0)^\lambda f(\mathbf{x}|\theta_1)^{1-\lambda}}{\sum_{\mathbf{a} \in \mathcal{X}} f(\mathbf{a}|\theta_0)^\lambda f(\mathbf{a}|\theta_1)^{1-\lambda}}$$

y λ^* es el valor de λ que verifica $D(f_{\lambda^*} \| f(\mathbf{x}|\theta_0)) = D(f_{\lambda^*} \| f(\mathbf{x}|\theta_1))$.

La cantidad $D(f_{\lambda^*} \| f(\mathbf{x}|\theta_0))$ se conoce como la información de Chernoff.

2.2.1.2. Contraste de hipótesis compuestas

Es frecuente que cada hipótesis represente un conjunto de estados, que equivale a decir que el estado puede tomar cualquier valor en dicho conjunto. En este caso las fdfs condicionadas a las hipótesis tienen en general parámetros desconocidos.

Test uniformemente más potentes

Definición 2.21 (Test uniformemente más potente). Un test de tamaño α se dice que es uniformemente más potente (UMP) cuando maximiza la potencia para todas las alternativas de H_1 , esto es, para todos los posibles valores de $\theta \in \Theta_1$.

En otras palabras, no hay ningún otro test de tamaño α que tenga una potencia mayor para ningún valor de $\theta \in \Theta_1$.

Definición 2.22 (Test unilateral/bilateral). Para $\Theta \subset \mathbb{R}$ un test unilateral es del tipo: $H_0 : \theta \leq \theta_0$ contra $H_1 : \theta > \theta_0$, o del tipo $H_0 : \theta \geq \theta_0$ contra $H_1 : \theta < \theta_0$. El test de tipo $H_0 : \theta = \theta_0$ contra $H_1 : \theta \neq \theta_0$ se denomina bilateral.

En general los test uniformemente más potentes no existen porque la potencia depende de θ .

Definición 2.23 (Distribución con cociente de verosimilitud monótono). La distribución unidimensional X tiene cociente de verosimilitud monótono siempre que para $\theta_1 < \theta_2$ el cociente de verosimilitud

$$\frac{f(x | \theta_2)}{f(x | \theta_1)}$$

es una función no decreciente de x .

La clase más importante de distribuciones para la que existe un test UMP está definida en \mathbb{R} (o un subconjunto) y tiene cociente de verosimilitud monótono. Un ejemplo es la familia exponencial de funciones (Andersen, 1970), que incluye las distribuciones normal, exponencial, gamma, χ^2 , beta, Dirichlet, Bernoulli, binomial, multinomial, Poisson, binomial negativa y geométrica. Los miembros de esta familia pueden escribirse como:

$$f(x | \theta) = h(x)e^{\eta(\theta)T(x) - A(\theta)},$$

donde T , h , η y A son funciones conocidas. El resultado principal para estas distribuciones es el siguiente (Berger, 1985; Scharf y Demeure, 1991; Lehmann, 1997):

Teorema 2.8 (Karlin-Rubin). *Sea X una distribución unidimensional con cociente de verosimilitud monótono, para el contraste de $H_0 : \theta \leq \theta_0$ y $H_1 : \theta > \theta_0$ los test de la forma:*

$$\delta(x) = \begin{cases} H_1 & \text{si } x > x_0 \\ H_1 \text{ con probabilidad } \gamma & \text{si } x = x_0 \\ H_0 & \text{si } x < x_0 \end{cases},$$

donde $-\infty \leq x_0 \leq \infty$ y $0 \leq \gamma \leq 1$, verifican:

1. La función potencia es no decreciente en θ .
2. Cualquiera de esos test es UMP de su tamaño.
3. Para cualquier $0 < \alpha \leq 1$ existe un test de tamaño α de esta forma que es UMP.

Volviendo a la teoría de la decisión, si asumimos que la función de coste $L(\theta, a_i)$, $i = 0, 1$, cumple:

$$\begin{aligned} L(\theta, a_1) - L(\theta, a_0) &\geq 0 & \text{si } \theta < \theta_0 \\ L(\theta, a_1) - L(\theta, a_0) &\leq 0 & \text{si } \theta > \theta_0 \end{aligned} \quad (2.3)$$

podemos enunciar el teorema de clase completa (Berger, 1985):

Teorema 2.9. *Si la distribución de X tiene cociente de verosimilitud monótono y la función de coste es del tipo (2.3), entonces:*

1. La clase de test del Teorema 2.8 es esencialmente completa.
2. Cualquier test de esa forma es admisible, siempre que el conjunto $\{x : f(x | \theta) > 0\}$ sea independiente de θ y existan $\theta_1, \theta_2 \in \Theta$, $\theta_1 \leq \theta_0 \leq \theta_2$ tales que

$$\begin{aligned} L(\theta_1, a_1) - L(\theta_1, a_0) &> 0 \\ L(\theta_2, a_1) - L(\theta_2, a_0) &< 0. \end{aligned}$$

También existen tests UMP para el contraste de hipótesis bilaterales de la forma:

$$\begin{aligned} H_0 : \theta \leq \theta_1 \text{ ó } \theta \geq \theta_2 & \quad (\theta_1 < \theta_2) \\ H_1 : \theta_1 < \theta < \theta_2 \end{aligned}$$

para la familia de distribuciones exponenciales en la recta real. Sin embargo, tales tests no existen (Lehmann, 1997) para las hipótesis bilaterales del tipo

$$H_0 : \theta = \theta_0 \text{ ó } H_0 : \theta_1 \leq \theta \leq \theta_2.$$

Si nos centramos en los tests insesgados existen más casos tanto en familias de distribuciones como en tipos de hipótesis para los que existen test UMP (Lehmann, 1997).

Test no UMP Cuando no existe un test UMP estamos obligados a emplear tests sub-óptimos. Un test comúnmente usado es el test de cociente de verosimilitud generalizado (*generalized likelihood ratio test*) (GLRT). Este test reemplaza los parámetros desconocidos de las fdps condicionadas a las hipótesis por sus estimas de máxima verosimilitud. Un GLRT decide H_1 si

$$\frac{f(\mathbf{x} | \hat{\theta}_1)}{f(\mathbf{x} | \hat{\theta}_0)} > \gamma,$$

donde $\hat{\theta}_1$ y $\hat{\theta}_0$ son respectivamente las estimas de máxima verosimilitud de $\theta \in \Theta_1$ y $\theta \in \Theta_0$. Esto es,

$$\hat{\theta}_i = \arg \max_{\theta \in \Theta_i} f(\mathbf{x} | \theta) \quad i = 0, 1.$$

Se puede demostrar que el GLRT es asintóticamente UMP entre todos los test que son invariantes (Lehmann, 1997). El concepto de invariancia se examina en (Berger, 1985; Lehmann, 1997; Scharf y Demeure, 1991). Zeitouni *et al.* (1992) analizan la optimalidad asintótica del GLRT, derivan una condición suficiente para que el GLRT sea asintóticamente óptimo y ponen un contraejemplo para mostrar que no siempre lo es.

Kay (1998, cap. 11) considera el contraste de $H_0 : \theta = \theta_0$ y $H_1 : \theta \neq \theta_0$ y presenta las distribuciones asintóticas del GLRT y de dos de sus alternativas: el test de Wald y el test de Rao. Este último es el más sencillo de los tres porque no requiere maximizar la verosimilitud. A veces las fdps condicionadas a las hipótesis tienen parámetros “molestos” (del inglés *nuisance parameters*) que son los parámetros de la fdp que no aportan ninguna información sobre la hipótesis y típicamente degradan las prestaciones del test. Kay También presenta el test localmente más potente para el contraste $H_0 : \theta = \theta_0$ y $H_1 : \theta > \theta_0$ para $\Theta \in \mathbb{R}$ y su distribución asintótica.

Hoeffding (1965) analiza el caso de una hipótesis simple contra otra compuesta para el caso de distribuciones discretas con observaciones iid y muestra que para cualquier test existe un test de cociente de verosimilitud que no tiene peores

prestaciones asintóticas. Levitan y Merhav (2002) extienden el Lema de Neyman-Pearson generalizado (Hoeffding, 1965) para variables con alfabeto finito relajando la condición sobre el error de Tipo I haciendo que dependa de $\theta_0 \in \Theta_0$ (también se puede aplicar a otras distribuciones como distribuciones exponenciales o gaussianas). De esta forma, proponen un test y obtienen las condiciones que deben cumplir las hipótesis y la familia de distribuciones para obtener una caída exponencial de la probabilidad de no detección. Establece, cuando es posible, para cada θ_0 el valor mínimo de la probabilidad de error de Tipo I para el que se puede obtener una caída exponencial en el error de Tipo II.

2.2.2. Enfoque bayesiano

El análisis bayesiano se realiza combinando, por medio del Teorema de Bayes, la información *a priori* sobre el estado con la información proporcionada por las observaciones para obtener la distribución *a posteriori* del estado dadas las observaciones. En el enfoque bayesiano el rol de la distribución *a priori* representa el conocimiento (o incertidumbre) del estado, esto es, las creencias del investigador sobre el estado. Estas creencias son actualizadas por las observaciones para dar la distribución *a posteriori* que representa cuál es el conocimiento del investigador, esto es, sus creencias una vez examinada la evidencia.

Como hemos visto, el contraste de hipótesis bayesiano conduce a reglas admisibles bajo condiciones generales. La decisión se obtiene de las probabilidades *a posteriori* de las hipótesis, $P(\Theta_0|\mathbf{x})$ y $P(\Theta_1|\mathbf{x})$.

$$P(\Theta_i | \mathbf{x}) = \frac{P(\mathbf{x} | \Theta_i)P(\Theta_i)}{P(\mathbf{x} | \Theta_0)P(\Theta_0) + P(\mathbf{x} | \Theta_1)P(\Theta_1)} \quad i = 0, 1,$$

donde $P(\Theta_i)$ $i = 0, 1$ es la probabilidad *a priori* de H_i . A diferencia del test de Neyman-Pearson, que maximiza la potencia del test para una probabilidad de error de Tipo I acotada, el enfoque Bayesiano proporciona directamente una probabilidad. Dicha probabilidad expresa la creencia (subjetiva) acerca de la veracidad de la hipótesis. Este enfoque no se limita a variables aleatorias sino que también podemos asignar creencias al valor de parámetros deterministas.

Definición 2.24 (Factor de Bayes). La cantidad

$$B = \frac{P(\Theta_0 | \mathbf{x})/P(\Theta_1 | \mathbf{x})}{P(\Theta_0)/P(\Theta_1)}$$

es el factor de Bayes en favor de H_0 .

Para hipótesis simples el factor de Bayes coincide con el cociente de verosimilitud. El contraste de hipótesis unilaterales es directo en la metodología bayesiana. Sin embargo, para contrastar la hipótesis $H_0 : \theta = \theta_0$, se asigna en la distribución *a priori* $\pi(\theta)$ la masa π_0 a θ_0 , y a los demás valores $\theta \in \Theta$, $\theta \neq \theta_0$ la masa $\pi_1 = 1 - \pi_0$ asumiendo para éstos una distribución $g(\theta)$:

$$\pi(\theta) = \begin{cases} \pi_0 & \theta = \theta_0 \\ (1 - \pi_0)g(\theta) & \theta \neq \theta_0 \end{cases} .$$

De este modo, la distribución *a posteriori* resulta (Berger, 1985):

$$\pi(\theta_0 | \mathbf{x}) = \frac{\pi_0 f(\mathbf{x} | \theta_0)}{\pi_0 f(\mathbf{x} | \theta_0) + (1 - \pi_0) \int_{\theta \neq \theta_0} f(\mathbf{x} | \theta) g(\theta) d\theta} .$$

Más razonable quizá, por evitar el uso de deltas en la distribución *a priori*, sería el contraste de la hipótesis $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$ donde $b > 0$ se escoge de manera que esta hipótesis sea indistinguible de la original. Si esto es aceptable, el análisis es directo.

2.2.2.1. Elección de la distribución *a priori*

Un aspecto importante del enfoque bayesiano es la selección de la distribución *a priori* (Berger, 1985, cap. 3). Las distribuciones *a priori* informativas representan nuestro conocimiento sobre el estado. Si no se dispone de conocimiento *a priori*, la distribución *a priori* debe ser no informativa. Una forma de expresar una distribución *a priori* no informativa es establecer un valor constante para todos los valores del estado. Esto, que puede resultar lógico para distribuciones discretas o distribuciones continuas acotadas, conduce a distribuciones *a priori* impropias (no tiene integral finita (Berger, 1985)) cuando el estado es una variable continua no acotada. Sin embargo, es posible hacer inferencias con estas distribuciones *a priori* siempre que la distribución *a posteriori* sea propia, esto es, $\int \pi(\theta | \mathbf{x}) d\theta = 1$.

El problema del diseño de distribuciones *a priori* no informativas que caractericen la situación de que no se dispone de ningún conocimiento *a priori* ha sido tratado en la literatura. En este sentido, Jeffreys (1961) escoge como distribución *a priori*

$$\pi(\theta) = [I(\theta)]^{1/2} ,$$

donde $I(\theta)$ es la información de Fisher (Cover y Thomas, 2006). Bernardo (1979) introduce el concepto de distribución *a priori* de referencia que se obtiene de maxi-

mizar la divergencia de Kullback-Leibler esperada entre la distribución *a priori* y la distribución *a posteriori*. La aplicación de esta clase de distribución *a priori* se examina en (Bernardo y Rueda, 2002). La definición de cualquier distribución *a priori* equivale a alguna asunción acerca del estado. En este sentido, se puede argumentar que no existen distribuciones *a priori* no informativas (Bernardo, 1997).

Cuando se dispone de información *a priori* parcial acerca del estado, es deseable emplear una distribución *a priori* lo más no informativa posible para las regiones del estado sobre las que no tenemos información. Las distribuciones *a priori* de máxima entropía son útiles para este propósito (Jaynes, 1968; Berger, 1985). Para más estrategias para seleccionar una distribución *a priori* véase (Berger, 1985).

Familias conjugadas En general la obtención de la distribución *a posteriori* no es sencilla. En ocasiones, sólo es posible evaluar numéricamente mediante aproximaciones las distribuciones *a posteriori*. Sin embargo, en otros casos la distribución *a priori* y la distribución *a posteriori* son de la misma familia, lo que resulta de particular interés por la sencillez de los cálculos involucrados en las inferencias.

Definición 2.25 (Distribución *a priori* conjugada). Sea \mathcal{F} una clase de fdps $f(\mathbf{x}|\theta)$ indexadas por θ . Una clase \mathcal{P} de distribuciones *a priori* $\pi(\theta)$ se dice que es la familia conjugada de \mathcal{F} si la distribución *a posteriori* $\pi(\theta|\mathbf{x})$ está en la clase \mathcal{P} para toda $f \in \mathcal{F}$ y $\pi \in \mathcal{P}$ (Berger, 1985).

En otras palabras, una distribución *a priori* es conjugada de una verosimilitud si dicha distribución por la verosimilitud entre la evidencia pertenece a la misma familia de la distribución *a priori*. Véase (Fink, 1997) para una lista de familias conjugadas. Una clase interesante de familias conjugadas es la combinación convexa finita de elementos de una familia conjugada (Dalal y Hall, 1983). Esto aumenta la flexibilidad para la elección de la distribución *a priori* mientras que preserva la simplicidad en los cálculos.

2.2.3. Reglas no paramétricas

Hasta ahora hemos supuesto que conocemos la familia de fdps que ha originado los datos bajo cada hipótesis, salvo quizá el valor de algunos de sus parámetros. Cuando tal conocimiento no está disponible debemos emplear una regla de decisión (o test) no paramétrico (Lehmann y D'Abbrera, 1975; Conover, 1998; Good, 2004; Wasserman, 2006).

Los procedimientos que describimos a continuación examinan cómo de extremo es el estadístico del contraste para los datos disponibles con respecto a la distribución

del estadístico bajo la hipótesis nula.

Definición 2.26 (No paramétrico). Un método estadístico es no paramétrico si satisface al menos uno de los siguientes criterios (Conover, 1998):

1. El método puede emplearse en datos con una escala nominal de medida, en la cual el número asignado a las observaciones es simplemente el nombre de la categoría a la que la observación pertenece.
2. El método puede emplearse en datos con una escala ordinal de medida, en las que sólo son relevantes las relaciones entre los números mayor, menor o igual que, esto es, el valor numérico solo sirve para ordenar los elementos.
3. El método puede emplearse en datos cuya función de distribución es desconocida o tiene un número infinito de parámetros desconocidos.

2.2.3.1. Contraste de Wald

Uno de los contrastes no paramétricos más simples es el contraste de Wald propuesto para parámetros escalares. Dada la hipótesis nula $H_0 : \theta = \theta_0$, si $\hat{\theta}$ es una estima de θ y σ la desviación típica de dicha estima, y se puede asumir que

$$\frac{\hat{\theta} - \theta_0}{\sigma} \longrightarrow N(0, 1) ;$$

donde la convergencia es en distribución cuando el número de muestras del estimador tiende a infinito. El test de Wald de tamaño α rechaza H_0 cuando $W = \left| \frac{\hat{\theta} - \theta_0}{\sigma} \right| > z_{\alpha/2}$ donde $z_{\alpha/2}$ es el percentil $\alpha/2$ de una variable aleatoria gaussiana de media cero y varianza uno. Otros resultados sobre este test pueden consultarse en Wasserman (2005). El contraste de Wald ha sido criticado y algunos autores no lo recomiendan por estar disponibles test más potentes en la mayoría de los casos (Fears *et al.*, 1996; Pawitan, 2000).

2.2.3.2. Contrastes basados en las distribuciones binomial y χ^2

Para datos que tienen escala nominal los test más importantes son los basados en la distribución binomial. El primero es el test binomial (Conover, 1998) que es útil para el contraste de la hipótesis bilateral $H_0 : p = p_0$ (H_1 es siempre la hipótesis complementaria, en este caso $H_1 : p \neq p_0$), o el contraste de las hipótesis unilaterales $H_0 : p \geq p_0$ y $H_0 : p \leq p_0$. El test analiza hipótesis relacionadas con el número de éxitos en N intentos independientes (denominamos “éxito” a un ensayo de Bernoulli que toma valor “1”). El estadístico T de este test es el número de éxitos. Para

el contraste $H_0 : p = p_0$ se obtienen t_l y t_h , los percentiles $\alpha/2$ y $1 - \alpha/2$ de la distribución binomial de parámetros p_0 y N . La hipótesis nula $H_0 : p = p_0$ se rechaza si $T \leq t_l$ o $T > t_h$. La hipótesis $H_0 : p \geq p_0$ se rechaza si $T \leq t$ donde t es el percentil α de la distribución binomial de parámetros p_0 y N . La hipótesis $H_0 : p \leq p_0$ se rechaza si $T > t$ donde t es el percentil $1 - \alpha$ de la distribución binomial de parámetros p_0 y N .

El test del signo, que es un caso particular del test binomial cuando $p_0 = 1/2$, consiste en la comparación pareada de dos muestras o la comparación de una muestra con un umbral obteniendo de cada par simplemente un signo: “+”, “-” (o “=” en caso de empate). Este test es sencillo y ha sido empleado en multitud de aplicaciones como: comunicaciones (Tantaratana y Thomas, 1977), estudios de tendencias o correlación en poblaciones, etc. (Conover, 1998).

Definición 2.27 (Tabla de contingencia). Una tabla de contingencia es una matriz donde se representa la relación entre dos o más variables. Las variables suelen ser de tipo nominal u ordinal. En las tablas de contingencia las observaciones se clasifican (separan) en función de los valores de las variables (Lehmann y D’Abrera, 1975; Conover, 1998; Fleiss *et al.*, 2004).

Los test más usados en tablas de contingencia son el test basado en la distribución χ^2 (o de Pearson) (Pearson, 1922; Conover, 1998) y el test “exacto” de Fisher (Fisher, 1935; Conover, 1998; Good, 2004). La primera aplicación de la distribución χ^2 en un test que contrasta el ajuste de una distribución multinomial a una muestra iid clasificada en c clases se debe a Pearson (Pearson, 1922; Wasserman, 2005; Conover, 1998). Bajo la hipótesis nula, la probabilidad de que la muestra pertenezca a la clase j es p_j para todas las clases. Este contraste emplea como tabla de contingencia una matriz $1 \times c$.

Un caso de tablas de contingencia que aparece con frecuencia en la práctica son las tablas 2×2 . En ellas se pueden realizar test de diferencias de probabilidades en las clases. La tabla se organiza de modo que las filas son poblaciones y las columnas clases. Las hipótesis nulas son del tipo: $H_0 : p_1 = p_2$, $H_0 : p_1 \geq p_2$ o $H_0 : p_1 \leq p_2$ donde p_1 es la probabilidad de pertenecer a la clase 1 y p_2 es la probabilidad de pertenecer a la clase 2. En el caso $r \times c$ de r poblaciones y c clases se puede contrastar la hipótesis $H_0 : p_{1j} = p_{2j} = \dots = p_{rj} \forall j$ y la hipótesis $H_1 : \text{existe algún } j \text{ y algún par de clases donde las proporciones no son iguales}$.

Las tablas de contingencia también resultan útiles para contrastar la independencia de dos criterios de clasificación. Se clasifica una muestra en r clases con respecto a un criterio y en c clases con respecto al otro criterio. A partir de ambas

clasificaciones se obtiene la tabla donde cada celda contiene el resultado de una combinación. La hipótesis nula representa que el evento: “una observación está en la fila i ” es independiente del evento “la misma observación está en la columna j ” $\forall i, j$. A partir de dicha tabla se construye este test sabiendo que la hipótesis nula sigue aproximadamente una distribución χ^2 con $(r - 1) \times (c - 1)$ grados de libertad (Conover, 1998).

Otro test basado en la distribución χ^2 es el test de mediana que examina si c poblaciones tienen la misma mediana. La escala de medida es al menos ordinal para poder comparar con la mediana. La tabla de contingencia es una matriz $2 \times c$ que se construye evaluando en cada población cuántos elementos superan la mediana y cuántos no (Conover, 1998). Las tablas de contingencia también proporcionan medidas de dependencia como el coeficiente de Cramér, la contingencia media al cuadrado de Pearson y el coeficiente Phi (Fleiss *et al.*, 2004; Conover, 1998).

2.2.3.3. Contrastes para escala ordinal

Si las observaciones disponen al menos de una escala ordinal, pueden emplearse contrastes que aprovechen esta información y no sólo la clase. Estos contrastes son más potentes que los vistos en el apartado anterior, es más, su pérdida en potencia con respecto a contrastes paramétricos es sorprendentemente pequeña (Conover, 1998).

Definición 2.28 (Problema de dos muestras). Sea $\{x_1, x_2, \dots, x_N\}$ una muestra iid de una variable aleatoria X (unidimensional²) con función de distribución F desconocida e $\{y_1, y_2, \dots, y_M\}$, adquirida independientemente de $\{x_1, x_2, \dots, x_N\}$, una muestra iid de una variable aleatoria Y con función de distribución G desconocida. El problema de dos muestras (del inglés *two-sample problem*) contrasta la hipótesis nula $H_0 : F = G$ (Lehmann, 1997).

Contraste de Wilcoxon-Mann-Whitney El contraste de Mann-Whitney (del inglés *Mann-Whitney rank test*), también conocido como contraste de Wilcoxon (Wilcoxon, 1945; Mann y Whitney, 1947), ataca el problema de dos muestras combinando ambas muestras en una única muestra ordenada de menor a mayor y asignando a cada elemento (x_i $1 \leq i \leq N$ o y_j $1 \leq j \leq M$) su número de orden en la muestra combinada sin considerar de qué población proviene. Esto es razonable puesto que bajo la hipótesis nula ambas muestras vienen de la misma distribución. El estadístico está basado en la suma de los órdenes de una de las poblaciones, esto es, las

²También hay métodos que consideran variables multidimensionales.

posiciones de los elementos pertenecientes a esta población en la citada muestra combinada. Este test es insesgado y consistente para contrastes de las hipótesis nulas $H_0 : F(x) = G(x)$, $H_0 : F(x) > G(x)$ o $H_0 : F(x) < G(x)$. Si asumimos además que se cumple siempre una de las alternativas $F(x) = G(x)$ o $F(x) = G(x + c)$ donde c es una constante, entonces este test también es insesgado y consistente para el contraste de las hipótesis nulas $H_0 : E(X) = E(Y)$, $H_0 : E(X) > E(Y)$ o $H_0 : E(X) < E(Y)$.

El contraste de Kruskal y Wallis (1952) generaliza el contraste de Wilcoxon-Mann-Whitney para k muestras, siendo la hipótesis nula que todas las muestras provienen de la misma distribución y la alternativa que al menos una de las muestras proviene de otra distribución. El test de Van der Waerden (1952) (*Normal scores*) mejora la eficiencia relativa asintótica del test de Kruskal-Wallis haciendo una transformación de los órdenes usando los percentiles de la distribución normal.

Conover (1998) también describe cómo hacer un test de varianza con estos contrastes así como medidas de correlación similares a las ya vistas en las tablas de contingencia pero empleando también el orden de las muestras.

Definición 2.29 (Variable simétrica). La distribución de una variable aleatoria X es simétrica con respecto a $x = c$, para alguna constante c , si la probabilidad de $X \leq c - x$ es igual a la probabilidad de $X \geq c + x$ para cada valor de x (Conover, 1998).

Contraste de los signos de Wilcoxon El contraste de los signos de Wilcoxon es una extensión del contraste del signo, discutido en el Apartado 2.2.3.2, para datos donde una diferencia esté definida, que equivale al menos a una escala de medida de intervalo, en la que es relevante no sólo el orden sino también la diferencia entre dos observaciones. Toma datos unidimensionales o la diferencia de datos bidimensionales. Asume que la distribución de los datos o de la diferencia es simétrica. La diferencia de dos muestras tomadas de la misma población tiene distribución simétrica. Es más, si son tomadas de las distribuciones $F(x)$ y $F(x - c)$ su diferencia tiene distribución simétrica (Good, 2004). Esto hace que la media y la mediana coincidan y por tanto cualquier conclusión sobre una de ellas es válida para la otra. El test está diseñado para contrastar hipótesis acerca de la media o mediana de una o dos poblaciones (Conover, 1998). Otra aplicación de este contraste es la determinación del intervalo de confianza para la diferencia de mediana (Conover, 1998).

Los test de Friedman (1937) y Quade (1979) hacen un análisis similar al test de los signos de Wilcoxon para el caso de varias muestras relacionadas. Estos test están diseñados para experimentos en bloques completos, esto es, cuando todas

las poblaciones están afectadas por los mismos tratamientos. El test de Durbin (1951) está diseñado para experimentos en bloques incompletos balanceados, donde la palabra balanceado informa de que todos los tratamientos se administran un número igual de veces.

2.2.3.4. Contrastes de permutación

Los test de permutación (Fisher, 1935; Lehmann, 1997; Conover, 1998; Good, 2004; Wasserman, 2005) son los test no paramétricos más potentes. Además, no necesitan que las muestras sean iid sino que es suficiente con que sean intercambiables (véase (Good, 2004)) para que el test sea insesgado y exacto, en cuanto al cálculo del nivel crítico. Desde el punto de vista de los contrastes basados en el orden de las muestras, estos procedimientos asignan a cada muestra su propio valor (Conover, 1998). La hipótesis nula de los test de permutación es que no hay diferencias entre las muestras. De este modo, bajo la hipótesis nula las etiquetas de las observaciones se pueden intercambiar. Así se puede obtener el nivel crítico del test sin más que contar las permutaciones que dan valores del estadístico más extremos que la muestra original y dividirlos por el número total de permutaciones posibles para las etiquetas.

Aunque este tipo de test puede emplearse para cualquier aplicación en la que se puedan aplicar los test anteriores, su principal desventaja es que el número de permutaciones, $N!$ (donde N es la suma del número de elementos de ambas poblaciones) aumenta rápidamente con el número de muestras. En ocasiones, no es necesario calcular el estadístico para más que unas pocas permutaciones obteniendo sólo los valores más extremos. Si todavía son demasiadas permutaciones es posible aproximar el nivel crítico usando un número razonable de permutaciones aleatorias (Wasserman, 2005).

2.2.3.5. Contrastes basados en la función de distribución

El contraste de Kolmogorov (1933) contrasta la hipótesis nula $F = G$ donde G es una función de distribución conocida y de F sólo se dispone de una muestra. Kolmogorov emplea como estadístico la distancia máxima entre la función de distribución G y la función de distribución empírica \hat{F} de los datos (véase Vapnik (1998) para métodos de construcción de la función empírica). El test de Kolmogorov también permite obtener un intervalo de confianza para la función de distribución real (Conover, 1998).

Los contrastes basados en la función de distribución son sensibles a cualquier diferencia entre las dos distribuciones y por tanto, son más adecuados para el pro-

blema de dos muestras que aquéllos que buscan diferencias en la media, la mediana, etc. El contraste de Smirnov (1939) contrasta la hipótesis nula de $F = G$ cuando ambas distribuciones son desconocidas y una muestra de cada una está disponible. Smirnov emplea como estadístico la máxima diferencia entre \hat{F} y \hat{G} , las funciones de distribución empíricas de las muestras provenientes de F y G . En este contraste sólo importa el orden de las muestras en la muestra conjunta y no sus valores concretos. Al igual que en el contraste de Mann-Whitney y en el test de permutación, asumimos en la hipótesis nula que las etiquetas de las muestras son intercambiables, esto permite construir la distribución de la hipótesis nula.

Un contraste más elaborado para el problema de dos muestras es el contraste de Cramér-Von Mises (Conover, 1998). A diferencia del contraste de Smirnov, este contraste usa los valores de ambas funciones de distribución empíricas en todas las muestras para construir el estadístico. La distribución asintótica de este estadístico fue obtenida por Anderson y Darling (1954).

2.2.3.6. Otras reglas no paramétricas

El problema de dos muestras ha sido estudiado desde hace tiempo y ha generado abundante literatura. Algunos trabajos recientes son (Einmahl y Khmaladze, 2001; Kim y Foutz, 1987; Cao y Van Keilegom, 2006). Gretton *et al.* (2007) y Borgwardt *et al.* (2006) presentan la versión basada en núcleos (del inglés *kernels*) del contraste de Fortet y Mourier (1953) con excelentes resultados tanto en potencia como en aplicabilidad al precio de una elevada carga computacional.

Si dos muestras provienen de la misma distribución su divergencia de Kullback-Leibler empírica tiende a cero cuando el número de muestras tiende a infinito. En esta línea, Perez-Cruz (2008) muestra que emplear la estima de la divergencia de Kullback-Leibler para el problema anterior y el problema de independencia entre dos muestras es más conveniente, especialmente en términos de eficiencia computacional, que las propuestas por Gretton *et al.* (2007) y Gretton *et al.* (2008).

2.2.4. Test secuenciales

Todos los contrastes vistos hasta el momento usan una muestra de tamaño fijo. Si se desea disminuir el tamaño (probabilidad de falsa alarma) de un test, en general se disminuye también su potencia (probabilidad de detección). La detección secuencial (Wald, 1947; Berger, 1985; Ghosh y Sen, 1991; Poor, 1994; Govindarajulu, 2004) fija las prestaciones deseadas (probabilidades de falsa alarma y no detección) y toma observaciones (muestras) hasta que dichas prestaciones son satisfechas.

El análisis secuencial es un problema de decisión que no sólo tiene en cuenta el coste de la decisión sino también el coste que supone obtener las observaciones. Desde este punto de vista, el problema que hay que resolver es el diseño del experimento y la toma de decisión de tal manera que el coste global esperado sea mínimo (Wald, 1947; Berger, 1985).

Berger (1985) introduce el análisis secuencial desde el punto de vista de la teoría de la decisión definiendo las funciones de coste, el riesgo y las reglas de decisión. Para contrastar las hipótesis simples $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$ el resultado más notable, debido a Wald (Wald, 1947), es el test secuencial de cociente de verosimilitud (*sequential probability ratio test*) (SPRT). Berger (1985) demuestra que el SPRT es un procedimiento bayesiano: define probabilidades *a priori* de las hipótesis, supone un coste C de adquisición de una observación y define la función de coste $L(\theta, a, C) = L(\theta, a) + nC$, donde nC es el coste de adquirir n muestras.

Se dispone de las observaciones iid $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$, $\mathbf{x}_i \in \mathcal{X}$, $i = 1, 2, \dots$ que son muestras (adquiridas de forma secuencial, tantas como sean necesarias) de una variable aleatoria \mathbf{X} cuya fdp es $f(\mathbf{x}|\theta_0)$ si H_0 es cierta o $f(\mathbf{x}|\theta_1)$ si H_1 es cierta.

Definición 2.30 (Regla de decisión secuencial). Una regla de decisión secuencial consiste en un par de secuencias (ϕ, δ) donde $\phi = \{\phi_j\}$, $j = 0, 1, 2, \dots$ se denomina regla de parada ($\phi_j: \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j\} \rightarrow \{0, 1\}$) y $\delta = \{\delta_j\}$, $j = 0, 1, 2, \dots$ se denomina regla de decisión final, δ_j toma una decisión sobre $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j\}$.

La regla de decisión secuencial (ϕ, δ) opera de la siguiente manera: para una secuencia $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$, la regla (ϕ, δ) toma la decisión $\delta_N(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$, donde $N = N(\phi)$ es la muestra de parada $N = \min_n \phi_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = 1$. Esto es, seguimos tomando muestras mientras $\phi_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = 0$, y cuando $\phi_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = 1$ dejamos de adquirir muestras y tomamos una decisión. Como hemos mencionado, el SPRT es un procedimiento bayesiano. Por tanto, las decisiones se toman a partir de la distribución *a posteriori* (Poor, 1994):

$$\begin{aligned} \pi(\theta_1 | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) &= \frac{\pi_1 \prod_{k=1}^n f(\mathbf{x}_k | \theta_1)}{\pi_1 \prod_{k=1}^n f(\mathbf{x}_k | \theta_1) + \pi_0 \prod_{k=1}^n f(\mathbf{x}_k | \theta_0)} \\ &= \frac{\pi_1 \lambda_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)}{\pi_1 \lambda_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) + \pi_0}, \end{aligned} \quad (2.4)$$

donde λ_n es el cociente de verosimilitud de n muestras dado por:

$$\lambda_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \frac{\prod_{k=1}^n f(\mathbf{x}_k | \theta_1)}{\prod_{k=1}^n f(\mathbf{x}_k | \theta_0)}.$$

Como (2.4) es monótonamente creciente en λ_n (Poor, 1994), el test SPRT con fron-

teras A y B , $SPRT(A, B)$ se puede escribir como:

$$\phi_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \begin{cases} 0 & \text{si } A < \lambda_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) < B \\ 1 & \text{en otro caso} \end{cases} .$$

$$\delta_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \begin{cases} 0 & \text{si } \lambda_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \leq A \\ 1 & \text{si } \lambda_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \geq B \end{cases} .$$

Para una regla secuencial (ϕ, δ) se denotará como $P_{FA}(\phi, \delta)$ su tamaño (o probabilidad de falsa alarma) y como $P_{ND}(\phi, \delta)$ su error de Tipo II (o probabilidad de no detección). Estos valores se definen como:

$$P_{FA}(\phi, \delta) = P(\delta_N(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = 1 \mid H_0)$$

$$P_{ND}(\phi, \delta) = P(\delta_N(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = 0 \mid H_1) .$$

Teorema 2.10 (Wald-Wolfowitz). *Sea (ϕ_0, δ_0) el $SPRT(A, B)$ y (ϕ, δ) cualquier otra regla de decisión secuencial para la cual*

$$P_{FA}(\phi, \delta) \leq P_{FA}(\phi_0, \delta_0)$$

$$P_{ND}(\phi, \delta) \leq P_{ND}(\phi_0, \delta_0) ,$$

entonces

$$E\{N(\phi) \mid H_i\} \geq E\{N(\phi_0) \mid H_i\} \quad i = \{0, 1\} ,$$

donde E es la esperanza respecto a cada hipótesis (Wald y Wolfowitz, 1948).

Por tanto, fijadas unas prestaciones, no hay ninguna regla de decisión secuencial cuyo tamaño muestral esperado sea menor. Un test de tamaño de muestra fijo es un caso particular de regla secuencial. Este teorema implica que en media el número de muestras que necesita un SPRT no es mayor que el de un test de tamaño fijo para las mismas prestaciones.

Vamos ahora con el cálculo de A y B para obtener las prestaciones deseadas. Sea $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$:

$$1 - P_{ND} = \int_{S_1} f(\mathbf{x} \mid \theta_1) d\mathbf{x} = \int_{S_1} \lambda_N(\mathbf{x}) f(\mathbf{x} \mid \theta_0) d\mathbf{x} \geq B \int_{S_1} f(\mathbf{x} \mid \theta_0) d\mathbf{x} = BP_{FA} ,$$

donde hemos usado el hecho de que en el momento de la decisión $\lambda_N(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \geq B$

y tenemos la cota $B \leq \frac{1-P_{ND}}{P_{FA}}$. De igual modo:

$$P_{ND} = \int_{S_0} f(\mathbf{x} | \theta_1) d\mathbf{x} = \int_{S_0} \lambda_N(\mathbf{x}) f(\mathbf{x} | \theta_0) d\mathbf{x} \leq A \int_{S_0} f(\mathbf{x} | \theta_0) d\mathbf{x} = A(1 - P_{FA}) ,$$

donde hemos usado el hecho de que en el momento de la decisión $\lambda_N(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \leq A$ y tenemos la cota $A \geq \frac{P_{ND}}{1-P_{FA}}$. Teniendo en cuenta que $A \leq B$ tenemos

$$\frac{P_{ND}}{1 - P_{FA}} \leq A \leq B \leq \frac{1 - P_{ND}}{P_{FA}} .$$

Las aproximaciones de Wald (1947) para unas restricciones de diseño en cuanto a probabilidad de falsa alarma α y de no detección γ consisten en tomar la igualdad en ambas cotas:

$$\begin{aligned} A &= \frac{\gamma}{1 - \alpha} \\ B &= \frac{1 - \gamma}{\alpha} . \end{aligned} \tag{2.5}$$

Con la selección de A y B de la Ecuación (2.5) obtenemos la siguiente cota para las prestaciones P_{ND} y P_{FA} obtenidas por el test cuando finaliza:

$$\begin{aligned} P_{FA} &\leq \frac{(1 - P_{ND})\alpha}{1 - \gamma} \leq \frac{\alpha}{1 - \gamma} = \frac{1}{B} \\ P_{ND} &\leq \frac{(1 - P_{FA})\gamma}{1 - \alpha} \leq \frac{\gamma}{1 - \alpha} = A . \end{aligned} \tag{2.6}$$

$$P_{FA} + P_{ND} \leq \alpha + \gamma$$

A la última ecuación de (2.6) se llega sumando $P_{FA}(1 - \gamma) \leq (1 - P_{ND})\alpha$ con $P_{ND}(1 - \alpha) \leq (1 - P_{FA})\gamma$ y simplificando. (2.6) indica que para probabilidades de diseño prácticas, las probabilidades reales de falsa alarma y no detección solo son ligeramente más grandes que su valor especificado. Es más, pueden fijarse valores A y B para garantizar cualquier prestación. En este sentido, cuando es absolutamente necesario garantizar γ y α Wald (1947) recomienda $A = \gamma$ y $B = 1/\alpha$ que garantiza su cumplimiento como fácilmente se comprueba en las desigualdades (2.6). Otro resultado interesante es que la suma de las probabilidades obtenidas P_{FA} y P_{ND} es menor que la suma de las probabilidades especificadas α y γ . Esto es, las probabilidades de falsa alarma y no detección obtenidas no pueden incrementarse a la vez.

Si hacemos $\gamma = \alpha$ Poor (1994) llega a la siguiente aproximación:

$$\begin{aligned} P_{FA} &\leq \alpha + O(\alpha^2) \\ P_{ND} &\leq \gamma + O(\gamma^2), \end{aligned}$$

que se demuestra sin más que partir de (2.6) y hacer el desarrollo en serie de Taylor en torno a cero.

Wald (1947) prueba que el $SPRT(A, B)$ termina con probabilidad 1. Para obtener el valor esperado de N empezamos definiendo la siguiente regla secuencial (Poor, 1994).

Definición 2.31 ($TS(a, b; g)$). Para cada $a < 0 < b$ y cada función $g: \mathcal{X} \rightarrow \mathbb{R}$ la regla secuencial $TS(a, b; g)$ está definida por el par (ϕ, δ) dado por

$$\begin{aligned} \phi_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) &= \begin{cases} 0 & \text{si } a < \sum_{i=1}^n g(\mathbf{x}_i) < b \\ 1 & \text{en otro caso} \end{cases} \\ \delta_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) &= \begin{cases} 0 & \text{si } \sum_{i=1}^n g(\mathbf{x}_i) \leq a \\ 1 & \text{si } \sum_{i=1}^n g(\mathbf{x}_i) \geq b. \end{cases} \end{aligned}$$

El $SPRT(A, B)$ es un caso particular de esta regla donde $a = \log A$, $b = \log B$ y $g(\mathbf{x}) = \log(f(\mathbf{x}|\theta_1)/f(\mathbf{x}|\theta_0))$.

Teorema 2.11 (Identidad fundamental del análisis secuencial). Sean $(\phi, \delta) = TS(a, b; g)$, $N = \min_n \phi_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = 1$, $S_n = \sum_{i=1}^n g(\mathbf{x}_i)$ y M_j la función generadora de momentos de la variable aleatoria $g(\mathbf{x}_1)$ bajo la hipótesis H_j , esto es,

$$M_j(t) = E\{e^{tg(\mathbf{x}_1)} | H_j\}, \quad j = 0, 1.$$

Si $P(g(\mathbf{x}_1) = 0 | H_j) \neq 1$ y $P(|g(\mathbf{x}_1)| < \infty | H_j) = 1$ entonces

$$E\{e^{tS_N} M_j(t)^{-N} | H_j\} = 1$$

para todo t real tal que $M_j(t) < \infty$ (Poor, 1994).

Corolario 2.1. Bajo las hipótesis del Teorema 2.11, sea $M_j(t) < \infty$ en un entorno de $t = 0$. Se define $\mu_j = E\{g(\mathbf{x}_1) | H_j\}$ y $\sigma_j^2 = \text{Var}(g(\mathbf{x}_1) | H_j)$ entonces:

$$E\{S_N | H_j\} = \mu_j E\{N | H_j\}$$

y

$$E\{(S_N - N\mu_j)^2 \mid H_j\} = \sigma_j^2 E\{N \mid H_j\} .$$

A partir de este corolario, llegamos a las aproximaciones de Wald para N , el valor esperado de la muestra en la que el SPRT termina:

$$E\{N \mid H_0\} \cong \frac{1}{\mu_0} \left[(1 - \alpha) \log \frac{\gamma}{1 - \alpha} + \alpha \log \frac{1 - \gamma}{\alpha} \right]$$

y

$$E\{N \mid H_1\} \cong \frac{1}{\mu_1} \left[\gamma \log \frac{\gamma}{1 - \alpha} + (1 - \gamma) \log \frac{1 - \gamma}{\alpha} \right] ,$$

(Wald, 1947; Poor, 1994) donde, a partir del Corolario 2.1,

$$\mu_j = E \left\{ \log \frac{f(\mathbf{x}_1 \mid \theta_1)}{f(\mathbf{x}_1 \mid \theta_0)} \mid H_j \right\} .$$

Berger (1985) puntualiza este resultado cambiando el símbolo \cong por \geq haciendo hincapié en que estas estimas de la esperanza de N son cotas inferiores. Al contrario que los resultados sobre las probabilidades de error alcanzadas (2.6), que eran independientes de la distribución, los resultados descritos en el Corolario 2.1 dependen de la distribución de los datos bajo cada una de las hipótesis.

