Dpto. de Teoría de la Señal y Comunicaciones
Universidad Carlos iii de Madrid

TESIS DOCTORAL

COST-SENSITIVE CLASSIFICATION BASED ON BREGMAN DIVERGENCES

Autor: Raúl Santos Rodríguez
Director: Dr. Jesús Cid Sueiro

Leganés, 2011

Tesis doctoral:
COST-SENSITIVE CLASSIFICATION BASED ON BREGMAN DIVERGENCES.

Autor:
RAÚL SANTOS RODRÍGUEZ

Director:
DR. JESÚS CID SUEIRO

El tribunal nombrado para juzgar la tesis doctoral arriba citada, compuesto por los doctores

Presidente:

Vocal:

Vocal:

Vocal:

Secretario:

acuerda otorgarle la calificación de

Leganés, a

RESUMEN EXTENDIDO

En este resumen se pretende dar una visión de conjunto del trabajo realizado durante la elaboración de la presente Tesis Doctoral. Tras introducir el objetivo general de la misma, describimos la organización y las aportaciones originales del trabajo de investigación para por último presentar las conclusiones que se consideran más relevantes.

Motivación y metodología

El principal objetivo de esta Tesis es la definición y estudio de nuevas *funciones de pérdida* que sean de utilidad a la hora de encontrar soluciones para el problema conocido dentro del aprendizaje máquina como *clasificación sensible a costes*.

La mayoría de los problemas de clasificación que se plantean en situaciones cotidianas son sensibles a costes. Desde un punto de vista abstracto, si hay una persona (o programa informático) que realiza acciones a partir de una decisión, las consecuencias de esas acciones pueden depender de la propia decisión, de la clase a la que pertenecen los datos, del valor de las observaciones o de otros factores que puede que ni siquiera sean observables antes de tomar la decisión. Cuantificar esas consecuencias es el primer paso para trabajar con máquinas sensibles al coste.

En la literatura de aprendizaje máquina se aprecia una preferencia por métodos que no tienen en cuenta los costes a la hora de clasificar. Esto no es extraño debido a que la evaluación de los costes entraña dificultades hasta en los problemas más simples. Como consecuencia, la mayoría de las bases de datos típicas en las que se evalúan algoritmos de clasificación, fundamentales para el trabajo experimental de los investigadores, no ofrecen información relativa al coste. Además, en general, realizar una predicción acertada sobre la clase a la que pertenecen los datos es menos costoso que errar (lo contrario llevaría a situaciones anómalas). Por lo tanto, diseñar sistemas que minimicen el error de la decisión parece una asunción razonable en

muchas situaciones, aunque no deja de ser subóptima. En cambio, nuestro objetivo se centra en minimizar el coste total en vez del número de errores. Es sencillo encontrar ejemplos en los que el aprendizaje sensible a costes resulta desde útil hasta vital. Por ejemplo, considere el problema de clasificar clientes en función de su riesgo de cara a conseguir créditos. La tarea consiste en separar a los posibles prestatarios en dos categorías: buenos o malos clientes. Desde el punto de vista del prestamista, el coste de clasificar un cliente de alto riesgo como bueno puede ser muy superior al coste de clasificar como malo un cliente de bajo riesgo. Este comportamiento se observa también en el campo del diagnóstico médico: un falso positivo (clasificar a un paciente como enfermo cuando está sano) puede ser menos costoso que clasificar un paciente enfermo como sano.

De entre los primeros trabajos rigurosos en esta dirección podemos destacar [Elkan, 2001a], que estableció una descripción del problema de aprendizaje con costes que sigue vigente en la actualidad. Nuestra propuesta encuentra su origen dentro del marco dispuesto por Charles Elkan, el cual se basa en la teoría de la decisión de Bayes a la hora de asignar las muestras a la clase con el mínimo coste esperado. Las reglas de decisión que siguen esta formulación se definen a través de la información disponible de los costes y de las probabilidades a posteriori de cada una de las clases. Si, como primera aproximación al problema, aceptamos que los costes tienen un valor determinista conocido, nuestro problema se centra en encontrar estimaciones precisas de las probabilidades a posteriori. Para llevar a cabo esta tarea debemos disponer de arquitecturas de aprendizaje adecuadas, conjuntos de datos que sean representativos del problema y una función de pérdidas para minimizar durante el proceso de aprendizaje. Asumimos que el conjunto de datos nos es dado y no tenemos control sobre él. Por el contrario, el diseño de la función de pérdidas nos da margen para proponer modelos que tengan en cuenta los costes: la teoría clásica de la decisión muestra que las matrices de costes definen las fronteras a través de las estimaciones de las probabilidades a posteriori de las mismas. Se puede concluir que, para tomar decisiones óptimas, sólo son necesarias estimaciones de las probabilidades que sean

precisas cerca de las fronteras de decisión. Esto nos lleva a pensar en estimas de probabilidad que sean más sensibles a las muestras que están cerca de la frontera que a las que se encuentran alejadas. Es importante destacar que la elección de la función de pérdidas es especialmente relevante cuando nuestra información a priori sobre el problema sea reducida, la arquitectura sea limitada o en el caso en el que los datos de entrenamiento de los que disponemos no sean adecuados en algún sentido. Es en estas situaciones donde el empleo de diferentes funciones de pérdida puede llevar a resultados drásticamente dispares.

Para nuestro estudio nos centraremos en funciones de pérdidas que pertenecen a la familia de las *divergencias de Bregman* [Bregman, 1967]. Estas medidas surgieron de la mano de L.M. Bregman en el ámbito de la obtención de probabilidades en estadística. Las razones de esta elección se pueden resumir en tres. En primer lugar, se trata de un conjunto de divergencias que incluye a algunas de las funciones de pérdidas más conocidas: la pérdida logística, la cuadrática o la exponencial son sólo algunos ejemplos de las posibilidades que ofrece. En segundo lugar, se trata de una familia con propiedades muy adecuadas para su manipulación con herramientas matemáticas sencillas. Por último, cabe destacar la relevancia que han adquirido en los últimos años en las comunidades de aprendizaje máquina y minería de datos [Banerjee et al., 2005b, Reid and Williamson, 2009a, Nock and Nielsen, 2009] debido a que ofrecen la posibilidad de reintrepretar, desde un punto de vista teórico, algoritmos tan extendidos como las *Máquinas de Vectores Soporte* o *Adaboost*. A pesar de este interés, no había trabajos recientes que empleasen las divergencias de Bregman en escenarios de aprendizaje sensible a costes.

Si retrocedemos hasta finales de la década de los noventa, en el campo de las redes neuronales, se encuentran estudios que presentan resultados relativos a las divergencias de Bregman, aunque bajo otras denominaciones. Se exploran posibles aplicaciones de divergencias alternativas a las clásicas. Finalmente, en [Cid-Sueiro and Figueiras-Vidal, 2001] se sugiere la posibilidad de aplicar estas divergencias a aprendizaje sensible a costes y aprendizaje semi-supervisado. En cuanto

a la primera idea, los resultados son prometedores en casos en los que arquitectura del estimador es limitada. Sin embargo se limita a trabajar con costes en errores y no se analiza el caso en el que los costes son dependientes de la muestra. Además, no se realiza un análisis asintótico del comportamiento de las divergencias. En cuanto al aprendizaje semi-supervisado, se realiza un trabajo experimental preliminar sobre el problema de sensado remoto y no se explora el uso combinado de muestras no etiquetadas y aprendizaje sensible a costes.