2.2.4.1. Hipótesis compuestas

Hemos presentado el test secuencial para hipótesis simples. La extensión de los test secuenciales al caso de hipótesis compuestas es considerablemente más compleja que la extensión de los test de tamaño fijo para los cuales existen test UMP para el contraste de hipótesis unilaterales (véase Apartado 2.2.1.2). Así, el test de Neyman-Pearson de nivel crítico α para la hipótesis compuesta $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ para familias de cociente de verosimilitud monótono en la recta real se reduce al contraste de las hipótesis simples $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1 > \theta_0$, que proporciona la solución óptima al no depender de θ_1 .

En el caso secuencial esta reducción a hipótesis simples no es posible ni siquiera en el caso que acabamos de exponer. Por ejemplo, para contrastar $H_0 : \theta \leq \theta_0$ y $H_1 : \theta \geq \theta_1 > \theta_0$ con probabilidades de error de Tipo I y II no mayores que α y γ es posible emplear un SPRT que contraste $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$ (Wald,

1947). Sin embargo, mientras el SPRT minimiza el número de muestras esperado para $\theta = \theta_0$ y $\theta = \theta_1$ esto no ocurre para otros valores de θ y su tamaño máximo esperado puede ser mayor que el del test óptimo de tamaño fijo, especialmente si $\theta_0 < \theta < \theta_1$ (Wijsman, 1991).

Sobel (1953) demuestra que para el contraste de hipótesis unilaterales en la familia exponencial de distribuciones de un parámetro los test de la forma

$$\phi_n(x_1, x_2, \dots, x_n) = \begin{cases} 0 & \text{si } A_n < \sum_{i=1}^n x_i < B_n \\ 1 & \text{en otro caso} \end{cases}$$

$$\delta_n(x_1, x_2, \dots, x_n) = \begin{cases} 0 & \text{si } \sum_{i=1}^n x_i \leq A_n \\ 1 & \text{si } \sum_{i=1}^n x_i \geq B_n \end{cases}$$

siendo $A_n < B_n$, forman una clase esencialmente completa de reglas de decisión. Para este caso, el test óptimo tiene esta forma. El problema, aún abierto, se reduce a encontrar las secuencias A_n y B_n .

Kiefer y Weiss (1957) consideran el problema de minimizar el tamaño muestral esperado en un punto θ^* sujeto a las restricciones de probabilidad de error en θ_0 y θ_1 en la familia exponencial de un parámetro real. Hoeffding (1960) derivó una cota inferior del número de muestras esperado para este caso. Lorden (1976) demostró que una solución asintótica a este problema es emplear 2-SPRT con una regla de parada del tipo

$$\phi_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \begin{cases} 1 & \text{si } \left\{ n : \prod_{i=1}^n \frac{f(\mathbf{x}_i|\theta^*)}{f(\mathbf{x}_i|\theta_0)} \geq A_0 \text{ o } \prod_{i=1}^n \frac{f(\mathbf{x}_i|\theta^*)}{f(\mathbf{x}_i|\theta_1)} \geq A_1 \right\} \\ 0 & \text{en otro caso} \end{cases} \quad (2.7)$$

El 2-SPRT como su nombre indica emplea dos SPRT unilaterales, donde uno de los umbrales se ha llevado a cero o a ∞ , y uno de los SPRT sirve para rechazar H_0 y el otro para rechazar H_1 . El valor θ^* se escoge como aquél donde el valor esperado del número de muestras es máximo y con ello se minimiza dicho valor. Una cualidad importante del 2-SPRT es que permite usar las aproximaciones de Wald (2.5) para establecer los umbrales de cada test unilateral A_0 y A_1 . Huffman (1983) estudia el 2-SPRT afinando las cotas del tamaño máximo esperado cuando las probabilidades de error tienden a cero.

El valor θ^* en el que se minimiza el tamaño máximo de muestras necesarias para el test debería ser el parámetro real, que es desconocido. Schwarz (1962) considera el contraste $H_0 : \theta \leq \theta_0$ y $H_1 : \theta > \theta_1 > \theta_0$ para la familia exponencial de distribuciones dentro de un enfoque bayesiano con una función de coste 0-1 y un coste por muestra

de C . Sustituyendo en (2.7) θ^* por el valor de máxima verosimilitud $\hat{\theta}_n$ en la etapa n cuando $C \rightarrow 0$ Schwarz derivó una solución asintótica a este problema. Chernoff y Ray (1965) se plantean el contraste $H_0 : \theta \leq \theta_0$ y $H_1 : \theta > \theta_1 > \theta_0$ siguiendo el mismo enfoque. Curiosamente, ambas soluciones asintóticas no coinciden si hacemos $\theta_1 = \theta_0$. Esta discrepancia fue resuelta por Lai (1988) quien propuso la siguiente regla de parada:

$$\phi_n(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \begin{cases} 1 & \text{si } \max \left\{ \prod_{i=1}^n \frac{f(\mathbf{x}_i|\hat{\theta}_n)}{f(\mathbf{x}_i|\theta_0)}, \prod_{i=1}^n \frac{f(\mathbf{x}_i|\hat{\theta}_n)}{f(\mathbf{x}_i|\theta_1)} \right\} \geq g(Cn) \\ 0 & \text{en otro caso} \end{cases} \quad (2.8)$$

La precisión de la estima $\hat{\theta}_n$ de θ varía con n . Así, la función $g(Cn)$ proporciona un umbral variante. Lai (1988) proporciona una expresión cerrada para g y argumenta con resultados analíticos y numéricos que el método es casi óptimo tanto desde un punto de vista frecuentista como desde un punto de vista bayesiano para una variedad de distribuciones *a priori*. Lai (1997) extiende este resultado para una clase general de funciones de coste y distribuciones *a priori*.

Existen varias extensiones a este resultado: Lai y Zhang (1994) generaliza este test al caso en que el estado es un vector y se contrastan las hipótesis $H_0 : \phi(\boldsymbol{\theta}) \leq \phi_0$ y $H_1 : \phi(\boldsymbol{\theta}) > \phi_1 \geq \phi_0$ para la familia de distribuciones exponencial multiparamétrica.

Cox (1963) analiza el caso de hipótesis compuestas con parámetros molestos, que es un caso particular en el que el estado es un vector. Propone una aproximación basada en máxima verosimilitud en la que realiza una aproximación del cociente de verosimilitud en serie de Taylor y en cada paso refina la aproximación del estado. Ghosh (1970) compila las aportaciones de test secuenciales con hipótesis compuestas en familias multiparamétricas. La mayoría de estas aportaciones usan la invariancia para reducir las hipótesis compuestas a hipótesis simples y a partir de ahí emplean el SPRT propuesto por Wald. Sin embargo, la invariancia reduce considerablemente la variedad de las hipótesis que es posible contrastar.

Darkhovskii (2006) ha propuesto recientemente un test minimax para el contraste secuencial de las hipótesis $H_0 : \boldsymbol{\theta} \in \Theta_0$ y $H_1 : \boldsymbol{\theta} \in \Theta_1$ ³ que coincide con el SPRT cuando las hipótesis compuestas se reducen a hipótesis simples.

2.2.4.2. Test secuenciales truncados

Cuando el número máximo de muestras está acotado, por ejemplo debido al coste de adquisición o al tiempo máximo de procesado, el test secuencial debe tomar una decisión cuando se alcanza el número máximo de muestras si no ha sido posible

³Las negritas aparecen porque las hipótesis definen regiones multidimensionales.

tomarla antes. En cuanto a la decisión tomada, una posibilidad es decidir H_1 si el logaritmo del cociente de verosimilitud es mayor que el umbral obtenido de partir a la mitad la distancia entre $\log A$ y $\log B$; otra es decidir H_1 siempre que el logaritmo del cociente de verosimilitud sea positivo. La elección de una u otra de estas alternativas depende de la aplicación.

Las combinaciones posibles de prestaciones garantizadas en ese punto pueden obtenerse usando las aproximaciones de Wald (2.6) y el valor final del cociente de verosimilitud. Govindarajulu (2004) acota las probabilidades de no detección y falsa alarma obtenidas en el momento de truncamiento N_0 siempre que éste sea suficientemente grande para que $\sum_i \log(f(\mathbf{x}_i|\theta_1)/f(\mathbf{x}_i|\theta_0))$ converja a una variable aleatoria normal. Así tenemos las desigualdades:

$$\begin{aligned} \alpha(N_0) &\leq \alpha + \frac{B-1}{B-A} \left[\Phi \left(\frac{\ln B - N_0 \mu_0}{\sqrt{N_0} \sigma_0} \right) - \Phi \left(\frac{-\sqrt{N_0} \mu_0}{\sigma_0} \right) \right] \\ \gamma(N_0) &\leq \gamma + \frac{B(1-A)}{B-A} \left[\Phi \left(\frac{-\sqrt{N_0} \mu_1}{\sigma_1} \right) - \Phi \left(\frac{\ln A - N_0 \mu_1}{\sqrt{N_0} \sigma_1} \right) \right], \end{aligned} \quad (2.9)$$

donde $\mu_j = E \left\{ \log \frac{f(\mathbf{x}_1|\theta_1)}{f(\mathbf{x}_1|\theta_0)} \mid H_j \right\}$, $\sigma_j^2 = var \left\{ \log \frac{f(\mathbf{x}_1|\theta_1)}{f(\mathbf{x}_1|\theta_0)} \mid H_j \right\}$, siendo *var* la varianza y Φ la función de distribución de una normal estándar. Estas últimas cotas son preferibles, siempre que N_0 sea suficientemente grande, a las obtenidas a partir del cociente de verosimilitud en N_0 y de las aproximaciones de Wald para los umbrales del test, que son más pesimistas.

Tantaratana y Thomas (1977) comparan el test de Neyman-Pearson con un test secuencial que usa solo el signo de las observaciones para la detección de señales antipodales. El resultado de la comparación en términos de muestras necesarias resultó favorable para el test secuencial a pesar de que usaba menos información que el test de Neyman-Pearson. La única desventaja del test secuencial es que el número máximo de muestras no está acotado. Tantaratana y Thomas (1977) proponen como alternativa truncar el test, esto es, limitar el máximo número de muestras, y argumentan que si la muestra en la que se trunca el test es tal que la probabilidad de necesitar más muestras es pequeña, las prestaciones del test secuencial apenas se ven degradadas con respecto a las prestaciones de diseño.

Tantaratana y Poor (1982) analizan las prestaciones asintóticas de los test secuenciales truncados. Desde su punto de vista, un test secuencial truncado es una mezcla entre un test secuencial y un test de Neyman-Pearson. Así, el test secuencial truncado bien diseñado necesita menos muestras en media que un test de Neyman-Pearson, y está cerca del número óptimo requerido por el test secuencial.

Anderson (1960) propone una modificación del SPRT con umbrales convergentes.

El momento en que ambos umbrales se cruzan es una cota superior del número de muestras de test, N . Este test también se puede truncar antes de dicho punto. Recientemente, Frazier y Yu (2007) han propuesto un método para adaptar los umbrales del test secuencial para que una decisión sea tomada en un tiempo limitado definido como la realización de una variable aleatoria.

2.2.5. Resumen

En esta sección hemos introducido los modelos paramétrico (frecuentista y bayesiano) y no paramétrico del contraste de hipótesis. Las fortalezas del modelo paramétrico son: los test UMP, que solo están disponibles en algunos casos, y test asintóticamente óptimos, como el GLRT, u óptimos en algún sentido, como el presentado por Levitan y Merhav (2002). En las ocasiones en que podamos formalizar el conocimiento *a priori* es razonable seguir la metodología bayesiana.

En algunas ocasiones la fdp o la familia paramétrica a la que pertenece es desconocida. El Teorema de Glivenko-Cantelli (Vapnik, 1998) demuestra la convergencia de la función de distribución empírica a la función de distribución real. Sin embargo, la estima de la fdp es un problema mal condicionado (Vapnik, 1998). Parzen (1962) propone un método no paramétrico de estima de fdp basado en núcleos. Vapnik (1998) discute cómo debe variar el ancho del kernel para la convergencia del estimador de Parzen y cuál es su tasa de convergencia para funciones de densidad suaves (véanse Silverman (1986) y Wasserman (2006) para un compendio de métodos de estima de fdp).

Los métodos bayesianos tienen la ventaja teórica de eliminar todos los parámetros que no interesan mediante integración, de forma que desaparecen del modelo. Un ejemplo son los parámetros molestos que aparecen frecuentemente en la metodología frecuentista. Sin embargo, no están exentos de problemas: en ocasiones no resulta sencillo formalizar el conocimiento *a priori* en una función, y la obtención de la distribución *a posteriori* es costoso si la distribución *a priori* no es conjugada de la verosimilitud. Es más, en algunas ocasiones no existe solución analítica y la distribución *a posteriori* debe ser aproximada con técnicas variacionales (MacKay, 2003), de Monte Carlo: *Population Monte Carlo* (Cappe et al., 2004), *Markov Chain Monte Carlo* (Fitzgerald, 2001) o los métodos descritos por Robert y Casella (2004). Los métodos de Monte Carlo proporcionan estimas puntuales, promedios y cualquier otra cosa que nos pueda interesar a cambio de un elevado coste computacional que cada día es más tolerable.

Por otro lado, la metodología no paramétrica es aplicable a cualquier situación. Los test basados en el orden de las muestras son una alternativa en términos de

complejidad y prestaciones a los test paramétricos. Sin embargo, la hipótesis de que las muestras son intercambiables bajo la hipótesis nula reduce el número de hipótesis que es posible realizar. Por ejemplo, para la hipótesis nula de medias iguales, las distribuciones deben ser iguales salvo por un desplazamiento. En este sentido, Conover (1998) y Lehmann y D'Abrera (1975) hacen énfasis en las asunciones de cada test. Los test de permutación superan las prestaciones de los contrastes basados en el orden de las muestras al precio de una considerable carga computacional si el número de muestras es grande. Otra limitación de los métodos no paramétricos es la dificultad de añadir información *a priori*.

Entre los modelos paramétricos y no paramétricos tenemos el enfoque de la verosimilitud empírica (del inglés *Empirical Likelihood*) (Owen, 2001; Einmahl y McKeague, 2003). Este enfoque define una verosimilitud empírica a partir de la función de distribución empírica. Hay algunos trabajos que han atacado el problema de dos muestras desde este punto de vista (Qin, 1994; Jing, 1995; Zhang, 2000; Cao y Van Keilegom, 2006).

Aunque sólo hemos presentado los test secuenciales paramétricos, en particular el SPRT, también es posible diseñar test secuenciales no paramétricos (Ghosh y Sen, 1991; Good, 2004). El SPRT permite limitar ambos tipos de error a cambio de no limitar el número de muestras y es el test que antes alcanza esas prestaciones en media. Lai (2001) revisa el análisis secuencial y sus aplicaciones a distintos campos que incluyen la medicina, la economía y la ingeniería.

2.3. Intervalos de confianza

Como hemos mencionado, existe una estrecha relación entre el contraste de hipótesis y los intervalos de confianza. Así, un intervalo de confianza $1 - \alpha$ bilateral para un parámetro θ no es otra cosa que la región del estado compuesta por todos los $\theta_0 \in \Theta$ para los que no se puede rechazar la hipótesis $H_0 : \theta = \theta_0$ a un nivel α con la muestra dada (Lehmann, 1997). Un intervalo de confianza, estimado a partir de una muestra, contiene el rango de valores en el que es probable que se encuentre un parámetro. Como la muestra se escoge aleatoriamente de la población, los extremos del intervalo son realizaciones de los estadísticos $a(\cdot)$ y $b(\cdot)$ que obtienen dichos extremos a partir de la muestra.

Definición 2.32 (Intervalo de confianza). Para una muestra \mathbf{x} obtenida de $f(\mathbf{x}|\theta)$ un intervalo de confianza $(1 - \alpha)$ para un parámetro $g(\theta)$ es un intervalo $[a(\mathbf{x}), b(\mathbf{x})]$ de modo que la probabilidad de que $g(\theta)$ esté dentro de ese intervalo es mayor o igual que $(1 - \alpha)$ para todo $\theta \in \Theta$.

En otras palabras, un intervalo de confianza es función de la muestra aleatoria e incluye el parámetro $g(\theta)$ con una probabilidad mínima de $(1 - \alpha)$. Esta definición de intervalo de confianza se relaja frecuentemente en cuanto a la confianza. Diremos que un intervalo de confianza es exacto⁴ cuando esta confianza se alcanza exactamente; diremos que es conservador cuando la confianza establecida es una cota inferior de la confianza alcanzada; finalmente, hablaremos de intervalos de confianza aproximados cuando la confianza fluctúa alrededor de la confianza establecida para distintos valores de θ .

Una interpretación de los intervalos de confianza desde la metodología frecuentista es que contienen al parámetro un $100 \times (1 - \alpha) \%$ del tiempo. En otras palabras, si estimamos el intervalo de confianza del parámetro a partir de K conjuntos de muestras independientes, aproximadamente $(1 - \alpha)K$ de dichos intervalos contiene el parámetro.

Existe una estrecha relación entre los intervalos de confianza y los contrastes de hipótesis (Lehmann, 1997). El siguiente teorema muestra esa relación (Good, 2004):

Teorema 2.12. *Sea \mathbf{x} una muestra de la distribución $f(\mathbf{x}|\theta)$. Para cada $\theta' \in \Theta$ sea $A(\theta')$ la región de aceptación de un test de nivel crítico α para la hipótesis nula $H_0 : \theta = \theta'$ y para cada \mathbf{x} sea $S(\mathbf{x}) = \{\theta : \mathbf{x} \in A(\theta), \theta \in \Theta\}$. Entonces, $S(\mathbf{x})$ forma una familia de intervalos de confianza para θ con un nivel de confianza $1 - \alpha$.*

Este resultado permite obtener intervalos de confianza exactos si conocemos la familia paramétrica que origina los datos. Hay resultados análogos para hipótesis nulas unilaterales que pueden verse como duales de intervalos unilaterales.

Entre las alternativas para obtener intervalos de confianza se encuentra el intervalo estándar (Wasserman, 2005) que asume que el estimador $g(\hat{\theta})$ se distribuye aproximadamente como una normal $N(\theta, \hat{s}e^2)$. El intervalo construido es de la forma:

$$\left(g(\hat{\theta}) - z_{\alpha/2} \hat{s}e, g(\hat{\theta}) + z_{\alpha/2} \hat{s}e \right),$$

donde $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$, siendo Φ la función de distribución de una normal estándar y $\hat{s}e$ la desviación típica de $g(\hat{\theta})$. Este intervalo es el más usado en la práctica, está apoyado por el teorema del límite central y funciona muy bien cuando la hipótesis de normalidad se cumple. Sin embargo, aunque un estimador sea asintóticamente normal, puede estar lejos de la normalidad con pocas muestras. En ese caso el intervalo estándar puede proporcionar coberturas muy alejadas del nivel de confianza especificado. La cobertura es la probabilidad de que el intervalo de un

⁴Otros autores identifican los términos exacto y conservador con el significado que hemos dado a conservador.

nivel de confianza dado capture el parámetro en realidad y depende en general de θ y del tamaño de la muestra. Que los intervalos de confianza proporcionen malas coberturas ocurre en algunas ocasiones, como ponen de manifiesto Brown *et al.* (2001).

Los métodos de permutación son exactos para el cálculo de intervalos de confianza, en el sentido que su cobertura es igual a su nivel de confianza. Sin embargo, no todas las coberturas son alcanzables en variables discretas (Lehmann, 1997; Good, 2004). El problema de estos métodos para el cálculo de intervalos de confianza es el mismo que para el contraste de hipótesis: hay que evaluar $N!$ permutaciones. Al igual que entonces, se puede aproximar el intervalo evaluando un subconjunto de las permutaciones posibles.

Otra familia de técnicas para el cálculo de intervalos de confianza aproximados es la denominada metodología bootstrap (Efron y Tibshirani, 1993; Shao y Tu, 1995; DiCiccio y Efron, 1996). Estos métodos son aplicables prácticamente a cualquier problema donde queramos obtener un intervalo de confianza, tanto usando estadística paramétrica como no paramétrica. Los métodos bootstrap no paramétricos crean B nuevas muestras del mismo tamaño que la muestra original a partir de la función de distribución empírica y obtienen una estima del parámetro para cada una de ellas. Los métodos bootstrap paramétricos estiman los parámetros de la función de distribución a partir de la muestra y mediante dicha distribución generan B muestras bootstrap. Efron y Tibshirani (1993) propone cuatro métodos para la estima de intervalos de confianza: el bootstrap_t, el método del percentil, el BCa (del inglés *Bias corrected and accelerated*) y el ABC (del inglés *Approximate Bootstrap Confidence*). En la práctica se recomiendan los métodos BCa y ABC, siendo éste último menos costoso computacionalmente.

Cabe preguntarse: ¿cuántas muestras bootstrap deben realizarse para obtener aproximaciones precisas? Efron y Tibshirani (1993) indican que hacen falta un mínimo de mil muestras bootstrap para un intervalo de confianza del 95%. Desde entonces han aparecido otros trabajos en la literatura considerando este problema desde un punto de vista menos heurístico (Davidson y MacKinnon, 2000; Andrews y Buchinsky, 2000, 2001, 2002). Estos trabajos buscan el número de muestras bootstrap que hacen que la potencia del test al que el intervalo de confianza es equivalente no diste de la potencia que se alcanza con infinitas muestras.

2.3.1. Enfoque bayesiano

Desde un punto de vista bayesiano toda la información acerca de una variable aleatoria está en su distribución *a posteriori*. De este modo, un intervalo de confianza

para una variable aleatoria se construye a partir de su distribución *a posteriori*. Un intervalo de confianza $(1 - \alpha)$ para el valor de una variable aleatoria se puede encontrar usando cualquier región en la que la distribución *a posteriori* integre $(1 - \alpha)$. Sin embargo, lo más usual es seleccionar la región más pequeña en la que la integral de la distribución *a posteriori* es mayor o igual que $(1 - \alpha)$. Esta región corresponde a la que selecciona los valores mayores de la distribución *a posteriori* en primer lugar (Berger, 1985).

Definición 2.33. El intervalo (de confianza) creíble con densidad *a posteriori* más alta (*highest posterior density*) (HPD) para θ es el subconjunto C de Θ de la forma

$$C = \{\theta \in \Theta : \pi(\theta | \mathbf{x}) \geq k(\alpha)\}$$

donde $k(\alpha)$ es la mayor constante que verifica

$$P(C | \mathbf{x}) \geq 1 - \alpha$$

Otra posibilidad, siguiendo el enfoque frecuentista, consiste en dejar $\frac{\alpha}{2}$ en cada cola de la función de densidad. El caso multidimensional es completamente análogo y en los casos en los que la distribución *a posteriori* no se conozca analíticamente o no sea fácil integrarla, los intervalos pueden obtenerse numéricamente (Chafaï y Concordet, 2009).

En los intervalos bayesianos, a diferencia de los frecuentistas, la confianza se corresponde con una probabilidad, entendiendo ésta como la creencia subjetiva que combina nuestra creencia *a priori* con la información proporcionada por la muestra.

Estos intervalos de confianza, también llamados conjuntos creíbles, son exactos, en cuanto a su confianza, si la distribución *a priori* usada se corresponde a la realidad. En caso contrario, son aproximados. Sin embargo, no hay grandes diferencias entre los intervalos bayesianos y los frecuentistas cuando las distribuciones *a priori* seleccionadas son no informativas (Berger, 1985).

2.3.2. Proporción de una distribución binomial

Un caso en el que el intervalo de confianza estándar funciona particularmente mal es la estima de la proporción⁵ de una distribución binomial (Agresti, 2003). Como hemos comentado, Brown *et al.* (2001) ponen de manifiesto el interés de este problema, complejo, aunque aparentemente sencillo, por sus aplicaciones prácticas y la

⁵Denominamos proporción al parámetro que describe la probabilidad de éxito (acierto) de una distribución binomial.

considerable literatura que ha generado.

El intervalo de confianza $(1 - \alpha)$ para la proporción de una distribución binomial consiste en todos los valores de p^* tal que los datos (N muestras) produzcan la aceptación de la hipótesis nula $H_0 : p = p^*$ a un nivel crítico α . Como estamos considerando una hipótesis bilateral, cada cola de la distribución binomial tiene probabilidad $\alpha/2$. El valor inferior del intervalo, p^*_l es aquel para el que $\alpha/2$ sea la probabilidad de que el número de aciertos sea el número observado (n_1) o mayor. p^*_l se obtiene de modo que

$$\alpha/2 = \sum_{i=n_1}^N \binom{N}{i} (p^*_l)^i (1 - p^*_l)^{N-i} .$$

El valor superior del intervalo, p^*_h se obtiene a partir de que $\alpha/2$ sea igual a la probabilidad de que el número de aciertos no sea mayor que el número observado.

$$\alpha/2 = \sum_{i=0}^{n_1} \binom{N}{i} (p^*_h)^i (1 - p^*_h)^{N-i} .$$

Los valores del intervalo se obtienen teniendo en cuenta la relación entre la distribución binomial y la distribución beta (Abramowitz y Stegun, 1964):

$$\sum_{i=n_1}^N \binom{N}{i} (p^*_l)^i (1 - p^*_l)^{N-i} = I_{p^*_l}(n_1, N - n_1 + 1) ,$$

donde

$$I_x(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1} (1-t)^{b-1} dt$$

es la función de distribución de una beta(a, b) que tiene la propiedad de simetría $I_x(a, b) = 1 - I_{1-x}(b, a)$ y $\Gamma(\cdot)$ es la función gamma. El extremo inferior del intervalo de confianza p^*_l es el percentil $\alpha/2 \times 100$ de la función de distribución de una beta($n_1, N - n_1 + 1$). Para p^*_h tenemos

$$\begin{aligned} \alpha/2 &= \sum_{i=0}^{n_1} \binom{N}{i} (p^*_h)^i (1 - p^*_h)^{N-i} \\ &= 1 - \sum_{i=n_1+1}^N \binom{N}{i} (p^*_h)^i (1 - p^*_h)^{N-i} \\ &= 1 - I_{p^*_h}(n_1 + 1, N - n_1) , \end{aligned}$$

de donde obtenemos que p^*_h es el percentil $(1 - \alpha/2) \times 100$ de la función de distribución de una beta($n_1 + 1, N - n_1$). Este intervalo se denomina en la literatura exacto (aunque nosotros lo llamamos conservador) y se debe a Clopper y Pearson

(1934). La cobertura del intervalo de Clopper-Pearson es muy superior en la práctica al nivel de confianza usado para su diseño. Este hecho se debe principalmente al carácter discreto de la distribución binomial (Agresti, 2003). Han ido apareciendo otros métodos para obtener el intervalo de confianza de la proporción de una distribución binomial también conservadores pero que proporcionan una cobertura más cercana al nivel de confianza. Cada uno de ellos ha refinando al anterior (Blyth y Still, 1983; Casella, 1986; Blaker, 2000; Cai y Krishnamoorthy, 2005). La ventaja de estos intervalos exactos (o conservadores) es que su cobertura es siempre mayor o igual que el nivel de confianza, lo que los hace recomendables para aplicaciones críticas. Esto es a costa de una mayor longitud del intervalo y de una cobertura considerablemente mayor que el nivel de confianza para valores cercanos a cero o a uno. Cai (2005) examina las malas prestaciones del intervalo unilateral estándar cuando se aplica a variables discretas y presenta un intervalo unilateral conservador, basado en el intervalo de Clopper-Pearson, y otro intervalo unilateral aproximado basado en el intervalo de Jeffreys (ver más abajo) para variables discretas.

El problema fundamental de los intervalos exactos es que su cobertura es mayor que su nivel de confianza y su longitud mayor de lo deseado. En las ocasiones en que se desee una cobertura aproximadamente igual al nivel de confianza, los intervalos aproximados son más adecuados porque su longitud es menor. Brown *et al.* (2001) examinan varias alternativas para construir intervalos aproximados y sus coberturas. Entre ellas, recomiendan el intervalo de Wilson (Wilson, 1927) o el intervalo bayesiano con la distribución *a priori* de Jeffreys, que equivale en este caso a una distribución $\text{beta}(1/2, 1/2)$, cuando el número de muestras es menor que 40, y el intervalo de Agresti y Coull (1998) que proporciona intervalos comparables a los anteriores y es más sencillo de calcular para el resto de los casos. Henderson y Meyer (2001) comparan la cobertura del intervalo de Wilson con el intervalo de Clopper-Pearson y con el intervalo bayesiano usando distintas distribuciones *a priori*.

DasGupta y Zhang (2005) discuten de manera detallada la estima puntual y los intervalos de confianza para la proporción y el número de intentos de variables aleatorias binomiales y multinomiales. Asimismo, presentan cotas inferiores analíticas de la cobertura de algunos de los citados intervalos de confianza.

En la Sección 2.2.3.2 hemos introducido las tablas de contingencia para contraste de hipótesis. Agresti y Hitchcock (2005) revisan la estima de parámetros e intervalos de confianza en tablas de contingencia desde un enfoque bayesiano y muestran algunas relaciones entre los estimadores frecuentistas y los estimadores bayesianos para ciertas distribuciones *a priori*. Por ejemplo, el estimador de máxima verosimilitud de una proporción lineal se corresponde con el estimador bayesiano para el que se

ha seleccionado una distribución *a priori* impropia.

2.3.3. Regiones de confianza para la proporción de una distribución multinomial

Las regiones de confianza son la extensión de los intervalos de confianza cuando el parámetro a estimar es un vector. Una región aleatoria \mathcal{R}_α es una región de confianza de nivel α para θ si $P(\theta \in \mathcal{R}_\alpha) = \alpha$. La estima de regiones de confianza para el vector de probabilidades de una distribución multinomial es la generalización de la estima de intervalos de confianza para una distribución binomial. Quesenberry y Hurst (1964) presentan un método para construir dichas regiones basándose en la aproximación asintótica a una χ^2 de

$$\sum_{i=1}^D (n_i - Np_i)^2 / Np_i$$

donde n_i es el número de observaciones cuyo valor es i , N es el número total de observaciones y p_i es la probabilidad de que una observación tome el valor i . Goodman (1965) refina estos intervalos construyendo otros más cortos que satisfacen la confianza exigida. Su método se basa en la aproximación normal de la probabilidad de una distribución binomial. Bailey (1980) propone una transformación que aproxima aún más dicha distribución a la normalidad.

Beran y Millar (1986) discuten las regiones de confianza desde un punto de vista general. Hall (1987) aproxima las regiones de confianza \mathcal{R}_α mediante bootstrap_t y estima no paramétrica de fdp. Esta aproximación es general y sirve para estimar la región de confianza de un parámetro de cualquier distribución, obteniendo regiones de confianza \mathcal{R}_α con una cobertura de $\alpha + O(N^{-1})$. Hall (1992) extiende dicho método y propone aproximaciones para evitar el remuestreo. A partir de este trabajo Glaz y Sison (1999) se centran en la estima de regiones de confianza para la proporción de la distribución multinomial.

Thompson (1987) presenta un procedimiento para determinar el tamaño muestral necesario para la estima de la proporción de la distribución multinomial asumiendo el caso peor para el valor del parámetro. Este problema está estrechamente relacionado con la determinación de una región de confianza para dicho parámetro. Chafaï y Concordet (2009) han presentado recientemente un método para estimar la región de confianza de la proporción de una multinomial válido para cualquier caso, especialmente aquéllos en los que las aproximaciones asintóticas presentadas arriba dan malos resultados. El precio es una elevada carga computacional: por ejemplo si

$D = 3$ hay que realizar una búsqueda para cada valor candidato de la proporción en la que en cada paso hay que evaluar una multinomial $(N + 1)(N + 2)/2$ veces.

2.4. Máquina de vectores soporte

Si queremos obtener la clase de una muestra de test para un problema de clasificación binario en el que las clases han sido caracterizadas mediante ejemplos, podríamos emplear los métodos paramétricos ya descritos mediante la estima previa de fdp, o emplear alguno de los métodos no paramétricos para comparar distribuciones. Sin embargo, si empleamos los primeros estamos resolviendo un problema de estima de fdp para después resolver un problema de clasificación y los segundos suelen emplear varias muestras de test y variables unidimensionales. Es por ello que presentamos un ejemplo del aprendizaje estadístico basado en muestras que funciona especialmente bien en casos donde la dimensión de los datos es alta y las clases están caracterizadas por relativamente pocas muestras.

La máquina de vectores soporte (*support vector machine*) (SVM) (Boser *et al.*, 1992; Cortes y Vapnik, 1995; Burges, 1998; Schölkopf y Smola, 2001) es una herramienta de clasificación lineal y no lineal inspirada en la minimización del riesgo estructural (Vapnik, 1995, 1998). En realidad, minimiza un funcional que pondera dos elementos: el riesgo empírico y un regularizador. La SVM también puede verse como una generalización del hiperplano óptimo de decisión (*optimal hyperplane decision rule*) (OHDR) (Vapnik, 1982). La SVM para problemas separables y lineales maximiza la mínima distancia entre las muestras y el plano que separa las clases. El caso no lineal se ataca proyectando las muestras en un espacio de características y separándolas en ese espacio mediante un hiperplano. Esto resulta equivalente a un clasificador no lineal en el espacio de las muestras. El teorema de Cover (1965) constata que la separabilidad mediante un hiperplano de un conjunto de muestras aumenta con la dimensión del espacio. Por otro lado, para problemas no separables se introducen variables auxiliares que relajan las condiciones de la formulación separable.

2.4.1. Formulación

Disponemos del conjunto de muestras iid $\{(\mathbf{x}_i, y_i)\}$, $i = 1, \dots, N$, donde $\mathbf{x}_i \in \mathbb{R}^d$ e $y_i \in \{\pm 1\}$, tomadas de una distribución desconocida $p(\mathbf{x}, y)$. La SVM minimiza el

siguiente problema denominado *primal*:

$$\min_{\mathbf{w}, \xi_i, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \quad (2.10)$$

sujeto a:

$$\begin{aligned} y_i(\phi(\mathbf{x}_i)\mathbf{w} + b) &\geq 1 - \xi_i && \forall i = 1, \dots, N \\ \xi_i &\geq 0 && \forall i = 1, \dots, N \end{aligned}$$

donde: $\phi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{\mathcal{H}}$ es la transformación, usualmente no lineal, del espacio de entrada al espacio de características; \mathbf{w} y b forman el clasificador lineal (hiperplano) en el espacio de características; C establece un compromiso entre reducir el error de entrenamiento y maximizar el margen, que es la distancia entre las muestras bien clasificadas de una clase y el hiperplano de separación; y ξ_i es la penalización que introducimos cuando una muestra no puede cumplir la restricción:

$$y_i(\phi(\mathbf{x}_i)\mathbf{w} + b) \geq 1 .$$

Aunque una muestra no cumple la restricción de margen siempre que $\xi_i > 0$, sólo está mal clasificada cuando $\xi_i \geq 1$. Para resolver el problema (2.10) hacemos uso de las condiciones de Karush-Kuhn-Tucker (KKT) (una generalización del teorema de los multiplicadores de Lagrange) (Fletcher, 1987) que nos lleva al siguiente problema:

$$\min_{\mathbf{w}, b, \xi_i} \max_{\alpha_i, \mu_i} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \mu_i \xi_i - \sum \alpha_i (y_i(\phi(\mathbf{x}_i)\mathbf{w} + b) - 1 + \xi_i) \quad (2.11)$$

en cuya solución es necesario que:

$$\frac{\partial L_P}{\partial \mathbf{w}} = \mathbf{w} - \sum_i \alpha_i y_i \phi(\mathbf{x}_i) = 0 \quad (2.12)$$

$$\frac{\partial L_P}{\partial b} = \sum_i \alpha_i y_i = 0 \quad (2.13)$$

$$\frac{\partial L_P}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \quad \forall i = 1, \dots, N \quad (2.14)$$

$$\alpha_i, \mu_i \geq 0 \quad \forall i = 1, \dots, N \quad (2.15)$$

$$\alpha_i (y_i(\phi(\mathbf{x}_i)\mathbf{w} + b) - 1 + \xi_i) = 0 \quad \forall i = 1, \dots, N \quad (2.16)$$

$$\mu_i \xi_i = 0 \quad \forall i = 1, \dots, N . \quad (2.17)$$

A partir de estas condiciones se deduce que el hiperplano de separación en el espacio de características se construye a partir de las muestras: $\mathbf{w} = \sum_i \alpha_i y_i \phi(\mathbf{x}_i)$. Sustituyendo dichas condiciones se llega al problema *dual*:

$$\max_{\{\alpha_i\}} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j), \quad (2.18)$$

en cuya solución es necesario que:

$$0 \leq \alpha_i \leq C \quad \forall i = 1, \dots, N \quad (2.19)$$

$$\sum_i \alpha_i y_i = 0, \quad (2.20)$$

donde $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ es la función núcleo (o *kernel*) que permite obtener el producto escalar de dos puntos en el espacio de características como una función no lineal en el espacio de entrada. El teorema de Mercer establece condiciones necesarias y suficientes para que una función $k(\mathbf{x}_i, \mathbf{x}_j)$ represente un producto escalar en un espacio de Hilbert (Vapnik, 1998).

Cuando tenemos más de dos clases se pueden emplear esquemas uno contra uno, o uno contra todos en los que la clasificación multiclase se apoya en la clasificación binaria (Schölkopf y Smola, 2001). Otra posibilidad es formular directamente el problema multiclase (Weston y Watkins, 1999; Crammer y Singer, 2002).

La formulación de la SVM puede generalizarse para resolver problemas de aprendizaje estructurado donde las etiquetas han de capturar la estructura de las muestras (Tsochantaridis *et al.*, 2005; Finley y Joachims, 2008; Sarawagi y Gupta, 2008). Esta extensión permite aplicar algoritmos tipo SVM a problemas complejos como el procesamiento del lenguaje natural (BakIr *et al.*, 2007a).

2.4.2. Reducción de complejidad

La SVM construye la frontera de decisión con un subconjunto disperso de muestras de entrenamiento denominadas vectores soporte:

$$f(\mathbf{x}) = \text{signo} \left(\sum_i \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b \right).$$

Sin embargo, todas las muestras mal clasificadas, en el sentido de incumplir su restricción sobre el margen, son vectores soporte e intervienen en la solución proporcionada por la SVM. La frontera de decisión así construida, puede expresarse de otras formas, a veces con menos muestras. En este sentido, Vapnik (1998) muestra

que el valor esperado del número N_v de vectores soporte está acotado por debajo por $(N - 1)B_K$, donde N es el número de muestras y B_K es el mínimo error de clasificación alcanzable con el *kernel* K . En caso de emplear un *kernel* universal, como el basado en funciones de base radial, este error es el error de Bayes. Recientemente, Steinwart (2004) indica que el número N_v de vectores soporte se incrementa linealmente con el número de muestras N . Específicamente:

$$N_v/N \longrightarrow 2B_K .$$

Por otro lado, el tiempo de cómputo necesario para clasificar una muestra mediante la SVM está determinado por el número de vectores soporte. Así, en algunas aplicaciones se simplifica la solución proporcionada por la SVM para reducir dicho tiempo.

Burges (1996) construye un conjunto reducido de nuevos vectores soporte $\{\mathbf{z}_\ell\}$, los cuales ni están necesariamente en la frontera ni tienen por qué ser muestras del conjunto de entrenamiento. Para ello, resuelve un problema de optimización, en general, no convexo. Si el punto del espacio de características que define la frontera de clasificación de la SVM es $\Psi = \sum_{i=1}^{N_s} \alpha_i y_i \phi(\mathbf{x}_i)$, Burges (1996) propone usar la frontera definida por $\Psi' = \sum_{\ell=1}^{N_z} \gamma_\ell \phi(\mathbf{z}_\ell)$ donde $N_z < N_s$ se define *a priori* y $\{\gamma_\ell, \mathbf{z}_\ell\}$ se obtienen de minimizar: $\|\Psi - \Psi'\|$.

La preimagen de un punto Ψ del espacio de características se define como la muestra \mathbf{z} del espacio de entrada que se proyecta en Ψ mediante $\Phi(\cdot)$

$$\mathbf{z} \rightarrow \Phi(\mathbf{z}) = \Psi .$$

En la práctica, no existe la preimagen en muchas ocasiones y suele emplearse su mejor aproximación. Schölkopf *et al.* (1998) proponen construir dicho conjunto reducido usando el método de la preimagen. Así, en el primer paso, obtienen el vector \mathbf{z}_1 como el que mejor aproxima Ψ y sucesivamente en el paso n el vector \mathbf{z}_n como el que mejor aproxima:

$$\Psi - \sum_{\ell=1}^{n-1} \gamma_\ell \mathbf{z}_\ell .$$

Así, proponen un algoritmo de punto fijo para encontrar iterativamente los nuevos vectores soporte como las preimágenes de la frontera original menos la aproximación actual de la misma:

$$\mathbf{z}_n = \frac{\sum_{i=1}^{N_s} \alpha_i y_i k(\|\mathbf{x}_i - \mathbf{z}_n\|^2) \mathbf{x}_i - \sum_{\ell=1}^{n-1} \gamma_\ell k(\|\mathbf{z}_\ell - \mathbf{z}_n\|^2) \mathbf{z}_\ell}{\sum_{i=1}^{N_s} \alpha_i y_i k(\|\mathbf{x}_i - \mathbf{z}_n\|^2) - \sum_{\ell=1}^{n-1} \gamma_\ell k(\|\mathbf{z}_\ell - \mathbf{z}_n\|^2)} .$$

Este algoritmo sólo es válido para *kernels* que verifiquen $k(\mathbf{x}, \mathbf{x}) = 1$, como, por ejemplo, el *kernel* gaussiano; y, aunque es más eficiente que la aproximación anterior, todavía sufre problemas numéricos por lo que necesita reiniciarse varias veces y escoger la mejor solución. Por otro lado, también presentan cómo obtener de forma óptima, en el sentido de error cuadrático, los coeficientes de la expansión $\boldsymbol{\gamma} = \{\gamma_\ell\}$; $\ell = 1, \dots, N_z$ que mejor aproxima la frontera original:

$$\boldsymbol{\gamma} = (\mathbf{K}^z)^{-1} \mathbf{K}^{zx} \boldsymbol{\alpha}, \quad (2.21)$$

donde $\mathbf{K}_{nm}^z = k(\mathbf{z}_n, \mathbf{z}_m)$; $n, m = 1, \dots, N_z$ es la matriz de *kernel* de los vectores $\{\mathbf{z}_\ell\}$; $\ell = 1, \dots, N_z$; $\mathbf{K}_{nm}^{zx} = k(\mathbf{z}_n, \mathbf{x}_m)$; $n = 1, \dots, N_z$ $m = 1, \dots, N_s$ y $\boldsymbol{\alpha} = \{\alpha_i\}$; $i = 1, \dots, N_s$.

Downs *et al.* (2001) proponen eliminar de la función que construye la frontera de decisión los puntos cuyas imágenes son combinación lineal de otras, manteniendo la misma frontera de decisión pero expresada usando menos vectores soporte (o aproximadamente la misma, si consideramos cero los autovalores pequeños de la matriz formada por los *kernels* de los vectores soporte). Schölkopf y Smola (2001, cap. 18) examinan el problema de la reducción de la complejidad y proponen, entre otros, un método de programación cuadrática usando la norma uno de los pesos para escoger un subconjunto de vectores soporte disperso para construir la frontera:

$$\left\| \sum_{i=1}^{N_s} \alpha_i y_i \Phi(\mathbf{x}_i) - \sum_{i=1}^{N_s} \gamma_i \Phi(\mathbf{x}_i) \right\|^2 - \lambda \sum_{i=1}^{N_s} c_i |\gamma_i|,$$

donde el valor de c_i no es importante y puede fijarse a 1 o a $\frac{\sum_i |\alpha_i|}{N|\alpha_i|}$ y $|\gamma_i|$ se reescribe como $\gamma_i = \gamma_i^+ - \gamma_i^-$ y la restricción $\gamma_i^+, \gamma_i^- \geq 0$. Una vez escogidos los vectores soporte con γ_i no nulos, la expansión óptima se obtiene de (2.21).

Kwok y Tsang (2004) proponen un método para obtener la preimagen que relaciona las distancias entre dos puntos en el espacio de características con las distancias entre sus preimágenes. Así, a partir de las distancias entre las imágenes de los puntos de entrenamiento y el punto del espacio de características que define la frontera de decisión, se consigue un conjunto de restricciones para aproximar la preimagen de dicha frontera a partir de los puntos de entrada. Este método tiene como ventaja la ausencia de problemas numéricos y es válido para *kernels* en los que se pueda relacionar las distancias entre el espacio de características y el espacio de entrada. Aun así, en los casos en los que el método propuesto por Schölkopf *et al.* (1998) puede emplearse, éste obtiene una mejor solución siendo inicializado con la preimagen del método de Kwok y Tsang (2004), como ha sido propuesto por Kim *et al.* (2005).

BakIr *et al.* (2004, 2007b) presentan la obtención de preimágenes como un problema de aprendizaje. Una vez obtenido el mapa que pasa del espacio de características al espacio de entrada la obtención de preimágenes es muy eficiente. Este método evita mínimos locales y otros problemas numéricos y es aplicable a espacios discretos. Sin embargo, este método es menos eficiente que otros de los citados métodos para construir conjuntos reducidos de vectores soporte cuando éstos son aplicables.

Nguyen y Ho (2005) proponen un método iterativo para reducir el número de vectores soporte reemplazando dos vectores cercanos de la misma clase por otro nuevo. Zheng y Lai (2006) proponen un método similar a Kwok y Tsang (2004) pero regularizando la forma de preservar la localidad. Zheng *et al.* (2006) emplean la información de la clase, en un sentido genérico de la palabra que depende de la aplicación, para construir no sólo preimágenes aproximadas sino también apropiadas a la aplicación. Li *et al.* (2007) proponen un método similar al propuesto por Downs *et al.* (2001) para construir una base que aproxime el espacio generado por las imágenes de los vectores soporte en el espacio de características.

Como hemos dicho, todas las muestras mal clasificadas intervienen en la construcción de la frontera de clasificación. Otra familia de métodos se centran en modificar el conjunto de entrenamiento para convertir el problema en otro, posiblemente separable, que dé lugar a la misma frontera pero construida por la SVM de modo más sencillo (una frontera que proporcione aproximadamente las mismas prestaciones). BakIr *et al.* (2005) propone un método que emplea técnicas de simplificación de clasificadores de k -vecinos más próximos para transformar el problema a resolver en otro en el que las distribuciones sean separables, sin alterar significativamente la frontera de clasificación obtenida por la SVM original. Si el problema es separable el número de vectores soporte ya no depende del número de muestras y los tiempos de entrenamiento y de clasificación disminuyen drásticamente. Zhan y Shen (2005) proponen un entrenamiento dividido en cuatro fases: en la primera, entrenan una SVM; en la segunda, excluyen del conjunto de entrenamiento los vectores soporte que más curvan la frontera de decisión; en la tercera, entrenan una SVM con las muestras restantes; y, finalmente emplean el método de Osuna y Girosi (1998) que consiste en aplicar una SVM para regresión a las muestras de entrenamiento y a su salida permitiendo expresar la frontera de clasificación con menos vectores y controlar el error cometido. Osuna y Girosi (1998) también proponen resolver el problema de la SVM en el primal (2.10) donde ninguna muestra va a formar parte de la construcción de la frontera debido a que su multiplicador de Lagrange esté limitado. Los autores hacen notar que ninguno de los dos métodos permite grandes mejoras si la mayoría de los multiplicadores de Lagrange de la solución original no

están saturados, esto es, si $\alpha_i < C$ en (2.19).

Otros métodos construyen iterativamente la frontera de clasificación sin necesidad de empezar por resolver primero la SVM. Los puntos elegidos para construir la frontera no tienen por qué tener ninguna relación con la frontera ni con el margen, y ni siquiera tienen por qué formar parte del conjunto de entrenamiento. Parrado-Hernández *et al.* (2003) proponen un algoritmo que va refinando la frontera de clasificación centrándose en las muestras mal clasificadas cerca de la frontera inspirados en el algoritmo *Boosting* (Freund y Schapire, 1997) que se concentra en las muestras cercanas al margen. La complejidad de este clasificador puede fijarse *a priori* o decidirse en función de las prestaciones requeridas. También proponen un algoritmo que combina distintos *kernels* gaussianos: inicialmente, construye la frontera con *kernels* de un ancho grande y, sucesivamente, va refinando la frontera con nuevas muestras que emplean un menor ancho de *kernel* mientras disminuya el error. Keerthi *et al.* (2006) presentan una aproximación similar formulando el problema en el primal usando un término cuadrático para los errores. La ventaja de este método es que la elección iterativa de las muestras está directamente relacionada con la función de coste.

Fuera de la formulación de la SVM existen otras aproximaciones que construyen la frontera de clasificación con un número reducido de muestras. Entre ellas tenemos métodos bayesianos como la *Relevance Vector Machine* (Tipping, 2001), la *Informative Vector Machine* (Lawrence *et al.*, 2003); o, la *Kernel Matching Pursuit* (Vincent y Bengio, 2002), un método discriminativo para la función de coste cuadrático.

2.4.2.1. Preimágenes para *kernel* polinómico

Si tenemos un *kernel* polinómico del tipo $k(\mathbf{x}, \mathbf{y}) = (\gamma \langle \mathbf{x}, \mathbf{y} \rangle + c)^d$ no podemos emplear el algoritmo de Schölkopf *et al.* (1998) para obtener las preimágenes. El algoritmo de Kwok y Tsang (2004) construye distancias indirectamente a partir del producto escalar y necesita obtener los vecinos más próximos, lo que resulta pesado computacionalmente. Un método más directo, inspirado en la construcción exacta de preimágenes cuando éstas existen (Schölkopf y Smola, 2001), es el siguiente: Supongamos que \mathbf{z} es la preimagen de $\Psi = \sum \alpha_i \Phi(\mathbf{x}_i)$. El *kernel* entre cualquier punto del espacio de entrada \mathbf{y} y \mathbf{z} se calcula como:

$$k(\mathbf{y}, \mathbf{z}) = \sum \alpha_i k(\mathbf{y}, \mathbf{x}_i) .$$

Para un *kernel* polinómico de grado impar puede obtenerse el producto escalar entre dos puntos en el espacio de entrada a partir de su *kernel* en el espacio de

características $\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{\gamma} (\sqrt[d]{k(\mathbf{x}, \mathbf{y})} - c)$. Una aproximación de la preimagen de Ψ se construye como sigue:

1. Sea \mathbf{Q} una matriz cuyas columnas forman una base en el espacio de entrada.
2. Obtener la matriz de *kernels* $\mathbf{K}_{\mathbf{Qz}}$ entre los elementos de la base \mathbf{q}_i y \mathbf{z} la preimagen de Ψ .
3. Obtener el producto escalar en el espacio de entrada $\mathbf{p}_i = \langle \mathbf{q}_i, \mathbf{z} \rangle$.
4. Una aproximación para \mathbf{z} puede obtenerse de resolver $\mathbf{Q}^T \mathbf{z} = \mathbf{p}$ o, como $\mathbf{z} = \sum \frac{\langle \mathbf{q}_i, \mathbf{z} \rangle}{\|\mathbf{q}_i\|} \mathbf{q}_i$ si las columnas de \mathbf{Q} forman una base ortogonal.

\mathbf{Q} puede ser cualquier base del espacio de entrada. También pueden emplearse las muestras de entrenamiento o cualquier conjunto de muestras, y resolver el sistema sobre-determinado.

Este método, aunque limitado a *kernels* de los que se pueda extraer el producto escalar en el espacio de entrada, es más rápido que el propuesto por Kwok y Tsang (2004) y no tiene parámetros que ajustar. Simplemente ha de elegirse la base o conjunto de muestras a emplear.

2.4.2.2. Clasificadores en cascada

En ocasiones, el tiempo de clasificación ha de ser menor que el requerido por el clasificador proporcionado por una algoritmia dada, la SVM en nuestro caso. Una posibilidad para acelerar el rendimiento del sistema son los clasificadores en cascada. Un clasificador en cascada está formado por varias etapas que rechazan las muestras de un tipo. Usualmente se rechaza la clase más probable o la contraria a la que desea detectar (véase Figura 2.1). Entre los trabajos publicados podemos citar a Viola y Jones (2001) que, en el contexto de la detección de caras, combinan AdaBoost (Freund y Schapire, 1997) para seleccionar las características visuales con una detección en cascada que descarta rápidamente el fondo. En este caso, los clasificadores de las etapas intermedias se ajustan para proporcionar una sensibilidad cercana a 1. De este modo, se consigue una sensibilidad para la cascada cercana a la obtenida por la máquina en bloque disminuyendo el tiempo requerido para clasificar una muestra.

Romdhani *et al.* (2001, 2004) consideran un esquema en cascada para la detección de caras con SVM después de reducir el número de vectores soporte.

LeCun *et al.* (1998) consideran varios métodos, entre los que están las redes neuronales convolucionales, para la detección de dígitos y caras.

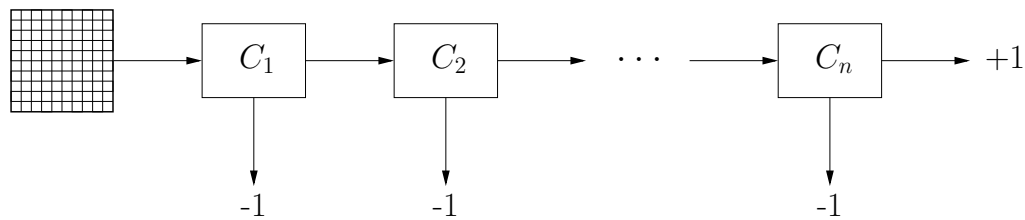


Figura 2.1: Ejemplo de clasificador en cascada diseñado para rechazar rápidamente la clase -1 . Sólo los parches que superan los clasificadores de las etapas intermedias pueden ser clasificados como $+1$.

2.5. Objetivos

Los objetivos de esta tesis son:

- Obtener test de hipótesis secuenciales que consideren la incertidumbre en las hipótesis, permitan saber cuáles son las máximas prestaciones posibles y proporcionen medidas de prestaciones. Estos test pueden ser conservadores en un sentido frecuentista o seguir la metodología bayesiana cuando haya información *a priori* disponible y/o no se desee una decisión conservadora.
- Obtener un método basado en aprendizaje estadístico equivalente a los cocientes de verosimilitud para clasificar un conjunto de muestras de test de la misma clase cuando las distribuciones condicionadas a las clases son desconocidas. Desarrollamos métodos para clasificar de manera conjunta un conjunto de muestras que pertenecen o bien a una hipótesis o bien a la otra.
- Aplicar el contraste de hipótesis inciertas al diagnóstico de tuberculosis implementando el sistema propuesto en el Capítulo 1.

Capítulo 3

Incertidumbre en las hipótesis

A la hora de realizar contrastes de hipótesis, hemos supuesto hasta el momento que los datos han sido generados por el modelo $f(\mathbf{x}|\theta)$ conocido salvo por algunos parámetros que dependen del estado θ que consideramos desconocido. Sin embargo, en algunas ocasiones no se conoce ni siquiera la familia según la cual se distribuyen los datos. Otras veces, dicha familia es conocida, pero la región de θ que define cada hipótesis es desconocida y difícil de caracterizar a partir de una muestra finita.

Esta es la situación que analizamos en este capítulo y constituye una de las aportaciones de esta tesis. Estamos interesados en el problema de la discriminación entre dos hipótesis caracterizadas por muestras a partir del examen de una muestra de test que viene de una de las dos distribuciones que caracterizan las hipótesis. Concretamente, se nos presentan tres conjuntos de muestras iid: $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$ que caracteriza la hipótesis H_0 , esto es, ha sido tomada cuando la hipótesis H_0 es cierta; $\mathbf{y} = \{y_1, y_2, \dots, y_M\}$ que caracteriza la hipótesis H_1 y la muestra de test $\mathbf{z} = \{z_1, z_2, \dots, z_K\}$ que proviene o bien de H_0 o de H_1 . Con esta información, ha de decidirse de qué hipótesis proviene \mathbf{z} y medir la calidad de esta decisión.

La metodología no paramétrica para el contraste de hipótesis examina un problema relacionado: el problema de dos muestras. En el capítulo anterior hemos considerado algunas de las aproximaciones a este problema, siendo quizá la más interesante para el problema de este capítulo la aproximación de Kolmogorov-Smirnov, porque compara las funciones de distribución a las que siempre se converge cuando el número de muestras que caracterizan las hipótesis tiende a infinito. Esta aproximación compara \mathbf{z} con \mathbf{x} y con \mathbf{y} y elige la hipótesis que proporciona menor valor del estadístico.

La aproximación paramétrica frecuentista considera también problemas relacionados. Si la familia de distribuciones que define la función de densidad de probabilidad es conocida, los parámetros que caracterizan cada hipótesis pueden estimarse

mediante máxima verosimilitud, y realizar un contraste de hipótesis simple. El inconveniente de esta postura es que sus resultados pueden ser muy optimistas porque no considera la incertidumbre en la caracterización de las hipótesis.

Lehmann (1997, cap. 9) analiza el problema de la incertidumbre en las hipótesis desde un punto de vista minimax. Divide el espacio de estados en tres regiones: Θ_0 , donde se penaliza decidir H_1 como cierta; Θ_1 , donde se penaliza decidir H_0 ; y, Θ_I , donde puede decidirse o bien H_0 o H_1 . A continuación busca el test que maximiza la potencia mínima entre los tests de nivel crítico α . En caso de distribuciones del tipo $P_i = (1 - \epsilon_i)Q_i + \epsilon_i G_i$, $i = 0, 1$ donde $\epsilon_i \in (0, 1)$ están dados y G_i son distribuciones desconocidas. La hipótesis que ha originado los datos puede decidirse por medio de un test de cociente de verosimilitud que censura el cociente de unas fdps construidas a partir de las fdps de Q_i (Huber y Strassen, 1973; Lehmann, 1997). En este problema hay dos fuentes de incertidumbre, por un lado, el estado que es una región, y por otro, las distribuciones, que están contaminadas con unas distribuciones desconocidas. Esta aproximación considera el problema del desconocimiento parcial de la familia de distribuciones mientras que nosotros, por otro lado, consideramos la región del estado que define las hipótesis desconocida salvo por \mathbf{x} e \mathbf{y} .

Nuestra aportación a la aproximación frecuentista transforma hipótesis simples desconocidas en hipótesis compuestas a las que asignamos la probabilidad de que el estado esté efectivamente contenido en la región que define cada hipótesis. Esta transformación se lleva a cabo por medio de la determinación de regiones de confianza para el valor del estado que define cada hipótesis.

Desde el punto de vista bayesiano modelamos este problema del siguiente modo: en primer lugar, modelamos nuestras creencias mediante una distribución *a priori* sobre el valor del estado bajo cada hipótesis; a continuación ese conocimiento se actualiza mediante las muestras \mathbf{x} e \mathbf{y} para formar las distribuciones *a posteriori* sobre el valor del estado bajo cada hipótesis. A partir de estas últimas se realizan las inferencias.

El resto del capítulo presenta el tratamiento de la incertidumbre en las hipótesis para variables aleatorias binarias, discretas no binarias y continuas desde los puntos de vista bayesiano y frecuentista. En principio asumiremos la familia de las distribuciones conocida y después sugerimos una posible extensión cuando la familia es desconocida. Además, a modo de ejemplo desarrollamos un test secuencial para este tipo de hipótesis y obtenemos cotas de sus prestaciones en función de las muestras \mathbf{x} e \mathbf{y} .

3.1. Variables aleatorias discretas binarias y no binarias

Introducimos el problema con variables aleatorias binarias, que nos permiten presentar uno de los casos más sencillos de incertidumbre. Esto es, el conjunto \mathbf{z} con $z_i \in \{0, 1\}$, ha sido tomado de una distribución de Bernoulli con probabilidad de éxito p^* o q^* . Debemos escoger entre las siguientes hipótesis:

H_0 : \mathbf{z} ha sido generado con una distribución Bernoulli(p^*).

H_1 : \mathbf{z} ha sido generado con una distribución Bernoulli(q^*).

Las probabilidades p^* y q^* son desconocidas y la única información sobre ellas está en los conjuntos \mathbf{x} e \mathbf{y} .

3.1.1. Enfoque de máxima verosimilitud

Una posibilidad para resolver este problema es obtener la estima de máxima verosimilitud de p^* y q^* a partir de las observaciones y aplicar la teoría para el contraste de hipótesis simples. El estimador de máxima verosimilitud para p^* es:

$$\hat{p}^* = n_1/N .$$

Este estimador es insesgado y su varianza es $\frac{1}{N}p^*(1 - p^*)$. La estima es mejor, en un sentido de error cuadrático medio, cuanto p^* sea más cercana a cero o a uno. Sin embargo, si vamos a emplear esta estima dentro de un cociente de verosimilitud, la sensibilidad del mismo aumenta cuando p^* está cerca de cero o uno. Errores en la estima cerca de cero o uno hacen que el cociente de verosimilitud crezca o decrezca mucho más rápido que un error de la misma magnitud en probabilidades cercanas a 0.5. Las fdps con parámetros estimados por máxima verosimilitud en cocientes de verosimilitud pueden producir sesgos si M y N son pequeños o p^* y q^* toman valores extremos, es decir, cuando no estamos cerca de la asunción de gaussianidad. Además, emplear fdps cuyos parámetros son estimas de máxima verosimilitud como si fuesen las fdps verdaderas conduce a tests que pueden errar en la hipótesis escogida, a la vez que proporcionan una alta confianza en su decisión. Esto se debe al crecimiento/decrecimiento exponencial del cociente de verosimilitud, y al decrecimiento exponencial del error en el test de cociente de verosimilitud para hipótesis simples (mostramos algunos ejemplos de este efecto en el Apartado 3.1.2.2 y la Figura 3.3).

Para el caso no binario el estimador de máxima verosimilitud es

$$\hat{\mathbf{p}}^* = [n_1/N, n_2/N, \dots, n_D/N],$$

donde D es el número de categorías y $n_j = \sum_{i=1}^N \delta_{x_i}^j$ $j = 1, 2, \dots, D$ y las deltas de Kronecker $\delta_{x_i}^j = 1$ si $x_i = j$ y cero en otro caso.

3.1.2. Enfoque bayesiano

Nuevamente, volvemos al caso binario. El enfoque bayesiano parte de las probabilidades *a priori* de las hipótesis y obtiene sus probabilidades *a posteriori* usando la regla de Bayes:

$$P(H_i | \mathbf{z}) = \frac{P(\mathbf{z} | H_i)P(H_i)}{P(\mathbf{z})} = \frac{P(\mathbf{z} | H_i)P(H_i)}{P(\mathbf{z} | H_0)P(H_0) + P(\mathbf{z} | H_1)P(H_1)}. \quad (3.1)$$

Aunque p^* , la probabilidad de acierto de la distribución de Bernoulli bajo la hipótesis H_0 , tiene un valor fijo, podemos atribuirle una distribución *a priori* basada en nuestra percepción de cuáles son sus posibles valores, y que da cuenta de nuestra incertidumbre acerca de su valor. Por ello, la denotaremos como p para enfatizar que la consideramos una variable aleatoria. La distribución *a posteriori* de p , que engloba todo nuestro conocimiento sobre p , se obtiene a partir de la combinación del conocimiento *a priori* y los datos \mathbf{x} .

$$P(p | \mathbf{x}) = \frac{P(\mathbf{x} | p)P(p)}{P(\mathbf{x})} = \frac{P(\mathbf{x} | p)P(p)}{\int P(\mathbf{x} | p)P(p) dp}, \quad (3.2)$$

donde la verosimilitud de p es:

$$P(\mathbf{x} | p) = \prod_{i=1}^N P(x_i | p) = p^{n_1}(1-p)^{N-n_1},$$

y $n_1 = \sum_{i=1}^N x_i$. De este modo, la verosimilitud de H_0 se obtiene integrando p como:

$$P(\mathbf{z} | H_0) = P(\mathbf{z} | \mathbf{x}) = \int P(\mathbf{z} | p)P(p | \mathbf{x}) dp, \quad (3.3)$$

donde

$$P(\mathbf{z} | p) = \prod_{k=1}^K P(z_k | p) = p^{k_1}(1-p)^{K-k_1}$$

y $k_1 = \sum_{k=1}^K z_k$. Escogemos como distribución *a priori* para p una distribución beta:

$$P(p) = \text{beta}(p \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1} \quad (3.4)$$

que es una distribución flexible y permite establecer una distribución *a priori* no informativa sobre p sin más que establecer el valor de los parámetros $a = b = 1^1$, para los que esta distribución es una uniforme entre cero y uno. La distribución beta es también conjugada de las verosimilitudes de una distribución de Bernoulli y de una distribución binomial, lo que justifica su elección como distribución *a priori* por simplificar considerablemente los cálculos. La verosimilitud de H_0 resulta:

$$P(\mathbf{z} \mid H_0) = \frac{\Gamma(N+a+b) \Gamma(K-k_1+N-n_1+b) \Gamma(k_1+n_1+a)}{\Gamma(N-n_1+b) \Gamma(n_1+a) \Gamma(a+K+N+b)}. \quad (3.5)$$

Análogamente, si definimos una distribución *a priori* beta para q con parámetros a' y b' , y $m_1 = \sum_{i=1}^M y_i$, la verosimilitud de H_1 resulta:

$$P(\mathbf{z} \mid H_1) = \frac{\Gamma(M+a'+b') \Gamma(K-k_1+M-m_1+b') \Gamma(k_1+m_1+a')}{\Gamma(M-m_1+b') \Gamma(m_1+a') \Gamma(a'+K+M+b')}. \quad (3.6)$$

3.1.2.1. Distribuciones no binarias.

La distribución categórica es la generalización de la distribución de Bernoulli a $D > 2$ categorías. Se define por un vector de probabilidades $\mathbf{p} = [p_1, p_2, \dots, p_D]$, donde p_ℓ , $\ell = 1, \dots, D$ es la probabilidad de que una muestra tome el valor ℓ . El resultado para el caso binario se generaliza fácilmente para distribuciones categóricas. Ahora la muestra de test es un conjunto de muestras iid \mathbf{z} con $z_i \in \{1, \dots, D\}$, con una distribución categórica definida por el vector $\mathbf{p}^* = [p_1, p_2, \dots, p_D]$ o por $\mathbf{q}^* = [q_1, q_2, \dots, q_D]$. Debe decidirse una de las siguientes hipótesis:

H_0 : \mathbf{z} ha sido generado con una distribución categórica de parámetro \mathbf{p}^* .

H_1 : \mathbf{z} ha sido generado con una distribución categórica de parámetro \mathbf{q}^* .

Como en el caso anterior, \mathbf{p}^* y \mathbf{q}^* son desconocidas y sólo tenemos los conjuntos de muestras \mathbf{x} y \mathbf{y} para caracterizarlas. Siguiendo el análisis previo, tomamos como distribución *a priori* para \mathbf{p}^* $P(\mathbf{p} \mid a_1, a_2, \dots, a_D)$ una distribución Dirichlet, que es conjugada a verosimilitudes categóricas y multinomiales, y obtenemos la verosimili-

¹Otra posibilidad ampliamente usada para distribuciones no informativas es la distribución de Jeffreys que consiste en $a=b=1/2$ (véase Apartado 2.2.2.1).

tud de H_0 como:

$$P(\mathbf{z} | H_0) = P(\mathbf{z} | \mathbf{x}) = \frac{\Gamma(N + \sum_{\ell=1}^D a_{\ell}) \prod_{\ell=1}^D \Gamma(k_{\ell} + a_{\ell} + n_{\ell})}{\Gamma(K + N + \sum_{\ell=1}^D a_{\ell}) \left(\prod_{\ell=1}^D \Gamma(a_{\ell} + n_{\ell}) \right)}, \quad (3.7)$$

donde $n_j = \sum_{i=1}^N \delta_{x_i}^j$, $k_j = \sum_{i=1}^K \delta_{z_i}^j$, $j = 1, 2, \dots, D$. De igual modo se llega a la expresión de $P(\mathbf{z} | H_1)$. La extensión a N_H hipótesis es directa sin más que obtener la verosimilitud de cada hipótesis y cambiar la Ecuación (3.1) por

$$P(H_i | \mathbf{z}) = \frac{P(\mathbf{z} | H_i)P(H_i)}{P(\mathbf{z})} = \frac{P(\mathbf{z} | H_i)P(H_i)}{\sum_{k=1}^{N_H} P(\mathbf{z} | H_k)P(H_k)} \quad i = \{0, 1, \dots, N_H\}. \quad (3.8)$$

Johansson y Olofsson (2007) han propuesto recientemente un método bayesiano de selección de modelo para modelos ocultos de Markov, modelos de Markov, y modelos categóricos. Estos autores interpretan el problema de selección de modelo como un problema de dos muestras. Aplicando su método a ambas hipótesis y comparando el resultado por medio de un cociente se llega al mismo resultado de la Ecuación (3.8), al que nosotros hemos llegado partiendo desde un punto de vista completamente distinto. La aproximación natural para el contraste de hipótesis es la presentada en este capítulo, no la resolución de dos problemas de dos muestras.

Sin embargo, puede criticarse la aplicación sucesiva de un modelo de dos muestras desde un punto de vista metodológico, porque no establece, o no está obligado a establecer, *a priori*, cuáles son los modelos posibles. Los modelos posibles han de escogerse antes de examinar las muestras (Gutiérrez-Peña y Walker, 2005; Walker y Gutiérrez-Peña, 2007). Por otro lado, en nuestro desarrollo del contraste no nos limitamos a distribuciones *a priori* no-informativas, sino que empleamos distribuciones Dirichlet que son flexibles para modelar el conocimiento *a priori* disponible.

3.1.2.2. Experimentos

Analizamos las prestaciones de los métodos bayesianos para variables aleatorias binarias. En el caso binario, comparamos tres métodos: el método bayesiano que hemos presentado; el método de máxima verosimilitud (ML), que usa las probabilidades estimadas por máxima verosimilitud como las reales; y el método de Wald (véase Apartado 2.2.3). Completamos el experimento mostrando las prestaciones del método exacto que consiste en un cociente de las verosimilitudes de las fdps de las distribuciones que originaron los datos. Comparamos los métodos en cuatro situaciones variando el tamaño de las secuencias que caracterizan cada hipótesis \mathbf{x} e \mathbf{y} : $N = M = 1000$ (ambas hipótesis están caracterizadas con 1000 muestras); $N = 1000$

y $M = 100$ (la hipótesis H_0 está caracterizada con 1000 muestras y la H_1 con 100 muestras); $N = 100$ y $M = 1000$; y, $N = M = 100$. Así mostramos el comportamiento de los métodos en situaciones con distinto balance en el número de muestras de entrenamiento. Hemos seleccionado $p^* = 0.1$ para H_0 y $q^* = 0.15$ para H_1 . Vamos a emplear una distribución *a priori* no informativa tanto para p como para q . Esto se corresponde con que $P(p) = P(q)$ y ambas iguales a una distribución beta(1, 1).

El experimento consta de las siguientes partes: primero, generamos muestras de \mathbf{x} e \mathbf{y} para caracterizar las hipótesis; a continuación, examinamos las muestras de \mathbf{z} bajo cada hipótesis y obtenemos las probabilidades de las hipótesis por medio de la Ecuación (3.1); y, por último, la precisión en clasificación es obtenida como el porcentaje de probabilidades que superan el umbral 0.5.

El promedio de 10^4 repeticiones de este experimento variando el número de muestras de test se representa en las Figuras 3.1 y 3.2. El primer efecto que apreciamos es el balance de incertidumbre entre las muestras de entrenamiento y la muestra de test: al principio, cuando se usan pocas muestras de test, la muestra de test es la fuente principal de incertidumbre; más tarde, con el crecimiento del número de muestras de test, la muestra de test caracteriza la hipótesis de la que proviene y el error restante se debe a la incertidumbre en el modelado de las hipótesis. El método exacto, con conocimiento estadístico perfecto, obtiene una caída exponencial en el error.

La diferencia más importante entre los métodos aparece en los escenarios desbalanceados, favoreciendo al método bayesiano. El método bayesiano tolera mucho mejor que el método de máxima verosimilitud la falta de muestras porque tiene en cuenta la incertidumbre en la estima. Sea cual fuere la calidad de la estima, el método ML asume que las estimas son perfectas y pequeños errores en la estima de las probabilidades pueden dar lugar a sesgos importantes en el valor del cociente de verosimilitud.

En este mismo experimento podemos preguntarnos acerca de las probabilidades predichas. La Figura 3.3 muestra las probabilidades predichas por los métodos de máxima verosimilitud y bayesiano para 10 y 1000 muestras de test en el caso de $N = M = 100$ muestras de entrenamiento. Las Subfiguras 3.3(a) y 3.3(b) muestran respectivamente las predicciones para muestras de test provenientes de H_0 y H_1 . En estas figuras se han ordenado dichas predicciones de la mejor a la peor, donde se acierta con probabilidad mayor que 0.5. Podemos observar, como cabía de esperar, que el método que emplea las estimas de máxima verosimilitud es mucho más extremo en sus predicciones. El método bayesiano, al contrario, es más conservador, como observamos por ejemplo en la Figura 3.3(a) para 1000 muestras de test, donde

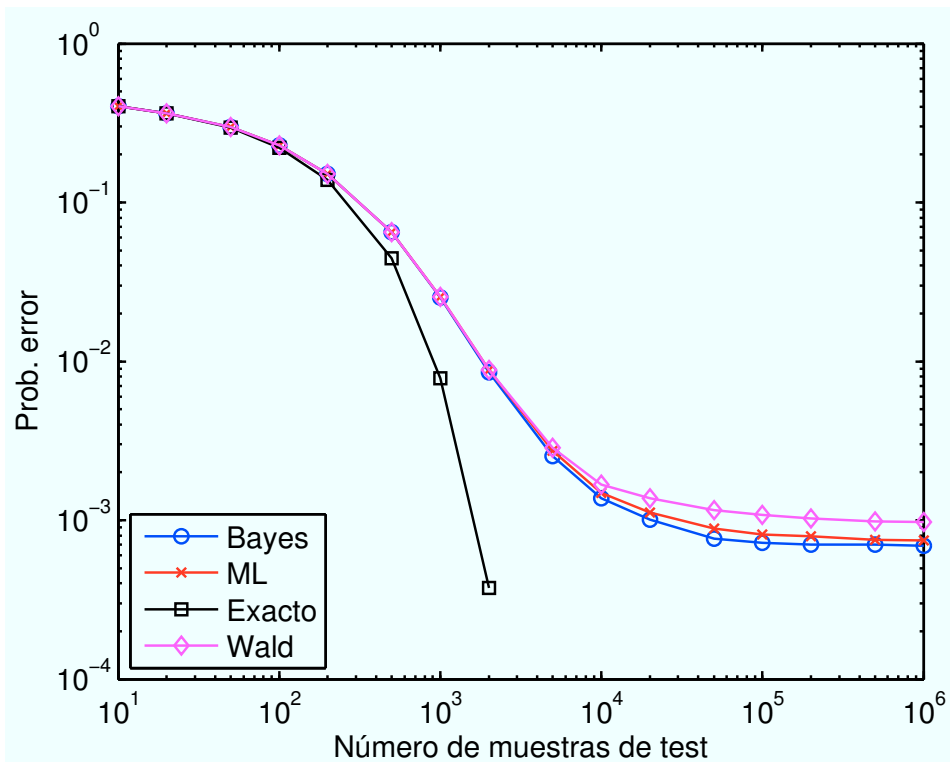
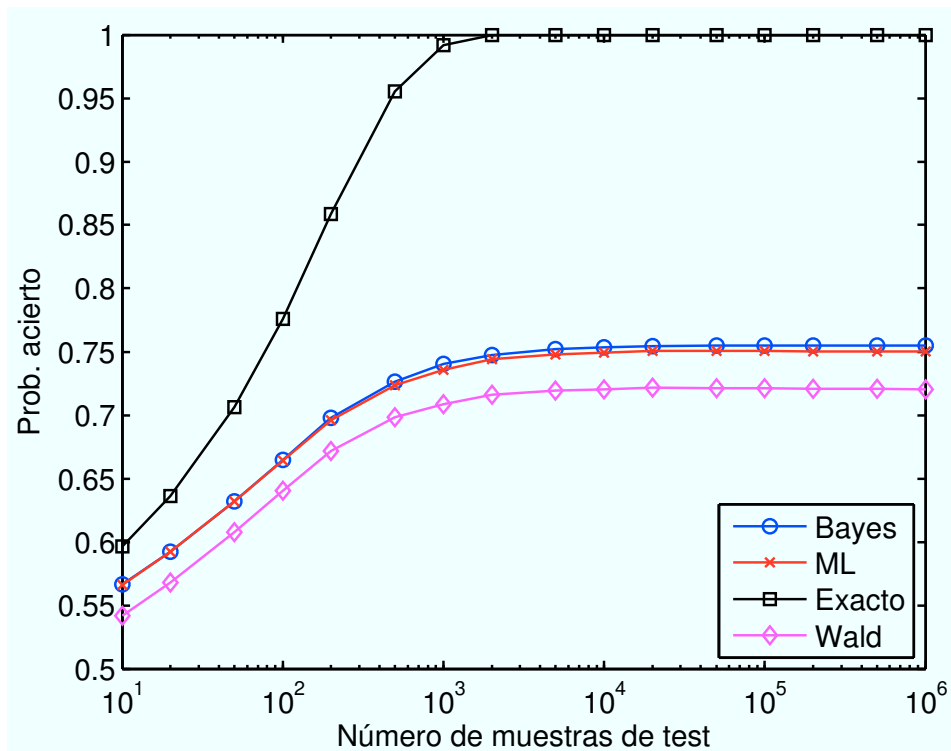
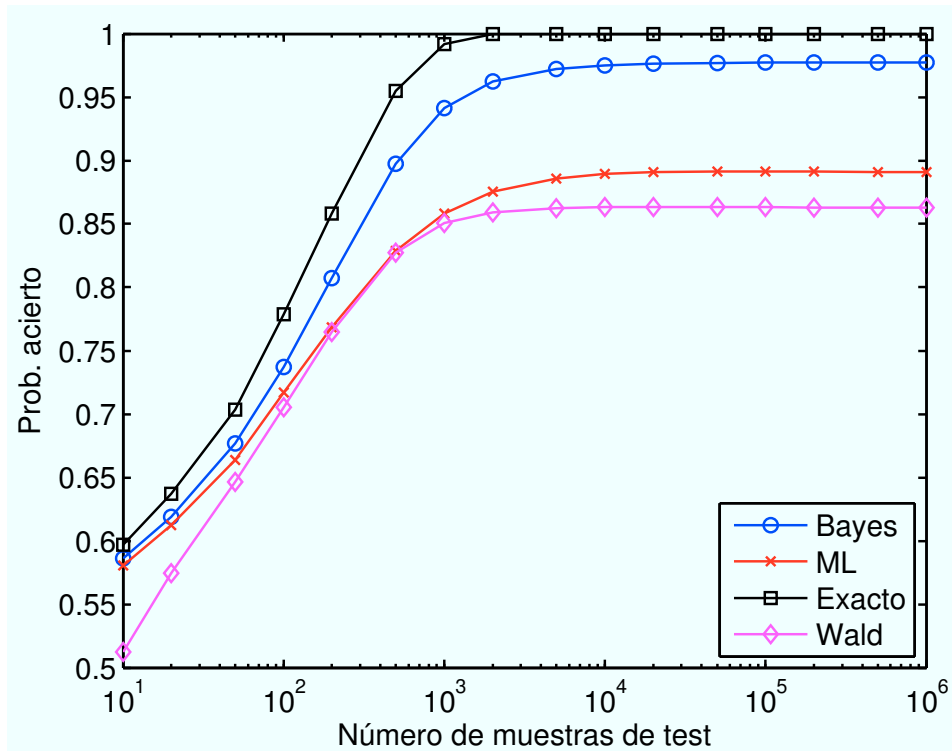
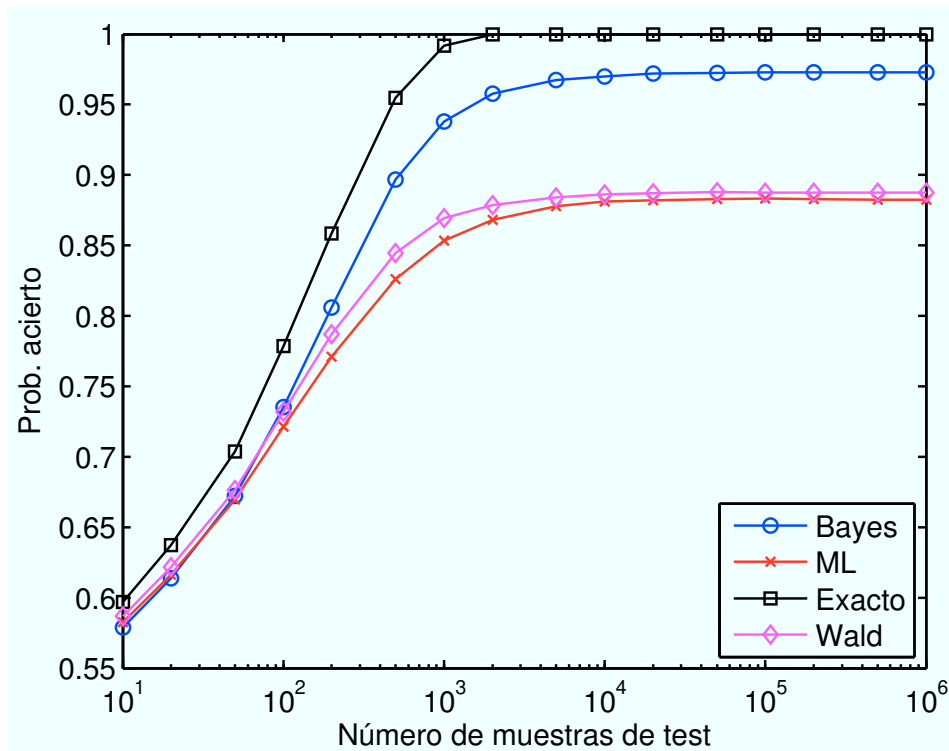
(a) $N = M = 1000$ (b) $N = M = 100$

Figura 3.1: Caso binario. Comparativa entre el método bayesiano y los métodos ML, Wald y exacto. Mostramos la precisión media (o el error (a)) obtenida para muestras de test generadas de las hipótesis H_0 y H_1 .



(a) $N = 1000, M = 100$



(b) $N = 100, M = 1000$

Figura 3.2: Caso binario. Comparativa entre el método bayesiano y los métodos ML, Wald y exacto. Mostramos la precisión media obtenida para muestras de test generadas de las hipótesis H_0 y H_1 .

el método de máxima verosimilitud pasa de acertar con probabilidades cercanas a uno a fallar con probabilidades cercanas a cero. El método basado en máxima verosimilitud ofrece mejores prestaciones que el método bayesiano para predecir las muestras de H_1 , mientras que el método bayesiano tiene mejores prestaciones para predecir H_0 y es globalmente mejor en ambas hipótesis como hemos visto en la Figura 3.1. Para 10 muestras, por otro lado, la fuente principal de incertidumbre son las muestras de test y aunque el método de máxima verosimilitud es ligeramente más extremo en sus predicciones ambos se comportan igual en lo que a clasificación se refiere.

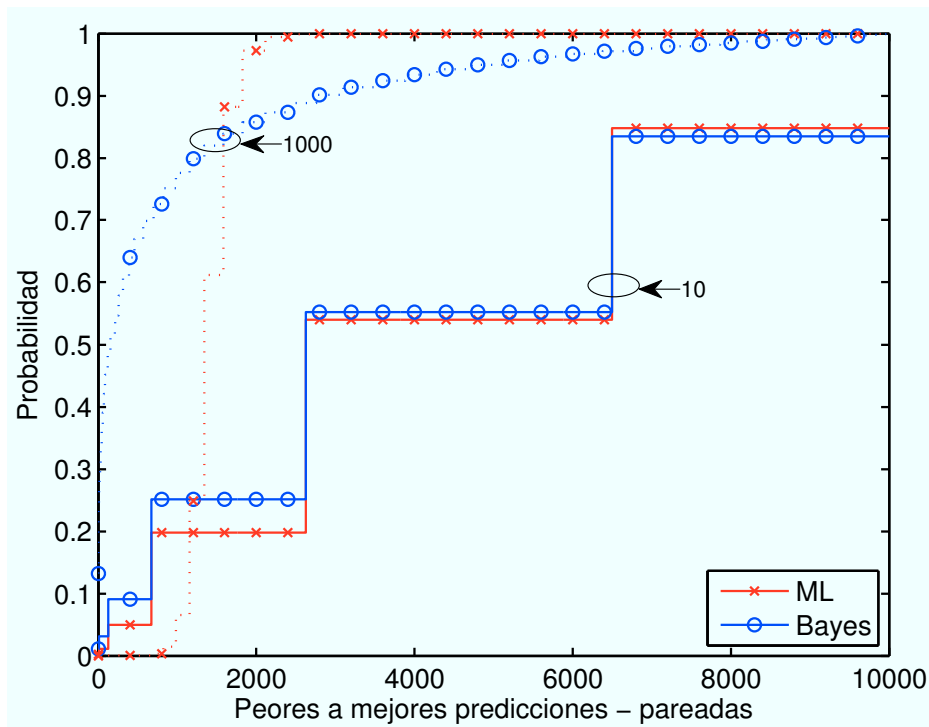
También se ha realizado un experimento con variables no binarias. En este caso, comparamos el método bayesiano con: el método de máxima verosimilitud, el método χ^2 de Pearson, el método *two-sample* de Johansson y Olofsson (2007) y el método exacto que usa las fdps de los datos. El experimento es similar al caso binario en el que hemos seleccionado $\mathbf{p}^* = [0.1, 0.4, 0.5]$ para H_0 y $\mathbf{q}^* = [0.15, 0.4, 0.45]$ para H_1 . El resultado se muestra en las Figuras 3.4 y 3.5. Como hemos comentado, los dos métodos bayesianos son equivalentes y extraemos las mismas conclusiones que en el caso binario.