En general, en la literatura no existe una metodología general de diseño de clasificadores sensibles a costes susceptible de aplicarse a problemas binarios y multiclase, supervisados o semi-supervisados, con costes dependientes e independientes de las muestras. Por el contrario, los métodos propuestos son en muchos casos especialmente difíciles de aplicar a modelos multiclase, o con datos no etiquetados, o con costes dependientes de las muestras, o con incertidumbre en los costes, o con conocimiento de los costes en test, ... Esto indica que, bajo nuestro punto de vista, aunque es posible diseñar clasificadores sensibles a costes empleando otros criterios de optimización, las divergencias de Bregman constituyen una aproximación natural a los diferentes problemas. En primer lugar, conducen a estimaciones de probabilidades que, de acuerdo con la teoría de la decisión, son estadísticos suficientes para clasificación con costes. En segundo lugar, las divergencias de Bregman tienen una estructura fácilmente interpretable, con un término de *entropía* independiente de las etiquetas y un término de error dependiente de las etiquetas, que permite adaptar el diseño a escenarios muy distintos.

Aportaciones originales de la Tesis

La Tesis tiene como punto de partida el diseño de una novedosa familia paramétrica de divergencias de Bregman. Uno de sus rasgos más destacados es que ofrece la posibilidad de crear una divergencia específica e individual para cada

problema concreto: la divergencia integra la información de la matriz de costes del problema en cuestión en su expresión analítica. Otro aspecto fundamental es que, debido a su formulación, resulta muy adecuada para problemas con más de dos clases. Los problemas multiclase con costes ha supuesto un reto hasta la fecha y la solución propuesta en esta Tesis ofrece una forma natural de abordarlos.

A partir de esta nueva familia de divergencias se desarrollan cuatro líneas fundamentales:

- **Clasificación supervisada sensible a costes**: En este sentido derivamos varios resultados asintóticos que caracterizan las divergencias propuestas. Un primer análisis garantiza que la divergencia de Bregman tiene tiene una sensibilidad máxima a cambios en vectores de probabilidad que se encuentran cerca de la frontera de decisión. Esto quiere decir que se cumple la propiedad que buscábamos y podemos garantizar que disponemos de una familia de medidas que da prioridad a obtener estimaciones de probabilidad más precisas en regiones que se encuentran cerca de zonas críticas: no tendremos dudas a la hora de clasificar puntos que se encuentran muy lejos de la frontera y por ello vale la pena intentar volcar todo el potencial de nuestras divergencias sobre las áreas más complicadas, las fronteras de decisión. Un segundo análisis garantiza que la optimización de nuestra divergencia resulta equivalente a una minimización del coste total en problemas donde los conjuntos de datos no son separables. Por último, establecemos relaciones entre la minimización de la divergencia propuesta y nociones relativas al *máximo margen*, siendo este el clasificador límite.

- **Clasificación semi-supervisada sensible a costes**: Habitualmente se plantean escenarios donde se recibe un conjunto escaso de datos etiquetados pero es posible encontrar datos no etiquetados en grandes cantidades. Normalmente, conseguir cada muestra etiquetada supone un esfuerzo mucho mayor que el que representan los datos no etiquetados. Es en estos casos en los que

se aplica el paradigma de aprendizaje semi-supervisado: se trata de buscar un buen clasificador para los datos etiquetados, modificando la solución para tener en cuenta de alguna modo los datos no etiquetados. En relación con nuestra familia de divergencias de Bregman, discutimos la posibilidad de emplear el principio de *Minimización de la Entropía*. Esta idea consiste en buscar fronteras de decisión que hagan que la entropía total de los datos sea baja. Para ello, se favorecen las soluciones en las que las fronteras de decisión atraviesan regiones de baja densidad de puntos. Para ello proponemos un planteamiento del problema que resulta estar relacionado con métodos muy relevantes de la literatura, incluyendo la *Regularización por Entropía* y las *Máquinas de Vectores Soporte Transductivas*. El resultado final es el primer algoritmo de aprendizaje semi-supervisado sensible a costes para problemas multiclase.

- **Definición y bases de las secuencias de divergencias de Bregman**: Se discute la transformación de las familias paramétricas de divergencias de Bregman en *secuencias* de divergencias de Bregman. En primer lugar tratamos de motivar la necesidad de este nuevo concepto a partir del estudio de las posibles relaciones entre las divergencias de Bregman y la convexidad. Para ello derivamos y redefinimos resultados sobre las propiedades de las funciones de activación y su papel en el aprendizaje. Posteriormente, intentamos responder a la pregunta de si es posible encontrar secuencias de divergencias de Bregman sensibles a costes que satisfagan propiedades similares a las descritas en los puntos anteriores, siendo la respuesta positiva. Bajo condiciones bastante generales es posible diseñar secuencias de divergencias cuya minimización lleve, asintóticamente, a soluciones de mínimo riesgo (sensible a costes) en problemas no separables y maximice un margen generalizado en problemas separables.

- **Aprendizaje sensible a costes cuando los costes son dependientes del ejemplo**: Así como considerar que un problema no es sensible a costes cuando intrínsecamente sí lo es, asumir que los costes son deterministas

puede ser incorrecto en muchas situaciones. En cualquiera de los ejemplos presentados anteriormente, decir que el coste es fijo para todos los casos es una aproximación que puede dar lugar a inconvenientes. Cada cliente o cada paciente son diferentes y eso puede verse reflejado en el coste que llevan asociado. En este sentido se propone una generalización de las secuencias de divergencias de Bregman para conseguir introducir esta nueva información de la forma más natural posible. Eso nos conduce a diferentes enfoques que tienen como resultado un algoritmo orientado a este problema, a veces olvidado en la literatura.

Conclusiones

A lo largo de esta Tesis exploramos la aplicación de divergencias de Bregman a problemas sensibles a costes. La flexibilidad de la formulación se demuestra en diferentes escenarios: aprendizaje supervisado, aprendizaje semi-supervisado y en escenarios donde los costes dependen de la muestra en cuestión. La idea clave es asociar cada muestra o conjunto de muestras del conjunto de entrenamiento con su propia divergencia, adaptada a sus costes. De esta manera, la función de pérdidas derivada de la divergencia refleja intrínsicamente la estructura del coste del problema. Optimizar la función resultante conduce a estimaciones de probabilidades de posteriori que son especialmente sensibles y precisas cerca de las fronteras de decisión óptimas. Este enfoque explota de forma natural la capacidad de las máquinas de aprendizaje para clasificación, poniendo énfasis en las zonas más relevantes del mapa de probabilidades a posteriori.

Cabe resaltar que el método presentado se beneficia de ventajas propias de métodos clásicos de clasificación discriminativa como las Máquinas de Vectores Soporte, así como las de métodos que se basan en estimar probabilidades. Ejemplos de ello se pueden encontrar a lo largo de los diferentes capítulos de forma explícita:

la relación con algoritmos de máximo margen, las resultados asintóticos de minimización del coste total o la posibilidad de utilizar los costes de las muestras en test (siempre que esa información esté disponible).

ABSTRACT

The main object of this PhD. Thesis is the identification, characterization and study of new loss functions to address the so-called *cost-sensitive classification*. Many decision problems are intrinsically cost-sensitive. However, the dominating preference for cost-insensitive methods in the machine learning literature is a natural consequence of the fact that true costs in real applications are difficult to evaluate. Since, in general, uncovering the correct class of the data is less costly than any decision error, designing low error decision systems is a reasonable (but suboptimal) approach. For instance, consider the classification of credit applicants as either being *good customers* (will pay back the credit) or *bad customers* (will fail to pay off part of the credit). The cost of classifying one risky borrower as *good* could be much higher than the cost of classifying a potentially good customer as *bad*.