3.1.2.3. Desarrollo secuencial

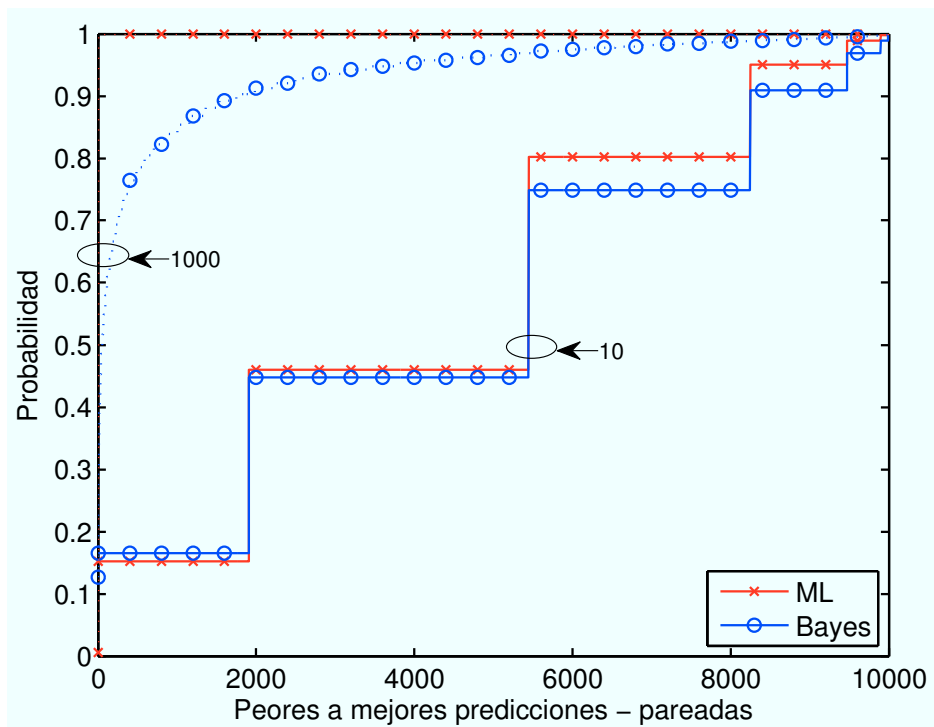
El test bayesiano puede emplearse dentro de un test secuencial cuya condición de parada sea que alguna hipótesis alcance una probabilidad mínima o se llegue a un límite de muestras. La expresión de la verosimilitud de H_0 dada por la Ecuación (3.5) puede escribirse de manera que pueda actualizarse para cada nueva muestra. Notamos que de los 6 términos en la expresión sólo tres dependen de la muestra de test.

$$\begin{aligned}
 & \blacksquare \Gamma(a + K + N + b) = (a + K - 1 + N + b)\Gamma(a + K - 1 + N + b). \\
 & \blacksquare \Gamma(K - k_{1,K} + N - n_1 + b) = \begin{cases} \Gamma(K - 1 - k_{1,K-1} + N - n_1 + b) & \text{si } z_K=1, \\ (K - 1 - k_{1,K-1} + N - n_1 + b) \\ \quad \times \Gamma(K - 1 - k_{1,K-1} + N - n_1 + b) & \text{si } z_K=0. \end{cases} \\
 & \blacksquare \Gamma(k_{1,K} + n_1 + a) = \begin{cases} (k_{1,K-1} + n_1 + a)\Gamma(k_{1,K-1} + n_1 + a) & \text{si } z_K=1, \\ \Gamma(k_{1,K-1} + n_1 + a) & \text{si } z_K=0. \end{cases}
 \end{aligned}$$

donde $k_{1,K} = \sum_{i=1}^K z_i$. Con lo cual, sólo tenemos que calcular la función Γ la primera iteración. A partir de ahí sólo hay que actualizar estos tres términos por el factor correspondiente a la vista de la siguiente muestra. Para evitar los problemas numéricos, es conveniente hacer los cálculos en escala logarítmica.



(a) H_0



(b) H_1

Figura 3.3: Caso binario. Probabilidades predichas por el método bayesiano propuesto y el método de máxima verosimilitud (ML) ordenadas de peor a mejor. (a) muestra las predicciones para H_0 y (b) para H_1 . Se muestran los resultados de 10^4 experimentos para 10 y 1000 muestras de test, representadas con diferente trazo. Se han empleado 100 muestras de entrenamiento para cada clase.

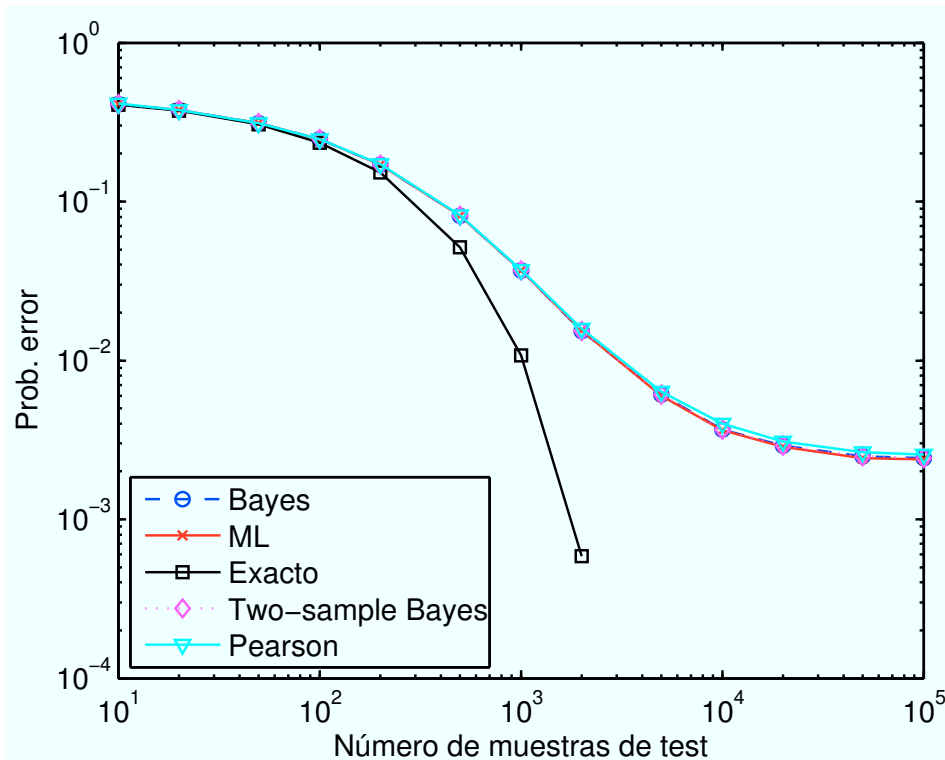
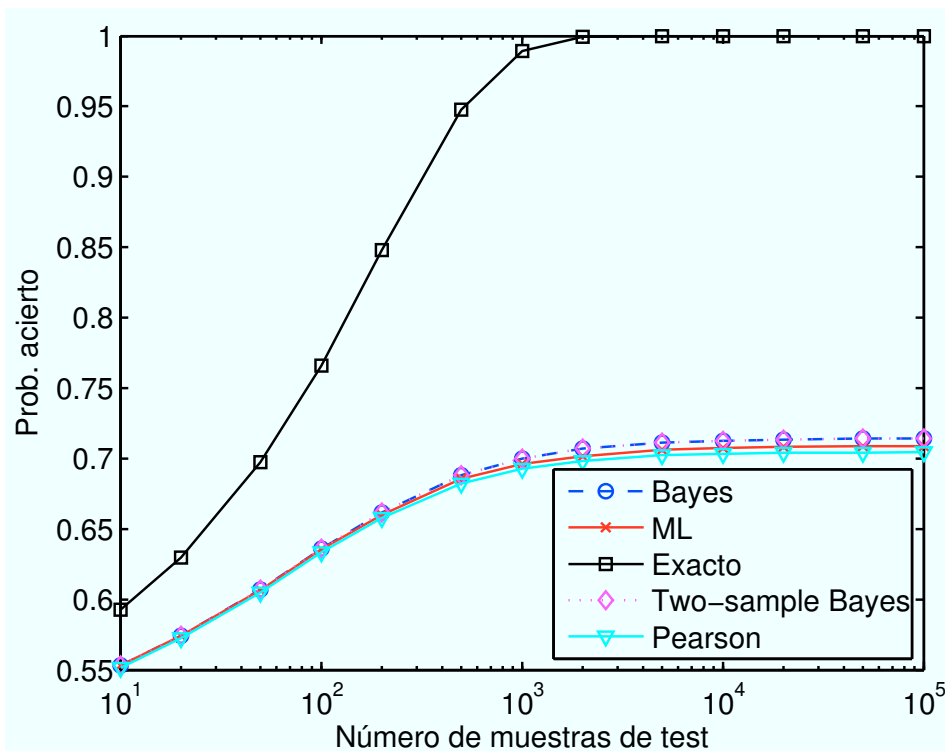
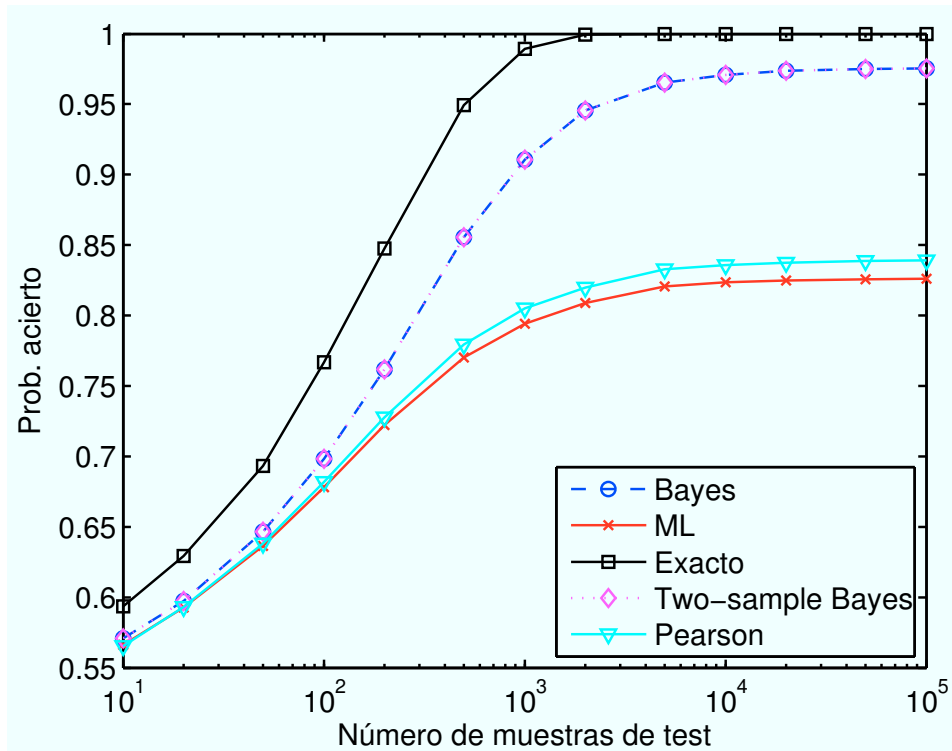
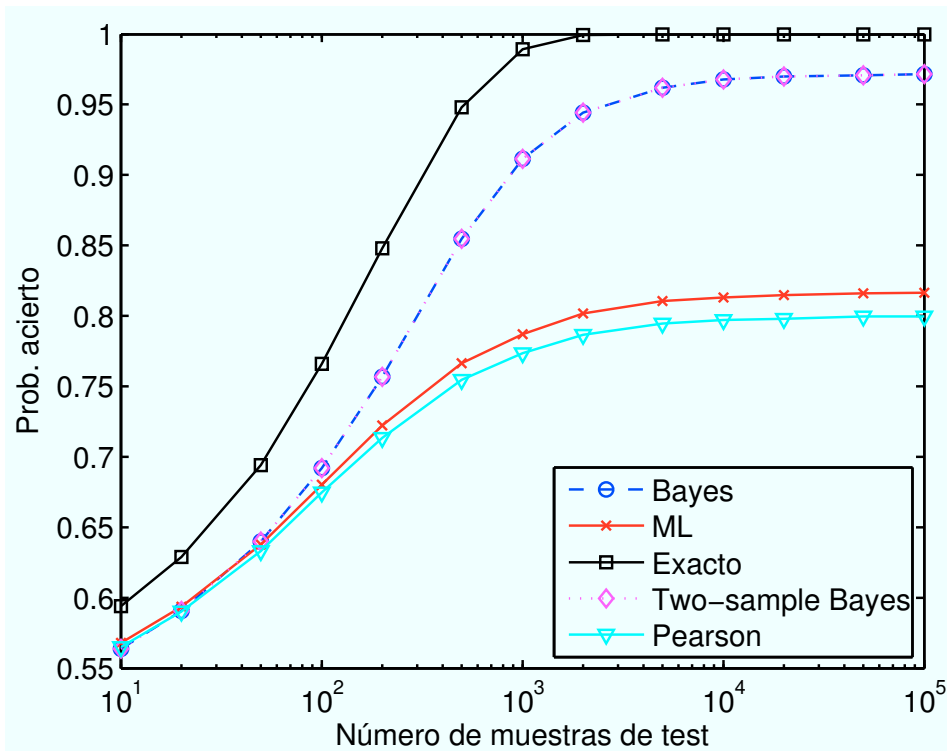
(a) $N = M = 1000$ (b) $N = M = 100$

Figura 3.4: Caso no binario. Comparativa entre el método bayesiano propuesto y los métodos ML, *two-sample* Bayes, Pearson y exacto. Mostramos la precisión media (o el error (a)) obtenida para muestras de test generadas de las hipótesis H_0 y H_1 .



(a) $N = 1000, M = 100$



(b) $N = 100, M = 1000$

Figura 3.5: Caso no binario. Comparativa entre el método bayesiano propuesto y los métodos ML, *two-sample* Bayes, Pearson y exacto. Mostramos la precisión media obtenida para muestras de test generadas de las hipótesis H_0 y H_1 .

El caso no binario es similar sin más que ver en la Ecuación (3.7) que tenemos dos tipos de términos que actualizar:

- $\Gamma\left(K + N + \sum_{\ell=1}^D a_{\ell}\right) = \left(K - 1 + N + \sum_{\ell=1}^D a_{\ell}\right) \Gamma\left(K - 1 + N + \sum_{\ell=1}^D a_{\ell}\right).$
- $\Gamma(k_{\ell,K} + n_{\ell} + a_{\ell}) = \begin{cases} (k_{\ell,K-1} + n_{\ell} + a_{\ell})\Gamma(k_{\ell,K-1} + n_{\ell} + a_{\ell}) & \text{si } z_K = \ell \\ \Gamma(k_{\ell,K-1} + n_{\ell} + a_{\ell}) & \text{si } z_K \neq \ell \end{cases}.$

donde $k_{\ell,K} = \sum_{i=1}^K \delta_{z_i}^{\ell}$. El resultado del test secuencial después de K muestras es exactamente el mismo que el del test bloque para las mismas K muestras. Los resultados del apartado anterior podrían haber sido obtenidos con un test secuencial.

3.1.2.4. Prestaciones asintóticas

Si suponemos, como efectivamente hacemos en el enfoque bayesiano, que nuestro conocimiento *a priori* modela la realidad, podemos establecer cotas de las prestaciones que podemos obtener. Empecemos con el caso binario: cuando el número de muestras en \mathbf{z} tiende a infinito, la verosimilitud $P(\mathbf{z}|p)$ tiende a $\delta(p - r)$, donde r es el valor real de la probabilidad de acierto de la distribución de Bernoulli que genera \mathbf{z} ($r = p^*$ si H_0 es cierta o $r = q^*$ si H_1 es cierta; ambas probabilidades son desconocidas). De este modo, la Ecuación (3.3) resulta:

$$P(\mathbf{z} | H_0)_{as} = \int P(\mathbf{z} | p)P(p | \mathbf{x}) dp = P(r | \mathbf{x}), \quad (3.9)$$

donde el subíndice “*as*” indica que es la probabilidad alcanzada asintóticamente. Sustituyendo estos valores en la Ecuación (3.1) llegamos a las probabilidades asintóticamente alcanzables de las hipótesis dados los datos. Dichas probabilidades son función de los conjuntos de entrenamiento y de r . Por ejemplo para H_0 tenemos:

$$P(H_0 | \mathbf{z})_{as} = \frac{P(r | \mathbf{x})P(H_0)}{P(r | \mathbf{x})P(H_0) + P(r | \mathbf{y})P(H_1)}. \quad (3.10)$$

Notamos que, debido a la incertidumbre, el valor de $P(H_0|\mathbf{z})$ está limitado en todos los casos por los conjuntos de entrenamiento, esto es, esta probabilidad no alcanza los valores 0 o 1 ni siquiera con un número infinito de muestras de test. El valor máximo alcanzable siempre será menor o igual que $\max_{r \in [0,1]} P(H_0|\mathbf{z})_{as}$. Además, es posible obtener un intervalo de confianza (en sentido bayesiano, véase (Henderson y Meyer, 2001)) para dicha probabilidad sin más que restringir la búsqueda del máximo y mínimo de $P(H_0|\mathbf{z})_{as}$ al intervalo en el que la distribución *a priori* $P(p)$ integra una confianza dada. El análisis bayesiano concluye que la incertidumbre de los conjuntos de entrenamiento \mathbf{x} e \mathbf{y} limita las probabilidades alcanzables para $P(H_0|\mathbf{z})_{as}$.

Para el caso no binario el análisis es idéntico puesto que $P(\mathbf{z}|\mathbf{p})$ vuelve a ser una delta cuando el número de muestras de \mathbf{z} tiende a infinito. La verosimilitud de H_0 resulta:

$$P(\mathbf{z} | H_0)_{as} = P(\mathbf{r} | \mathbf{x}) ,$$

donde \mathbf{r} es la distribución de probabilidad real.

La extensión a múltiples hipótesis es inmediata siguiendo estos pasos para cada una de ellas y sustituyendo las verosimilitudes en (3.8).

3.1.3. Enfoque frecuentista

Nuevamente partimos de \mathbf{z} con $z_i \in \{0, 1\}$, que proviene de una distribución de Bernoulli con probabilidad de éxito p^* o q^* , ambas fijas y desconocidas. Para caracterizar estas probabilidades disponemos de un número finito de muestras en \mathbf{x} e \mathbf{y} . Debemos decidir entre las siguientes hipótesis:

H_0 : \mathbf{z} ha sido generado con una distribución Bernoulli(p^*).

H_1 : \mathbf{z} ha sido generado con una distribución Bernoulli(q^*).

Este contraste de hipótesis puede formalizarse de la siguiente manera: el estado Θ es el conjunto $[0, 1]$ y las hipótesis compuestas son $H_0 : \theta \in \Theta_0$ y $H_1 : \theta \in \Theta_1$. Ahora bien, para que este contraste esté bien definido, $\Theta_0 \cap \Theta_1 = \emptyset$. Justamente es en esta última expresión donde aparece el problema. Asumiendo que $p^* \in \Theta_0$ y $q^* \in \Theta_1$ no podemos asegurar por culpa de la incertidumbre que Θ_0 y Θ_1 son disjuntos. Supongamos, a modo de ilustración, que H_0 se caracteriza por una moneda en la que se obtienen 10 caras después de 10 tiradas. Lo único que podemos decir para la probabilidad de cara es $p_{cara} \in (0, 1]$. Lo más probable es que sea un valor cercano a uno, sin embargo, no es descartable que sea un valor pequeño como 10^{-6} . La hipótesis H_1 se caracteriza por otra moneda en la que se obtienen 8 caras en 9 tiradas. Lo único que podemos decir de esta otra moneda es que su probabilidad de cara $q_{cara} \in (0, 1)$.

Nuestra propuesta para este problema es establecer hipótesis en las que las regiones del espacio de estados sean disjuntas para una confianza dada. Esto es, hay que encontrar dos regiones Θ'_0 y Θ'_1 tales que $p^* \in \Theta'_0$ con una probabilidad c_0 , $q^* \in \Theta'_1$ con probabilidad c_1 y $\Theta'_0 \cap \Theta'_1 = \emptyset$. De este modo el problema de contraste de hipótesis simples se transforma en un problema de contraste de hipótesis compuestas. La incertidumbre en la descripción de las hipótesis limita los valores alcanzables para c_0 y c_1 a aquéllos en los que los intervalos no se solapan.

Vamos a desarrollar, a modo de ejemplo, un test secuencial que cumpla estas propiedades para dos distribuciones de Bernoulli. El primer paso es encontrar intervalos

de confianza disjuntos para p^* y q^* . Sean estos intervalos $[p^*_l, p^*_h]$ con confianza c_0 y $[q^*_l, q^*_h]$ con confianza c_1 .

Podemos asumir, sin pérdida de generalidad, que $p^*_h < q^*_l$. Wald (1947) propuso que para contrastar las hipótesis compuestas:

H'_0 : z sigue una distribución Bernoulli cuyo parámetro está en $[p^*_l, p^*_h]$

H'_1 : z sigue una distribución Bernoulli cuyo parámetro está en $[q^*_l, q^*_h]$

es suficiente emplear un SPRT que contraste:

H'_0 : z viene de una distribución Bernoulli cuyo parámetro es p^*_h .

H'_1 : z viene de una distribución Bernoulli cuyo parámetro es q^*_l .

Cuando el SPRT termina, alcanza unas prestaciones P_{FA} y P_D descritas en la Ecuación (2.6). Sin embargo, debido a la incertidumbre, esas probabilidades sólo son representativas con una probabilidad $P_c = c_0 \times c_1$ que es el producto de las confianzas que tenemos en que los parámetros que describen las hipótesis simples caigan dentro de los intervalos que definen las hipótesis compuestas. Por otro lado, con probabilidad $1 - P_c$ las prestaciones obtenidas pueden ser acotadas por el caso peor que consiste en $P_{D,n} = 0$ y $P_{FA,n} = 1$. Así, combinando ambas situaciones, las prestaciones obtenidas por el test secuencial, $P_{D,end}$ y $P_{FA,end}$ cumplen las desigualdades:

$$P_{D,end} \geq P_c P_D + (1 - P_c) P_{D,n} = P_c P_D \leq P_c \quad (3.11)$$

$$P_{FA,end} \leq P_c P_{FA} + (1 - P_c) P_{FA,n} = P_c P_{FA} + (1 - P_c) \geq 1 - P_c. \quad (3.12)$$

3.1.3.1. Prestaciones asintóticas

Las últimas desigualdades de cada expresión de las Ecuaciones (3.11) y (3.12) alcanzan la igualdad con infinitas muestras de test, cuando el SPRT tiene absoluta confianza en su decisión: $P_D = 1$ y $P_{FA} = 0$. En este punto, las cotas que hemos propuesto no garantizan una probabilidad de detección $P_{D,end}$ mayor que P_c ni una probabilidad de falsa alarma $P_{FA,end}$ menor que $1 - P_c$. Vemos, por tanto, que la incertidumbre en las hipótesis limita las prestaciones alcanzables.

La relación entre la incertidumbre y las prestaciones alcanzables por el test queda explícitamente reflejada en las cotas que hemos propuesto. Las prestaciones con infinitas muestras no pueden ser arbitrariamente grandes porque las regiones que definen las hipótesis, Θ_0 y Θ_1 , no pueden solaparse, lo que limita el máximo de P_c . Los valores de las confianzas c_0 y c_1 también afectan a la terminación del test secuencial: cuanto mayores sean, más distantes estarán los umbrales del SPRT para que sus prestaciones, P_D y P_{FA} , sean despreciables frente a P_c ; y, por otro lado,

cuánto menores sean, más distantes estarán las regiones que definen cada hipótesis, lo que hará que el test termine antes para el caso peor, que consiste en que el valor del parámetro de una hipótesis esté en el punto de la frontera de su región de confianza más cercano a la región de la otra hipótesis.

3.1.3.2. Variables discretas no binarias

Del mismo modo que en el caso de variables binarias, podemos emplear los métodos descritos en el Apartado 2.3.3 para obtener las regiones de confianza c_0 y c_1 para \mathbf{p} y \mathbf{q} de modo que dichas regiones no se solapen. La elección de c_0 y c_1 para evitar el solape, al igual que en el caso binario, pasa por que $P_c = c_0 \times c_1$ sea lo mayor posible.

Una vez determinadas las regiones de confianza obtenemos el máximo de la verosimilitud en cada región. El cociente de estas dos verosimilitudes compara dos hipótesis simples y por tanto se compara con los umbrales de Wald (2.5). El test termina cuando el cociente de estas verosimilitudes cruza alguno de los umbrales.

Aunque el test se realice de modo secuencial, hay que recalcular las verosimilitudes de la muestra completa $\{z_1, z_2, \dots, z_k\}$ con cada nueva muestra z_k porque el valor de θ para el que se obtiene el máximo bajo cada hipótesis puede cambiar.

3.1.3.3. Experimentos

Empezamos por examinar la bondad de las cotas (3.11) y (3.12) mediante un experimento similar al que hemos propuesto para el caso binario bayesiano. Para empezar, obtenemos los intervalos bilaterales, $[p^*_l, p^*_h]$ y $[q^*_l, q^*_h]$, de confianzas c_0 y c_1 para p^* y q^* de modo que no se solapen. Como estos intervalos no definen una partición del espacio del parámetro, podemos extenderlos hacia el lugar que no solapan y así aumentar su confianza de manera sencilla. Si asumimos, sin pérdida de generalidad que $p^*_h < q^*_l$, la confianza c'_0 del intervalo $[0, p^*_h]$ para p^* se calcula como

$$c'_0 = 1 - \sum_{i=0}^{n_1} \binom{N}{i} (p^*_h)^i (1 - p^*_h)^{N-i} ,$$

donde $n_1 = \sum_{i=1}^N x_i$. La confianza c'_1 del intervalo $[q^*_l, 1]$ se calcula como

$$c'_1 = 1 - \sum_{i=m_1}^M \binom{M}{i} (q^*_l)^i (1 - q^*_l)^{M-i} ,$$

donde $m_1 = \sum_{i=1}^M y_i$. De este modo recalculamos el valor de $P_c = c'_0 \times c'_1$ de forma menos conservadora. Esta metodología es equivalente al cálculo de intervalos uni-

laterales en el que el primer paso simplemente determina hacia dónde se extiende cada uno de los intervalos unilaterales. Para otros intervalos bilaterales como el de Jeffreys o el de Agresti-Coull se procede de forma similar.

El experimento consiste en lo siguiente:

- Obtenemos N muestras de \mathbf{x} y $M = N$ muestras de \mathbf{y} .
- Estimamos los intervalos de confianza y P_c .
- Obtenemos las prestaciones que demandamos al test secuencial de modo que P_c domine las cotas para $P_{FA,end}$ y $P_{D,end}$. En concreto, empleamos las siguientes probabilidades de diseño: falsa alarma $\alpha = \frac{1-P_c}{1000P_c}$ y no detección $\gamma = \frac{1e-6}{P_c}$ con las que nos garantizamos más que de sobra dicho comportamiento.
- Generamos tantas muestras como sean necesarias para que el test termine y estimamos P_{FA} y P_D a partir de las probabilidades de acierto y error de las hipótesis.

Hemos seleccionado $p^* = 0.1$ para H_0 y $q^* = 0.15$ para H_1 . Se han limitado los valores de las confianzas de los intervalos que definen las hipótesis, c_0 y c_1 , a un valor mínimo de 0.1 y un valor máximo de 0.9999 en el intervalo de Clopper-Pearson (0.99995 para Jeffreys y Agresti-Coull). Esto hace que haya intervalos que solapen en algunas iteraciones, en cuyo caso la decisión se toma al azar, que equivale a $P_{FA} = P_D = 0.5$, y se fija $P_c = 0.01$, que es el producto de las confianzas mínimas establecidas para los intervalos. De cada par de muestras que caracterizan las hipótesis hacemos 20 test distintos: 10 extraídos de H_0 y 10 de H_1 ; este experimento se promedia 10^4 veces.

En el Cuadro 3.1 presentamos, para distintos valores de N , los resultados del promedio de los valores de P_c , P_{FA} y P_D cuando se emplea el intervalo de Clopper-Pearson, que es conservador. Estos promedios se obtienen a partir de los resultados del experimento cuando los intervalos no solapan o a partir de los valores mencionados arriba en caso contrario. El Cuadro 3.1 muestra los valores de la cota $P_{D,cota} \approx P_c$ y $P_{FA,cota} \approx 1 - P_c$; los valores obtenidos del promedio del experimento P_{FA} y P_D ; y la probabilidad de acierto estimada P_a . Apreciamos que la cota es bastante conservadora, lo que es un resultado esperable y heredado del comportamiento conservador del intervalo de Clopper-Pearson.

Los Cuadros 3.2 y 3.3 muestran, respectivamente, los resultados para los intervalos de confianza de Agresti-Coull y de Jeffreys. Estos intervalos de confianza proporcionan aproximadamente la confianza nominal aunque no garantizan en todos los casos que su cobertura sea mayor que el valor de confianza establecido. Las cotas

resultantes de su uso son mucho menos conservadoras en el caso de pocas muestras que las cotas que proporciona el intervalo de Clopper-Pearson. Cuando el número de muestras aumenta, los tres intervalos proporcionan prácticamente la misma cota. Con 5000 muestras todas las cotas alcanzan sus mejores valores que dependen de P_c , que como máximo puede tomar el valor $P_c = 0.9998$ para el caso de Clopper-Pearson y $P_c = 0.9999$ para el caso de Jeffreys y Agresti-Coull. Este hecho queda reflejado en el valor de la cota para la probabilidad de falsa alarma para 5000 muestras de entrenamiento.

N	# no solape	$P_{D,cota}$	P_D	$P_{FA,cota}$	P_{FA}	P_a
10	7371	0.273	0.396	0.727	0.334	0.531
20	8282	0.354	0.584	0.646	0.368	0.608
50	9091	0.454	0.689	0.546	0.319	0.685
100	9511	0.544	0.774	0.456	0.197	0.788
200	9805	0.657	0.893	0.343	0.108	0.892
500	9983	0.835	0.985	0.165	0.014	0.985
1000	10000	0.950	1.000	0.050	6.00e-4	0.999
2000	10000	0.994	1	6.00e-3	0	1
5000	10000	1.000	1	2.06e-4	0	1

Cuadro 3.1: Comprobación de las cotas (3.11) y (3.12) para el caso binario empleando el intervalo de confianza de Clopper-Pearson.

N	# no solape	$P_{D,cota}$	P_D	$P_{FA,cota}$	P_{FA}	P_a
10	7294	0.458	0.535	0.542	0.462	0.537
20	8333	0.497	0.585	0.503	0.441	0.572
50	9092	0.543	0.666	0.457	0.359	0.653
100	9504	0.599	0.767	0.401	0.237	0.765
200	9432	0.682	0.879	0.318	0.124	0.878
500	9935	0.847	0.983	0.153	0.016	0.984
1000	9991	0.951	0.999	0.049	8.50e-4	0.999
2000	10000	0.994	1	5.97e-3	0	1
5000	10000	1.000	1	1.09e-4	0	1

Cuadro 3.2: Comprobación de las cotas (3.11) y (3.12) para el caso binario empleando el intervalo de confianza de Agresti-Coull.

Para el caso de variables discretas no binarias seleccionamos, de igual modo que en el caso bayesiano, los parámetros $\mathbf{p} = [0.1, 0.4, 0.5]$ para H_0 y $\mathbf{q} = [0.15, 0.4, 0.45]$ para H_1 . Para la determinación de las regiones de confianza, hemos escogido el método de Bailey (1980) en lugar del método propuesto por Chafaï y Concordet (2009) por la simplicidad del primero y la elevada carga computacional del último. Entre estos dos está el método de Glaz y Sison (1999) cuya carga computacional

N	# no solape	$P_{D,cota}$	P_D	$P_{FA,cota}$	P_{FA}	P_a
10	7293	0.440	0.501	0.560	0.470	0.516
20	8325	0.488	0.550	0.512	0.402	0.574
50	9129	0.541	0.657	0.459	0.305	0.676
100	9522	0.602	0.764	0.398	0.230	0.767
200	9422	0.679	0.875	0.321	0.118	0.879
500	9934	0.850	0.984	0.150	0.016	0.984
1000	9995	0.953	0.999	0.047	5.50e-4	0.999
2000	10000	0.995	1	5.39e-3	0	1
5000	10000	1.000	1	1.08e-4	0	1

Cuadro 3.3: Comprobación de las cotas (3.11) y (3.12) para el caso binario empleando el intervalo de confianza de Jeffreys.

es aun importante. Una primera diferencia a la hora de determinar regiones de confianza es que ya no resulta tan sencillo hacer que no solapen y a continuación extenderlas de modo que su confianza aumente. El experimento consiste en repetir 1000 veces la conformación de las regiones de confianza de cada parámetro. Para cada par de regiones se realizan 200 test: en 100 se extrae la muestra de test de la distribución que caracteriza H_0 y en 100 de H_1 . Al igual que en el caso binario hemos fijado la confianza mínima de una región a 0.1. El Cuadro 3.4 muestra la bondad de la cota propuesta para este experimento. Apreciamos que la cota es más conservadora que en los experimentos del caso binario debido a que no hemos aumentado la confianza de las regiones por medio de extenderlas dónde no se solapa con la otra hipótesis. En el cuadro se aprecia que en los experimentos que emplean menos de 1000 muestras para caracterizar las hipótesis muchas de las regiones de confianza solapan. Esto hace que la decisión se tome al azar y con ello los malos resultados tanto de la cota como del experimento si los comparamos con las Figuras 3.5 y 3.4. Es, por tanto, vital escoger un método capaz de obtener el menor volumen que contenga una confianza dada, para determinar las regiones de confianza y extenderlas hacia dónde no haya solape para evitar el exceso de conservadurismo. Como también cabe esperar con suficientes muestras, 10^4 en este caso, la cota se ve limitada por el valor máximo de la confianza que hemos usado en este experimento, $1 - 1 \times 10^{-4}$ en este caso que equivale a una $P_c = 0.9998$.

3.2. Variables aleatorias continuas

Una posibilidad para tratar datos que provienen de una variable aleatoria continua es cuantificar y emplear alguno de los procedimientos discretos. Si preferimos no cuantificar y la familia es conocida podemos aplicar el procedimiento bayesiano o

N	# no solape	$P_{D,cota}$	P_D	$P_{FA,cota}$	P_{FA}	P_a
50	392	0.152	0.561	0.848	0.458	0.551
100	462	0.210	0.605	0.790	0.417	0.594
200	621	0.303	0.707	0.697	0.301	0.703
500	846	0.533	0.873	0.467	0.135	0.869
1000	958	0.771	0.964	0.229	0.045	0.960
2000	992	0.938	0.995	0.062	5.02e-03	0.995
5000	1000	0.997	1	2.59e-03	0	1
10000	1000	1.000	1	2.06e-04	0	1

Cuadro 3.4: Comprobación de las cotas (3.11) y (3.12) para el caso discreto no binario empleando la región de confianza de Bailey.

el procedimiento frecuentista que hemos aplicado para las variables discretas. Para este último caso, obtenemos las regiones de confianza de cada parámetro bajo las hipótesis, por ejemplo mediante el método de Hall (1987), y procedemos como en el caso discreto no binario.

Si la familia es desconocida podemos emplear el enfoque frecuentista por medio de bootstrap y la estima no paramétrica de fdp. Para la estima no paramétrica de fdp seguimos el método de Parzen (1962) que es consistente y converge en media cuadrática a la fdp real cuando el número de muestras tiende a infinito. Vapnik (1998) recomienda un ancho para el *kernel* igual $\frac{\log \log K}{K}$, donde K es el número de muestras disponibles para la estima. Sheather y Jones (1991) proponen un método para la estima de dicho ancho que funciona mejor en la práctica para el caso unidimensional. La carga computacional de este método ha sido reducida de $O(KM)$ a $O(K+M)$ (M es el número de puntos a estimar de la fdp) por Raykar y Duraiswami (2006). Botev *et al.* (2009) han propuesto recientemente un método ligeramente más rápido.

Empleamos bootstrap para obtener un intervalo de confianza de la verosimilitud para la hipótesis H_0 de la siguiente manera:

1. Obtenemos B muestras Bootstrap de \mathbf{x} usando la metodología no paramétrica.
2. Para cada una de ellas obtenemos la verosimilitud de H_0 .
3. Ordenamos los B valores de verosimilitud.
4. Para una confianza c_0 despreciamos $(1 - c_0)B/2$ valores de cada uno de los extremos del intervalo formado por los valores de verosimilitud ordenados.
5. Nos quedamos con el máximo de los valores del intervalo de verosimilitud conservado.

Análogamente se hace para H_1 . Los valores de c_0 y c_1 se establecen de forma que los intervalos no solapen. Los umbrales del test secuencial se fijan de forma que sus prestaciones finales tengan un efecto despreciable respecto a P_c . De manera similar al caso discreto no binario este método ha de recalcular las verosimilitudes y las confianzas con cada nueva muestra, y, comparar el cociente de las verosimilitudes máximas con los umbrales de Wald (Ecuación (2.5)).

Si hay muchas muestras la estima de Parzen puede resultar muy costosa. En lugar de ella podemos, por ejemplo, realizar una estima de fdp basada en mezclas de gaussianas cuyo número puede estimarse siguiendo por ejemplo a Roberts *et al.* (1998). Para la determinación del intervalo de confianza de la verosimilitud el procedimiento es análogo al anterior: cada muestra bootstrap proporciona unos parámetros de la mezcla de gaussianas y a partir de este modelo un valor de la verosimilitud.

Ilustramos el caso de familias desconocidas mediante el siguiente experimento: \mathbf{x} proviene de una distribución gaussiana de media cero y varianza 4; \mathbf{y} proviene de una gaussiana de media 1 y varianza 4. La familia se supone desconocida. Obtenemos una muestra aleatoria \mathbf{x} para caracterizar H_0 y una muestra aleatoria \mathbf{y} para caracterizar H_1 . Suponemos que disponemos de tantas muestras de test como sean necesarias para tomar una decisión cuya calidad dependa esencialmente de P_c . Las muestras de test se generan de una de las distribuciones. Para obtener el intervalo de confianza de la verosimilitud de cada hipótesis obtenemos las verosimilitudes de la estima de Parzen de cada uno de los B remuestreos de las muestras que caracterizan cada hipótesis. P_c se determina de modo que estos intervalos no solapen. Por cuestiones de carga computacional hemos limitado B a 10^4 lo que permite estimar intervalos de confianzas menores o iguales a 0.99 lo que limita el valor máximo de $P_c = 0.9801$. Las prestaciones del test secuencial las hemos establecido como: probabilidad de falsa alarma $\alpha = \min(0.01, \frac{1-P_c}{100P_c})$ y probabilidad de no detección $\gamma = \min(0.01, \frac{1e-4}{P_c})$. Repetimos este experimento 1000 veces variando el número de muestras de entrenamiento y obtenemos la probabilidad de acierto y de error para cada clase.

Los resultados de este experimento se muestran en el Cuadro 3.5. Apreciamos que la cota resulta bastante conservadora para este experimento. Este hecho puede deberse al uso del método de percentil en lugar de otros métodos para el cálculo de intervalos bootstrap como el BCa, que es menos conservador aunque resulta mucho más intensivo computacionalmente. Otra fuente de conservadurismo es que nos quedamos con la P_c obtenida con el número de muestras con las que el test termina. Además estamos imponiendo que los intervalos de las verosimilitudes no se solapen, lo que es en sí mismo bastante más conservador que imponer que las

regiones de parámetros no se solapen. Esta última condición puede relajarse puesto que la condición necesaria para la terminación del test secuencial es que el máximo de las verosimilitudes de ambas hipótesis se vaya separando más y más con el número de muestras.

N	# no solape	$P_{D,cota}$	P_D	$P_{FA,cota}$	P_{FA}	P_a
10	999	0.214	0.674	0.793	0.373	0.650
20	1000	0.259	0.777	0.742	0.250	0.764
50	1000	0.432	0.941	0.593	0.080	0.930
100	1000	0.660	0.992	0.334	0.016	0.988
200	1000	0.861	0.999	0.113	2.00e-03	0.999
500	1000	0.977	1	0.024	0	1
1000	1000	0.980	1	0.020	0	1

Cuadro 3.5: Comprobación de las cotas (3.11) y (3.12) para el caso continuo asumiendo variables desconocidas y obteniendo los intervalos de confianza de la verosimilitud mediante el método bootstrap del percentil.

3.3. Resumen

La metodología bayesiana proporciona soluciones cuya exactitud depende, en general, de lo acertado en la elección de la distribución *a priori*, cuya importancia descende con el aumento del número de muestras de \mathbf{x} y de \mathbf{y} . Formalizado el conocimiento *a priori*, la incertidumbre en las hipótesis se maneja perfectamente en el enfoque bayesiano. La única dificultad, si acaso, consiste en la necesidad de tener que evaluar numéricamente alguna integral, cosa que hemos evitado por medio de distribuciones *a priori* conjugadas. Los resultados que proporciona esta metodología son quizá los más deseables para la mayoría de las aplicaciones.

Por otro lado, la metodología frecuentista es conservadora, lo que puede resultar útil en situaciones en las que se desea protegerse del caso peor. Hemos presentado métodos para acotar las prestaciones alcanzables para el caso binario, discreto no binario y continuo. El método binario aprovecha la precisión de las estimas frecuentistas de intervalos de confianza (DasGupta y Zhang, 2006). Para el método discreto no binario, las cotas son más conservadoras debido a la mayor dificultad en extender las regiones de manera que sigan siendo disjuntas. Finalmente, el método para variables continuas desconocidas, que solamente presentamos a modo de extensión, proporciona cotas muy conservadoras.

En conclusión, la incertidumbre en la caracterización de las hipótesis, debido al número finito de muestras, acarrea un límite en las prestaciones alcanzables que se ha de tener en cuenta a la hora de proporcionar la calidad, medida en términos de

probabilidad de detección y probabilidad de falsa alarma, de una decisión. En este capítulo hemos propuesto cotas inferiores de esta calidad tanto desde el enfoque bayesiano como del enfoque frecuentista.

Capítulo 4

Extensión del espacio de entrada

En muchas aplicaciones las prestaciones proporcionadas por una decisión tomada con una sola muestra nos son aceptables y es necesario combinar varias muestras de test. En redes de sensores, por ejemplo, un sólo sensor puede resultar insuficiente para discriminar si un objetivo está presente; Pero, combinar varios puede reducir la tasa de falsa alarma al nivel deseado (Varshney, 1996). Para detectar si un paciente está infectado con tuberculosis por medio del análisis de imágenes microscópicas de su esputo se necesita procesar muchas imágenes para inferir con una alta confianza cuando está sano (Veropoulos *et al.*, 1999). En radar y sonar es habitual procesar varias muestras del entorno antes de poner un objeto en escena y clasificarlo (Van Trees, 1992). En todos estos casos existe un coste asociado al procesado de más muestras (tiempo, dinero, computación, ...), pero se recopila más información para decidir porque se espera reducir la probabilidad de error. En general, siempre aparece un equilibrio entre coste y prestaciones.

Si las distribuciones condicionadas a las clases son conocidas, el clasificador óptimo, conocido como clasificador de Bayes, se obtiene como la solución de un problema de decisión que considera las distribuciones y el riesgo de cada decisión (Berger, 1985). En problemas de clasificación binarios comparamos el cociente de las verosimilitudes de las clases con un umbral para clasificar cada conjunto de observaciones (Neyman y Pearson, 1933). La probabilidad de error de este clasificador tiene una caída asintóticamente exponencial con el número de observaciones, y el mejor exponente de error para la probabilidad de no detección es la divergencia de Kullback-Leibler entre las fdps de las distribuciones condicionadas a las clases (Cover y Thomas, 2006). Pero, en muchas situaciones dichas distribuciones son desconocidas y dependemos de un conjunto de muestras para describir cada clase.