Our proposal relies on Bayes decision theory where the goal is to assign instances to the class with minimum expected cost. The decision is made involving both costs and posterior probabilities of the classes. Obtaining calibrated probability estimates at the classifier output requires a suitable learning machine, a large enough representative data set as well as an adequate loss function to be minimized during learning. The design of the loss function can be aided by the costs: classical decision theory shows that cost matrices define class boundaries determined by posterior class probability estimates. Strictly speaking, in order to make optimal decisions, accurate probability estimates are only required near the decision boundaries. It is key to point out that the election of the loss function becomes especially relevant when the prior knowledge about the problem is limited or the available training examples are somehow unsuitable. In those cases, different loss functions lead to dramatically different posterior probabilities estimates. We focus our study on the set of *Bregman divergences*. These divergences offer a rich family of *proper* losses that has recently become very popular in the machine learning community [Nock and Nielsen, 2009, Reid and Williamson, 2009a].

The first part of the Thesis deals with the development of a novel parametric

family of multiclass Bregman divergences which captures the information in the cost matrix, so that the loss function is adapted to each specific problem. Multiclass cost-sensitive learning is one of the main challenges in cost-sensitive learning and, through this parametric family, we provide a natural framework to successfully overcome binary tasks. Following this idea, two lines are explored:

- **Cost-sensitive supervised classification**: We derive several asymptotic results. The first analysis guarantees that the proposed Bregman divergence has maximum *sensitivity* to changes at probability vectors near the decision regions. Further analysis shows that the optimization of this Bregman divergence becomes equivalent to minimizing the overall cost regret in non-separable problems, and to maximizing a margin in separable problems.

- **Cost-sensitive semi-supervised classification**: When labeled data is scarce but unlabeled data is widely available, semi-supervised learning is an useful tool to make the most of the unlabeled data. We discuss an optimization problem relying on the minimization of our parametric family of Bregman divergences, using both labeled and unlabeled data, based on what is called the *Entropy Minimization principle*. We propose the first multiclass cost-sensitive semi-supervised algorithm, under the assumption that inter-class separation is stronger than intra-class separation.

The second part of the Thesis deals with the transformation of this parametric family of Bregman divergences into a *sequence* of Bregman divergences. Work along this line can be further divided into two additional areas:

- **Foundations of sequences of Bregman divergences**: We generalize some previous results about the design and characterization of Bregman divergences that are suitable for learning and their relationship with convexity. In addition, we aim to broaden the subset of Bregman divergences that are interesting for cost-sensitive learning. Under very general conditions, we find sequences of

(cost-sensitive) Bregman divergences, whose minimization provides minimum (cost-sensitive) risk for non-separable problems and some type of maximum margin classifiers in separable cases.

- **Learning with example-dependent costs**: A strong assumption is widespread through most cost-sensitive learning algorithms: misclassification costs are the same for all examples. In many cases this statement is not true. We claim that using the example-dependent costs directly is more natural and will lead to the production of more accurate classifiers. For these reasons, we consider the extension of cost-sensitive sequences of Bregman losses to example-dependent cost scenarios to generate finely tuned posterior probability estimates.

AGRADECIMIENTOS

Yo no sé muchas cosas, es verdad.
Digo tan sólo lo que he visto.

León Felipe

Ha sido duro. El camino ha estado lleno de altibajos, de decepciones y de alguna pequeña victoria. Seguramente esta Tesis no será la más innovadora ni la mejor escrita. Pero si por algo debería ser recordada es por el número de personas a las que debo expresar mi gratitud.

La casualidad me dio la oportunidad de trabajar con mi director, Jesús Cid. Ha sido una experiencia inmejorable. Cada día he tenido su apoyo. Es emocionante sentir como, incluso en este último año y medio lejos de Leganés, siempre puedo contar con él, incluso para trasnochar por algún deadline. Por encima de todo, creo que su trabajo es una verdadera proeza de flexibilidad e inteligencia. Su motivación, ética y empatía son contagiosas. Por todo ello y más, no podré agradecerte lo suficiente.

El Grupo de Métodos Probabilísticos para Registros Temporales (últimamente G2PI) es un lugar perfecto para desarrollar un doctorado. Un grupo de investigación no sería nada sin sus miembros. Me gustaría dar las gracias a Fran González, Sara y en especial a Rocío. Para ella no hay problema demasiado grande ni pregunta que no sepa contestar a quien lo necesite. Y no sólo he estado acompañado por los miembros de mi grupo. Durante estos años he compartido el laboratorio 4.2.A.01 con Edu y con Iván, que me han soportado desde que estaba en tercero de carrera.

Rubén me dió los mejores consejos y le hice poco caso. Sigo pensando lo mismo que cuando escribí los agradecimientos de mi proyecto fin de carrera, es de las personas más buenas que he conocido. No me puedo olvidar del grupo de la una y media con sus comidas interminables. Comer un sandwich en veinte minutos nunca será lo mismo... Y se echa mucho en falta a Oscar del Ama. Era una gran excusa para parar de vez en cuando y dar una vuelta por el 4.2.A.03 (a pesar de su demagogia e incapacidad para ver fútbol de forma objetiva). Con él y con Juanjo el Máster se hizo mucho más soportable.

En un departamento tan grande como el de Teoría de la Señal y Comunicaciones hacen falta pilares que, con su buena voluntad, trabajen para mejorar el funcionamiento del grupo. Admiro la visión aguda de Jero y la compresión de Fernando Pérez-Cruz como jefe de departamento. Siempre que lo he necesitado, he podido contar con su ayuda. La labor de las personas que participan en tareas de infraestructura y administración merece también un elogio, en especial Harold. Además, trabajar a distancia durante tantos meses hubiera sido imposible sin las personas con las que he coincidido en la docencia: José Miguel Leiva, Ángel Bravo o Vanessa. Gracias por ponerme las cosas un poco más fáciles.

Sin ninguna duda, este doctorado habría sido muy distinto sin Emilio Parrado. Emilio, el impacto de tus consejos en mi vida profesional ha sido enorme. Me enseñaste que había vida fuera de Leganés y que era un mundo que merecía la pena descubrir. Nadie debería empezar el doctorado sin hablar contigo antes.

Visits played a key role in my PhD. I am indebted to John Shawe-Taylor, who was such a friendly host during my visit to University College London. I had a fantastic time in your group. While in CSML, you gave me the chance to attend your group meetings, seminars and reading groups which were a revealing experience to me. It changed my (view on) research in a drastic way. I learned a lot from Pascal Germain,

François Laviolette, Zakria Hussain, and many other. The always helpful Rebecca Martin deserves a special mention. Tijl, I admire the professional yet personal way you head the MIR group. You devote yourself to your collaborators with complete dedication. I really appreciate you gave me the opportunity of enjoying this year in the Intelligent Systems Lab. I had a lot of fun jumping from structured prediction to complexity, from curriculum learning to Group Lasso, from chord transcription to ranking with label proportions. I can't wait to see what comes next. Thanks to Matt McVicar and Yizhao Ni for putting up with me during the most hectic year of my life.

Lo que nunca me hubiera imaginado es que me tocaría sentarme en frente del doctorando más brillante que he conocido. Y que, siendo de grupos diferentes, acabaríamos trabajando juntos. Y que, por si fuera poco, se convirtiría en uno de mis mejores amigos. Tantos vuelos, ciudades, congresos y siempre es interesante. Darío, ha sido una suerte y un placer. Las grandes ligas estarán donde tú vayas.

Gracias a mis más queridísimos amigos: Ugarte, Pellón, Juan, Ana, Alex, Elena. A Iciar, gracias por animarme siempre, sin que nunca tenga que decirte nada.

Docemente adicado ós meus pais e ó meu irmán. Quérovos.

Bristol, 2011

Contents

List of Figures

List of Tables

List of Symbols

The most important symbols used in this text are (other specific notational conventions will be made at the beginning of the relevant chapter):

- General notation:

matrices, vectors, scalars, sets	All matrices ($\mathbf{C}$, $\mathbf{W}$, ...) are bold face upper case. Vectors ($\mathbf{x}$, $\mathbf{w}$, ...) are bold face lower case. They are column vectors unless stated otherwise. Scalar variables (n, R, ...) are standard lower case or upper case. The element at the i-th row and j-th column of $\mathbf{C}$ is denoted by c_{ij} , the i-th element of a vector $\mathbf{c}$ is denoted by c_i. Sets ($\mathcal{P}$, $\mathcal{X}$, $\mathcal{S}$, ...) are denoted with calligraphic letters.
i, j, k, m, n, ...	Scalar integers, of which i and j are preferentially used as indices in vectors and matrices. Where a sample size is meant, k is used.
$\mathbf{1}$, $\mathbf{I}$	The vector containing all ones is denoted by $\mathbf{1}$. The identity matrix is denoted by $\mathbf{I}$.
$\mathbb{E}$	The expectation operator.

- Notation specific for (multiclass, binary) decision problems:

$\mathbf{x}$	Input sample.
$\mathbf{y}, y$	Class label.
$\hat{\mathbf{y}}, \hat{y}$	Class label estimate.
$\mathbf{p}, p$	True posterior probability.
$\mathbf{z}, z$	Posterior probability estimate.
$K, ()^k$	Number of samples in the training set, k-th sample of the training set.
$\mathbf{W}, \mathbf{w}$	Parameter (weight) matrix/vector.
L	Loss function or number of classes.
$\mathbf{C}$	Cost matrix.
q	Normalized cost, unless stated otherwise.
$I_{\mathcal{A}}$	Indicator function that indicates membership of an element in a subset $\mathcal{A}$, having the value 1 for all elements of $\mathcal{A}$ and the value 0 otherwise.
δ	Kronecker delta.
u	Unit step function.

- Notation specific for (multiclass, binary) Bregman divergences:

D	Divergence.
D_h, D_ϕ, D_R	Bregman divergence.
$h(\mathbf{z}), h(z)$	Concave Bregman generator (Generalized entropy).
$\phi(\mathbf{z}), \phi(z)$	Convex Bregman generator.
$\nabla_{\mathbf{z}} h(\mathbf{z}), \psi(z)$	Gradient, first derivative of the generator.
$\mathbf{H}_{\mathbf{zz}}(\mathbf{z}), g(z)$	Hessian matrix, second derivative of the generator (weighting function).
s	Sensitivity of a Bregman divergence, unless stated otherwise.
P	Potential function.
f^*	Legendre conjugate of f.

Part I

Introduction

Chapter 1

Cost-sensitive learning

The one where cost-sensitive learning is presented as a well-established and relevant area in machine learning. Many (likely, most) decision problems are cost-sensitive: if there is a (human or machine) actuator after a decision maker, the consequences of the actions taken after any decision may depend on the decision itself, on the class of the data, on the values of the observations and even on other unobserved factors that may be unpredictable before deciding. The dominating preference for cost-insensitive methods in the machine learning literature is likely a natural consequence of the fact that true costs in real applications are difficult to evaluate, and cost information is usually not available in benchmark databases. Also, it is a consequence of the fact that the cost is often a non-homogeneous measure and dealing with a mixture of -economic, social, personal, ...- costs is a non-trivial issue. Since, in general, deciding the actual class of the data is less costly than any decision error, designing low error decision systems is a reasonable (but suboptimal) approach. This chapter presents a qualitative description of what is called cost-sensitive learning.

1.1 Why cost-sensitive learning?

Let us commence this Thesis with an example that is far from machine learning. Everyday, living beings face problems that involve decision making. Thomson's gazelles live on dry, grassy plains in Sudan, Tanzania and the Serengeti areas of Kenya. The gazelle is a main food item of many savanna predators such as lions, leopards, hyenas, hunting dogs and cheetahs. When a gazelle spots a stalking predator, it will pronk to alert other gazelles to the danger and escape if it is the objective of the predator. It lives on alert because it is continuously running into a situation where a decision has to be made: have I sensed the presence of a predator or not? A clear answer to this question is crucial. It is easy to see that failing to recognize a predator and hence not fleeing is far most costly than fleeing from a non-predator.

Getting closer to computer science, in applications related with text retrieval or image retrieval, failing to display a relevant could be more/less costly than displaying a completely irrelevant one. In general, any situation of life that involves decision making and where costs can be quantified will lead to cost-sensitive learning. Appropriately, The Encyclopedia of Machine Learning [Ling and Sheng, 2008] defines cost-sensitive learning as the type of learning that takes the misclassification costs (and possibly other types of cost) into consideration. The goal of this type of learning is to minimize the total cost. The key difference between cost-sensitive learning and cost-insensitive learning is that cost-sensitive learning incorporates the different cost information to the game. Cost-insensitive learning does not take the costs into consideration even under circumstances where obviating the costs could be potentially fatal.

In the International Conference on Data Mining (ICDM) 2005, Qiang Yang and Xindong Wu Yang [Yang and Wu, 2006] started an initiative to identify 10 challenging problems in data mining research, by consulting some of the most active researchers in data mining and machine learning for their opinions on what are considered important and worthy topics for future research in both communities.

Problem number 10 was stated as *Dealing with Non-Static, Unbalanced and Cost-Sensitive Data*. They argue the following

"...how to deal with unbalanced and cost-sensitive data is a major challenge in research. Charles Elkan made the observation in an invited talk at ICML 2003 Workshop on Learning from Imbalanced Data Sets. First, in previous studies, it has been observed that UCI datasets are small and not highly unbalanced. In a typical real-world dataset, there are at least 10^5 *examples and* $10^{2.5}$ *features, without single well-defined target class. Interesting cases have a frequency of less than* 0.01*. There is much information on costs and benefits, but no overall model of profit and loss. There are different cost matrices for different examples. However, most cost matrix entries are unknown. Furthermore, the costs of different outcomes are dependent on the examples; for example, the false negative cost of direct marketing is directly proportional to the amount of a potential donation. Traditional methods for obtaining these costs relied on sampling methods. However, sampling methods can easily give biased results."*

This passage reflects the relevance of cost-sensitive learning, as one of the most active and important research areas in machine learning. It plays an important role in real-world machine learning and data mining applications. This was also supported by the Technological Roadmap of the MLnetII project (European Network of Excellence in Machine Learning, [Saitta and Lavrač, 2001]), stating that "the inclusion of costs into learning and classification is one of the most relevant topics of future machine learning research".

1.2 Types of costs

In business, retail, and accounting, a cost is the value of money that has been used up to produce something, and hence is not available for use anymore. In economics, a cost is an alternative that is given up as a result of a decision. In business, the cost

may be one of acquisition, in which case the amount of money expended to acquire an item is counted as cost. In general, the cost is a variable to be minimized. In the following, *cost* should be interpreted in its most abstract sense. Cost may be measured in many different units, such as monetary units (dollars), temporal units (seconds), or abstract units of utility. Turney [Turney, 2000] provides a comprehensive survey of a large variety of different types of costs in data mining and machine learning, including misclassification costs and data acquisition costs. In this section we revise the most representative types of cost.

1.2.1 Misclassification Costs

The misclassification cost is singled out as the most important cost because it is the only relevant cost when there is a perfect knowledge of data distributions, and it has also been mostly studied in recent years. Different types of misclassification errors usually involve different costs. They can either be deterministic or example-dependent.

It is common practice the representation of the misclassification costs in a matrix. The cost matrix is organized so that c_{ij} is the i-th row, j-th column element of the cost matrix $\mathbf{C}$, and contains the cost of classifying as class i when the true class is j [Elkan, 2001a, O'Brien et al., 2008]. Table 1.1 shows the structure of a cost matrix for two classes.