Es este caso, podemos emplear un clasificador basado en aprendizaje, como por ejemplo: un perceptrón multicapa (Bishop, 1995); una red de funciones de base

radial (Haykin, 1994); una SVM (véase Sección 2.4); o un proceso gaussiano para clasificación (*gaussian process for classification*) (GPC) (Rasmussen y Williams, 2006), entre otros. También necesitamos una regla para combinar la salida del clasificador para cada observación a fin de proporcionar una sola decisión. En el caso de la SVM, podemos sumar los márgenes de las distintas muestras y decidir la clase en función de si dicha suma es mayor o menor que cero. También podemos emplear el método de Platt (2000) para obtener una estima de la probabilidad *a posteriori* a partir del margen y combinarlas (multiplicarlas) para decidir la clase del conjunto de muestras. Sin embargo, la SVM no ha sido diseñada para proporcionar probabilidades y estos métodos pueden proporcionar resultados indeseados. Otra posibilidad es emplear un clasificador que proporcione probabilidades *a posteriori* directamente como un GPC. Las probabilidades proporcionadas por un GPC para cada observación se multiplican para formar una predicción conjunta, pero la precisión del método depende de cómo de buenas sean las predicciones de dichas probabilidades.

También podemos ver este problema como un conjunto de problemas de dos muestras (véase Apartado 2.2.3) en los cuáles contrastamos si las observaciones vienen de la misma distribución que la descrita por el conjunto de muestras de entrenamiento de una clase. Sin embargo, los test más potentes de dos muestras emplean toda la muestra de entrenamiento (Gretton *et al.*, 2007), haciendo el proceso costoso computacionalmente. Además, el problema de dos muestras no tiene en cuenta el número de clases ni la posible asimetría entre el número de observaciones de test y el número de muestras de entrenamiento de cada clase. Por tanto, de alguna manera, no emplean toda la información disponible para decidir la clase correcta de las observaciones de test.

4.1. Motivación y método

Supongamos que un banco de sangre nos asigna la construcción de un test de alguna enfermedad infecciosa. Para ello, nos facilita un conjunto de muestras iid $(\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}_i, y_i)\} \quad i = 1, \dots, n$ donde $\mathbf{x}_i \in \mathbb{R}^d$ es un vector de características que describe la sangre del donante e $y_i \in \{\pm 1\}$ es la etiqueta asociada a cada muestra que indica si la sangre está infectada. Construimos un clasificador mediante nuestro algoritmo favorito de aprendizaje, el cuál, escoge la función $f(\cdot)$ entre una familia \mathcal{F} que minimiza algún funcional de riesgo sobre $(\mathcal{X}, \mathcal{Y})$. Predecimos la clase y^* de una nueva muestra de sangre mediante $y^* = f(\mathbf{x}^*)$, donde \mathbf{x}^* es el vector de características para la sangre.

El gerente del banco de sangre considera la probabilidad de error proporcionada

por una muestra de test es demasiado alta y recomienda repetir el test L veces de forma iid. Por tanto, obtenemos un conjunto de L observaciones $\{\mathbf{x}_1^*, \dots, \mathbf{x}_L^*\}$ de cada donante de sangre para reducir la probabilidad de error. Una posibilidad para obtener una predicción global sobre la clase de dichas observaciones es combinar las salidas individuales del clasificador $f(\cdot)$ para cada observación. Si el clasificador proporciona salidas probabilísticas, $p(y = 1 | \mathbf{x}_i^*)$ $i = 1, \dots, L$, multiplicamos sus salidas y consideramos las probabilidades *a priori* de las clases $p(y = 1)$, $p(y = -1)$ para obtener una estima de $p(y = 1 | \mathbf{x}_1^*, \dots, \mathbf{x}_L^*)$ como:

$$\begin{aligned}
p(y = 1 | \mathbf{x}_1^*, \dots, \mathbf{x}_L^*) &= \frac{p(\mathbf{x}_1^*, \dots, \mathbf{x}_L^* | y = 1)p(y = 1)}{p(\mathbf{x}_1^*, \dots, \mathbf{x}_L^*)} \\
&= \frac{p(y = 1) \prod_{i=1}^L p(\mathbf{x}_i^* | y = 1)}{p(y = 1)p(\mathbf{x}_1^*, \dots, \mathbf{x}_L^* | y = 1) + p(y = -1)p(\mathbf{x}_1^*, \dots, \mathbf{x}_L^* | y = -1)} \\
&= \frac{p(y = 1) \prod_{i=1}^L p(\mathbf{x}_i^* | y = 1)}{p(y = 1) \prod_{i=1}^L p(\mathbf{x}_i^* | y = 1) + p(y = -1) \prod_{i=1}^L p(\mathbf{x}_i^* | y = -1)} \\
&= \frac{\frac{p(y=1) \prod_{i=1}^L (p(y=1|\mathbf{x}_i^*)p(\mathbf{x}_i^*))}{p(y=1)^L}}{\frac{p(y=1) \prod_{i=1}^L (p(y=1|\mathbf{x}_i^*)p(\mathbf{x}_i^*))}{p(y=1)^L} + \frac{p(y=-1) \prod_{i=1}^L (p(y=-1|\mathbf{x}_i^*)p(\mathbf{x}_i^*))}{p(y=-1)^L}} \\
&= \frac{\frac{p(y=1) \prod_{i=1}^L p(y=1|\mathbf{x}_i^*)}{p(y=1)^L}}{\frac{p(y=1) \prod_{i=1}^L p(y=1|\mathbf{x}_i^*)}{p(y=1)^L} + \frac{p(y=-1) \prod_{i=1}^L p(y=-1|\mathbf{x}_i^*)}{p(y=-1)^L}} \\
&= \frac{\frac{\prod_{i=1}^L p(y=1|\mathbf{x}_i^*)}{p(y=1)^{L-1}}}{\frac{\prod_{i=1}^L p(y=1|\mathbf{x}_i^*)}{p(y=1)^{L-1}} + \frac{\prod_{i=1}^L (1-p(y=1|\mathbf{x}_i^*))}{(1-p(y=1))^{L-1}}} . \tag{4.1}
\end{aligned}$$

En los casos en los que sólo nos importa la clase con mayor probabilidad *a posteriori*, basta con escoger la clase +1 cuando:

$$\frac{\prod_{i=1}^L p(y = 1 | \mathbf{x}_i^*)}{p(y = 1)^{L-1}} > \frac{\prod_{i=1}^L p(y = -1 | \mathbf{x}_i^*)}{p(y = -1)^{L-1}} = \frac{\prod_{i=1}^L (1 - p(y = 1 | \mathbf{x}_i^*))}{(1 - p(y = 1))^{L-1}} .$$

Sin embargo, los clasificadores discriminativos pueden proporcionar estimas pobres de probabilidad en regiones del espacio alejadas de la frontera entre las clases que podrían empeorar las prestaciones del test global. Por tanto, proponemos obtener una sola salida para todas las observaciones de test y construir un clasificador $f_L(\cdot)$ que clasifique el vector de L observaciones como entrada. Este clasificador deberá entrenarse con vectores de L muestras que se generarán a partir del conjunto de entrenamiento $(\mathcal{X}, \mathcal{Y})$ disponible. Al nuevo conjunto de entrenamiento lo llamamos extendido porque se forma a partir de la combinación de muestras del conjunto original.

Construimos el conjunto de entrenamiento extendido de la siguiente manera: primero, separamos las muestras de entrenamiento por clases, de forma que cada clase es una matriz con una muestra por fila; luego, para cada clase, creamos las muestras del conjunto de entrenamiento extendido concatenando horizontalmente L permutaciones de las filas de dicha matriz. Supongamos que las muestras (una por fila) de clase +1 de \mathcal{X} son:

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \cdot \\ \mathbf{x}_{N_1} \end{bmatrix} .$$

Un posible ejemplo para K_1 muestras (una por fila) de un conjunto de entrenamiento extendido para $L = 3$ sería:

$$\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{z}_3 \\ \cdot \\ \mathbf{z}_{K_1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_1 \\ \mathbf{x}_6 & \mathbf{x}_4 & \mathbf{x}_2 \\ \mathbf{x}_5 & \mathbf{x}_7 & \mathbf{x}_5 \\ \dots & \dots & \dots \\ \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_6 \end{bmatrix} .$$

Repetimos los pasos de permutar y concatenar tantas veces como sea necesario. En otras palabras, cada muestra extendida $\mathbf{z}_k \in \mathbb{R}^{d \times L}$ etiquetada como $y'_k \in \{\pm 1\}$ del conjunto de entrenamiento extendido, $(\mathcal{Z}, \mathcal{Y}') = \{(\mathbf{z}_k, y'_k)\} \quad k = 1, \dots, K$, está compuesta por L muestras de la clase y'_k . Si N_1 es el número de muestras de la clase +1, existen un total de $\binom{N_1}{L} L!$ posibles muestras diferentes¹ del conjunto extendido de clase +1.

Asumamos que el conjunto de entrenamiento $(\mathcal{X}, \mathcal{Y})$ está ordenado por clase, $y_i = 1 \quad i = 1, \dots, N_1$. El método de la permutación construye el conjunto de entrenamiento extendido usando diferentes permutaciones de todas las muestras $\{\mathbf{x}_1, \dots, \mathbf{x}_{N_1}\}$. El método *bootstrap* (Efron y Tibshirani, 1993) muestrea $\{\mathbf{x}_1, \dots, \mathbf{x}_{N_1}\}$ en cada concatenación. Este último método toma muestras de la distribución empírica de cada clase. Si la clase tiene N_1 muestras, cada muestreo toma aproximadamente $0.66N_1$ muestras diferentes (Efron y Tibshirani, 1993). Hay dos diferencias fundamentales entre ambos métodos: la primera es el número de muestras del conjunto original empleadas en cada concatenación: el método de permutación usa todas las muestras cada vez y bootstrap aproximadamente $0.66N_1$. Por otro lado, bootstrap

¹Si consideramos que una muestra original no puede aparecer dos veces en la misma muestra extendida. En caso contrario, tenemos N_1^L muestras diferentes.

muestra de manera independiente la función de distribución empírica mientras que en el método de permutación no está clara la independencia de las muestras resultantes por el requisito de emplearlas todas. En cuanto a prestaciones podemos esperar prestaciones similares de ambos métodos aunque el método de permutación debería de funcionar un poco mejor con pocas muestras porque el requisito de usarlas todas construye un conjunto ligeramente más variado.

La función de distribución empírica de \mathcal{X} converge a la función de distribución de \mathcal{X} cuando el número de muestras de entrenamiento n crece a infinito. Si L es finito, tanto el método de permutación como el método bootstrap producen distribuciones empíricas de \mathcal{Z} que convergen a la distribución de la extensión L -ésima de \mathcal{X} .

Cualquier conocimiento *a priori* de las muestras puede usarse para construir el conjunto extendido. Por ejemplo, si las muestras $\{\mathbf{x}_1, \mathbf{x}_{17}\}$ pertenecen a la misma clase, y añadimos al conjunto extendido el punto $\mathbf{z}_1 = [\mathbf{x}_1, \mathbf{x}_{17}]$ también podemos añadir el punto $\mathbf{z}_2 = [\mathbf{x}_{17}, \mathbf{x}_1]$, esto es, es posible añadir al conjunto extendido muestras que forman colecciones de puntos simétricos. Si añadimos para cada conjunto de L muestras todas sus simetrías, tendremos simetrías en la frontera de decisión. En particular para la SVM, esto dará lugar en (2.10) a que los multiplicadores de Lagrange, α_i , correspondientes a las muestras \mathbf{z}_i que se forman a partir de las simetrías de L muestras \mathbf{x}_k tengan el mismo valor. Esto simplifica considerablemente la carga computacional del entrenamiento del método extendido.

Para ilustrar este punto, vamos a suponer que $L = 3$ y que añadimos todas las permutaciones de cada conjunto de 3 muestras al conjunto de entrenamiento. Por comodidad, ilustramos este hecho con la clase +1 y las muestras $\{\mathbf{x}_1, \dots, \mathbf{x}_6\}$ nos sirven para construir.

$$\begin{bmatrix} \mathbf{z}_1^1 \\ \mathbf{z}_1^2 \\ \mathbf{z}_1^3 \\ \mathbf{z}_1^4 \\ \mathbf{z}_1^5 \\ \mathbf{z}_1^6 \\ \mathbf{z}_2^1 \\ \mathbf{z}_2^2 \\ \mathbf{z}_2^3 \\ \mathbf{z}_2^4 \\ \mathbf{z}_2^5 \\ \mathbf{z}_2^6 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \\ \mathbf{x}_1 & \mathbf{x}_3 & \mathbf{x}_2 \\ \mathbf{x}_2 & \mathbf{x}_1 & \mathbf{x}_3 \\ \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_1 \\ \mathbf{x}_3 & \mathbf{x}_1 & \mathbf{x}_2 \\ \mathbf{x}_3 & \mathbf{x}_2 & \mathbf{x}_1 \\ \mathbf{x}_4 & \mathbf{x}_5 & \mathbf{x}_6 \\ \mathbf{x}_4 & \mathbf{x}_6 & \mathbf{x}_5 \\ \mathbf{x}_5 & \mathbf{x}_4 & \mathbf{x}_6 \\ \mathbf{x}_5 & \mathbf{x}_6 & \mathbf{x}_4 \\ \mathbf{x}_6 & \mathbf{x}_4 & \mathbf{x}_5 \\ \mathbf{x}_6 & \mathbf{x}_5 & \mathbf{x}_4 \end{bmatrix}.$$

Las muestras $\{\mathbf{z}_1^1, \dots, \mathbf{z}_1^6\}$ forman todas las permutaciones de las muestras $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$

y las muestras $\{\mathbf{z}_2^1, \dots, \mathbf{z}_2^6\}$ forman a su vez las permutaciones de $\{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$. Ahora bien, si añadimos todas las permutaciones de todas las muestras \mathbf{z}_i sabemos por la simetría impuesta en la construcción que $\alpha_1^1 = \alpha_1^2 = \dots = \alpha_1^6 = \alpha_1$ y que $\alpha_2^6 = \alpha_2^2 = \dots = \alpha_2^6 = \alpha_2$ por lo que el problema de optimización a resolver (véase (2.19)):

$$\max_{\alpha_i} \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \mathbf{1}$$

no tiene tantas incógnitas como puntos en el conjunto de entrenamiento sino que dicho número se reduce por $L!$. El término $\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}$ en nuestro caso resulta

$$\begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \end{bmatrix}_{2 \times 12} \begin{bmatrix} K_{1,1}^{1,1} & \dots & K_{1,1}^{1,6} & K_{1,2}^{1,1} & \dots & K_{1,2}^{1,6} \\ K_{1,1}^{2,1} & \dots & K_{1,1}^{2,6} & K_{1,2}^{2,1} & \dots & K_{1,2}^{2,6} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{1,1}^{6,1} & \dots & K_{1,1}^{6,6} & K_{1,2}^{6,1} & \dots & K_{1,2}^{6,6} \\ K_{2,1}^{1,1} & \dots & K_{2,1}^{1,6} & K_{2,2}^{1,1} & \dots & K_{2,2}^{1,6} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{2,1}^{6,1} & \dots & K_{2,1}^{6,6} & K_{2,2}^{6,1} & \dots & K_{2,2}^{6,6} \end{bmatrix}_{12 \times 12} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

donde los elementos de la matriz de *kernel* se corresponden a $K_{i,j}^{n,m} = k(\mathbf{z}_i^n, \mathbf{z}_j^m)$. La matriz de *kernel* equivalente resulta:

$$\begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^6 \sum_{j=1}^6 K_{1,1}^{i,j} & \sum_{i=1}^6 \sum_{j=1}^6 K_{1,2}^{i,j} \\ \sum_{i=1}^6 \sum_{j=1}^6 K_{2,1}^{i,j} & \sum_{i=1}^6 \sum_{j=1}^6 K_{2,2}^{i,j} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}.$$

De modo que el número de variables del problema de optimización no aumenta por forzar la simetría sino que simplemente el cálculo de la matriz de *kernels* se ve afectado por ésta.

Consideramos ahora dos posibilidades para el *kernel* usado en el espacio extendido. Siguiendo con el ejemplo anterior consideramos:

$$k(\mathbf{z}_1, \mathbf{z}_7) = \boldsymbol{\varphi}(\mathbf{z}_1)^T \boldsymbol{\varphi}(\mathbf{z}_7) = \boldsymbol{\phi}([\mathbf{x}_1 \ \mathbf{x}_2 \ \mathbf{x}_3])^T \boldsymbol{\phi}([\mathbf{x}_4 \ \mathbf{x}_5 \ \mathbf{x}_6])$$

Para $k(\mathbf{z}_1, \mathbf{z}_7) = e^{-\gamma \|\mathbf{z}_1 - \mathbf{z}_7\|^2}$ este *kernel* puede escribirse en función de *kernels* del

espacio de entrada como

$$\begin{aligned} k(\mathbf{z}_1, \mathbf{z}_7) &= e^{-\gamma\|\mathbf{z}_1-\mathbf{z}_7\|^2} = e^{-\gamma\|\mathbf{x}_1-\mathbf{x}_4\|^2} e^{-\gamma\|\mathbf{x}_2-\mathbf{x}_5\|^2} e^{-\gamma\|\mathbf{x}_3-\mathbf{x}_6\|^2} \\ &= k(\mathbf{x}_1, \mathbf{x}_4)k(\mathbf{x}_2, \mathbf{x}_5)k(\mathbf{x}_3, \mathbf{x}_6) . \end{aligned}$$

También podemos construir el *kernel* en el espacio original, lo que para un *kernel* genérico resulta

$$k(\mathbf{z}_1, \mathbf{z}_7) = \boldsymbol{\psi}(\mathbf{z}_1)^T \boldsymbol{\psi}(\mathbf{z}_7) = [\boldsymbol{\phi}(\mathbf{x}_1) \ \boldsymbol{\phi}(\mathbf{x}_2) \ \boldsymbol{\phi}(\mathbf{x}_3)] \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_4) \\ \boldsymbol{\phi}(\mathbf{x}_5) \\ \boldsymbol{\phi}(\mathbf{x}_6) \end{bmatrix} = k(\mathbf{x}_1, \mathbf{x}_4) + k(\mathbf{x}_2, \mathbf{x}_5) + k(\mathbf{x}_3, \mathbf{x}_6) .$$

De este modo, los *kernel* $k(\mathbf{z}_1, \mathbf{z}_7)$ para todas las permutaciones de \mathbf{z}_1 y \mathbf{z}_7 pueden obtenerse a partir de los *kernel*s de $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ y $\{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$. Esto reduce el número de *kernel*s a calcular directamente de 36 a 9.

Combinando las propiedades que acabamos de mencionar y usando un *kernel* adecuado podemos argumentar que emplear simetrías en el clasificador no aumenta considerablemente la complejidad del problema de optimización que resuelve la SVM.

Consideramos a continuación el efecto de estos dos tipos de *kernel* en el *primal* (2.10) de la SVM. Vamos a ver como $\boldsymbol{\psi}(\mathbf{z}_i)$ permite imponer las simetrías sin necesidad de generar las permutaciones de todas las muestras \mathbf{x}_j que forman cada \mathbf{z}_i . El problema de optimización es:

$$\min_{\mathbf{w}, \xi_i, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i$$

sujeto a:

$$\begin{aligned} y_i(\mathbf{w}^T \boldsymbol{\psi}(\mathbf{z}_i) + b) &\geq 1 - \xi_i & \forall i = 1, \dots, N \\ \xi_i &\geq 0 & \forall i = 1, \dots, N \end{aligned} \quad (4.2)$$

Si ahora definimos $\mathbf{w}^T = [\mathbf{w}_1^T \ \mathbf{w}_2^T \ \mathbf{w}_3^T]/3$ y tenemos en cuenta cómo es $\boldsymbol{\psi}$, podemos escribir (4.2) como:

$$y_i \left(\frac{1}{3} \sum_{k=1}^3 \mathbf{w}_k^T \boldsymbol{\phi}(\mathbf{x}_i^k) + b \right) \geq 1 - \xi_i ,$$

donde \mathbf{x}_i^k es la muestra que ocupa el lugar k -ésimo en \mathbf{z}_i . La salida para $\boldsymbol{\psi}(\mathbf{z}_1)$ ha de ser la misma que para cualquiera de sus permutaciones y, por tanto, $\mathbf{w}_1^T = \mathbf{w}_2^T = \mathbf{w}_3^T$,

con lo que podemos escribir la ecuación anterior como:

$$y_i \left(\frac{1}{3} \mathbf{w}_1^T \sum_{k=1}^3 \phi(\mathbf{x}_i^k) + b \right) \geq 1 - \xi_i .$$

Esta transformación no lineal obtiene la media de todos los \mathbf{x}_i^k de un \mathbf{z}_i en el espacio de características y aplica una SVM.

Para el kernel φ necesitamos generar todas las permutaciones de cada \mathbf{z}_i para imponer la simetría. En nuestro ejemplo tendríamos $3!$ permutaciones para cada muestra \mathbf{z}_i que denotaremos como \mathbf{z}_i^j con $j = 1, \dots, 6$. Ahora bien, debido a las simetrías esperamos que la salida para todas las permutaciones sea la misma, esto es, $\mathbf{w}^T \varphi(\mathbf{z}_i^1) = \dots = \mathbf{w}^T \varphi(\mathbf{z}_i^6)$. De modo que tenemos una única variable ξ_i y un único multiplicador de Lagrange α_i que comparten todas las muestras \mathbf{z}_i^j . La restricción del margen para todas las muestras es:

$$y_i \left(\mathbf{w}^T \varphi(\mathbf{z}_i^j) + b \right) \geq 1 - \xi_i ,$$

y el funcional a minimizar es

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i - \sum_i \alpha_i \left(y_i \left(\mathbf{w}^T \sum_j \varphi(\mathbf{z}_i^j) + b \right) - 1 + \xi_i \right) - \sum_i \mu_i \xi_i .$$

Ahora en lugar de calcular la media de \mathbf{x}_i^k obtenemos la media de todas las posibles permutaciones de \mathbf{z}_i en el espacio de características.

Pasando a la fase de test, clasificamos un vector de L muestras por medio del clasificador extendido. Si hemos impuesto simetrías en el conjunto de entrenamiento, las $L!$ permutaciones de las muestras de test proporcionarán la misma salida. En caso contrario, tenemos varias alternativas: la primera, consiste en hacer el test sobre todas las permutaciones y escoger la muestra que proporcione mejor probabilidad o margen. La segunda, es evaluar más de una permutación y combinar las salidas. La forma de combinarlas no es evidente porque las distintas permutaciones no dan lugar a muestras de test independientes. Aun así, puede resultar ventajoso combinarlas sumando los márgenes, multiplicando las probabilidades o haciendo una votación.

Combinar las salidas de distintas permutaciones del vector de test, proporciona una decisión más robusta y puede disminuir la probabilidad de error. Por otra parte, esto puede ser una manera de intercambiar complejidad entre las fases de entrenamiento y test, que puede resultar conveniente cuando el test es rápido.

Comparamos los métodos de permutación y bootstrap en la próxima sección; en la que también mostramos como las simetrías mejoran los resultados de clasificación;

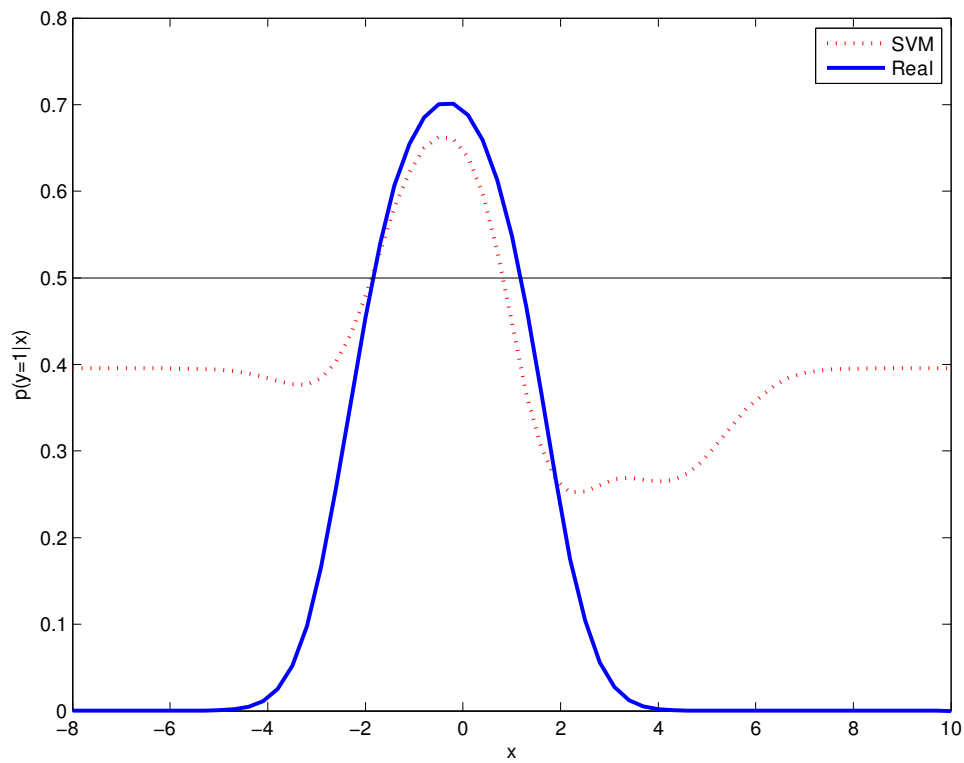
además, presentamos la mejora de combinar las salidas de las permutaciones del vector de test.

4.2. Ilustración

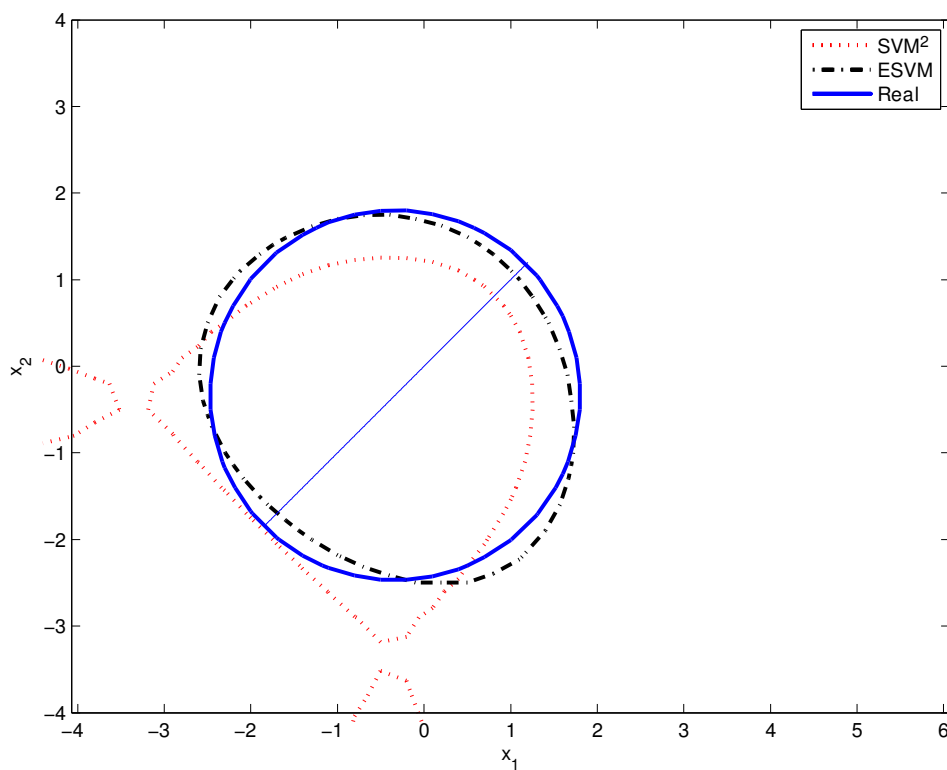
Ilustramos el funcionamiento del método propuesto con un ejemplo que nos permitirá mostrar que extender la dimensión de entrada puede ser una buena idea cuando la probabilidad de error es alta. Escogemos un problema de clasificación binaria donde las muestras de la clase $+1$ han sido tomadas de una gaussiana de media cero y varianza unidad y las muestras de la clase -1 han sido tomadas de una gaussiana de media uno y varianza 4. El error de Bayes en este problema para una muestra de test es 0.305. Tomamos dos observaciones de cada una de las clases y comparamos la probabilidad de error obtenida por la combinación de las dos salidas del clasificador unidimensional con la salida del clasificador bidimensional entrenado en el conjunto extendido. Realizamos la comparación para SVM y GPC. Las salidas de la SVM se convierten a probabilidades siguiendo el método de Platt (Platt, 2000). Combinamos las probabilidades por medio de la regla de Bayes descrita en la Ecuación (4.1). El conjunto de entrenamiento $(\mathcal{X}, \mathcal{Y})$ contiene 100 muestras y el conjunto de entrenamiento extendido contiene 200 muestras creadas usando los métodos propuestos donde hemos incrementado el número de muestras del conjunto extendido por $L = 2$, esto es, el factor en que hemos incrementado la dimensión. Los parámetros de la SVM y del GPC se ajustan respectivamente para cada conjunto de entrenamiento por medio de validación cruzada y maximizando su verosimilitud marginal. En los cuadros y figuras denominamos la combinación de las dos salidas probabilísticas como SVM² y GPC² y la salida del clasificador extendido como EGPC y ESVM, donde SVM o GPC hacen referencia al clasificador empleado.

La Figura 4.1(a) muestra las predicciones probabilísticas de la SVM para una muestra de test para la clase $+1$ contra la probabilidad *a posteriori* obtenida de las distribuciones reales de las que se obtuvieron las muestras. La SVM predice muy bien la frontera, esto es, la probabilidad 0.5. Sin embargo, las predicciones lejos de la frontera no son precisas.

La Figura 4.1(b) muestra la frontera de decisión de la clase $+1$ para la SVM² *versus* la ESVM; también mostramos la frontera del decisor bayesiano para comparación. Observamos en la Figura 4.1(b) que la ESVM funciona mejor que la combinación SVM² en la región superior izquierda, que coincide con la región donde más masa tienen las gaussianas. Además, la ESVM estima mejor que SVM² la casi totalidad de la frontera. La frontera predicha por la SVM² es precisa en el extremo inferior



(a)



(b)

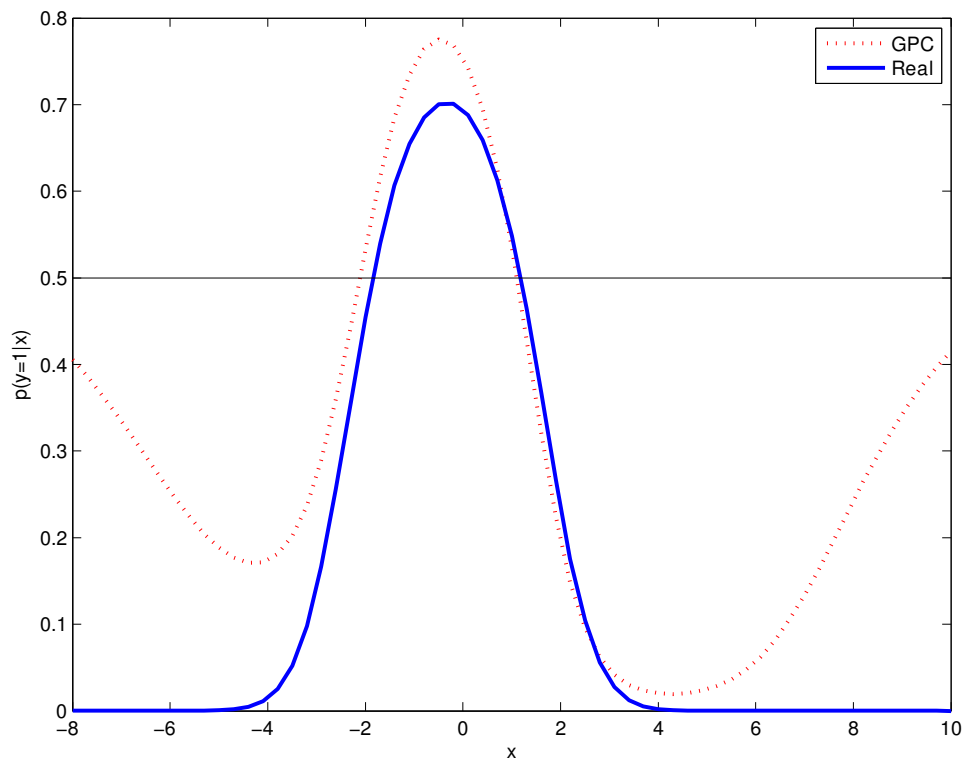
Figura 4.1: (a) $p(y = 1|x)$ predicha por la SVM *versus* $p(y = 1|x)$ real (clasificador bayesiano). (b) Frontera de la combinación de dos predicciones de la SVM $p(y = 1|x_1^*, x_2^*) = 0.5$ (SVM²) *versus* la predicción de la ESVM (usando el método de permutación) *versus* la frontera real.

de la bisección, al igual que la SVM lo hacía determinando la frontera izquierda en la Figura 4.1(a). La SVM² trata de aproximar los extremos de la bisección, pero su precisión es baja para probabilidades en otros puntos del plano.

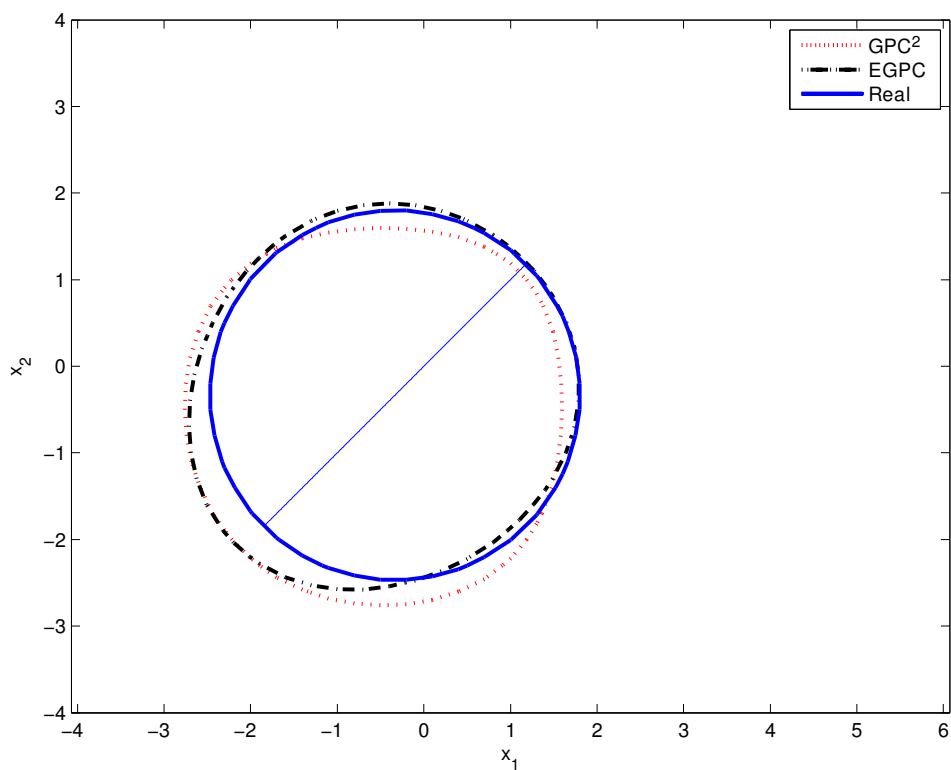
Hemos repetido el mismo experimento usando un GPC y observamos en la Figura 4.2(a) que las probabilidades estimadas por el GPC son mejores que las estimadas por la SVM. Sin embargo, el GPC falla con las probabilidades extremas, lo que permite una mejora para el clasificador entrenado en el conjunto extendido, como podemos observar en la Figura 4.2(b).

Para mostrar la diferencia cuantitativa entre la SVM² y el GPC², y la ESVM y el EGPC, repetimos el experimento de clasificar vectores de dos muestras de test con conjuntos de entrenamiento aleatorios variando el número de muestras de entrenamiento. Los resultados se resumen en el Cuadro 4.1, donde mostramos la media y la desviación típica del error para ambos métodos. Observamos que las prestaciones de la ESVM son muy superiores a los de los demás métodos para 40 muestras de entrenamiento. Esperamos que la ESVM obtenga buenas prestaciones donde hay pocas muestras por dimensión. Mostramos, además, la diferencia entre los métodos para construir el conjunto extendido. El método de permutación (subíndice “*p*”) proporciona mejores prestaciones que el método bootstrap (subíndice “*b*”) en todos los casos como hemos conjeturado en la sección anterior. Para evaluar el uso de simetrías, hemos añadido conjuntos de datos simétricos como hemos explicado más arriba (subíndice “*sim*”) que claramente beneficia a la ESVM. También hemos evaluado la combinación probabilística de las salidas de distintas permutaciones (2 en realidad) del vector de test (subíndice “*tp*”) que también beneficia a la ESVM y proporciona los mejores resultados para 100 y 200 muestras de entrenamiento. El error de Bayes para este problema es de 0.2201 que es prácticamente alcanzado por el EGPC con 400 muestras de entrenamiento. Queremos hacer hincapié en que el error de clasificadores basados en aprendizaje entrenados mediante conjuntos extendidos converge al error de Bayes, como hemos presentado en la sección anterior.

En la Figura 4.3 comparamos la salida de la SVM contra la ESVM, con los métodos bootstrap y permutación, cuando el número de muestras en el vector de test aumenta. El número de muestras de entrenamiento en el conjunto extendido aumenta su tamaño un factor de L . Observamos que la ESVM siempre funciona mejor que la combinación de las salidas de la SVM y que los métodos propuestos para generar el conjunto extendido son prácticamente indistinguibles. Las prestaciones de la ESVM están limitadas por el tamaño de entrenamiento porque algunos conjuntos son pobres y no se gana nada por emplear más muestras de test. Este efecto disminuye, como cabía de esperar, con el aumento del número de muestras de entrenamiento.



(a)



(b)

Figura 4.2: (a) $p(y = 1|x)$ predicha por el GPC *versus* $p(y = 1|x)$ real (clasificador bayesiano). (b) Frontera de la combinación de dos predicciones del GPC $p(y = 1|x_1^*, x_2^*) = 0.5$ (GPC²) *versus* la predicción del EGPC (usando el método de permutación) *versus* la frontera real.

muestras	40	100	200	400
SVM ²	0.333 ± 0.154	0.249 ± 0.030	0.238 ± 0.018	0.240 ± 0.020
GPC ²	0.315 ± 0.079	0.246 ± 0.033	0.228 ± 0.016	0.223 ± 0.006
ESVM _b	0.276 ± 0.074	0.243 ± 0.021	0.231 ± 0.010	0.226 ± 0.006
ESVM _p	0.269 ± 0.071	0.238 ± 0.017	0.229 ± 0.010	0.225 ± 0.007
EGPC _b	0.298 ± 0.061	0.246 ± 0.028	0.230 ± 0.011	0.225 ± 0.007
EGPC _p	0.277 ± 0.055	0.240 ± 0.026	0.227 ± 0.008	0.223 ± 0.006
ESVM _{b,sim}	0.269 ± 0.050	0.240 ± 0.019	0.229 ± 0.011	0.225 ± 0.007
ESVM _{p,sim}	0.264 ± 0.040	0.237 ± 0.016	0.227 ± 0.009	0.225 ± 0.006
EGPC _{b,sim}	0.298 ± 0.059	0.251 ± 0.033		
EGPC _{p,sim}	0.286 ± 0.050	0.244 ± 0.028		
ESVM _{b,tp}	0.273 ± 0.066	0.239 ± 0.019	0.229 ± 0.010	0.225 ± 0.007
ESVM _{p,tp}	0.269 ± 0.057	0.236 ± 0.017	0.227 ± 0.009	0.225 ± 0.006

Cuadro 4.1: Media y desviación típica del error con dos muestras de test en el ejemplo de dos gaussianas variando el número de muestras de entrenamiento para SVM, GPC, ESVM_b, ESVM_p, EGPC_b, EGPC_p (métodos extendidos bootstrap y permutación), ESVM_{b,sim}, ESVM_{p,sim}, GPC_{b,sim}, GPC_{p,sim} (métodos extendidos bootstrap y permutación con restricción de simetría) ESVM_{b,tp}, ESVM_{p,tp}, (métodos extendidos bootstrap y permutación combinando las salidas de las permutaciones del vector de test).

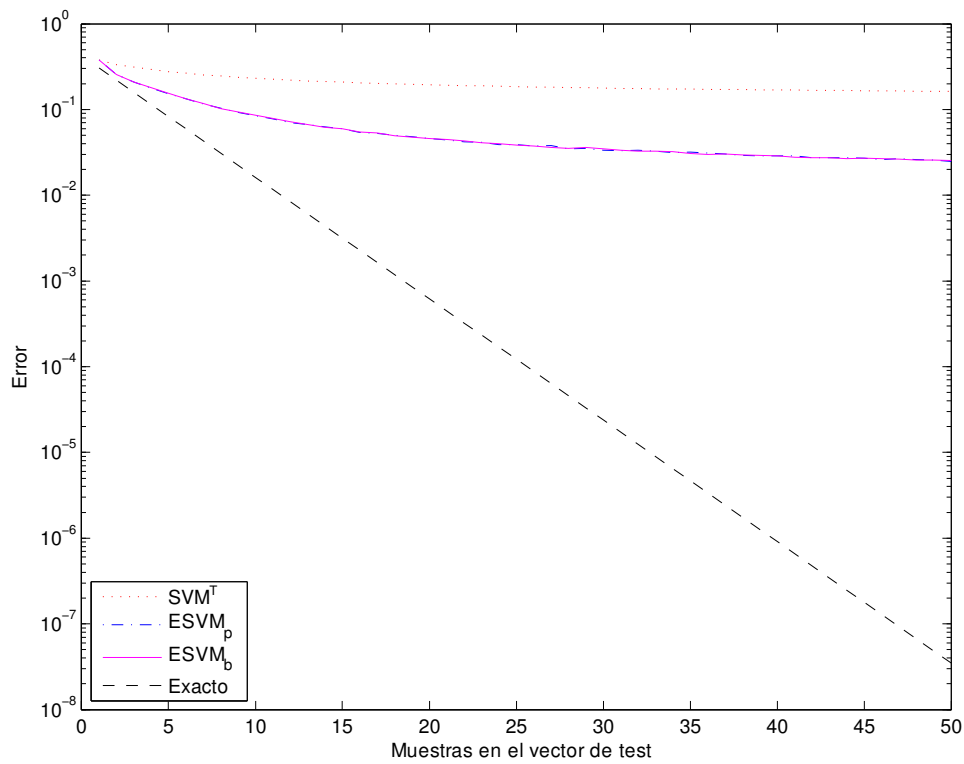
La Figura 4.3(a) muestra los resultados para 40 muestras de entrenamiento y la Figura 4.3(b) para 100, donde se alcanzan menores probabilidades de error. También representamos en ambas figuras el error de Bayes, que cae exponencialmente con L .

En la Figura 4.4 comparamos la diferencia entre SVM, ESVM y ESVM clasificando todas las permutaciones del vector de test y combinando sus salidas probabilísticas. La comparación se realiza hasta vectores de test de 6 muestras y observamos que los métodos que usan las permutaciones del vector de test (subíndice “ b,tp ” y “ p,tb ”) siempre funcionan mejor aunque esta diferencia es pequeña en este caso.