In our notation, the cost of a false positive is c_{10} while the cost of a false negative is c_{01}. Conceptually, the cost of labeling an example incorrectly should always be greater than the cost of labeling it correctly. Mathematically, it should always be the case that $c_{10} > c_{00}$ and $c_{01} > c_{11}$. These conditions are called *reasonableness conditions* [Elkan, 2001a]. Suppose that the first reasonableness condition is violated, so $c_{10} < c_{00}$ but still $c_{01} > c_{11}$. In this case the optimal policy is to label all examples positive. Note that, even though these are reasonable requirements, all the results in this Thesis would apply to situations where these conditions do not hold.

Given a cost matrix, the decisions that are optimal are unchanged if each entry in

	actual negative	actual positive
predict negative	$\mathbf{C}(0,0) = c_{00}$	$\mathbf{C}(0,1) = c_{01}$
predict positive	$\mathbf{C}(1,0) = c_{10}$	$\mathbf{C}(1,1) = c_{11}$

Table 1.1: Example of cost matrix.

the matrix is multiplied by a positive constant. This scaling corresponds to changing the unit of account for costs. Similarly, the decisions that are optimal are unchanged if a constant is added to each entry in the matrix. This shifting corresponds to changing the baseline away from which costs are measured. We will discuss further results about the cost matrix in later chapters.

1.2.2 Other common costs

The literature studies a wide variety of costs. We just describe the most relevant ones. For a detailed list we refer the reader to [Turney, 2000].

- Test costs: Each test (i.e., attribute, measurement, feature) may have an associated cost. For example, in medical diagnosis, a blood test has a cost. If the misclassification costs surpass the test costs greatly, then all tests should be performed. If the test costs are much more than the misclassification costs, then it is rationale not to do any tests.

- Cost of acquiring new examples: There is often a cost associated with acquiring cases (i.e., examples, feature vectors). Typically a machine learning researcher is given a small set of cases, and acquiring further cases is either very expensive or practically impossible.

- Cost of teacher (labeling cost): Suppose we have a practically unlimited supply of unclassified examples (i.e., cases, feature vectors), but it is expensive to determine the correct class of an example. For example, every human is a potential case for medical diagnosis, but we require a physician to determine

the correct diagnosis for each person. A learning algorithm could seek to reduce the cost of teaching by actively selecting cases for the teacher. A wise learner would classify the easy cases by itself and reserve the difficult cases for its teacher.

- Cost of computation: Computers are a limited resource, so it is meaningful to consider the cost of computation. The various types of computational complexity are essentially different forms of cost that we may wish to take into account.

- Cost of learning: This cost includes finding the right features for describing the cases, finding the right parameters for optimizing the performance of the learning algorithm, converting the data to the format required by the learning algorithm, analyzing the output of the learning algorithm, and incorporating domain knowledge into the learning algorithm or the learned model.

It is important to note that all the presented types of costs (including misclassification costs and additional costs the reader could think of) differ on nature, i.e., they might potentially be non-homogeneous measures and therefore difficult to combine.

1.3 Classical cost-sensitive problems

Based on the existing types of costs and the typical machine learning tasks, some scenarios have become more popular in cost-sensitive learning. Nevertheless, the scope of cost-sensitive learning goes beyond the restricted list presented in this section. In fact, most machine learning problems are susceptible to be modified in order to take into account cost information [Santos-Rodriguez and Garcia-Garcia, 2010].

1.3.1 Cost-sensitive classification

Consider a standard classification task. A learner can be trained from a set of training examples with class labels, and can be used to predict the class labels of new exam-

ples. The class label is usually discrete and finite. Many effective algorithms and techniques have been developed, such as Decision Trees, Neural Networks, Support Vector Machines or Gaussian Processes. However, most of the algorithms pursue to minimize the error rate: the percentage of the incorrect prediction of class labels. They ignore the difference between types of errors. In particular, they implicitly assume that all misclassification errors cost are equal. In many real-world applications, this assumption is not true. The differences can be dramatic. The first examples a cost-sensitive learning practitioner comes across when starting in cost-sensitive learning are related to classification tasks in medical diagnosis. It is easy to see that a false negative prediction (failing to detect an existing disease) can be extremely delicate, while a false positive (treating a patient that does not have a certain disease) can be less severe. Another application is the classification of credit applicants to a bank as either being a *good customer* (the person will pay back the credit) or a *bad customer* (the person will not pay back part of the credit loan). We suggest the reader [Elkan, 2001a] as one of the first major attempts to address cost-sensitive classification in a rigorous way. The Thesis mainly focuses in this task, so it will be studied in detail in subsequent chapters.

1.3.2 Cost-sensitive learning for data acquisition

In this setting, the goal of cost-sensitive learning is to minimize data acquisition costs while maximizing the accuracy of the learner/predictor [Ji and Carin, 2007]. Many fields in machine learning attempt to address this task. For example, in semi-supervised learning, class-labels are assumed to be expensive and features are implicitly assumed to have zero cost. In active learning, labels are again assumed to be expensive; however the learner may ask an oracle to reveal a label for unlabeled data for selected examples. Active feature acquisition assumes that obtaining features is expensive (but typically all features are assumed to be equally expensive), and the learner identifies instances for which complete information is most informative to classify a particular test sample. Inductive transfer learning and domain adaptation

methods assume that training data for a particular task is expensive or but other data from other domains may be cheaper (although relative costs are usually not explicitly modeled). Cascaded classifier architectures are primarily designed in order to reduce the cost of acquiring features to classify a sample (a sample may be classified the moment the available data is sufficient to provide sufficient classification confidence, without waiting for all features to be obtained).

There is an underlying common thread linking all of these different learning methods: the need to minimize the cost of data acquisition in many different application domains such as computer-aided medical diagnosis, computational linguistics, computational biology, and computer vision. Although all of these areas have felt the need for a principled solution to the problem, the partial solutions that have tried to solve the problem rarely model the cost explicitly, and very little effort has been expended on modeling application specific characteristics. Few successful ideas have been proposed in this direction. In [Ji and Carin, 2007] Ji and Carin describe this scenario. In their classification problem the features (sensing results) are not given a priori; the algorithm determines which features to acquire next, as well as when to stop sensing and make a classification decision based on previous observations (accounting for the costs of various types of errors, as well as the rewards of being correct). They define the cost-sensitive classification problem using a partially observable Markov decision process. Recently, [Settles et al., 2008] analyzes the novel problem of performing active learning on spatial data where label acquisition costs are proportional to distance traveled. It is motivated by the following example: consider the task of classification of land-cover using hyperspectral data. Then, acquiring labels may involve traveling to a particular location and performing some sort of test such as determining the type of land at that point or collecting various samples, such as soil, water, or foliage samples, that require physical access. Traveling to this point incurs some type of cost (e.g., gas or time) proportional to the distance traveled. The distance traveled also depends on the order in which one visits the points that need to be labeled, meaning that the label acquisition cost for a particular point is

dependent on other, previously visited points.

1.3.3 Cost-sensitive learning for class-imbalance

Another widely addressed problem is class-imbalance. In many applications, the categories that we particularly want to model are rare [Elkan, 2001b, Liu and Zhou, 2006]. Given a training set with a small number of members of rare categories, it is pointless to apply excessively complicated learning methods, or to use an excessively time-consuming model search method.

Take into account that, if the goal of a classifier is to maximize the accuracy (or minimize the error rate) instead of minimizing the cost, predicting everything as member of the majority class for a highly imbalanced dataset is often the solution provided by cost-insensitive methods.

Nonetheless, note that sometimes the number of examples of the minority class is too small for classifiers to learn adequately. This is the problem of *insufficient training data*, different from that of the imbalanced datasets.

1.4 Creating cost-sensitive algorithms

Broadly speaking, cost-sensitive learning can be categorized into two categories. The first one is to design classifiers that are cost-sensitive in themselves. We call them the direct method. The other category is to design a wrapper that converts any existing cost-insensitive (or cost-blind) classifiers into cost-sensitive ones. The wrapper method is also called cost-sensitive meta-learning method. These sets of algorithms are not disjoint: the boundary between then is sometimes diffuse and combinations of both approaches are possible.

1.4.1 Direct method

The main idea of building a direct cost-sensitive learning algorithm is to directly introduce and utilize misclassification costs into the learning algorithms. There are several works on direct cost-sensitive learning algorithms. A classical example is ICET [Turney, 1995], a method that incorporates misclassification costs in the fitness function of genetic algorithms. Note that, as ICET directly takes costs into model building, the algorithm could also take easily attribute costs (and perhaps other costs) directly into consideration, while cost-sensitive meta-learning algorithms generally cannot.

1.4.2 Cost-sensitive meta-leaning method

Cost-sensitive meta-learning converts existing cost-insensitive classifiers into cost-sensitive ones without modifying them. Thus, it can be regarded as a middleware component that pre-processes the training data, or post-processes the output, from the cost-insensitive learning algorithms.

Cost-sensitive meta-learning can be further classified into two main categories: thresholding and sampling.

Thresholding methods

Thresholding uses as a threshold to classify examples into positive or negative if the cost-insensitive classifiers can produce probability estimations. The classical algorithm MetaCost [Domingos, 1999] is a thresholding method. It first uses bagging on decision trees to obtain reliable probability estimations of training examples, relabels the classes of training examples, and then uses the relabeled training instances to build a cost-insensitive classifier. In general, thresholding-based meta-learning methods relies on accurate probability estimations. To achieve this, Zadrozny and Elkan propose several methods to improve the calibration of probability estimates [Zadrozny and Elkan, 2001a].

Sampling methods

Sampling first modifies the class distribution of training data, and then applies cost-insensitive classifiers on the sampled data directly. There is no need for the classifiers to produce probability estimations, as long as it can classify positive or negative examples accurately. [Zadrozny et al., 2003] shows that proportional sampling with replacement produces duplicated cases in the training, which in turn produces overfitting in model building. Therefore, [Zadrozny et al., 2003] proposes to use rejection sampling to avoid duplication. More specifically, each instance in the original training set is drawn once, and accepted into the sample with the accepting probability related to the cost.

1.5 Motivation of the Thesis

Following [Elkan, 2001a], this Thesis addresses the problem of cost-sensitive classification. Note that, from now on, we will assume homogeneous costs (measured in the same units) and also additive (the total cost of a series of decisions is the sum of the cost of each individual decision).

Our approach relies on Bayes decision theory, where the goal is to assign instances to the class with minimum expected cost. Therefore, we aim to obtain a calibrated probability estimates at the classifier output. For that reason, we need a suitable learning machine, a large enough representative dataset and an adequate loss function to be minimized during learning. The so-called Bregman divergences offer a rich family of *proper* losses [Reid and Williamson, 2009a] that have recently become very popular. Bregman divergences are named after L. M. Bregman, who introduced the concept in 1967 in the framework of a relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming [Bregman, 1967]. In line with Bregman's work, the same concept arose in the literature of probability elicitation, i.e. [Savage, 1971]. There are two ways in which Bregman divergences are important. Firstly, they generalize

squared Euclidean distance to a class of distances that all share similar properties. Secondly, they bear a strong connection to exponential families of distributions. In modern machine learning, Bregman divergences have attracted interest because they allow to re-interpret well-known algorithms such as the Support Vector Machine [Cristianini and Shawe-Taylor, 2000] or Adaboost [Freund and Schapire, 1995] as minimizers of certain *surrogate losses* and permit to develop theoretical foundations of the consistency of those methods [Zhang, 2003], as an alternative to explanations based on the VC dimension vapnik1998. Despite the interest in Bregman divergences, recent applications to cost-sensitive learning go no further than a few words in [Shirazi and Vasconcelos, 2008].

In the field of Neural Networks different authors rediscovered Bregman divergences [Miller et al., 1991, Cid-Sueiro et al., 1999] without explicitly naming them Bregman divergences. They preliminary explored possible applications of different divergences, as alternatives to the classical cross-entropy or the $L2$ norm. In particular, [Cid-Sueiro and Figueiras-Vidal, 2001] pointed the possibility of applying these divergences to cost-sensitive learning or semi-supervised learning. Regarding cost-sensitive learning, both in binary and multiclass problems, the initial results supported the idea that Bregman divergence might be useful when dealing with limited architectures. In relation to semi-supervised learning, an application of cross-entropy to remote sensing was presented.

Although the idea of applying Bregman divergences to cost-sensitive learning is not new, it is not explored enough. Previous works do not consider the possibility of combining unlabeled samples and cost-sensitive learning. Also, they do not examine alternative divergences other than the cross-entropy. The analysis is limited to label-dependent costs instead of considering example-dependent costs. Lastly, the state of the art provides no study of the asymptotic properties of cost-sensitive Bregman divergences.

The main objective of this Thesis is to investigate the possibilities of Bregman divergences in cost-sensitive learning in depth and provide with a framework to design

cost-sensitive divergences. Specifically, we aim to:

- Explore novel (cost-sensitive) divergences and analyze their asymptotic properties in multiclass problems (Chapter 2).
- Describe a general cost-sensitive semi-supervised learning method for multiclass tasks (Chapter 3).
- Analyze in detail the asymptotic properties of Bregman divergences in binary experiments (Chapter 4), establishing links between the minimization of certain sequences of Bregman divergences and some margin maximization.
- Extend the previous analysis to the case where the costs are example-dependent instead of just class-dependent (Chapter 5).

1.6 Outline

The structure of the Thesis remains as follows. In Chapter 2 we propose a general procedure to train multiclass classifiers for particular cost-sensitive decision problems, which is based on estimating posterior probabilities using Bregman divergences. Chapter 3 introduces a general procedure to train multiclass semi-supervised classifiers, establishing an optimization problem relying on the empirical risk minimization of a Bregman loss together with what is called Entropy Minimization principle. Chapter 4 broads the approach to uncover the links between Bregman divergences and convexity and also looks for general conditions to define a richer family of cost-sensitive sequences of Bregman divergences with nice properties. In Chapter 5 we extend the cost-sensitive sequences of Bregman divergences to tackle example-dependent costs (non-deterministic cost matrices). Finally, Chapter 6 subsumes the main conclusions of this dissertation and the on-going and future research lines.

Additionally, the reader may seek advice from Appendix A to consult some useful properties of Bregman divergences. Appendix B briefly discusses other families of divergences and the relationships among them.

We do not devote an entire chapter to revise the state of the art. Instead, we discuss the state of the art regarding each individual aspect of the Thesis in the introductory section of the different chapters.

1.7 Summary

This chapter tries to motivate the relevance of a dissertation on cost-sensitive learning. Cost-sensitive learning is sometimes forgotten or obviated in many machine learning applications even if the problem is inherently cost-sensitive. Several methods have been proposed since the late 90s in the machine learning and data mining communities. Looking back in time is nice to see an evolution from genetic algorithms and tree-based methods towards nowadays methods, relying on advanced learning techniques.

We just scratched the surface of the state of the art of cost-sensitive learning. For a comprehensive list of well-known cost-sensitive learning algorithms, please refer to [Qin et al., 2010, Zhou and Liu, 2010, Ling and Sheng, 2008].

Next chapters will insist on some of the concepts already described in this review. In particular, this Thesis will be focused on the broad problem of cost-sensitive learning classification, approached from Bayes decision theory. We will deeply study multiclass and binary problems while dealing with supervised and semi-supervised learning or deterministic and example-dependent cost matrices.