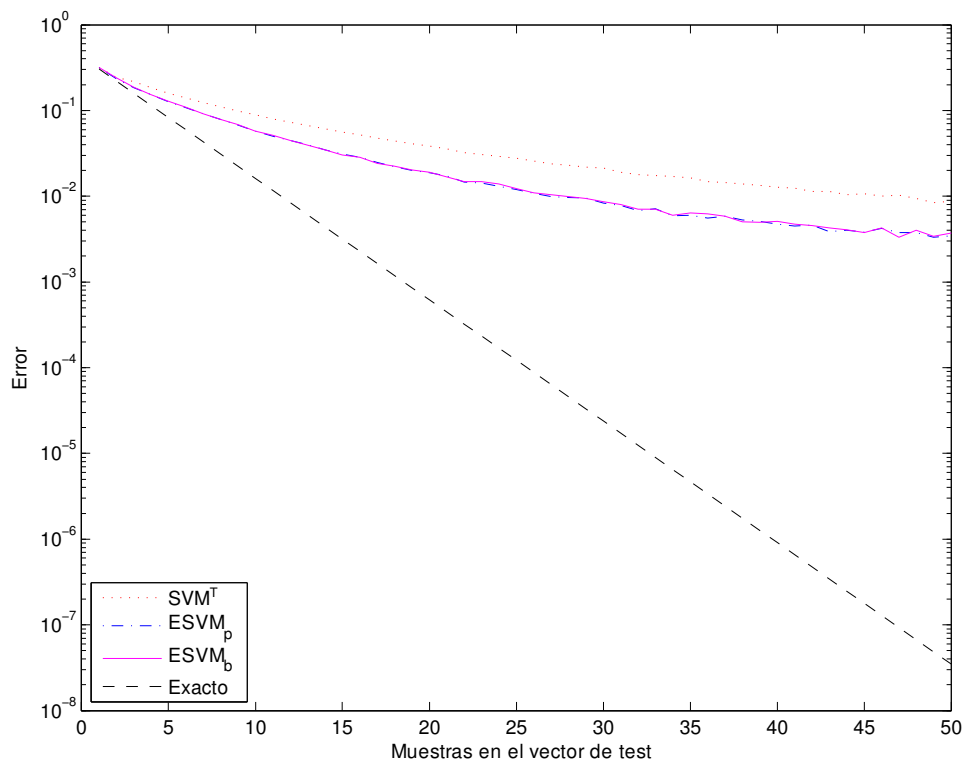
4.3. Experimentos

En esta sección comparamos el error obtenido por la combinación de L salidas de SVM (SVM ^{L}) con el error obtenido por el método extendido: ESVM ^{L} . La comparación se lleva a cabo con las bases de datos empleadas en (Mika *et al.*, 1999). Cada base de datos contiene 100 realizaciones (20 para splice e image) de los conjuntos de entrenamiento y test. El Cuadro 4.2 contiene una descripción de las bases de datos: “Dim” representa el número de dimensiones de los datos; “#entrenamiento” es el número de muestras de entrenamiento; “#test” es el número de muestras de test; y “SVM” se corresponde con la media de su probabilidad de error.

Los experimentos se realizan para vectores de $L = 3$ muestras de test de la misma

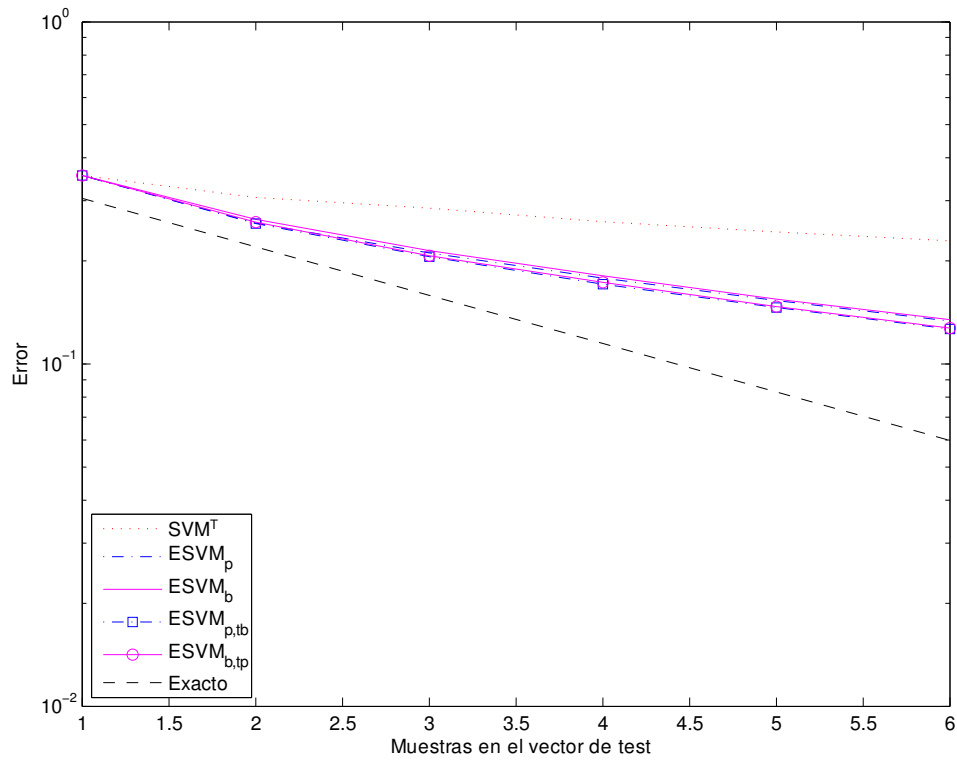


(a) 40 muestras de entrenamiento

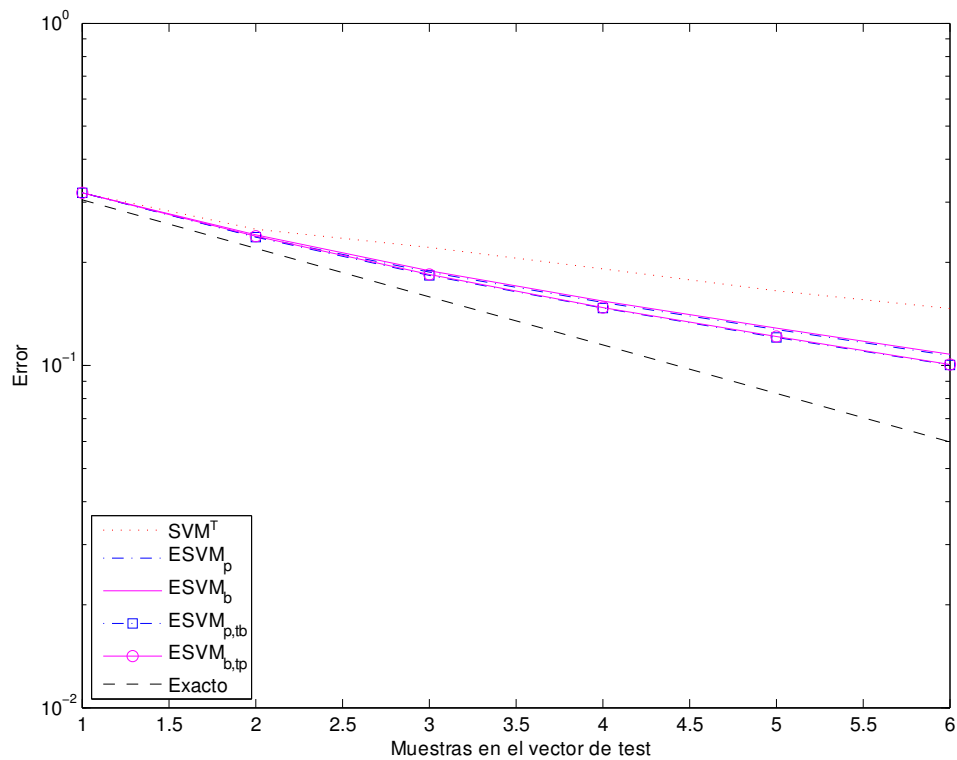


(b) 100 muestras de entrenamiento

Figura 4.3: Probabilidad de error de la combinación de las salidas de la SVM *versus* ESVM, métodos bootstrap y permutación, como función del número de muestras L en el vector de test.



(a) 40 muestras de entrenamiento



(b) 100 muestras de entrenamiento

Figura 4.4: Probabilidad de error de la combinación de las salidas de la SVM *versus* ESVM, métodos bootstrap y permutación, *versus* ESVM, métodos bootstrap y permutación combinando las $L!$ permutaciones del vector de test, como función del número de muestras L en el vector de test.

BD	Dim	#entrenamiento (n)	#test	SVM
breast-cancer	9	200	77	0.252 ± 0.045
diabetis	8	468	300	0.232 ± 0.017
flare-solar	9	666	400	0.323 ± 0.018
german	20	700	300	0.241 ± 0.022
titanic	3	150	2051	0.228 ± 0.012
waveform	21	400	4600	$0.098 \pm 4.35e-3$
thyroid	5	140	75	0.046 ± 0.021
twonorm	20	400	7000	$0.024 \pm 1.4e-3$
ringnorm	20	400	7000	$0.015 \pm 9.46e-4$
heart	13	170	100	0.155 ± 0.034
banana	2	400	4900	$0.109 \pm 5.55e-3$
splice	60	1000	2175	$0.108 \pm 7.42e-3$
image	18	1300	1010	$0.032 \pm 6.08e-3$

Cuadro 4.2: Descripción de las bases de datos y la media y desviación típica de las probabilidades de error obtenidas por la SVM.

clase. Las salidas probabilísticas de la SVM $p(y = 1|x_i^*)$ se combinan para obtener $p(y = 1|x_i^*, \dots, x_{i+L-1}^*)$. La extensión del conjunto de entrenamiento se realiza como se ha explicado en la Sección 4.1. La SVM emplea un *kernel* gaussiano y sus parámetros se ajustan por validación cruzada en los conjuntos de entrenamiento. Las salidas probabilísticas se obtienen con el método de Platt (2000). Para los experimentos hemos empleado la implementación de la SVM del paquete LibSVM (véase Chang y Lin, 2001).

El primer experimento comparara SVM³ y ESVM³ cuando se emplea un subconjunto de las muestras de entrenamiento para ajustar los clasificadores. Los resultados se resumen en el Cuadro 4.3. En este experimento hemos empleado el método de permutación para generar el conjunto extendido. En concreto, hemos generado $24 \cdot n'$ muestras donde n' es el número de muestras de dicho subconjunto. Como apreciamos en el cuadro, el método extendido es claramente superior para las bases de datos: breast-cancer, flare-solar, titanic y ringnorm (menos cuando se emplea el 10% en esta última). Es también superior cuando se emplean el 10% de las muestras en las bases: german, thyroid y heart. Funciona peor en las bases: german, thyroid, image, twonorm (en la que los parámetros escogidos por validación cruzada pueden ir mal) y heart. No hay diferencias estadísticamente relevantes en las bases: diabetes, waveform, banana y splice. A la vista de estos resultados podemos decir que extender el conjunto de entrenamiento es ventajoso en algunos casos, especialmente, si hay pocas muestras.

A continuación comparamos tres formas de extender el conjunto de entrenamiento: el método bootstrap, el método de permutación y el método de permutación

BD	Método	Porcentaje de muestras de entrenamiento					
		10 %	20 %	50 %	66 %	75 %	100 %
breast-cancer	ESVM ³	0.272	0.239	0.203	0.200	0.187	0.187
	SVM ³	0.318	0.264	0.223	0.212	0.207	0.189
diabetis	ESVM ³	0.191	0.154	0.133	0.128	0.126	0.122
	SVM ³	0.195	0.142	0.129	0.125	0.124	0.123
flare-solar	ESVM ³	0.244	0.216	0.201	0.195	0.192	0.189
	SVM ³	0.426	0.269	0.230	0.230	0.231	0.231
german	ESVM ³	0.222	0.188	0.170	0.162	0.161	0.155
	SVM ³	0.232	0.178	0.158	0.155	0.152	0.146
titanic	ESVM ³	0.254	0.192	0.159	0.151	0.144	0.139
	SVM ³	0.297	0.214	0.199	0.191	0.184	0.185
waveform	ESVM ³	0.051	0.019	0.014	0.012	0.012	0.011
	SVM ³	0.059	0.022	0.015	0.013	0.013	0.011
thyroid	ESVM ³	0.022	3.27e-03	2.48e-03	8.33e-04	1.25e-03	1.25e-03
	SVM ³	0.059	5.32e-03	2.07e-03	2.45e-03	8.33e-04	4.00e-04
twonorm	ESVM ³	0.114	9.40e-03	2.26e-03	1.25e-03	8.06e-04	5.53e-04
	SVM ³	0.011	6.47e-04	3.43e-04	4.54e-04	3.90e-04	3.64e-04
ringnorm	ESVM ³	1.62e-03	1.76e-04	7.72e-05	9.00e-05	6.43e-05	5.57e-05
	SVM ³	6.00e-04	2.19e-04	3.04e-04	3.47e-04	3.99e-04	3.52e-04
heart	ESVM ³	0.164	0.102	0.085	0.073	0.073	0.069
	SVM ³	0.476	0.084	0.056	0.056	0.051	0.053
banana	ESVM ³	0.129	0.057	0.030	0.024	0.023	0.020
	SVM ³	0.116	0.040	0.024	0.021	0.020	0.018
splice	ESVM ³	0.097	0.053	0.033	0.028	0.027	0.023
	SVM ³	0.078	0.043	0.026	0.022	0.021	0.017
image	ESVM ³	0.033	0.017	0.012	8.04e-03	7.74e-03	6.40e-03
	SVM ³	0.029	6.99e-03	4.76e-03	3.42e-03	1.19e-03	1.04e-03

Cuadro 4.3: Probabilidades de error medias de ESVM³ y SVM³ variando el porcentaje empleado de las muestras de entrenamiento. ESVM mediante permutación.

imponiendo la simetría. Tanto para el método bootstrap como para el método de permutación, se obtienen las salidas de $L!$ permutaciones de cada vector y se decide a partir de la más extrema. El método de permutación empleando las simetrías considera todas las permutaciones de cada muestra de entrenamiento del conjunto extendido. Los resultados se muestran los Cuadros 4.4, 4.5 y 4.6. Apreciamos que el método de la permutación y el método bootstrap funcionan aproximadamente de la misma forma: el primero gana ligeramente en las bases: breast-cancer, diabetis, flare-solar, german y waveform; y el segundo gana en titanic, twonorm e image. La comparación ente el método de permutación (solamente considerando la tabla con $N_p = 3$) y el método de permutación imponiendo las simetrías (considerando para tener el mismo número de muestras de entrenamiento $N_p \leq 9$) encontramos que imponer las simetrías funciona mejor en las bases: titanic, thyroid, twonorm, bana-

na, splice e image; funciona peor en las bases: breast-cancer, diabetis, flare-solar, german, waveform, ringnorm y heart. Así y todo las diferencias no son muy significativas en ningún caso. A la vista de estos resultados comprobamos la conjetura de que el método de permutación funciona mejor que el método de bootstrap con menos muestras. La simetría funciona peor en 7/13 bases y mejor en 6/13 con lo que su ventaja se presenta en la fase del test porque no es necesario evaluar todas las permutaciones de la muestra para escoger la mejor. En muchos casos el mejor resultado no se obtiene cuando el conjunto extendido se construye a partir del mayor número de permutaciones. Conjeturamos que esto se debe a un efecto de saturación combinado con que la elección de los hiperparámetros se realizó para un menor número de muestras de entrenamiento que las empleadas aquí.

BD	ESVM ³				
	$N_p = 3$	$N_p = 8$	$N_p = 13$	$N_p = 18$	$N_p = 23$
breast-cancer	0.188	0.181	0.182	0.185	0.186
diabetis	0.119	0.121	0.121	0.123	0.126
flare-solar	0.188	0.190	0.191	0.193	0.195
german	0.148	0.152	0.158	0.161	0.163
titanic	0.139	0.138	0.135	0.136	0.137
waveform	0.011	0.011	0.011	0.011	0.011
thyroid	1.65e-03	1.23e-03	1.25e-03	8.33e-04	4.17e-04
twonorm	6.56e-04	5.79e-04	5.40e-04	5.10e-04	4.97e-04
ringnorm	6.00e-05	5.14e-05	5.14e-05	6.00e-05	7.29e-05
heart	0.058	0.066	0.069	0.073	0.078
banana	0.020	0.017	0.015	0.015	0.015
splice	0.025	0.023	0.021	0.020	0.021
image	7.44e-03	5.80e-03	5.21e-03	4.91e-03	3.87e-03

Cuadro 4.4: Probabilidades de error medias de la ESVM entrenada mediante $3 * N_p * n$ muestras generadas empleando el método de *bootstrap*.

Ahora comparamos en el Cuadro 4.7 los resultados de SVM³, ESVM³ extendiendo el conjunto de entrenamiento mediante el método de la permutación, y el método máxima discrepancia en media (*maximum mean discrepancy*) (MMD) para el problema de dos muestras (que aparece en el cuadro como MMD³)(Gretton *et al.*, 2007). Para el algoritmo MMD hemos empleado el código de los autores² y seleccionado un *kernel* gaussiano cuyo ancho es ajustado automáticamente por el método para cada vector de test. Para predecir la clase de cada vector de test el MMD emplea las muestras de entrenamiento de las clases y selecciona la clase más verosímil. Podemos apreciar en el cuadro que la ESVM³ es mucho más ventajosa que los otros métodos

²<http://www.kyb.mpg.de/~arthur>.

BD	ESVM ³				
	$N_p = 3$	$N_p = 8$	$N_p = 13$	$N_p = 18$	$N_p = 23$
breast-cancer	0.188	0.180	0.183	0.185	0.186
diabetis	0.118	0.120	0.122	0.123	0.126
flare-solar	0.187	0.190	0.191	0.193	0.194
german	0.147	0.153	0.156	0.160	0.161
titanic	0.138	0.136	0.136	0.136	0.137
waveform	0.011	0.010	0.011	0.011	0.012
thyroid	1.65e-03	4.17e-04	8.33e-04	4.17e-04	8.17e-04
twonorm	6.56e-04	5.57e-04	5.27e-04	5.14e-04	5.06e-04
ringnorm	5.14e-05	5.57e-05	6.86e-05	6.86e-05	7.29e-05
heart	0.058	0.068	0.072	0.073	0.077
banana	0.020	0.017	0.015	0.015	0.015
splice	0.025	0.023	0.022	0.020	0.020
image	7.74e-03	5.21e-03	4.46e-03	4.46e-03	4.17e-03

Cuadro 4.5: Probabilidades de error medias de la ESVM entrenada mediante $3*N_p*n$ muestras generadas empleando el método de permutación.

en las bases de datos: flare-solar, titanic y ringnorm. Además obtiene el mejor resultado en siete de las trece bases de datos. Por otro lado, SVM³ gana claramente en la base de datos image y gana en cinco de las trece bases de datos. El método MMD³ obtiene las mejores prestaciones en twonorm pero la ventaja no es estadísticamente significativa y funciona sensiblemente peor en la mayoría de las otras bases de datos.

4.4. Conclusiones

Hemos considerado la clasificación de conjuntos de muestras de la misma clase desde el punto de vista de aprendizaje. Hemos visto que entrenar clasificadores para esta tarea no es excesivamente complejo y que la incorporación de simetrías no añade variables al problema de optimización en el caso de la SVM. Los experimentos que hemos realizado indican que en ocasiones puede ser ventajoso emplear la ESVM en lugar de combinar las salidas individuales de cada muestra de test. La ESVM solo necesita predecir correctamente la frontera, algo que los clasificadores basados en aprendizaje hacen bien, mientras que para combinar las salidas probabilísticas de cada muestra la SVM necesita predecir con precisión todo el rango de probabilidades.

BD	ESVM ³			
	$N_p = 3$	$N_p = 6$	$N_p = 9$	$N_p = 12$
breast-cancer	0.189	0.192	0.192	0.192
diabetis	0.123	0.124	0.128	0.127
flare-solar	0.188	0.189	0.191	0.192
german	0.155	0.157	0.163	0.166
titanic	0.141	0.137	0.139	0.140
waveform	0.015	0.016	0.017	0.018
thyroid	8.17e-04	4.17e-04	8.33e-04	4.17e-04
twonorm	3.73e-04	3.47e-04	3.77e-04	4.16e-04
ringnorm	6.00e-05	6.43e-05	9.00e-05	8.15e-05
heart	0.070	0.074	0.075	0.079
banana	0.024	0.020	0.019	0.018
splice	0.028	0.025	0.023	0.021
image	7.14e-03	5.80e-03	4.46e-03	4.32e-03

Cuadro 4.6: Probabilidades de error medias de la ESVM simétrica entrenada con $N_p * n$ muestras empleando el método de permutación para construir el conjunto extendido.

BD	SVM ³	ESVM ³	MMD ³
breast-cancer	0.189 ± 0.057	0.180 ± 0.065	0.22 ± 0.075
diabetis	0.122 ± 0.031	0.118 ± 0.028	0.158 ± 0.033
flare-solar	0.231 ± 0.057	0.187 ± 0.030	0.251 ± 0.034
german	0.146 ± 0.032	0.147 ± 0.034	0.174 ± 0.036
titanic	0.185 ± 0.041	0.136 ± 0.023	0.196 ± 0.034
waveform	0.011 ± 4.42e-3	0.010 ± 2.79e-03	0.039 ± 7.82e-3
thyroid	4.00e-4 ± 0.010	4.17e-04 ± 4.17e-03	0.012 ± 0.024
twonorm	3.64e-4 ± 3.50e-4	5.06e-04 ± 5.63e-04	3.34e-4 ± 3.2e-4
ringnorm	3.51e-4 ± 4.65e-4	5.14e-05 ± 1.40e-04	0.016 ± 4.44e-3
heart	0.052 ± 0.037	0.058 ± 0.040	0.058 ± 0.038
banana	0.017 ± 4.52e-3	0.015 ± 3.31e-03	0.126 ± 0.029
splice	0.016 ± 4.77e-3	0.020 ± 4.14e-03	0.086 ± 0.014
image	1.04e-3 ± 1.4e-3	4.17e-03 ± 4.57e-03	0.092 ± 0.015

Cuadro 4.7: Error medio y desviación típica para los métodos SVM³, ESVM³ empleando el método de la permutación para construir el conjunto extendido y MMD³.

Capítulo 5

Diagnóstico automático de la tuberculosis

En este capítulo proponemos un sistema automático para el diagnóstico de la tuberculosis. Este sistema consta de dos partes: un clasificador de bacilos que implementaremos con una SVM que analiza parches que provienen de una ventana que se desliza sobre la imagen y un test secuencial que fusiona la información proveniente del clasificador de bacilos y continúa la adquisición de imágenes mientras no se alcance la certidumbre deseada sobre la decisión. En lo que resta de capítulo veremos por un lado, la arquitectura del clasificador local, formado por una extracción lineal de características y una SVM implementada en cascada; y por otro, las alternativas, desde el enfoque frecuentista y bayesiano, del test secuencial.

5.1. El problema

Se dispone de un conjunto de imágenes microscópicas del esputo de un paciente. El objetivo es el diseño de un sistema capaz de proporcionar una alta fiabilidad en el diagnóstico de un paciente de tuberculosis. El sistema adquirirá y analizará tantas imágenes como sea necesario para alcanzar la citada fiabilidad. En caso de terminar las imágenes disponibles o alcanzar un número de imágenes máximo, el sistema proporcionará una decisión y la calidad alcanzada. El protocolo médico recomienda analizar unas 100 imágenes por paciente. El sistema automático puede diseñarse para analizar 300 imágenes antes de diagnosticar al paciente. Si se trata de un paciente claramente positivo, no hacen falta muchas imágenes para confirmar.

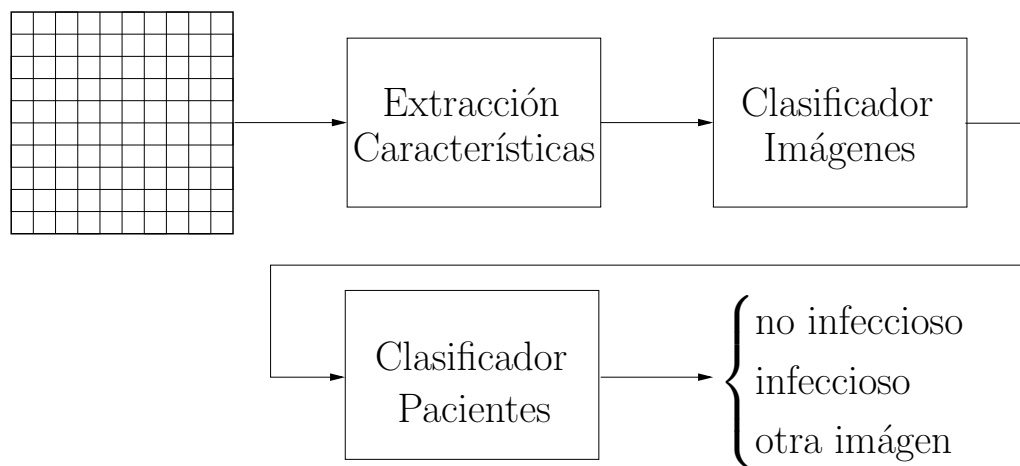


Figura 5.1: Sistema automático de diagnóstico de tuberculosis.

5.2. Sistema propuesto

El sistema mostrado en la Figura 5.1 sigue los siguientes pasos:

1. La imagen es dividida en parches como muestra la Figura 5.2. Para evitar perder los bacilos que puedan aparecer en la intersección de dos parches, hacemos varias pasadas de la imagen desplazando el origen de la rejilla.
2. Cada parche es examinado por el clasificador de bacilos que envía su salida al clasificador de pacientes.
3. El clasificador de pacientes actualiza su certidumbre sobre el estado del paciente con cada nuevo parche. Si la certidumbre es suficiente para tomar una decisión sobre el paciente el algoritmo termina. En caso contrario es necesario examinar más parches.

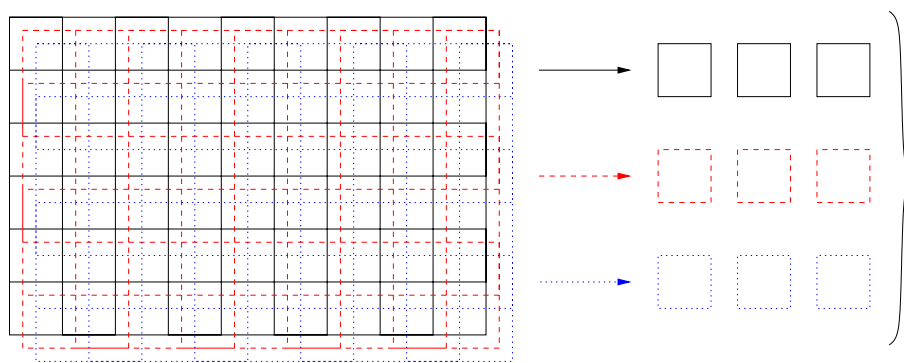


Figura 5.2: División de una imagen en parches.

5.3. Clasificador de parches

El clasificador de bacilos se encarga de determinar si en cada trozo de imagen está presente un bacilo. La arquitectura seleccionada para el clasificador de bacilos es la máquina de vectores soporte. Para el clasificador de parches se ha seleccionado un tamaño de ventana de 37x37 pixels con el fin de que un bacilo centrado en una ventana entre cómodamente si se emplea un objetivo de 20 aumentos en el microscopio. Se usan imágenes RGB aunque la mayor cantidad de información se espera en el verde que es el color predominante de la tinción mediante auramina.

5.3.1. Datos

Los pacientes que forman la base de datos se dividen de la siguiente manera:

- Secuencias de imágenes de pacientes infecciosos de entrenamiento (9): 31087, 32741, 46989, 48940, 51859, 52707, 31404, 32742, 51861.
- Secuencias imágenes de pacientes infecciosos de test (7): 32743, 40279, 50304, 51862, 55165, 55798, 56036.
- Secuencias de imágenes de pacientes no infecciosos de entrenamiento (34): 32550, 30663, 29994, 31060, 30261, 30262, 30619, 30214, 30855, 30881, 31547, 31934, 32497, 31522, 30207, 30304, 30783, 30825, 30880, 30911, 31230, 31549, 32111, 32748, 30986, 30998, 30877, 31245, 32642, 32688, 32750, 32955, 51523, 38841.
- Secuencias de imágenes de pacientes no infecciosos de entrenamiento (15): 30257, 30266, 30295, 30665, 30725, 31228, 31523, 31534, 31684, 31819, 32240, 32633, 32781, 32956, 36856.

Para los ejemplos positivos, se han añadido parches centrados en el bacilo declarados como tales por un experto. El centrado se obtiene mediante técnicas de segmentación estándar (véase Figura 5.3). Para los ejemplos negativos se han añadido parches escogidos aleatoriamente de imágenes de pacientes sanos y también parches escogidos aleatoriamente de pacientes sanos cuya energía supere un umbral con el fin de evitar entrenar con muchos parches demasiado oscuros que son la mayoría y aportan escasa información.

5.3.2. Entrenamiento

De los pacientes negativos disponibles, 15 se dedican a test y el resto (34) a entrenamiento. De los pacientes positivos disponibles, 7 se dedican a test y el resto

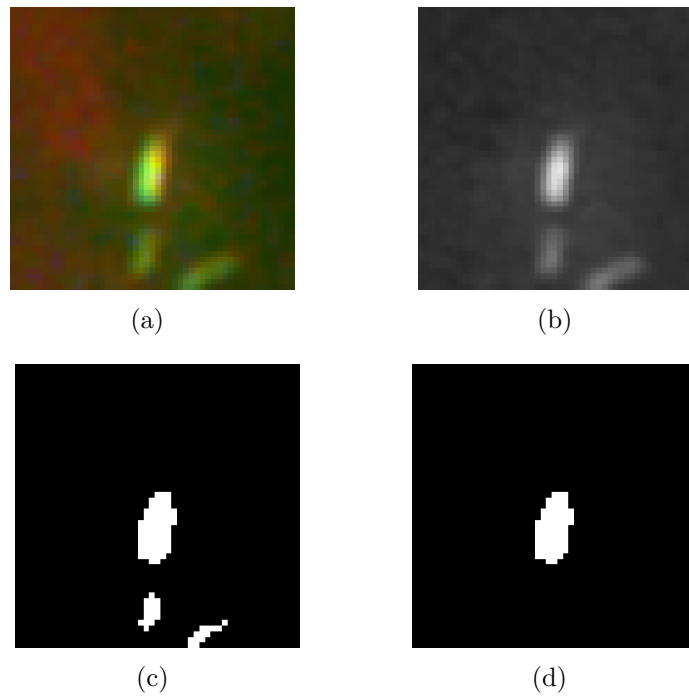


Figura 5.3: Segmentación del bacilo: (a) parche original RGB, (b) imagen en escala de grises, (c) imagen binaria de para posibles candidatos, (d) región más probable del bacilo. De la imagen en (d) se obtiene el centro del bacilo, el parche de este bacilo se obtiene como la ventana de 37x37 pixels de la imagen centrada en él.

(9) a entrenamiento. Se hace notar que la base de datos de pacientes negativos es mucho mayor que la de los pacientes positivos ya que lo que perseguimos es una alta especificidad.

5.3.2.1. División del conjunto de entrenamiento

Dada la escasez de ejemplos positivos hemos dividido las muestras positivas por un lado y las negativas por otro. Las secuencias de pacientes negativos de entrenamiento y test se dividen como indica la Figura 5.4.

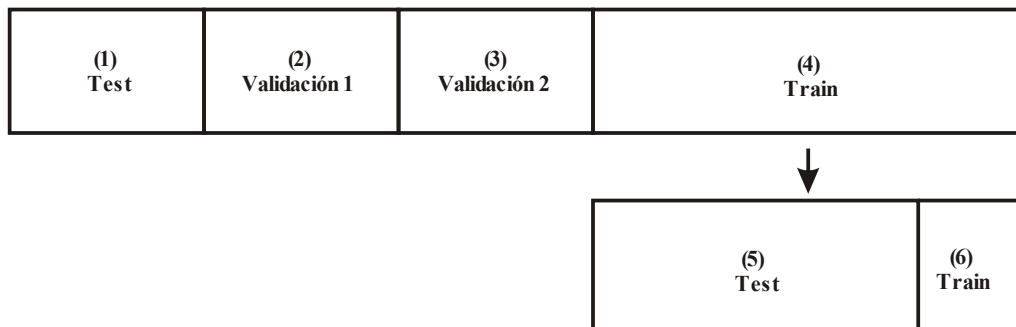


Figura 5.4: División del conjunto de bacilos disponibles.

Se han generado 4 grupos:

Test (1). Para evaluar las prestaciones de nuestro clasificador una vez concluido.

Validación 1 (2). Para ajustar el clasificador.

Validación 2 (3). Para ajustar el clasificador con preimagen y/o cascada.

Entrenamiento (4). Conjunto de entrenamiento con el que se obtiene la extracción de características y se hace una búsqueda preliminar de los hiperparámetros de la SVM. Para ello se divide de forma aleatoria a su vez en dos subconjuntos:

Entrenamiento Búsqueda (5). Subconjunto con el 20% de las muestras para entrenar.

Test Búsqueda (6). Subconjunto con el resto de las muestras para validar la selección.

Por otro lado, los 945 bacilos se han repartido en:

Test (1): 170.

Validación 1 (2): 170.

Validación 2 (3): 170.

Entrenamiento (4): 435.

Además, crearemos muestras virtuales a partir de las muestras de bacilo a fin de enriquecer la base de datos¹ por medio de traslaciones y rotaciones teniendo cuidado que el resultado sean bacilos enteros dentro del parche. De este modo se generan 56154 muestras de bacilos que se reparten estando cada bacilo con sus muestras virtuales del siguiente modo:

Test (1): 10177.

Validación 1 (2): 10167.

Validación 2 (3): 10082.

Entrenamiento (4): 25728.

Entrenamiento Búsqueda (5): 5000.

¹Decoste y Schölkopf (2002) recomiendan añadir sólo vectores soporte virtuales en lugar de muestras virtuales. Es más, recomiendan generar vectores soporte virtuales de ambas clases.

Test Búsqueda (6): 20728.

En total se tienen 63438 muestras de fondo que se reparten:

Test (1): 11227.

Validación 1 (2): 10348.

Validación 2 (3): 10857.

Entrenamiento (4): 31006.

5.4. Clasificador de pacientes

Como hemos dicho, el clasificador de pacientes combina la salida del clasificador local hasta que se ha acumulado suficiente información para tomar una decisión acerca del paciente. La herramienta que hemos elegido para esta tarea es un test secuencial. El entrenamiento del test secuencial se realiza con la salida blanda de la SVM para todos los pacientes de entrenamiento.

5.4.1. Test secuencial clásico para variables binarias con incertidumbre

El Algoritmo 1 muestra la modificación que hemos propuesto del test secuencial clásico para variables binarias. Esto equivale a establecer un modelo binario para ambos tipos de pacientes. Los pasos que damos son los siguientes:

1. Pasamos la salida blanda por un umbral buscando la mayor discriminación entre pacientes sanos y enfermos. Este umbral se encuentra empíricamente a partir de los pacientes de entrenamiento.
2. Estimamos intervalos de confianza para las probabilidades de aparición de bacilo para ambos tipos de pacientes. Esto es una aproximación de la tasa de falsa alarma para un paciente sano y se corresponde con la probabilidad de aparición de bacilo más la tasa de falsa alarma para un paciente enfermo. Cada intervalo se ajusta de modo que tengan una confianza parecida y que proporcionen el producto de confianzas P_c más alto posible sin solaparse. Hemos seleccionado el método de Cai y Krishnamoorthy (2005) para el cálculo de los intervalos de confianza.

Los pacientes son muy diferentes tanto con los de la otra clase como entre sí mismos. De hecho, es posible que en pacientes no contagiosos se detecten más

bacilos que en pacientes contagiosos. Este hecho trae consigo las siguientes alternativas para la selección de los intervalos de confianza que caracterizan las clases:

- Usar todos los pacientes positivos como un sólo paciente y lo mismo para los pacientes negativos. Esta posibilidad evita la elección del paciente prototipo de cada clase. Sin embargo, esta opción puede causar la mala clasificación de pacientes que se desvíen del comportamiento medio hacia la otra clase.
 - Seleccionar el paciente de una clase que sea más parecido a la otra. Esta es la opción que nos parece más adecuada si las clases no se solapan. En caso de solape entre las clases, debemos escoger como pacientes prototipo aquéllos que satisfagan los requisitos de precisión o sensibilidad o especificidad que deseemos.
3. El test secuencial se ajusta teniendo en mente que no se detenga antes de analizar el número de imágenes estipulado para declarar un paciente sano y que termine en cuanto sea posible para declararlo enfermo. Si no hubiese incertidumbre, esto equivaldría a una probabilidad de no detección muy baja y a una probabilidad de falsa alarma razonable. Sin embargo, al considerar la incertidumbre, llega con hacer dichas probabilidades despreciables frente a P_c . El umbral superior del test secuencial B está relacionado con la probabilidad de falsa alarma, de (3.12) obtenemos:

$$P_{FA,end} \leq P_c P_{FA} + (1 - P_c)$$

Si fijamos $\epsilon = P_c P_{FA}$ despreciable con respecto a $(1 - P_c)$ obtenemos un valor razonable de la P_{FA} que debemos pedir al test secuencial

$$P_{FA} = \frac{\epsilon}{P_c}$$

de donde, teniendo en cuenta (2.6), calculamos el valor del umbral superior del test secuencial para el ϵ que nos parezca razonable como:

$$B = \frac{P_c}{\epsilon}.$$

Por otro lado, para el valor de A (que declara al paciente sano) queremos ser conservadores. Una posibilidad consiste en establecer un valor que asegure el análisis del número de imágenes requerido antes de declarar un paciente sano.

Otra posibilidad consiste en seguir un procedimiento análogo al anterior: si partimos de (3.11) tenemos

$$P_{D,end} \geq P_c P_D$$

Deseamos que $P_c P_D \approx P_c$, que equivale a que su diferencia sea un valor pequeño ϵ' . Por tanto, $\epsilon' = P_c(1 - P_D)$ y

$$P_D = 1 - \frac{\epsilon'}{P_c}.$$

El valor de A se obtiene de (2.6) como:

$$A = \frac{\epsilon'}{P_c}.$$

4. Si el test no termina antes del fin de las imágenes, proporcionamos la decisión final y la calidad de dicha decisión. La decisión se toma a partir de qué umbral está más cercano a ser cruzado.

Algoritmo 1 Test secuencial binario con incertidumbre

Entradas: $\{x_1, x_2, \dots, x_N\}$ y $\{y_1, y_2, \dots, y_M\}$; muestra de test $\{z_1, z_2, \dots, z_K\}$ (adquirida secuencialmente); requisitos: P_{FA} y P_D

Salidas: decisión del paciente, $P_{FA,end}$ y $P_{D,end}$

- 1: Convertir las muestras a binario
 - 2: Obtiene los intervalos de confianza para p^* y q^*
 - 3: Obtiene el producto de confianzas P_c
 - 4: Obtiene los umbrales A y B del test secuencial
 - 5: $k \leftarrow 1$
 - 6: **Mientras** ($k < K$) y ($A < SPRT(p^*, q^*, P_{FA}, P_D) < B$) {
 - 7: Obtiene la muestra z_k
 - 8: Actualiza el cociente de verosimilitud
 - 9: $k \leftarrow k + 1$
 - 10: }
 - 11: **Si** Se han alcanzado los requisitos {
 - 12: Devolver la decisión y las prestaciones alcanzadas
 - 13: }
 - en otro caso** {
 - 14: Obtener la decisión a partir del cociente actual
 - 15: Devolver las prestaciones moviendo los umbrales hasta que alguno de ellos toque el valor final del cociente de verosimilitud
 - 16: }
-

Estamos suponiendo que las probabilidades que se piden al test son alcanzables. Una vez obtenida P_c ya sabemos a qué prestaciones podemos aspirar.

El caso discreto no binario es análogo al binario sin más que cuantificar las señales en distintos niveles. Para obtener los niveles óptimos es necesario conocer la fdp de los datos. Por tanto, se puede seguir el criterio que más nos satisfaga como: hacer todos los niveles iguales, hacer que tengan las mismas muestras, usar recursivamente la estrategia del caso binario, *etc.*

5.4.1.1. Estima del número de muestras necesarias

El número de muestras necesarias para tomar una decisión, dependerá de la muestra de test y de P_c , o lo que es lo mismo, de lo solapados que se encuentren los intervalos que definen las hipótesis. La estima que proponemos aproxima el número de muestras por su valor esperado si realizásemos un contraste de hipótesis simples caracterizadas por p^*_h y q^*_l .

Si contrastamos:

H_0 : z viene de una distribución Bernoulli cuyo parámetro es p^*_h .

H_1 : z viene de una distribución Bernoulli cuyo parámetro es q^*_l .

la función característica de operación, $L(p)$, (*del inglés operating characteristic function*), que proporciona la probabilidad de no rechazar H_0 en función del valor real de p para la muestra de test, se puede obtener a partir de las siguientes expresiones (Wald, 1947)

$$L(p) = \frac{\left(\frac{1-\gamma}{\alpha}\right)^h - 1}{\left(\frac{1-\gamma}{\alpha}\right)^h - \left(\frac{\gamma}{1-\alpha}\right)^h}$$

$$p = \frac{1 - \left(\frac{1-q^*_l}{1-p^*_h}\right)^h}{\left(\frac{q^*_l}{p^*_h}\right)^h - \left(\frac{1-q^*_l}{1-p^*_h}\right)^h},$$

dando valores a h . Para $h = \infty, 1, -1, -\infty$ se obtienen los valores para $p = 0, p^*_h, q^*_l, 1$. Dado el punto $[p, L(p)]$ correspondiente a h , el punto $[p', L(p')]$ correspondiente a $-h$ se obtiene como $p' = \left(\frac{q^*_l}{p^*_h}\right)^h p$ y $L(p') = \left(\frac{\gamma}{1-\alpha}\right)^h L(p)$.

Finalmente, la estima del valor esperado del número de muestras de test en función de p resulta:

$$E_p[K] = \frac{L(p) \log \frac{\gamma}{1-\alpha} + (1 - L(p)) \log \frac{1-\gamma}{\alpha}}{p \log \frac{q^*_l}{p^*_h} + (1 - p) * \log \frac{1-q^*_l}{1-p^*_h}}.$$

De dónde obtenemos la estima del número de muestras como la media de dicho valor para q_l^* y p_h^* .

5.4.2. Test secuencial bayesiano para variables binarias con incertidumbre

Planteamos ahora el test bayesiano. El test parte de las muestras de entrenamiento y el conocimiento *a priori* del que se disponga y devuelve la probabilidad de que el paciente esté sano. El Algoritmo 2 resume este test para el caso de dos pacientes prototipo.

Algoritmo 2 Test secuencial bayesiano binario caracterizando las hipótesis con dos pacientes prototipo.

Entradas: $\{x_1, x_2, \dots, x_N\}$ y $\{y_1, y_2, \dots, y_M\}$; muestra de test $\{z_1, z_2, \dots, z_K\}$ (adquirida secuencialmente); probabilidad mínima P_{min} para terminar el test; distribuciones *a priori* de las clases y de la probabilidad de bacilo.

Salidas: decisión del paciente, probabilidad de la decisión.

- 1: Convertir las muestras a binario
 - 2: $k \leftarrow 1$
 - 3: **Mientras** ($k < K$) y ($P(H_0|\mathbf{z}) < P_{min}$) {
 - 4: Obtiene la muestra z_k
 - 5: Actualiza el *posterior* de las hipótesis.
 - 6: $k \leftarrow k + 1$
 - 7: }
 - 8: Devolver la decisión y la probabilidad $P(H_0|\mathbf{z})$
-

Los pasos que seguimos en este test son los siguientes:

1. En caso de caracterizar las clases con un paciente prototipo, como hicimos en el caso frecuentista, escogemos dichos pacientes y las distribuciones *a priori*. Aunque esperamos que la probabilidad de aparición de bacilo sea pequeña podemos escoger distribuciones *a priori* no informativas tanto para los pacientes sanos como para los enfermos.