Part II

Cost-sensitive Bregman divergences

Chapter 2

Cost-sensitive learning based on Bregman divergences

The one where a parametric family of Bregman divergences is designed for multiclass cost-sensitive learning. This chapter analyzes the application of a particular class of Bregman divergences to design cost-sensitive classifiers for multiclass problems. We show that these divergence measures can be used to estimate posterior probabilities with maximal accuracy for the probability values that are close to the decision boundaries. Asymptotically, the proposed divergence measures provide classifiers minimizing the sum of decision costs in non-separable problems, and maximizing a margin in separable MAP problems. The chapter subsumes the joint work with Jesus Cid-Sueiro, Rocio Alaiz-Rodriguez, Alicia Guerrero-Curieses and Dario Garcia-Garcia [Santos-Rodriguez et al., 2009b, Santos-Rodriguez et al., 2009a, Santos-Rodriguez et al., 2009c].

2.1 Introduction

As we mentioned in Chapter 1, the general problem of cost-sensitive learning consists in designing decision or regression machines that take into account the costs involved

in the whole decision/estimation process: this includes the cost of data acquisition, which may depend on the attributes, the cost of labeling training samples, and the cost of each possible decision error. This chapter is focused in the latter case, though we believe that the proposed approach to the problem could be extended to some more general situations, like those where the cost may depend on the selected features.

Three main general approaches have been proposed to deal with multiclass cost-sensitive problems:

1. Data-based methods: these methods are based on modifying the training dataset. The most popular technique lies in rescaling the original class distribution of the training dataset according to the cost decision matrix by means of subsampling/oversampling, modifying decision thresholds or assigning instance weights. These modifications have shown to be effective in many binary problems and can also be applied to any cost insensitive learning algorithm [Zadrozny et al., 2003, Liu and Zhou, 2006].

2. Training-based methods: these methods change the learning process in order to build a binary cost-sensitive classifier, such as those proposed for neural networks [Kukar and Kononenko, 1998] decision trees [Bradford et al., 1998] or boosting-based ensemble machines like AdaCost [Fan et al., 1999]. Finally,

3. Decision-based methods: these methods based on the Bayes decision theory that assign instances to the class with minimum expected cost. Obtaining calibrated probability estimates at the classifier output requires a suitable learning machine, a large enough representative dataset as well as an adequate loss function to be minimized during learning. Nonetheless, real-valued scores from any classifier can also be transformed into calibrated probabilities by methods like Platt Scaling [Platt, 1999] or Isotonic Regression [Zadrozny and Elkan, 2002]. Though less popular [Zadrozny and Elkan, 2001a], this is the approach that uses a natural way to cope with multiclass cost-sensitive problems.

Cost-sensitive learning in multiclass domains becomes a challenging task due to the number of misclassification costs involved the decision making process. Abe et al. [Abe et al., 2004] propose an iterative method for these problems that can be used by any binary classification algorithm. Other works tackle this issue by decomposing the original problem into multiple two-class classification tasks [Marrocco and Tortorella, 2004, Lozano and Abe, 2008] or converting the cost matrix with $L \times L$ elements (where L is the number of classes) into a cost vector [Kukar and Kononenko, 1998, Liu and Zhou, 2006] with L components.[1]

Our proposal belongs to the third category (based on Bayes decision theory) and focus on the unequal costs that result from the different misclassification errors. Classical decision theory shows that cost matrices define class boundaries determined by posterior class probability estimates. So, accurate posterior class probabilities estimates should be achieved to optimize decisions.

In a binary problem, a empirical threshold can be found with the ROC (Receiver Operating Characteristics) curve plotted for different thresholds [Provost and Fawcett, 2001]. Recently, it has also been extended for multiclass problems [O'Brien and Gray, 2005, O'Brien et al., 2008] using a greedy optimization approach that may lead in some cases to local optima. Another alternative is to improve the overall quality of the probability estimates. Zadrozny and Elkan propose several post-processing methods to transform classifier scores into calibrated probability estimates for binary [Zadrozny and Elkan, 2001b] and multiclass problems (through a decomposition into binary classification problems) [Zadrozny and Elkan, 2002].

Strictly speaking, in order to make optimal decisions, accurate probability estimates are only required near the decision boundaries. This chapter is grounded on some previous works [Miller et al., 1993, Cid-Sueiro et al., 1999] on the analysis and description, in the context of machine learning, of proper loss functions, which are those minimized at calibrated probabilities. The idea of designing proper loss

[1] Note, however, that its effectiveness depends on the cost information which is lost with the transformation.

functions to increase the estimation accuracy for some pre-defined probability values was initially suggested in [Cid-Sueiro and Figueiras-Vidal, 2001], further explored in [Guerrero-Curieses et al., 2004] for binary classification, and extended to multiclass problems in [Guerrero-Curieses et al., 2005].

In this chapter, we reformulate some of these previous results by using Bregman divergences [Bregman, 1967]. Our first purpose is to establish some links between several results published in the machine learning literature, concerning the estimation of posterior class probabilities, with some general results on the problem of probability elicitation, which has been widely studied in the context of subjective probability: general conditions on proper loss functions can be dated back to [Savage, 1971], and it is also well known (see [Gneiting and Raftery, 2007] and the references therein) that any proper loss function is essentially characterized by a Bregman divergence.

Bregman divergences have attracted recent attention in the machine learning literature [Banerjee et al., 2005b]. The utility of these measures to define tailored loss functions for cost-sensitive classification has been explored in [Buja et al., 2005] for binary problems. The application of Bregman divergences (though under the name of *strict sense Bayesian* divergences) was also proposed in [Guerrero-Curieses et al., 2005], which is, up to our knowledge, the first published work on the multiclass case.

In this chapter, we propose a novel parametric family of Bregman divergences that may be used to train cost-sensitive classifiers in multi-class situations. The proposed divergence measures are in general non-convex functions of the model parameters, but we show some connections between the minimization of the divergence measures and some kind of large margin classifiers, which opens the door to some convex optimization algorithms.

The structure of this chapter is as follows: Section 2.2 states the learning and decision problem and shows the fundamentals of entropy and divergence measures. Section 2.3 presents a new family of entropy functions used to design a Bregman divergence that achieves maximal sensitivity near the decision boundaries defined by

unequal costs. The asymptotic behavior of this divergence measure is analyzed in Section 2.4. Its application to some different real datasets is exposed in Section 2.5. Finally, we summarize the main conclusions in Section 2.6.

2.2 Decision and learning

2.2.1 Cost-sensitive decision problems

Let $\mathcal{X}$ be an observation space and $\mathcal{U}_L$ a finite set of L classes or labels. For mathematical convenience, we assume that the i-th class in $\mathcal{U}_L$ is a binary unit vector $\mathbf{u}_i$ with components $u_{ij} = \delta_{i-j}$ (that is, a unique "1" at the i-th position).

In a general classification problem, a pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{U}_L$ is generated according to a probability model $p(\mathbf{x}, \mathbf{y})$. The goal is to predict class vector $\mathbf{y}$ when only $\mathbf{x}$ is observed.

In a general setting, a cost $c(\widehat{\mathbf{y}}, \mathbf{y}, \mathbf{x})$ can be associated with deciding in favor of class $\widehat{\mathbf{y}}$ when the true class is $\mathbf{y}$ and the observation is $\mathbf{x}$. The general decision problem consists in making decisions minimizing the mean risk $\mathbb{E}\{c(\widehat{\mathbf{y}}, \mathbf{y}, \mathbf{x})\}$.