Dado que no tenemos un único paciente de cada clase, seguimos otra posibilidad que la metodología bayesiana ofrece para tratar las diferencias entre pacientes de la misma clase. Representamos al paciente k por el par (N^k, N_1^k) que nos da respectivamente la información de cuántos parches se han analizado y cuántos han sido declarados como bacilo en ese paciente. A partir de estas cantidades podemos aproximar su probabilidad de bacilo $p^k \approx N_1^k/N^k$ y asumir que la probabilidad de bacilo de cada paciente es una realización

de la distribución *a posteriori* de la probabilidad de bacilo $P(p|\mathbf{x})$ que aproximaremos por una distribución beta(a, b) (el equivalente a (3.2) para varios pacientes) cuyos parámetros estimamos por máxima verosimilitud a partir de los pacientes de entrenamiento.

2. Pasamos las muestras por un umbral para obtener una muestra binaria como en el caso frecuentista.
3. Una vez escogida la forma del test, procesamos muestras hasta que la probabilidad de alguna hipótesis cumpla los requisitos o se terminen las muestras. A partir del conocimiento *a priori* y de los datos de entrenamiento que caracterizan las hipótesis pueden obtenerse qué prestaciones máximas podemos esperar de (3.10).

En el caso discreto no binario seguimos teniendo, al igual que para el método frecuentista, la complicación de cómo establecer los escalones de cuantificación. Por otro lado, toda la metodología es idéntica que la del caso binario estableciendo las distribuciones *a priori* correspondientes.

5.5. Experimentos

En esta sección mostramos los experimentos que hemos realizado para ajustar el sistema. En cuanto al detector de bacilos hemos dividido los experimentos en la extracción de características, el ajuste del clasificador y su reducción de complejidad. En cuanto a la fusión de datos hemos evaluado test secuenciales que tengan en cuenta la incertidumbre siguiendo el método frecuentista y el método bayesiano.

5.5.1. Extracción de características

La extracción de características tiene un propósito doble: disminuir la dimensión de los datos y con ello, aliviar la carga computacional y facilitar la tarea del clasificador. La extracción de características es un paso clave en esta aplicación por el requisito de tiempo real y la necesidad de mantener la información discriminante.

Aquí hemos barajado dos alternativas:

- Emplear una extracción de características no lineal y a continuación emplear un clasificador lineal. Un ejemplo de esto sería usar análisis de componentes principales basado en núcleos (*kernel principal component analysis*) (KPCA)

como extracción de características y una SVM lineal como clasificador (Schölkopf y Smola, 2001). Esta opción lleva la complejidad a la extracción de características y deja la clasificación como un proceso rápido que consiste simplemente en la evaluación de un vector soporte.

- La otra opción es emplear una extracción de características lineal, en la que la matriz de transformación se obtiene durante el entrenamiento. Así la extracción de características es simplemente el producto de dos matrices. Siguiendo esta opción podemos permitirnos un clasificador más complejo como puede ser una SVM con un *kernel* no lineal.

Nosotros hemos apostado por la segunda alternativa debido a su sencillez y a que posteriormente es posible reducir la complejidad del clasificador mediante técnicas de reducción del conjunto de vectores soporte y la disposición de clasificadores en cascada.

La selección de un buen algoritmo de extracción de características es clave para obtener buenas prestaciones en la posterior clasificación. Hemos probado los métodos análisis discriminante lineal (*linear discriminant analysis*) (LDA) (Fukunaga, 1990), análisis de componentes independientes (*independent component analysis*) (ICA) (Hyvarinen *et al.*, 2001), maximización de la información mutua (*maximization of mutual information*) (MMI) (Leiva-Murillo y Artés-Rodríguez, 2004) y análisis de componentes principales (*principal component analysis*) (PCA) (Bishop, 1995) para esta tarea y finalmente hemos escogido PCA por ser el que mejores prestaciones de clasificación proporciona en este problema. Partimos de $37 \times 37 \times 3$ dimensiones y, a la vista de los autovalores de la matriz de covarianzas de los datos (véase Figura 5.5), nos quedamos con la mayor parte de la energía conservando las 200 direcciones principales más importantes y proyectando sobre ellas los datos.

5.5.2. Clasificador de bacilos

El clasificador escogido es una SVM. Los parámetros de este clasificador se ajustan en el conjunto de entrenamiento por medio de validación cruzada. Como hemos visto, no todas las muestras de los pacientes negativos designados para el entrenamiento del clasificador de parches se emplean en el mismo por razones de capacidad computacional. Por ello, una vez entrenado el clasificador con la base de entrenamiento, clasificamos los pacientes negativos de entrenamiento para añadir a dicha base aquellos parches mal clasificados, puesto que sabemos que la etiqueta de todos los parches de un paciente sano es -1 . De este modo empleamos toda la información disponible para mejorar el clasificador y mejoramos la especificidad.

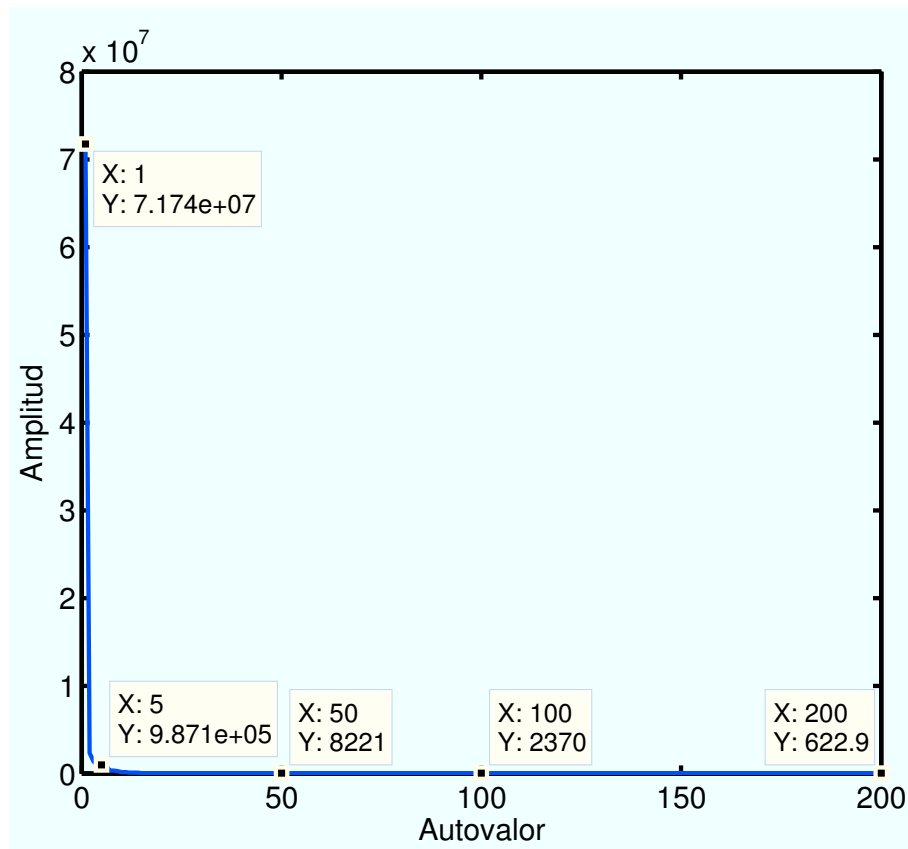


Figura 5.5: Los 200 primeros autovalores de la matriz covarianza de los datos.

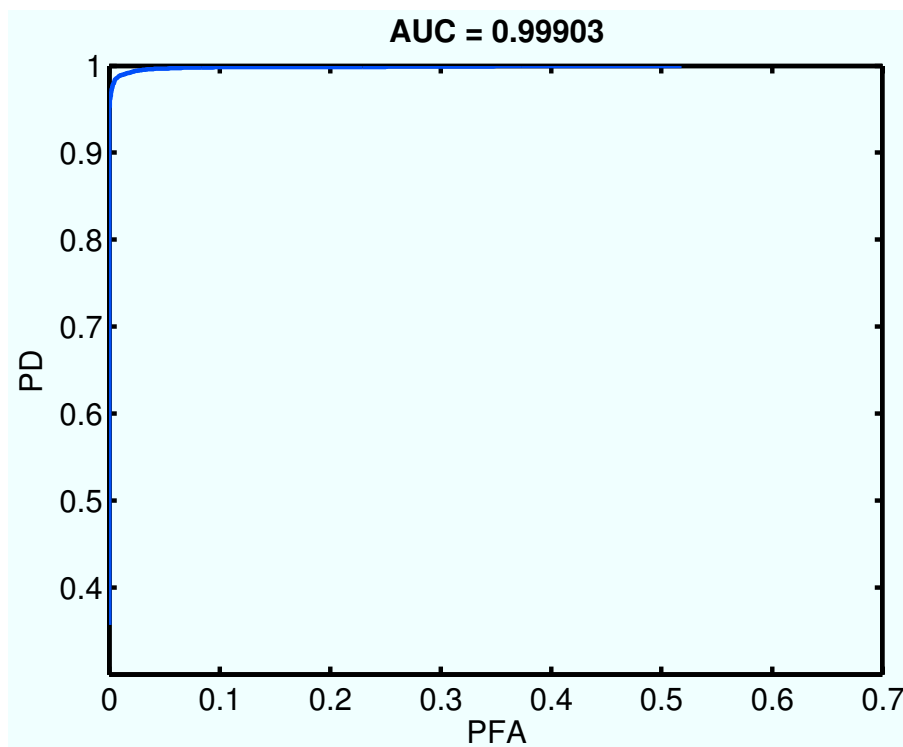


Figura 5.6: Curva ROC para PCA 100. AUC es el acrónimo inglés para área bajo la curva ROC.

Las figuras 5.6 y 5.7 muestran la curva característica operativa del receptor (*receiver operating characteristic*) (ROC) (más bien pseudo-roc, puesto que sólo hemos variado el umbral de la SVM para su elaboración) de un clasificador con kernel gaussiano para una reducción a 100 y 200 dimensiones por medio de PCA.

La Tabla 5.1 resume los resultados para otras dimensiones tras la extracción de características. A la vista de la sensibilidad, especificidad y probabilidad de error se obtienen mejores resultados en test realizando una reducción a 100 dimensiones. Sin embargo, con 200 dimensiones se obtiene una mayor área bajo la curva ROC (*area under the ROC curve*) (AUC). Esta pequeña diferencia se debe a que hay alguna/s muestra/s que necesitan bajar mucho el umbral para clasificarse correctamente en el caso de 100 dimensiones y que se clasifican correctamente con un umbral más elevado en el caso de 200 dimensiones. Este efecto puede apreciarse en las figuras 5.6 y 5.7 donde la curva se dibuja hasta que se alcanza una probabilidad de detección unidad. Escoger la reducción a 200 dimensiones parece una opción más resistente al caso peor a costa de una ligera pérdida en prestaciones.

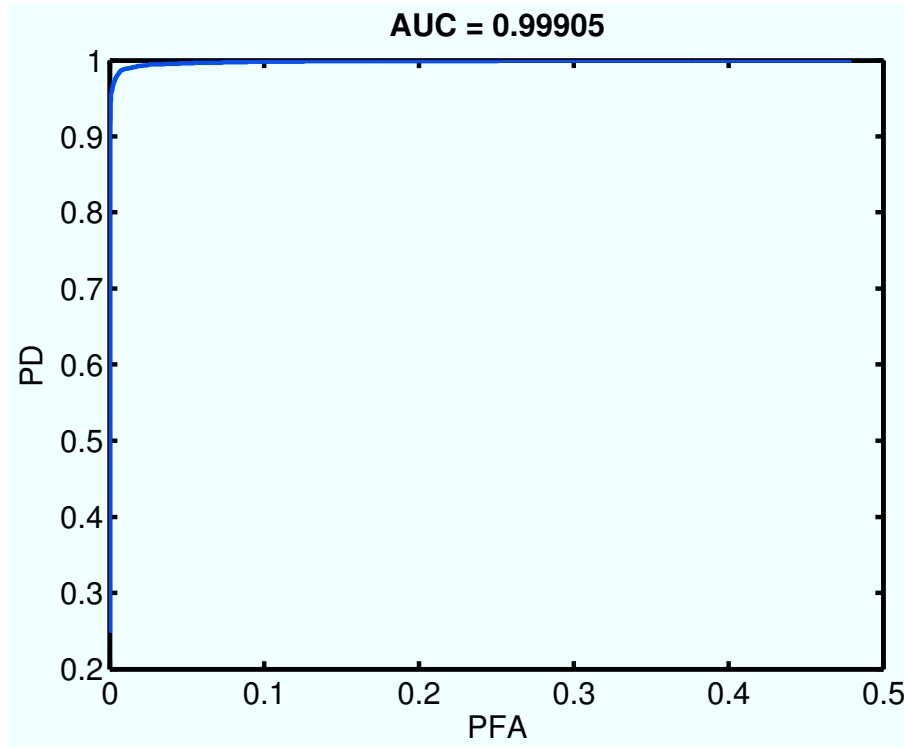


Figura 5.7: Curva ROC para PCA 200.

	nSV	Sens. (%)	Espec. (%)	P_e train	P_e test	AUC
PCA10	11996	93.69	99.46	0.0218	0.0332	0.99033
PCA20	9144	96.42	99.77	0.0115	0.0184	0.99792
PCA50	9332	97.24	99.75	0.0049	0.0146	0.9984
PCA100	8485	97.43	99.71	0.0060	0.0138	0.99903
PCA200	9074	97.32	99.66	0.0025	0.0147	0.99905

Cuadro 5.1: Tabla resumen del entrenamiento del clasificador de parches.

5.5.2.1. Implementación en tiempo real

Para permitir la ejecución en tiempo real hemos reducido la complejidad del clasificador disminuyendo el número de vectores soporte y hemos implementado el clasificador mediante una estructura en cascada. Tenemos varias alternativas para reducir la complejidad. La primera, consiste en emplear los *reduced set methods* como por ejemplo los distintos métodos para obtener preimágenes. Hemos observado que esto funciona bien para nuestro problema si el kernel utilizado para la SVM es el gaussiano; sin embargo, para otros kernels como el kernel polinómico el método de la preimagen no funciona demasiado bien. La ventaja de esta opción es que evita el entrenamiento de la SVM de nuevo. La otra opción consiste en entrenar la máquina de nuevo construyendo iterativamente la frontera de clasificación (Parrado-Hernández *et al.*, 2003; Keerthi *et al.*, 2006). Ambas alternativas cuentan con la ventaja adicional de devolver los vectores soporte en orden aproximado de importancia lo que es particularmente conveniente para la implementación en cascada.

La Tabla (5.2) muestra un resumen de la combinación de la reducción de dimensionalidad mediante PCA con la reducción del número de vectores soporte mediante el algoritmo de la preimagen. Las mejores prestaciones se obtienen para PCA100 y 200 vectores soporte. En este experimento apreciamos como el restringir el número de vectores soporte nos penaliza el aumento de la dimensión del espacio de entrada. Así encontramos en 50 y 100 dimensiones mejores probabilidades de error y mejores AUC que para 200 dimensiones. Es más, la mejor AUC aparece para 50 dimensiones. Por otro lado, en 10 y 20 dimensiones el clasificador sin reducir tiene una estructura compleja. Esto hace que los errores de los clasificadores con un número reducido de vectores soporte estén casi un orden de magnitud por encima del clasificador sin reducir. Así encontramos que los clasificadores que hacen mejor balance entre complejidad y prestaciones se entrenan con datos entre 50 y 100 dimensiones.

La otra dirección para acelerar el clasificador de bacilos es la implementación en cascada. En nuestra aplicación la mayoría de los parches pertenecientes a pacientes sanos y enfermos no son bacilos, por tanto el clasificador en cascada puede diseñarse de modo que sólo los parches declarados como bacilo pasen al siguiente elemento. La Figura 5.8 muestra un diagrama de éste clasificador. Para que la sensibilidad del clasificador global disminuya lo menos posible ajustamos los parámetros (el sesgo (o *bias*) en caso de una SVM) de cada clasificador de las etapas intermedias para obtener una sensibilidad igual a uno o muy cercana a uno. Hacer la sensibilidad igual a 1 es un requisito que puede relajarse si lo imponen las restricciones temporales según se avanza en las etapas de la cascada.

Nuestra implementación de una SVM $f(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b$ sigue los siguientes

	Sensib (%)	Especif (%)	P_e test	AUC
PCA10RED50	71.55	99.75	0.1227	0.97068
PCA10RED100	68.37	99.80	0.1514	0.97068
PCA10RED200	72.82	99.76	0.1305	0.96239
PCA20RED50	75.09	95.31	0.1430	0.98422
PCA20RED100	56.47	95.55	0.2303	0.96629
PCA20RED200	82.07	97.15	0.1002	0.98171
PCA50RED50	64.55	99.87	0.1693	0.96654
PCA50RED100	83.02	99.90	0.0812	0.99023
PCA50RED200	91.85	99.60	0.0408	0.99445
PCA100RED50	67.53	99.96	0.1546	0.96549
PCA100RED100	99.96	99.96	0.0426	0.98446
PCA100RED200	93.73	99.65	0.0316	0.99379
PCA200RED50	70.35	99.98	0.1410	0.9543
PCA200RED100	80.15	99.99	0.0944	0.97255
PCA200RED200	87.00	99.92	0.0622	0.98858

Cuadro 5.2: Resumen de la evaluación de los clasificadores en el conjunto de test para distintos valores de la reducción de dimensionalidad mediante PCA y número de vectores soporte obtenidos por el método de la preimagen. El número que sigue a “PCA” indica el número de dimensiones a las que se reduce y el número que sigue a “RED” indica el número de vectores soporte usados por la aproximación.

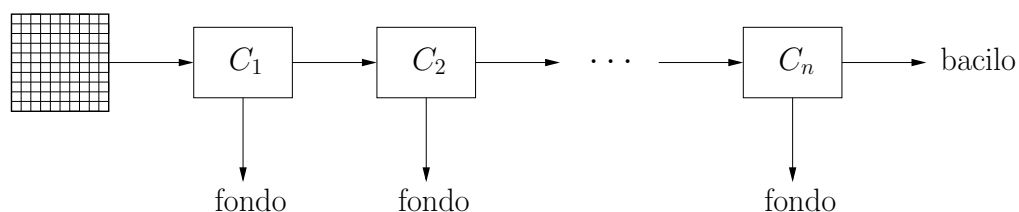


Figura 5.8: Ejemplo de clasificador en cascada diseñado para rechazar rápidamente la clase fondo. Sólo los parches que superan los clasificadores de las etapas intermedias pueden ser clasificados como bacilo.

pasos:

- Reducimos la SVM por el método de Downs *et al.* (2001).
- Como primer elemento de la cascada elegimos una SVM lineal que puede escribirse como un producto escalar por un único vector soporte.

$$f(\mathbf{x}) = \sum_i \beta_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b = \langle \mathbf{x}, \mathbf{x}' \rangle + b$$

donde $\mathbf{x}' = \sum \beta_i \mathbf{x}_i$. De este modo, si el problema tiene una buena aproximación lineal sacamos partido de ella y de lo contrario es poco costoso.

El resto de los elementos implementan una SVM no lineal del siguiente modo:

- En la etapa ℓ se obtiene

$$f(x) = \sum_{i=1}^{N_\ell} \alpha_i^\ell k(\mathbf{x}, \mathbf{x}_i) + b_\ell$$

donde cada uno de los kernel $k(\mathbf{x}, \mathbf{x}_i)$ sólo se calcula en la primera etapa que aparece y se almacena mientras la muestra siga en la cascada. El conjunto de coeficientes α_i^ℓ que multiplica a los kernels se optimiza para cada etapa bien optimizando algún funcional de riesgo como en (Keerthi *et al.*, 2006) o minimizando la distancia con la frontera original (Schölkopf y Smola, 2001). El coeficiente b_ℓ se obtiene a partir de la sensibilidad deseada.

- En la última etapa usamos los coeficientes y el umbral que resultaron del entrenamiento de la SVM y es ahí donde discriminamos los bacilos del fondo aproximadamente con la probabilidad de error obtenida en el conjunto de test.

Ilustramos este procedimiento para un paciente positivo y uno negativo con una implementación en cascada del clasificador PCA200RED200. En esta implementación hemos prescindido de la SVM lineal y hemos hecho que cada etapa evalúe un vector soporte más que la anterior. Las figuras 5.9 y 5.10 muestran el porcentaje de parches descartados en cada etapa para un paciente positivo y uno negativo.

5.5.3. Clasificador de pacientes

En este apartado vamos a examinar las prestaciones del clasificador de pacientes. Seguimos ambas aproximaciones, la frecuentista y la bayesiana.

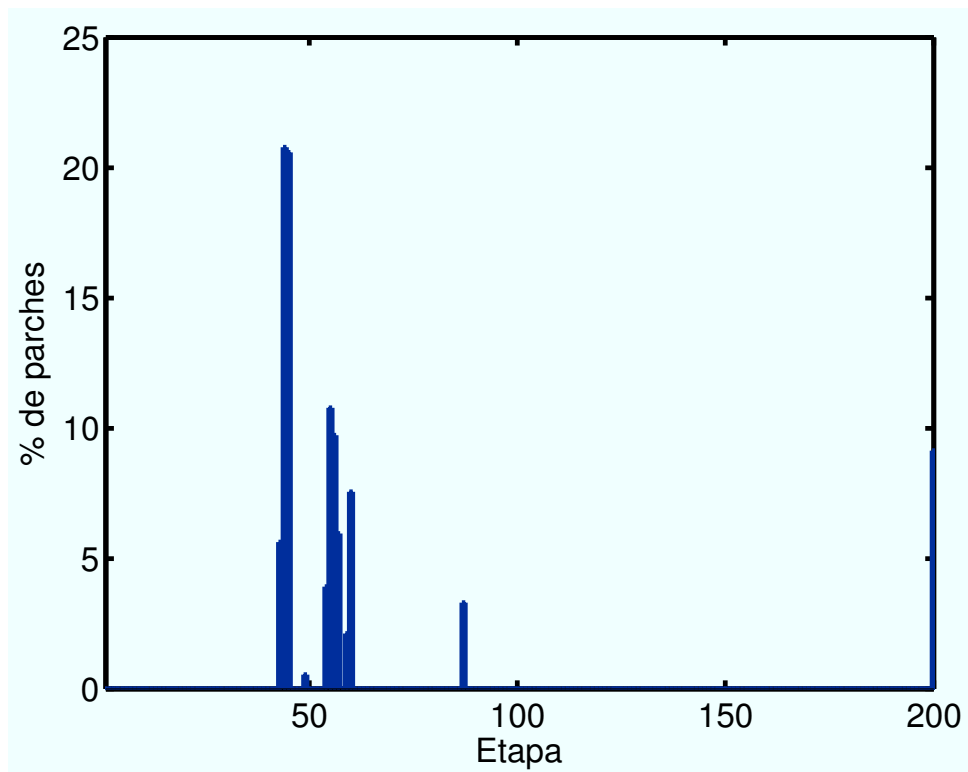


Figura 5.9: Implementación en cascada del clasificador PCA200RED200. Porcentaje de parches evaluados en cada una de las etapas para un paciente positivo.

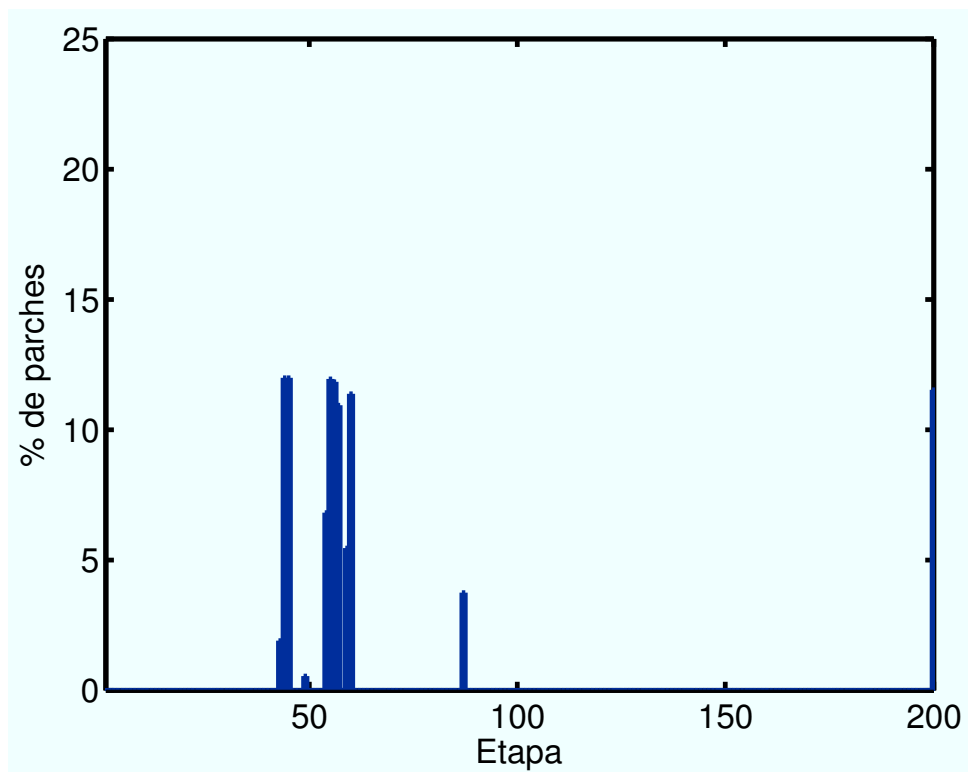


Figura 5.10: Implementación en cascada del clasificador PCA200RED200. Porcentaje de parches evaluados en cada una de las etapas para un paciente negativo.

ID paciente	#muestras	#bacilos	p_{bacilo}
32742	378115	0	0
48940	337316	0	0
52707	375236	0	0
51861	447186	1	2.24e-06
31404	374107	1	2.67e-06
51859	316095	1	3.16e-06
31087	395889	4	1.01e-05
32741	367358	5	1.36e-05
46989	304341	71	2.33e-04

Cuadro 5.3: Pacientes enfermos. Número de parches declarados como bacilos y probabilidad de bacilo en cada pacientes después de pasar la salida blanda de la SVM por el umbral que maximiza la capacidad de discriminación.

5.5.3.1. Aproximación frecuentista

Nos centramos en el caso binario e ilustramos las distintas etapas que hemos descrito arriba.

1. Encontrar el umbral que maximice la capacidad de discriminar pacientes. En este paso hemos seleccionado el umbral que maximiza la probabilidad de clasificar un paciente correctamente siempre que no se clasifique erróneamente toda una clase y se clasifiquen la mayor cantidad de pacientes enfermos. Si hubiese varios umbrales que igualasen dichos criterios se escoge aquél que discrimine mejor a nivel de parche.

El resultado de aplicar el criterio anterior a los pacientes de entrenamiento se muestra en el Cuadro 5.3 para los enfermos y en el Cuadro 5.4 para los sanos. La columna “ID paciente” muestra el identificador del paciente, la columna “#muestras” contiene el número de parches que tiene el paciente, la columna “#bacilos” contiene el número de parches declarados como bacilo y la columna “ p_{bacilo} ” es la probabilidad media de declaración de bacilo.

2. Intervalos de confianza. A la vista de los Cuadros 5.3 y 5.4 parece evidente que debemos tomar como prototipo de un paciente sano aquél en el que se detectan 0 bacilos y como un paciente enfermo aquél en el que se detectan al menos 4 bacilos. En el Cuadro 5.5 mostramos los intervalos de confianza para la probabilidad de bacilo condicionada a paciente sano y a enfermo si escogemos los pacientes prototipo “51523” para los sanos y “31087” para los enfermos. La columna “ N_b ” contiene el número de bacilos asignados al prototipo de los enfermos; “ P_c ” es la confianza de que las verdaderas probabilidades estén en sus intervalos; $[p_{l,\text{sano}}, p_{h,\text{sano}}]$ forman el intervalo de la probabilidad de aparición de

ID paciente	#muestras	#bacilos	p_{bacilo}
30214	345242	0	0
30261	420377	0	0
30262	389929	0	0
30304	394930	0	0
30619	423600	0	0
30663	371216	0	0
30825	398745	0	0
30855	378676	0	0
30877	403796	0	0
30880	415717	0	0
30881	434528	0	0
30911	383563	0	0
30986	406695	0	0
30998	381192	0	0
31245	409800	0	0
31522	373237	0	0
31547	383463	0	0
31934	387449	0	0
32111	390678	0	0
32497	383379	0	0
32550	427327	0	0
32642	423447	0	0
32748	404043	0	0
32750	393425	0	0
32955	376269	0	0
38841	213689	0	0
51523	450316	0	0
31060	387925	1	2.58e-06
30783	355001	2	5.63e-06
31549	350956	2	5.70e-06
29994	394365	3	7.61e-06
31230	369991	15	4.05e-05
30207	381579	20	5.24e-05
32688	397627	102	2.57e-04

Cuadro 5.4: Pacientes sanos. Número de parches declarados como bacilos y probabilidad de bacilo en cada pacientes después de pasar la salida blanda de la SVM por el umbral que maximiza la capacidad de discriminación.

N_b	P_c	$p_{l,sano}$	$p_{h,sano}$	$p_{l,enf}$	$p_{h,enf}$
3	0.663	0	3.74e-06	3.74e-06	1
4	0.774	0	4.71e-06	4.72e-06	1
5	0.85	0	5.66e-06	5.68e-06	1
6	0.901	0	6.61e-06	6.63e-06	1
7	0.933	0	7.51e-06	7.60e-06	1

Cuadro 5.5: Intervalos de confianza para la probabilidad de aparición de bacilo considerando el paciente “51523” como el prototipo para los sanos y el “31087” como el prototipo de los enfermos. Se ha variado el número de bacilos del prototipo positivo, N_b , para escoger una calidad, P_c , aceptable.

P_c	$p_{l,sano}$	$p_{h,sano}$	$p_{l,enf}$	$p_{h,enf}$	ε	$E_{p_{h,sano}}[K]$	$E_{p_{l,enf}}[K]$
0.901	0	6.61e-06	6.63e-06	1	2.56e-08	9.09674e+10	9.085e+10
0.874	0	6.07e-06	7.07e-06	1	1.00e-06	6.06372e+07	5.76332e+07
0.846	0	5.61e-06	7.48e-06	1	1.87e-06	1.75862e+07	1.5979e+07
0.819	0	5.23e-06	7.85e-06	1	2.62e-06	9.07999e+06	7.9316e+06
0.794	0	4.92e-06	8.16e-06	1	3.24e-06	5.99913e+06	5.06997e+06
0.769	0	4.65e-06	8.46e-06	1	3.80e-06	4.39827e+06	3.60597e+06

Cuadro 5.6: Relación entre P_c , la distancia entre intervalos, ε , y una estima del número esperado de muestras necesario, $E_{p_{h,sano}}[K]$ y $E_{p_{l,enf}}[K]$, si se realizase un contraste entre las hipótesis simples caracterizadas por $p_{h,sano}$ y $p_{l,enf}$ si el paciente prototipo positivo tiene 6 bacilos.

bacilos para los pacientes sanos y $[p_{l,enf}, p_{h,enf}]$ dicho intervalo para los enfermos. Para la elaboración de estos intervalos nos hemos restringido a confianzas iguales para cada uno de ellos. Si consideramos los cuadros anteriores junto con el Cuadro 5.5 vemos que resulta razonable fijar en 6 bacilos el límite inferior para los pacientes positivos puesto que esto proporciona una $P_c > 0.9$ y no eleva demasiado el umbral para declarar a un paciente como positivo.

También podemos preguntarnos cómo se relacionan P_c , el valor estimado del número de muestras necesarias para terminar y la distancia, ε , que separa $p_{h,sano}$ y $p_{l,enf}$. Para ello, mostramos el Cuadro 5.6 que se obtiene variando ε para 6 bacilos en el paciente positivo prototipo. En el cuadro $E_{p_{h,sano}}[K]$ y $E_{p_{l,enf}}[K]$ representan el valor esperado del número de muestras de test necesarias para finalizar para las probabilidades $p_{h,sano}$ y $p_{l,enf}$ suponiendo un contraste de estas dos probabilidades como hipótesis simples. Apreciamos que una ligera pérdida de calidad, puede proporcionar un test que necesite muchas menos muestras. Vemos, por ejemplo, que pasar de una calidad de 0.901 a una calidad de 0.874 produce un descenso en el número esperado de muestras de tres órdenes de magnitud.

ID paciente	dec.	fin	LLR	#muestras	$1 - P_D$	P_{FA}	#bacilos
32743	0	0	-6.22e-03	393931	0.994	0.994	1
40279	1	0	5.34e-03	395551	0.995	0.995	4
50304	1	0	0.559	361000	0.614	0.614	147
51862	1	0	1.20e-03	406313	0.999	0.999	3
55165	1	0	0.0432	427629	0.962	0.962	14
55798	1	0	6.20e-03	361878	0.994	0.994	4
56036	0	0	-2.60e-03	403750	0.998	0.998	2

Cuadro 5.7: Pacientes enfermos de test. Mostramos la clasificación, las cotas de la probabilidad de no detección y de falsa alarma y el número de bacilos detectados. $P_c = 0.901$.

- Ahora aplicamos el test secuencial a los pacientes de test. Los resultados para $P_c = 0.901$ se muestran en los Cuadros 5.8 y 5.7 en las que “dec.” muestra la decisión donde 0 es para sano y 1 para enfermo; “fin” muestra si el test ha terminado; “LLR” muestra el valor del cociente de verosimilitud; “#muestras” muestra el número de parches empleados en la decisión; “ $1 - P_D$ ” y “ P_{FA} ” son cotas superiores a la probabilidad de error y “#bacilos” muestra el número de parches declarados como bacilo. Los resultados de este experimento clasifican correctamente a 5/7 pacientes sanos y a 12/15 pacientes enfermos. Las cotas de la probabilidad de falsa alarma y no detección están muy lejos de $1 - P_c$. Esto se debe a que los intervalos que definen las hipótesis están muy cercanos y al uso del SPRT que, aunque decide correctamente, sobrestima dichas probabilidades. Claramente, la regla de parada de Lai, si se considera un coste por muestra, o las de Lorden y Hubber, en caso contrario, harían que el test se detuviese antes, sin embargo, en estos casos es más difícil estimar qué probabilidades de error se han alcanzado si se acaban las muestras y el test no termina.

Si seleccionamos los intervalos para los que $P_c = 0.846$ los resultados se muestran en los Cuadros 5.10 y 5.9 en las que apreciamos las mismas decisiones pero con probabilidades de no detección y falsa alarma mucho menos conservadoras que en el caso anterior. El compromiso entre P_c y el número de muestras necesarias para finalizar el test se aprecia con claridad. En este caso, el test secuencial finaliza antes de terminar las muestras para el caso de 147 bacilos en los pacientes positivos y 399 en los pacientes negativos. Así y todo los pacientes negativos necesitarían muchas más muestras que las disponibles para terminar. Para los números que manejamos, las decisiones para pacientes sanos son muy conservadoras.

ID paciente	dec.	fin	LLR	#muestras	$1 - P_D$	P_{FA}	#bacilos
30257	0	0	-2.77e-03	410090	0.998	0.998	2
30266	0	0	-0.0105	409178	0.991	0.991	0
30295	0	0	-0.0101	392553	0.991	0.991	0
30665	0	0	-1.33e-03	353973	0.999	0.999	2
30725	1	0	1.67e-03	387746	0.998	0.998	3
31228	1	0	0.0258	351546	0.977	0.977	9
31523	0	0	-9.34e-03	364598	0.992	0.992	0
31534	0	0	-6.55e-03	406909	0.994	0.994	1
31684	0	0	-9.87e-03	385355	0.991	0.991	0
31819	0	0	-9.60e-03	374866	0.991	0.991	0
32240	0	0	-0.0107	418109	0.990	0.990	0
32633	0	0	-9.90e-03	386712	0.991	0.991	0
32781	1	0	1.53228	433886	0.294	0.294	399
32956	0	0	-9.51e-03	371446	0.991	0.991	0
36856	0	0	-9.46e-03	369371	0.992	0.992	0

Cuadro 5.8: Pacientes sanos de test. Mostramos la clasificación, las cotas de la probabilidad de no detección y de falsa alarma y el número de bacilos detectados. $P_c = 0.901$.

ID paciente	dec.	fin	LLR	#muestras	$1 - P_D$	P_{FA}	#bacilos
32743	0	0	-0.449	393931	0.694	0.694	1
40279	1	0	0.411	395551	0.715	0.715	4
50304	1	1	41.6158	361000	0.162	0.162	147
51862	1	0	0.103	406313	0.917	0.917	3
55165	1	0	3.22817	427629	0.187	0.187	14
55798	1	0	0.474	361878	0.680	0.680	4
56036	0	0	-0.179	403750	0.861	0.861	2

Cuadro 5.9: Pacientes enfermos de test. Mostramos la clasificación, las cotas de la probabilidad de no detección y de falsa alarma y el número de bacilos detectados. $P_c = 0.846$.

ID paciente	dec.	fin	LLR	#muestras	$1 - P_D$	P_{FA}	#bacilos
30257	0	0	-0.191	410090	0.853	0.853	2
30266	0	0	-0.765	409178	0.547	0.547	0
30295	0	0	-0.734	392553	0.560	0.560	0
30665	0	0	-0.0864	353973	0.930	0.930	2
30725	1	0	0.138	387746	0.891	0.891	3
31228	1	0	1.93196	351546	0.276	0.276	9
31523	0	0	-0.682	364598	0.582	0.582	0
31534	0	0	-0.473	406909	0.681	0.681	1
31684	0	0	-0.720	385355	0.565	0.565	0
31819	0	0	-0.701	374866	0.574	0.574	0
32240	0	0	-0.782	418109	0.541	0.541	0
32633	0	0	-0.723	386712	0.564	0.564	0
32781	1	1	113.978	433886	0.162	0.162	399
32956	0	0	-0.694	371446	0.576	0.576	0
36856	0	0	-0.691	369371	0.578	0.578	0

Cuadro 5.10: Pacientes sanos de test. Mostramos la clasificación, las cotas de la probabilidad de no detección y de falsa alarma y el número de bacilos detectados. $P_c = 0.846$.

Resulta evidente que las prestaciones del test dependen de las prestaciones del clasificador de bacilos. La mayor cantidad de bacilos detectados, aparecen en un paciente negativo lo que nos hace pensar que el clasificador de bacilos no está haciendo un trabajo muy bueno (esta aplicación requiere una probabilidad de falsa alarma muy baja) y por tanto las prestaciones globales se resienten por ello.

5.5.3.2. Aproximación bayesiana

El primer experimento consiste en emplear dos pacientes prototipo para discriminar entre las clases. Se escogen los mismos pacientes que en la aproximación frecuentista y se escoge como distribución *a priori* una uniforme entre 0 y 1. Este experimento se lleva a cabo con la probabilidad *a priori* de un paciente sano, $P(H_0) = 0.92$, (Cuadros 5.11 y 5.12) y también para comparar con las decisiones frecuentistas con una probabilidad *a priori* de un paciente sano, $P(H_0) = 0.5$ (Cuadros 5.13 y 5.14). La columna " $P(H_0|\mathbf{z})$ " es la probabilidad *a posteriori* de que el paciente sea sano. Del examen de los cuadros anteriores vemos influencia en la decisión de la probabilidad *a priori* de paciente sano, puesto que cambiarla, varía el resultado de la decisión en 4 pacientes. También queremos resaltar que las predicciones se vuelven muy extremas en cuanto aumenta el número de bacilos detectados, y son muy altas para pacientes negativos sin bacilos. Este comportamiento, es esperable puesto que estamos contrastando dos hipótesis simples y por tanto, solo estamos buscando a

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
30257	0.965	0	410090	2
30266	0.999	0	409178	0
30295	0.999	0	392553	0
30665	0.946	0	353973	2
30725	0.877	0	387746	3
31228	0.0554	1	351546	9
31523	0.998	0	364598	0
31534	0.991	0	406909	1
31684	0.999	0	385355	0
31819	0.998	0	374866	0
32240	0.999	0	418109	0
32633	0.999	0	386712	0
32781	1.71e-21	1	433886	399
32956	0.998	0	371446	0
36856	0.998	0	369371	0

Cuadro 5.11: Pacientes sanos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior uniforme y $P(H_0) = 0.92$.

cuál se parecen más los datos.

Para considerar la influencia de la distribución *a priori* vamos a repetir el experimento con la distribución *a priori* de Jeffreys que equivale a una $\text{beta}(0.5, 0.5)$. Los Cuadros 5.15 a 5.14 muestran el resultado de este experimento en donde se aprecia que el método se vuelve más conservador para declarar un paciente como sano y más extremo para declararlo como enfermo. Esto, si comparamos con lo que habíamos obtenido con la distribución *a priori* uniforme, produce decisiones distintas en dos pacientes enfermos cuando $P(H_0) = 0.92$; y en dos pacientes sanos y uno enfermo cuando $P(H_0) = 0.5$.

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
32743	0.990	0	393931	1
40279	0.740	0	395551	4
50304	1.36e-12	1	361000	147
51862	0.892	0	406313	3
55165	0.0104	1	427629	14
55798	0.684	0	361878	4
56036	0.963	0	403750	2

Cuadro 5.12: Pacientes enfermos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior uniforme y $P(H_0) = 0.92$.

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
30257	0.704	0	410090	2
30266	0.987	0	409178	0
30295	0.985	0	392553	0
30665	0.603	0	353973	2
30725	0.384	1	387746	3
31228	5.07e-03	1	351546	9
31523	0.982	0	364598	0
31534	0.908	0	406909	1
31684	0.984	0	385355	0
31819	0.983	0	374866	0
32240	0.988	0	418109	0
32633	0.984	0	386712	0
32781	1.49e-22	1	433886	399
32956	0.983	0	371446	0
36856	0.982	0	369371	0

Cuadro 5.13: Pacientes sanos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior uniforme y $P(H_0) = 0.5$.

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
32743	0.900	0	393931	1
40279	0.199	1	395551	4
50304	1.18e-13	1	361000	147
51862	0.419	1	406313	3
55165	9.10e-04	1	427629	14
55798	0.158	1	361878	4
56036	0.694	0	403750	2

Cuadro 5.14: Pacientes enfermos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior uniforme y $P(H_0) = 0.5$.

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
30257	0.919	0	410090	2
30266	0.999	0	409178	0
30295	0.999	0	392553	0
30665	0.880	0	353973	2
30725	0.725	0	387746	3
31228	0.0162	1	351546	9
31523	0.998	0	364598	0
31534	0.983	0	406909	1
31684	0.999	0	385355	0
31819	0.998	0	374866	0
32240	0.999	0	418109	0
32633	0.999	0	386712	0
32781	3.77e-22	1	433886	399
32956	0.998	0	371446	0
36856	0.998	0	369371	0

Cuadro 5.15: Pacientes sanos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior de Jeffreys y $P(H_0) = 0.92$.

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
32743	0.982	0	393931	1
40279	0.492	1	395551	4
50304	3.04e-13	1	361000	147
51862	0.753	0	406313	3
55165	2.72e-03	1	427629	14
55798	0.424	1	361878	4
56036	0.916	0	403750	2

Cuadro 5.16: Pacientes enfermos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior de Jeffreys y $P(H_0) = 0.92$.

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
30257	0.498	1	410090	2
30266	0.986	0	409178	0
30295	0.985	0	392553	0
30665	0.389	1	353973	2
30725	0.187	1	387746	3
31228	1.43e-03	1	351546	9
31523	0.981	0	364598	0
31534	0.838	0	406909	1
31684	0.984	0	385355	0
31819	0.982	0	374866	0
32240	0.987	0	418109	0
32633	0.984	0	386712	0
32781	3.27e-23	1	433886	399
32956	0.982	0	371446	0
36856	0.982	0	369371	0

Cuadro 5.17: Pacientes sanos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior de Jeffreys y $P(H_0) = 0.5$.

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
32743	0.824	0	393931	1
40279	0.0776	1	395551	4
50304	2.64e-14	1	361000	147
51862	0.210	1	406313	3
55165	2.37e-04	1	427629	14
55798	0.0601	1	361878	4
56036	0.486	1	403750	2

Cuadro 5.18: Pacientes enfermos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior de Jeffreys y $P(H_0) = 0.5$.

ID paciente	$P(H_0 z)$	dec.	#muestras	#bacilos
30257	0.910	0	410090	2
30266	0.943	0	409178	0
30295	0.943	0	392553	0
30665	0.909	0	353973	2
30725	0.906	0	387746	3
31228	0.889	0	351546	9
31523	0.943	0	364598	0
31534	0.916	0	406909	1
31684	0.943	0	385355	0
31819	0.943	0	374866	0
32240	0.943	0	418109	0
32633	0.943	0	386712	0
32781	0.0382	1	433886	399
32956	0.943	0	371446	0
36856	0.943	0	369371	0

Cuadro 5.19: Pacientes sanos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior extraído de todos los pacientes y $P(H_0) = 0.92$.

Finalmente, empleamos toda la información de los pacientes de entrenamiento de los cuadros 5.3 y 5.4 para construir una distribución *a posteriori* de la probabilidad de aparición de bacilo para cada clase. Para ello, aproximamos para cada paciente el valor de dicha probabilidad por su estima de máxima verosimilitud. Asumimos que esta distribución *a posteriori* tiene forma de beta y la estimamos sus parámetros por máxima verosimilitud a partir de las estimas de las probabilidades de aparición de bacilo.

Los resultados de este experimento se muestran en los Cuadros 5.19 y 5.20 para la probabilidad de sano $P(H_0) = 0.92$ y en los Cuadros 5.21 y 5.22 para $P(H_0) = 0.5$. En el primer caso, sólo un paciente enfermo y uno sano son declarados enfermos, lo que muestra el dominio de las probabilidades *a priori* de las clase. Si hacemos $P(H_0) = 0.5$ tenemos que todos los pacientes con al menos un bacilo son declarados enfermos. Aparte de la clasificación destacamos que el test es claramente conservador en sus predicciones, sólo separándose sensiblemente de 0.5 en un caso negativo con 399 bacilos y en un caso positivo con 147 bacilos. En los demás, la evidencia es escasa a favor de ninguna hipótesis.

A modo de resumen de las distintas aproximaciones bayesianas, mostramos el Cuadro 5.23. Este cuadro contiene el número de pacientes que han sido declarados como enfermos para cada una de las distribuciones *a priori*. Notamos que los resultados varían entre una aproximación y otra, especialmente, en el caso que construye

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
32743	0.916	0	393931	1
40279	0.903	0	395551	4
50304	0.434	1	361000	147
51862	0.907	0	406313	3
55165	0.883	0	427629	14
55798	0.902	0	361878	4
56036	0.910	0	403750	2

Cuadro 5.20: Pacientes enfermos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con prior extraído de todos los pacientes y $P(H_0) = 0.92$.

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
30257	0.469	1	410090	2
30266	0.591	0	409178	0
30295	0.590	0	392553	0
30665	0.466	1	353973	2
30725	0.456	1	387746	3
31228	0.410	1	351546	9
31523	0.589	0	364598	0
31534	0.487	1	406909	1
31684	0.590	0	385355	0
31819	0.589	0	374866	0
32240	0.591	0	418109	0
32633	0.590	0	386712	0
32781	3.44e-03	1	433886	399
32956	0.589	0	371446	0
36856	0.589	0	369371	0

Cuadro 5.21: Pacientes sanos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con con distribución *a priori* extraída de todos los pacientes y $P(H_0) = 0.5$.

ID paciente	$P(H_0 \mathbf{z})$	dec.	#muestras	#bacilos
32743	0.487	1	393931	1
40279	0.448	1	395551	4
50304	0.0626	1	361000	147
51862	0.457	1	406313	3
55165	0.395	1	427629	14
55798	0.445	1	361878	4
56036	0.469	1	403750	2

Cuadro 5.22: Pacientes enfermos de test. Mostramos la clasificación, la probabilidad de que sea sano y el número de bacilos declarados. Test bayesiano con distribución *a priori* extraída de todos los pacientes y $P(H_0) = 0.5$.

Distribución <i>a priori</i>		uniforme		Jeffreys		conjunta	
$P(H_0)$		0.92	0.5	0.92	0.5	0.92	0.5
Enfermos	Declarados enfermos	2	5	4	6	1	7
	Declarados sanos	5	2	3	1	6	0
Sanos	Declarados enfermos	2	3	2	5	1	6
	Declarados sanos	13	12	13	10	14	9

Cuadro 5.23: Tabla de confusión obtenida con las distribuciones *a priori* uniforme, Jeffreys y la obtenida de conjuntamente con todos los pacientes de entrenamiento por máxima verosimilitud.

la distribución *a priori* de manera conjunta.

5.6. Resumen

Hemos presentado un sistema para la clasificación de pacientes de tuberculosis con un reconocedor de patrones basado en la SVM y un centro de fusión basado en un test secuencial. A la vista de los resultados obtenidos tanto en la aproximación frecuentista, que es mejorable sustituyendo el SPRT por la regla de parada de Lai o la de Hubber; como en la aproximación bayesiana, cuando incluimos en la distribución *a priori* toda la información de las clases; podemos concluir que hay que hacer mayor esfuerzo en el clasificador, especialmente en la extracción de características.

Dicho esto, podemos decir que la metodología desarrollada en la tesis, se adapta perfectamente a esta aplicación y que la incertidumbre queda claramente reflejada por los métodos propuestos. Los resultados de los Cuadros 5.9 y 5.10 para el caso frecuentista y los Cuadros 5.21 y 5.22 para el bayesiano muestran claramente que hacen falta muchas más muestras de test o que hay que mejorar el clasificador de bacilos.

A la vista de los resultados, quizá el método bayesiano con la distribución *a priori* conjunta sobre todos los pacientes de su clase sea la opción más fácil. Aquí, por supuesto, justificamos el uso de la beta por comodidad analítica y su estima por máxima verosimilitud quizá no sea lo más ortodoxo. Esta distribución puede seguir otro modelo como una mezcla de betas o establecer distribuciones *a priori* sobre los parámetros de la beta y resolver, pero esto complica analítica y computacionalmente el test que ha de funcionar prácticamente en tiempo real (1 segundo por imagen).

El enfoque frecuentista evita todas estas justificaciones, al precio de decisiones más conservadoras. En ocasiones se desean estas decisiones, como podría ser la aplicación presente donde un paciente sano se va a su casa y uno enfermo se queda días en el hospital.

Capítulo 6

Conclusiones y Líneas Futuras

6.1. Conclusiones

En los capítulos anteriores hemos considerado la aplicación de la teoría de la decisión al problema de diagnóstico de tuberculosis basado en imágenes. Para ello, hemos introducido el contraste de hipótesis inciertas y métodos para llevarlo a cabo. También hemos considerado la clasificación de un conjunto de muestras de la misma clase como el equivalente desde el punto de vista de aprendizaje máquina a un cociente de verosimilitud.

En el contraste de hipótesis inciertas hemos formalizado de manera novedosa el problema desde el punto de vista frecuentista. La conclusión más importante tanto si seguimos el camino frecuentista como el bayesiano es que no es posible alcanzar prestaciones arbitrariamente buenas, ni aún con un número infinito de muestras, si las hipótesis están definidas con incertidumbre. Los resultados más relevantes son las cotas superiores a las probabilidades de error en el caso frecuentista y las cotas superiores de la probabilidad *a posteriori* alcanzable desde el punto de vista bayesiano. Algunos resultados preliminares de esta línea se han publicado en (Santiago-Mozos y Artés-Rodríguez, 2006a,b; Santiago-Mozos *et al.*, 2008) y otro se ha enviado recientemente (Santiago-Mozos *et al.*, 2009b).

En la clasificación de conjuntos de muestras de la misma clase desde el punto de vista de aprendizaje hemos visto que entrenar clasificadores en el conjunto extendido no es excesivamente complejo y que la incorporación de simetrías no añade variables al problema de optimización en caso de la SVM. Los experimentos que hemos realizado indican que en ocasiones puede ser ventajoso entrenar un clasificador que proporcione una única salida al conjunto de test en lugar de combinar las salidas individuales de cada muestra de test. Este trabajo se ha enviado recientemente (Santiago-Mozos *et al.*, 2009a).

Finalmente, hemos considerado la implementación de un sistema de diagnóstico de tuberculosis que funciona en tiempo real, adquiere imágenes de un paciente mientras su certidumbre no es suficiente para tomar una decisión y finalmente informa de la calidad de dicha decisión. La ventaja de este sistema es que es posible conocer *a priori* cuáles son las mejores prestaciones que podemos alcanzar a partir de los datos proporcionados para entrenar el sistema. Las conclusiones más importantes en este desarrollo son dos, la primera que los métodos desarrollados son útiles para medir la prestaciones del sistema y la segunda, que hay que dedicar mayor esfuerzo al clasificador local para poder alcanzar buenas prestaciones con un número razonable de muestras. Este trabajo se enviará próximamente a una revista tipo Transactions on Medical Imaging o Transactions on Biomedical Engineering y a alguna revista médica.

6.2. Líneas futuras

Como líneas futuras proponemos:

- En la línea del contraste de hipótesis inciertas:
 - Reemplazar el SPRT por el test secuencial propuesto por Lai para evitar el exceso de conservadurismo de la aproximación frecuentista y hacer que el test secuencial termine antes.
 - Encontrar mejores métodos para la determinación de regiones de confianza en los casos discreto y continuo, haciendo especial énfasis en cómo extender las regiones de modo que no solapen para ocupar la mayor cantidad posible del espacio de estados y así evitando el exceso de conservadurismo.
 - Cómo medir la incertidumbre en el caso continuo con distribuciones desconocidas ha sido sólo brevemente considerado en este trabajo y merece un estudio más profundo.
- En la línea de clasificación de conjuntos de muestras de la misma clase, podría continuarse el trabajo haciendo un estudio teórico de cuáles son las condiciones en las que es mejor “extender” el conjunto de entrenamiento, y cuál es la forma más eficiente de hacerlo.
- Si nos centramos en la aplicación que tenemos entre manos, mejorar el detector local es absolutamente necesario para la viabilidad del sistema. Una parte fundamental del detector de bacilos es la extracción de características, aquí

se podrían probar algún tipo de descriptor local como SIFT (Lowe, 2004), isomap (Tenenbaum *et al.*, 2000), LLE (Roweis y Saul, 2000) u otros, pero siempre teniendo en cuenta el funcionamiento del sistema en tiempo real.

- Relacionado con el punto anterior, sería interesante comparar las redes neuronales con la SVM. Las primeras “suelen” proporcionar arquitecturas más sencillas y probablemente más rápidas.

Bibliografía

- Abramowitz, M. y Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing^a edición. ISBN 0-486-61272-4.
- Agresti, A. (2003). Dealing with discreteness: making “exact” confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical Methods in Medical Research*, 12(1), p. 3.
- Agresti, A. y Coull, B. A. (1998). Approximate is Better than “Exact” for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2), pp. 119–126.
- Agresti, A. y Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14(3), pp. 297–330.
- Andersen, E. B. (1970). Sufficiency and Exponential Families for Discrete Sample Spaces. *Journal of the American Statistical Association*, 65(331), pp. 1248–1255. ISSN 01621459.
<http://www.jstor.org/stable/2284291>
- Anderson, T. W. (1960). A Modification of the Sequential Probability Ratio Test to Reduce the Sample Size. *The Annals of Mathematical Statistics*, 31(1), pp. 165–197.
- Anderson, T. W. y Darling, D. A. (1954). A Test of Goodness of Fit. *Journal of the American Statistical Association*, 49(268), pp. 765–769.
- Andrews, D. W. K. y Buchinsky, M. (2000). A Three-step Method for Choosing the Number of Bootstrap Repetitions. *Econometrica*, 68(1), pp. 23–51.
- (2001). Evaluation of a three-step method for choosing the number of bootstrap repetitions. *Journal of Econometrics*, 103(1-2), pp. 345–386.

- (2002). On the number of bootstrap repetitions for BCa confidence intervals. *Econometric Theory*, 18(04), pp. 962–984.
- Bailey, B. J. R. (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based on transformations of the cell frequencies. *Technometrics*, pp. 583–589.
- BakIr, G.; Bottou, L. y Weston, J. (2005). Breaking SVM Complexity with Cross-Training. En: *Advances in Neural Information Processing Systems 17: Proceedings Of The 2004 Conference*, MIT Press.
- BakIr, G.; Hofmann, T.; Schölkopf, B.; Smola, A. J. y Taskar, B. (2007a). *Predicting structured data*. The MIT Press.
- BakIr, G. H.; Schölkopf, B. y Weston, J. (2007b). On the Pre-Image Problem in Kernel Methods. En: *Kernel Methods in Bioengineering, Signal and Image Processing*, pp. 284–302. Idea Group Publishing.
- BakIr, G. H.; Weston, J. y Schölkopf, B. (2004). Learning to Find Pre-Images. En: *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, .
- Bauquerez, R.; Blanc, L.; Bierrenbach, A.; Brands, A.; Ciceri, K.; Falzon, D.; Floyd, K.; Glaziou, P.; Gunneberg, C.; Hiatt, T.; Hosseini, M.; Pantoja, A.; Uplekar, M.; Watt, C. y Wright, A. (2009). *Global tuberculosis control : epidemiology, strategy, financing : WHO report 2009*. World Health Organization.
http://www.who.int/entity/tb/publications/global_report/2009/pdf/full_report.pdf
- Beran, R. y Millar, P. (1986). Confidence sets for a multivariate distribution. *The Annals of Statistics*, pp. 431–443.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York, Inc., 2ª edición.
- Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B*, 41(2), pp. 113–147.
- (1997). Noninformative priors do not exist: A discussion. *J. Statistics Planning and Inference*, 65, pp. 159–189.
- Bernardo, J. M. y Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70, pp. 351–372.

- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.
- Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics= Revue Canadienne de Statistique*, 28(4), p. 783.
- Blyth, C. R. y Still, H. A. (1983). Binomial Confidence Intervals. *Journal of the American Statistical Association*, 78(381), pp. 108–116. ISSN 0162-1459.
- Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H. P.; Schölkopf, B. y Smola, A. J. (2006). Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14), pp. 49–57. doi: 10.1093/bioinformatics/btl242. <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/14/e49>
- Boser, B. E.; Guyon, I. M. y Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. En: *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152. ACM, New York, NY, USA.
- Botev, Z. I.; Grotowski, J. F. y Kroese, D. P. (2009). Kernel Density Estimation via Diffusion. *Annals of Statistics*.
- Brown, L. D.; Cai, T. y DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16, pp. 101–133.
- Burges, C. J. C. (1996). Simplified Support Vector Decision Rules. En: *International Conference on Machine Learning*, pp. 71–77.
- (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), pp. 121–167. citeseer.ist.psu.edu/burges98tutorial.html
- Cai, T. (2005). One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference*, 131(1), pp. 63–88.
- Cai, Y. y Krishnamoorthy, K. (2005). A simple improved inferential method for some discrete distributions. *Computational Statistics and Data Analysis*, 48(3), pp. 605–621.
- Cao, R. y Van Keilegom, I. (2006). Empirical likelihood tests for two-sample problems via nonparametric density estimation. *The Canadian Journal of Statistics*, 34(1), pp. 61–77.

- Cappe, O.; Guillin, A.; Marin, J. M. y Robert, C. P. (2004). Population Monte Carlo. *J. Comput. Graph. Statist*, 13(4), pp. 907–929.
- Cascina, A.; Fietta, A. y Casali, L. (2000). Is a Large Number of Sputum Specimens Necessary for the Bacteriological Diagnosis of Tuberculosis? *Journal of Clinical Microbiology*, 38(1), p. 466.
- Casella, G. (1986). Refining binomial confidence intervals. *The Canadian Journal of Statistics*, 14(2), pp. 113–129.
- Chafaï, D. y Concordet, D. (2009). Confidence regions for multinomial parameter with small sample size. *Journal of the American Statistical Association*.
- Chang, C. C. y Lin, C. J. (2001). *LIBSVM: a library for support vector machines*. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Chaturvedi, N. y Cockcroft, A. (1992). Tuberculosis screening in health service employees: who needs chest X-rays? *Occupational Medicine*, 42(4), pp. 179–182.
- Chernoff, H. y Ray, S. N. (1965). A Bayes sequential sampling inspection plan. *The Annals of Mathematical Statistics*, pp. 1387–1407.
- Clopper, C. J. y Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4), pp. 404–413.
- Conover, W. J. (1998). *Practical Nonparametric Statistics*. John Wiley & Sons. ISBN 0471160687.
- Cortes, C. y Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), pp. 273–297.
- Cover, T. M. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern recognition. *IEEE Transactions on Electronic Computers*, 14, pp. 326–334.
- Cover, T. M. y Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, 2^a edición. ISBN 0-471-24195-4.
- Cox, D. R. (1963). Large Sample Sequential Tests for Composite Hypotheses. *Sankhy: The Indian Journal of Statistics, Series A*, 25(1), pp. 5–12. ISSN 0581572X.
<http://www.jstor.org/stable/25049244>

- Crammer, K. y Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2, pp. 265–292.
- Dalal, S. R. y Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society. Series B. Methodological*, 45(2), pp. 278–286.
- Darkhovskii, B. S. (2006). Sequential testing of two composite statistical hypotheses. *Autom. Remote Control*, 67(9), pp. 1485–1499. ISSN 0005-1179. doi: <http://dx.doi.org/10.1134/S0005117906090104>.
- DasGupta, A. y Zhang, T. (2005). *Encyclopedia of Statistical Sciences*. capítulo Inference for binomial and multinomial parameters. Wiley.
<http://www.stat.purdue.edu/~tlzhang/binomialeneycl.pdf>
- (2006). On the false discovery rates of a frequentist: Asymptotic expansions. *Recent Developments in Nonparametric Inference and Probability*, 50, pp. 190–212.
- Davidson, R. y MacKinnon, J. G. (2000). Bootstrap tests: how many bootstraps? *Econometric Reviews*, 19(1), pp. 55–68.
- Decoste, D. y Schölkopf, B. (2002). Training Invariant Support Vector Machines. *Machine Learning*, 46(1), pp. 161–190.
- DiCiccio, T. J. y Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, 11, pp. 189–211.
- Dinnes, J.; Deeks, J.; Kunst, H.; Gibson, A.; Cummins, E.; Waugh, N.; Drobniewski, F. y Lalvani, A. (2007). A systematic review of rapid diagnostic tests for the detection of tuberculosis infection. *Health Technol Assess*, 11(3), pp. 1–196.
- Downs, T.; Gates, K. E. y Masters, A. (2001). Exact simplification of support vector solutions. *The Journal of Machine Learning Research*, 2, pp. 293–297.
- Durbin, J. (1951). Incomplete blocks in ranking experiments. *British Journal of Psychology*, 4, pp. 85–90.
- Edwards, W.; Miles, R. F. y Von Winterfeldt, D. (Eds.) (2007). *Advances in Decision Analysis: From Foundations to Applications*. Cambridge University Press.

- Efron, B. y Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.
- Einmahl, J. H. J. y Khmaladze, E. V. (2001). The two-sample problem in \mathbb{R}^m and measure-valued martingales. En: M. de Gunst; C. Klaassen y A. van de Vaart (Eds.), *State of the Art in Probability and Statistics. IMS Lecture Notes - Monograph Series*, volumen 36, pp. 434–463. IMS, Beachwood, Ohio.
- Einmahl, J. H. J. y McKeague, I. W. (2003). Empirical likelihood based hypothesis testing. *Bernoulli-London-*, 9(2), pp. 267–290.
- Fears, T. R.; Benichou, J. y Gail, M. H. (1996). A reminder of the fallibility of the wald statistic. *The American statistician*, 50(3), pp. 226–227.
- Fink, D. (1997). A Compendium of Conjugate Priors.
<http://www.people.cornell.edu/pages/df36/CONJINTRnew%20TEX.pdf>
- Finley, T. y Joachims, T. (2008). Training structural SVMs when exact inference is intractable. En: *Proceedings of the 25th international conference on Machine learning (ICML 2008)*, pp. 304–311.
- Fisher, R. A. (1935). The design of experiments. *New York: Hafner*.
- Fitzgerald, W. J. (2001). Markov chain Monte Carlo methods with applications to signal processing. *Signal Processing*, 81(1), pp. 3–18.
- Fleiss, J. L.; Levin, B. y Paik, M. C. (2004). *Statistical methods for rates and proportions*. Wiley-Interscience.
- Fletcher, R. (1987). *Practical methods of optimization; (2nd ed.)*. Wiley-Interscience, New York, NY, USA.
- Forero, M. G.; Cristóbal, G. y Alvarez-Borrego, J. (2003). Automatic identification techniques of tuberculosis bacteria. En: *SPIE proceedings of the applications of digital image processing XXVI*, volumen 5203, pp. 71–81.
- Forero, M. G.; Cristobal, G. y Desco, M. (2006). Automatic identification of Mycobacterium tuberculosis by Gaussian mixture models. *Journal of Microscopy*, 223(2), pp. 120–132. doi: 10.1111/j.1365-2818.2006.01610.x.
<http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-2818.2006.01610.x>

- Forero, M. G.; Sroubek, F. y Cristóbal, G. (2004). Identification of tuberculosis bacteria based on shape and color. *Real-Time Imaging*, 10(4), pp. 251–262.
<http://dx.doi.org/10.1016/j.rti.2004.05.007>
- Fortet, R. y Mourier, E. (1953). Convergence de la répartition empirique vers la répartition théorique. *Annales scientifiques de l'École Normale Supérieure Sér. 3*, 70(3), pp. 267–285.
- Frazier, P. I. y Yu, A. J. (2007). Sequential Hypothesis Testing under Stochastic Deadlines. En: *Neural Information Processing Systems*, .
- French, S. (1986). *Decision theory: an introduction to the mathematics of rationality*. Ellis Horwood Series In Mathematics And Its Applications.
- Freund, Y. y Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1), pp. 119–139.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, pp. 675–701.
- Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. New York: Academic Press.
- Ghosh, B. K. (1970). *Sequential tests of statistical hypotheses*. Addison-Wesley Educational Publishers Inc., US.
- Ghosh, B. K. y Sen, P. K. (1991). *Handbook of sequential analysis*. Marcel Dekker.
- Glaz, J. y Sison, C. P. (1999). Simultaneous confidence intervals for multinomial proportions. *Journal of Statistical Planning and Inference*, 82(1-2), pp. 251–262.
- Good, P. (2004). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer-Verlag New York.
- Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics*, pp. 247–254.
- Govindarajulu, Z. (2004). *Sequential statistics*. World Scientific.
- Gretton, A.; Fukumizu, K.; Teo, C. H.; Song, L.; Schölkopf, B. y Smola, A. J. (2008). A Kernel Statistical Test of Independence. En: J. C. Platt; D. Koller; Y. Singer

- y S. Roweis (Eds.), *21th Neural Information Processing Systems Conference*, pp. 585–592. MIT Press, Cambridge, MA, USA.
http://books.nips.cc/papers/files/nips20/NIPS2007_0730_slide.pdf
- Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B. y Smola, A. J. (2007). A Kernel Method for the Two-Sample-Problem. En: B. Schölkopf; J. Platt y T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19*, pp. 513–520. MIT Press, Cambridge, MA.
- Gutiérrez-Peña, E. y Walker, S. G. (2005). Statistical Decision Problems and Bayesian Nonparametric Methods. *International Statistical Review*, 73(3), pp. 309–330.
- Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer Verlag.
- (1987). On the Bootstrap and Likelihood-Based Confidence Regions. *Biometrika*, 74(3), pp. 481–493. ISSN 00063444.
<http://www.jstor.org/stable/2336687>
- Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall.
- Henderson, M. y Meyer, M. C. (2001). Exploring the Confidence Interval for a Binomial Parameter in a First Course in Statistical Computing. *The American Statistician*, 55(4), pp. 337–344. ISSN 0003-1305.
- Hoeffding, W. (1960). Lower bounds for the expected sample size and the average risk of a sequential procedure. *The Annals of Mathematical Statistics*, pp. 352–368.
- (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist*, 36, pp. 369–408.
- Howard, R. (2000). Decisions in the face of Uncertainty. En: C. Alexander (Ed.), *Visions of Risk*, London: Pearson Education Limited.
- Huber, P. J. y Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, pp. 251–263.
- Huffman, M. D. (1983). An Efficient Approximate Solution to the Kiefer-Weiss Problem. *The Annals of Statistics*, 11(1), pp. 306–316. ISSN 00905364.
<http://www.jstor.org/stable/2240484>
- Hyvarinen, A.; Karhunen, J. y Oja, E. (2001). *Independent Component Analysis*. Wiley & Sons.

- Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4, pp. 227–41.
- Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.
- Jing, B. Y. I. (1995). Two-sample empirical likelihood method. *Statistics & probability letters*, 24(4), pp. 315–319.
- Johansson, M. y Olofsson, T. (2007). Bayesian Model Selection for Markov, Hidden Markov, and Multinomial Models. *IEEE Signal Processing Letters*, 14(2), pp. 129–132.
- Jordan, J. (2006). *Pascal's Wager: Pragmatic Arguments and Belief in God*. Oxford University Press.
- Kay, S. M. (1998). *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. Prentice Hall PTR.
- Keerthi, S. S.; Chapelle, O. y DeCoste, D. (2006). Building Support Vector Machines with Reduced Classifier Complexity. *The Journal of Machine Learning Research*, 7, pp. 1493–1515.
- Kiefer, J. y Weiss, L. (1957). Some Properties of Generalized Sequential Probability Ratio Tests. *The Annals of Mathematical Statistics*, 28(1), pp. 57–74. ISSN 00034851.
<http://www.jstor.org/stable/2237023>
- Kim, K. I.; Franz, M. O. y Schölkopf, B. (2005). Iterative Kernel Principal Component Analysis for Image Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1351–1366.
- Kim, K. K. y Foutz, R. V. (1987). Tests for the Multivariate Two-Sample Problem Based on Empirical Probability Measures. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 15(1), pp. 41–51. ISSN 0319-5724.
- Kolmogorov, A. N. (1933). Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, 4, pp. 1–11.
- Kruskal, W. y Wallis, W. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, pp. 583–621.
- Kumar, V. y Cotran, R. S. (2003). *Robbins basic pathology*. Saunders Philadelphia.

- Kwok, J. T. Y. y Tsang, I. W. H. (2004). The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*, 15(6), pp. 1517–1525.
- Lai, T. L. (1988). Nearly optimal sequential tests of composite hypotheses. *The Annals of Statistics*, pp. 856–886.
- (1997). On optimal stopping problems in sequential hypothesis testing. *Statistica Sinica*, 7, pp. 33–52.
- (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, 11(2), pp. 303–350.
- Lai, T. L. y Zhang, L. (1994). A modification of Schwarz’s sequential likelihood ratio tests in multivariate sequential analysis. *Sequential Analysis*, 13(2), pp. 79–96.
- Lawrence, N.; Seeger, M. y Herbrich, R. (2003). Fast Sparse Gaussian Process Methods: The Informative Vector Machine. *Advances in Neural Information Processing Systems*, pp. 625–632.
- LeCun, Y.; Bottou, L.; Bengio, Y. y Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), pp. 2278–2324.
- Lehmann, E. L. (1997). *Testing statistical hypotheses*. Springer, 2ª edición.
- Lehmann, E. L. y D’Abrera, H. J. M. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day.
- Leiva-Murillo, J. M. y Artés-Rodríguez, A. (2004). A Gaussian Mixture Based Maximization of Mutual Information for Supervised Feature Extraction. En: *5th International Conference ICA 2004*, pp. 271–278. Granada, Spain.
- Levitán, E. y Merhav, N. (2002). A competitive Neyman-Pearson approach to universal hypothesis testing with applications. *IEEE Transactions on Information Theory*, 48(8), pp. 2215–2229. ISSN 0018-9448. doi: 10.1109/TIT.2002.800478.
- Li, Q.; Jiao, L. y Hao, Y. (2007). Adaptive simplification of solution for support vector machine. *Pattern Recognition*, 40(3), pp. 972–980.
- Lorden, G. (1976). 2-SPRT’s and the modified Kiefer-Weiss problem of minimizing an expected sample size. *The Annals of Statistics*, pp. 281–291.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), pp. 91–110.

- MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.
- Mann, H. B. y Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pp. 50–60.
- Mika, S.; Ratsch, G.; Weston, J.; Schölkopf, B. y Muller, K. R. (1999). Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pp. 41–48. doi: 10.1109/NNSP.1999.788121.
- Nelson, S. M.; Deike, M. A. y Cartwright, C. P. (1998). Value of Examining Multiple Sputum Specimens in the Diagnosis of Pulmonary Tuberculosis. *Journal of Clinical Microbiology*, 36(2), pp. 467–469.
- Neyman, J. y Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, pp. 289–337. ISSN 0264-3952.
- Nguyen, D. y Ho, T. (2005). An efficient method for simplifying support vector machines. En: *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pp. 617–624.
- North, D. W. (1968). A Tutorial Introduction to Decision Theory. *IEEE Transactions on Systems Science and Cybernetics*, 4(3), pp. 200–210.
- Organización Mundial de la Salud (2003). Nota descriptiva OMS N°104. <http://www.who.int/mediacentre/factsheets/fs104/es/index.html>
- Osuna, E. y Girosi, F. (1998). Reducing the run-time complexity of support vector machines. En: *ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition*, .
- Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC.
- Parrado-Hernández, E.; Mora-Jiménez, I.; Arenas-García, J.; Figueiras-Vidal, A. R. y Navia-Vázquez, A. (2003). Growing support vector classifiers with controlled complexity. *Pattern Recognition*, 36(7), pp. 1479–1488.
- Parzen, E. (1962). On the Estimation of Probability Density Function and the Mode. *Annals of Mathematical Statistics*, 33, pp. 1065–1076.

- Pawitan, Y. (2000). A reminder of the fallibility of the Wald statistic: Likelihood explanation. *The American statistician*, 54(1), pp. 54–56.
- Pearson, K. (1922). On the χ^2 test of Goodness of Fit. *Biometrika*, 14(1-2), pp. 186–191.
- Perez-Cruz, F. (2008). Estimation of Information Theoretic Measures for Continuous Random Variables. En: *Advances on Neural Information Processing (NIPS)*, Vancouver (Canada).
- Platt, J. (2000). Probabilities for SV Machines. En: A. J. Smola; P. L. Bartlett; B. Schölkopf y D. Schuurmans (Eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, Cambridge, MA.
- Poor, H. V. (1994). *An introduction to signal detection and estimation*. Springer-Verlag New York, Inc., New York, NY, USA, 2ª edición. ISBN 0-387-94173-8.
- Qin, J. (1994). Semi-empirical likelihood ratio confidence intervals for the difference of two sample means. *Annals of the Institute of Statistical Mathematics*, 46(1), pp. 117–126.
- Quade, D. (1979). Using weighted rankings in the analysis of complete blocks with additive block effects. *Journal of the American Statistical Association*, pp. 680–683.
- Quesenberry, C. y Hurst, D. (1964). Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, pp. 191–195.
- Rasmussen, C. E. y Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.
- Raykar, V. C. y Duraiswami, R. (2006). Fast optimal bandwidth selection for kernel density estimation. En: J. Ghosh; D. Lambert; D. Skillicorn y J. Srivastava (Eds.), *Proceedings of the sixth SIAM International Conference on Data Mining*, pp. 524–528.
- Robert, C. P. y Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.
- Roberts, S. J.; Husmeier, D.; Rezek, I. y Penny, W. (1998). Bayesian Approaches to Gaussian Mixture Modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), pp. 1133–1142.

- Robledo, J. A.; Mejia, G. I.; Morcillo, N.; Chacon, L.; Camacho, M.; Luna, J.; Zurita, J.; Bodon, A.; Velasco, M.; Palomino, J. C.; Martin, A. y Portaels, F. (2006). Evaluation of a rapid culture method for tuberculosis diagnosis: a Latin American multi-center study. *International Journal of Tuberculosis and Lung Disease*, 10(6), p. 613.
- Romdhani, S.; Torr, P.; Schölkopf, B. y Blake, A. (2001). Computationally Efficient Face Detection. En: *Proceedings of the International Conference on Computer Vision*, volumen 2, pp. 695–700.
- Romdhani, S.; Torr, P.; Schölkopf, B. y Blake, A. (2004). Efficient face detection by a cascaded support-vector machine expansion. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 460(2051), pp. 3283–3297.
- Roweis, S. y Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), pp. 2323–2326.
- Santiago-Mozos, R.; Fernández-Lorenzana, R.; Pérez-Cruz, F. y Artés-Rodríguez, A. (2008). On the Uncertainty in Sequential Hypothesis Testing. En: *Fifth IEEE International Symposium on Biomedical Imaging (ISBI '08)*, pp. 1123–1126. Paris, France.
- Santiago-Mozos, R.; Pérez-Cruz, F. y Artés-Rodríguez, A. (2009a). Extended Input Space Support Vector Machine. Enviado a Advances on Neural Information Processing (NIPS).
- (2009b). Sequential decision under uncertain hypotheses. Enviado a IEEE Signal Processing Letters.
- Santiago-Mozos, R. y Artés-Rodríguez, A. (2006a). Distributed Hypothesis Testing Using Local Learning based classifiers. En: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volumen 4, pp. 861–864.
- (2006b). Uncertainty-based Censoring Scheme in Distributed Detection Using Learning Techniques. En: *Proceedings of the Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pp. 2027–2034.
- Sarawagi, S. y Gupta, R. (2008). Accurate max-margin training for structured output spaces. En: *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pp. 888–895.
- Scharf, L. L. y Demeure, C. (1991). *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley Pub. Co..

- Schölkopf, B.; Knirsch, P.; Smola, A. y Burges, C. (1998). Fast approximation of support vector kernel expansions, and an interpretation of clustering as approximation in feature spaces. En: *Mustererkennung*, volumen 20, pp. 124–132.
- Schölkopf, B. y Smola, A. (2001). *Learning with Kernels*. MIT Press, Cambridge, MA, USA.
- Schwarz, G. (1962). Asymptotic shapes of Bayes sequential testing regions. *The Annals of Mathematical Statistics*, pp. 224–236.
- Shao, J. y Tu, D. (1995). *The jackknife and bootstrap*. Springer.
- Sheather, S. J. y Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 683–690.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.
- Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Moscow University Mathematics Bulletin*, 2, pp. 3–14.
- Sobel, M. (1953). An Essentially Complete Class of Decision Functions for Certain Standard Sequential Problems. *The Annals of Mathematical Statistics*, 24(3), pp. 319–337. ISSN 00034851.
<http://www.jstor.org/stable/2236284>
- Steingart, K.; Henry, M.; Ng, V.; Hopewell, P.; Ramsay, A.; Cunningham, J.; Urbaniczik, R.; Perkins, M.; Aziz, M. y Pai, M. (2006). Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. *Lancet Infect Dis*, 6(9), pp. 570–81.
- Steinwart, I. (2004). Sparseness of Support Vector Machines—some asymptotically sharp bounds. *Advances in Neural Information Processing Systems*, 16, pp. 1069–1076.
- Tantaratana, S. y Poor, H. (1982). Asymptotic efficiencies of truncated sequential tests. *IEEE Transactions on Information Theory*, 28(6), pp. 911–923.
- Tantaratana, S. y Thomas, J. (1977). On sequential sign detection of a constant signal. *IEEE Transactions on Information Theory*, 23(3), pp. 304–315.

- Tenenbaum, J. B.; de Silva, V. y Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500), pp. 2319–2323.
- Thompson, S. K. (1987). Sample Size for Estimating Multinomial Proportions. *The American Statistician*, 41(1), pp. 42–46. ISSN 00031305.
<http://www.jstor.org/stable/2684318>
- Tipping, M. E. (2001). Sparse bayesian learning and the Relevance Vector Machine. *The Journal of Machine Learning Research*, 1, pp. 211–244.
- Tsochantaridis, I.; Joachims, T.; Hofmann, T. y Altun, Y. (2005). Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research*, 6(2), pp. 1453–1484.
- Van der Waerden, B. L. (1952). Order tests for two-sample problem and their power. *Indagationes Mathematicae*, 14(253), p. 458.
- Van Trees, H. L. (2001). *Detection, Estimation, and Modulation Theory: Part I*. Wiley New York.
- (1992). *Detection, Estimation, and Modulation Theory: Radar-Sonar Signal Processing and Gaussian Signals in Noise*. Krieger Publishing Co., Inc., Melbourne, FL, USA. ISBN 0894647482.
- Vapnik, V. (1982). *Estimation of dependences based on empirical data*. Springer.
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag, New York.
- (1998). *Statistical Learning Theory*. John Wiley & Sons, New York.
- Varshney, P. K. (1996). *Distributed Detection and Data Fusion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. ISBN 0387947124.
- Veropoulos, K.; Learmonth, G.; Campbell, C.; Knight, B. y Simpson, J. (1999). The Automated Identification of Tubercle Bacilli in Sputum: A Preliminary Investigation. *Analytical and Quantitative Cytology and Histology*, 21(4), pp. 277–281.
- Vincent, P. y Bengio, Y. (2002). Kernel Matching Pursuit. *Machine Learning*, 48(1), pp. 165–187.
- Viola, P. y Jones, M. (2001). Robust Real-time Object Detection. En: *Second International Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing, and Sampling*, .

- Wald, A. (1947). *Sequential Analysis*. John Wiley and Sons, New York.
- Wald, A. y Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, 19(3), pp. 326–339. ISSN 0003-4851.
- Walker, S. G. y Gutiérrez-Peña, E. (2007). Bayesian parametric inference in a nonparametric framework. *TEST*, 16(1), pp. 188–197.
- Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.
- (2005). *All of Statistics*. Springer, New York.
- Weston, J. y Watkins, C. (1999). Multi-class Support Vector Machines. En: *Proceedings of the Seventh European Symposium On Artificial Neural Networks*, .
- Wijsman, R. A. (1991). *Handbook of Sequential Analysis*. capítulo Stopping times: Termination, moments, distribution. Marcel Dekker.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, pp. 80–83.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of The American Statistical Association*, 22, pp. 209–212.
- Yu, B.; Dazzo, F. B.; Srivatsa, R.; Zhang, T. y Jain, A. K. (1997). A Computer-aided System for Image Analysis of Morphological Diversity, Abundance and Spatial Distribution. *Informe técnico MSU-CPS-97-24*, Department of Computer Science, Michigan State University, East Lansing, Michigan.
- Zeitouni, O.; Ziv, J. y Merhav, N. (1992). When is the generalized likelihood ratio test optimal? *IEEE Transactions on Information Theory*, 38(5), pp. 1597–1602.
- Zhan, Y. y Shen, D. (2005). Design efficient support vector machine for fast classification. *Pattern Recognition*, 38(1), pp. 157–161.
- Zhang, B. (2000). Estimating the treatment effect in the two-sample problem with auxiliary information. *Journal of Nonparametric Statistics*, 12(3), pp. 377–389.
- Zheng, W. S. y Lai, J. (2006). Regularized Locality Preserving Learning of Pre-Image Problem in Kernel Principal Component Analysis. En: *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, volumen 2, pp. 456–459.

Zheng, W. S.; Lai, J. H. y Yuen, P. C. (2006). Weakly Supervised Learning on Pre-image Problem in Kernel Methods. En: *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, volumen 2, pp. 711–715.

Agradecimientos

Mi primer agradecimiento es para Dios que provee de todo lo necesario para que este trabajo se pueda realizar: vida, salud, capacidad, directores e incluso el trabajo mismo que Dios ha provisto para los hombres nos ocupemos en él. También quiero dar gracias a Dios por Jesucristo que murió en una cruz en nuestro lugar para que por medio de la fe y el arrepentimiento recibamos perdón en ese Juicio por el que todos vamos a pasar una vez muertos.

También quiero agradecer a Blaise Pascal (1670) por La Apuesta de Pascal (Jordan, 2006). Aquí Pascal nos presenta uno de los primeros ejemplos formales de problema de decisión. Este problema establece un escenario donde el estado es binario: Dios existe o no existe; y las acciones son dos: vivir como si Dios existe o como si no existe. Pascal propuso el siguiente cuadro de decisión en la que se muestra el beneficio obtenido de cada decisión dependiendo de la existencia de Dios y nuestra decisión donde K_1 y K_2 son constantes finitas y concluye que creer en Dios y vivir como si existiese es la apuesta más segura.

	Dios existe	Dios no existe
Vivir como si Dios existiese	∞ (cielo)	K_1
Vivir como si Dios no existiese	$-\infty$ (infierno)	K_2

Que Dios existe, como dice el Apóstol Pablo a los romanos, es claramente visible por medio de las cosas hechas. Sin embargo, me permito mostrar dicho cuadro para invitar a una reflexión al lector y dar crédito a este eminente matemático cuya lectura me animó en los comienzos de la redacción de la tesis.

Quiero agradecer a mis tutores, el Dr. Antonio Artés Rodríguez y el Dr. Fernando Pérez Cruz, su dirección, su apoyo, su confianza y su estímulo sin los cuáles este trabajo no habría visto la luz. También quiero agradecer a Manuel Desco los datos y el problema que dió lugar a los desarrollos mostrados en este trabajo.

Quiero agradecer tantos buenos ratos a mis compañeros del Chivi-Lab José Miguel Leiva, Mario de Prado, Sancho Salcedo y Ricardo Torres, su amistad y su amplio sentido del humor, también a José Emilio que se exilió por un tiempo a Ginebra. Del mismo modo, quiero agradecer a Luca Martino y a Eduardo Masó por nuestras “profundas” conversaciones y a Joaquín Escudero mi compañero de fatigas en mi nueva etapa como ayudante.

Estos años he disfrutado la amistad y las sonrisas de Matilde Sánchez, María Julia Fernandez-Getino, Ana Belén Rodríguez y Marcelino Lázaro. Agradezco a estos dos últimos su entrenamiento en las tareas de trasegar cerveza, hablar correctamente, poner las comas en su sitio y muchás más entrañables experiencias que han ocupado mi tiempo libre.

Gracias a Harold Molina por su paciencia con mis extrañas peticiones de recursos informáticos, a Saúl Blanco por su profesionalidad, a Emilio Parrado por sus amables sobremesas de pizarra, a Ascensión Gallardo por sus ánimos y su simpatía, a Jerónimo Arenas por la Eurocopa de 2008 y por sus, algunas veces controvertidos, puntos de vista, a Luis de Inclán por su ayuda con los trámites de la tesis y a Eva Rajo por su carácter gallego y porque dice no cuando quiere decir no.

A mi madre María del Carmen, mis hermanos Jacobo y Salvador y mis tías Joaquina y María por su apoyo incondicional y sus preguntas de para cuándo la tesis.

Y a Beatriz, por su ánimo, paciencia y comprensión, por su alegría y su voz y por todos los motivos que cada día me da para seguir adelante.

Mis más sinceras gracias.