It is well-known that such minimum is reached by taking, for every sample $\mathbf{x}$, class $\widehat{\mathbf{y}}^*$ such that

$$\widehat{\mathbf{y}}^* = \arg\min_{\widehat{\mathbf{y}}} \left\{ \sum_{j=1}^{L} \mathbb{E}\{c(\widehat{\mathbf{y}}, \mathbf{u}_j, \mathbf{x})|\mathbf{x}\} p_j \right\} \tag{2.1}$$

where $p_j = P\{\mathbf{y} = \mathbf{u}_j \mid \mathbf{x}\}$ is the posterior probability of class j given sample $\mathbf{x}$. In this chapter we assume that c is deterministic, and it does not depend on the observation, so that, defining $c_{ij} = c(\widehat{\mathbf{u}}_i, \mathbf{u}_j, \mathbf{x})$, we can write the optimal decision as $\widehat{\mathbf{y}}^* = \mathbf{u}_{i^*}$ such that

$$i^* = \arg\min_{i} \left\{ \sum_{j=1}^{L} c_{ij} p_j \right\} \tag{2.2}$$

In particular, taking $c_{ij} = (1 - \delta_{i-j})$, we get $i^* = \arg\max_i\{p_i\}$, which is the decision rule of the *Maximum A Posteriori* (MAP) classifier. The reader can compare

the general cost-sensitive setting and the MAP classifier in Figure 2.1. The triangle is the simplex containing all possible posterior probability values. The figure illustrates that the cost values shift the decision boundaries with respect to those of the MAP classifier.

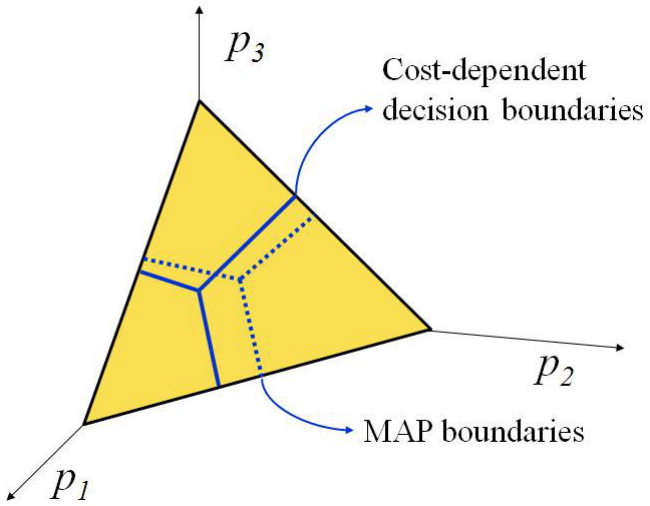


Figure 2.1: Simplex containing all possible posterior probabilities values. Cost-sensitive decision boundaries and MAP boundaries.

2.2.2 Posterior probability estimation

In a general learning problem, the probability model $p(\mathbf{x}, \mathbf{y})$ is unknown, and only a training set $\mathcal{S} = \{(\mathbf{x}^k, \mathbf{y}^k), k = 1, \ldots, K\}$ of statistically independent samples (drawn from model p) is available. The classical discriminative approach to the problem consists in estimating a posterior probability map $\mathbf{z} = \mathbf{f}_\mathbf{w}(\mathbf{x})$, where $\mathbf{f}_\mathbf{w} : \mathcal{X} \to \mathcal{P}_L$ is a function with parameters $\mathbf{w}$, transforming every element of the observation space into an element of the set of probability vectors $\mathcal{P}_L = \{\mathbf{p} : 0 \leq p_i \leq 1, \sum_{i=1}^{L} p_i = 1\}$, and replace the true probabilities p_i in (2.1) by their estimates z_i.

Estimating posterior probabilities may be inefficient. If the goal is to optimize decisions, accurate estimates of posterior probabilities far from the decision boundaries are actually not needed, and focusing learning on these estimates may be suboptimal.

Some previous definitions are required. Following [Kapur and Kesavan, 1993], we define generalized entropy and divergence measures as follows.

Definition (Entropy)

Function $h : \mathcal{P}_L \rightarrow \mathbb{R}$, is an *entropy* if $h(\mathbf{u}_i) = 0$, for every $\mathbf{u}_i \in \mathcal{U}_L$ and h is strictly concave[2] in $\mathcal{P}_L$

Note that any entropy verifies $h(\mathbf{p}) \geq 0$, for every $\mathbf{p} \in \mathcal{P}_L$.

Definition (Divergence)

Function $D : \mathcal{P}_L \times \mathcal{P}_L \rightarrow \mathbb{R}$, is a divergence among probability vectors $\mathbf{p}$ and $\mathbf{z}$ if it satisfies the following properties:

1. Non-negativity: $D(\mathbf{p}, \mathbf{z}) \geq 0$;

2. Identity: $D(\mathbf{p}, \mathbf{z}) = 0$ iff $\mathbf{p} = \mathbf{z}$;

3. Convexity: $D(\mathbf{p}, \mathbf{z})$ is a strictly convex function of $\mathbf{p}$.

Our approach in this chapter is based on the estimation of posterior class probabilities by minimizing divergence sums (Empirical Risk Minimization principle [Devroye et al., 1996]) given by

$$R(\mathbf{w}) = \sum_{k=1}^{K} D(\mathbf{y}^k, \mathbf{z}^k) \tag{2.3}$$

where $\mathbf{z}^k = f_{\mathbf{w}}(\mathbf{x}^k)$. One may wonder if parameters $\mathbf{w}^*$ minimizing $R(\mathbf{w})$ provide an estimate of posterior probabilities $\mathbf{p}$. The answer is positive for a particular class of divergence measures.

[2]Since concavity and convexity are not unanimously defined in the literature, let us make clear that, in this chapter, a function is strictly concave (convex) if its Hessian matrix is negative definite (positive).

Definition (Bregman Divergence [Bregman, 1967])

Given entropy $h : \mathcal{P}_L \to \mathbb{R}$, the Bregman divergence $D : \mathcal{P}_L \times \mathcal{P}_L \to \mathbb{R}$ relative to h is defined as

$$D_h(\mathbf{p}, \mathbf{z}) = h(\mathbf{z}) - h(\mathbf{p}) + (\mathbf{p} - \mathbf{z})^T \nabla_{\mathbf{z}} h(\mathbf{z}) \tag{2.4}$$

where $\nabla_{\mathbf{z}} h(\mathbf{z})$ represents the gradient vector of h evaluated at $\mathbf{z}$.

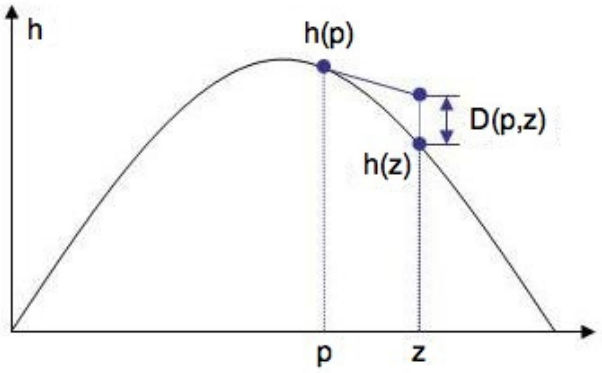


Figure 2.2: Bregman divergence.

An example of the definition is shown in Figure 2.2. The main result is the following

Theorem 2.2.1 *Let $(\mathbf{x}, \mathbf{y}) \subset \mathcal{X} \times \mathcal{U}_L$ a pair of random variables with arbitrary joint distribution $p(\mathbf{x}, \mathbf{y})$, and let $\mathbf{p}$ be the posterior probability map given by $p_i = P\{\mathbf{y} = \mathbf{u}_i|\mathbf{x}\}$. The divergence measure $D : \mathcal{P}_L \times \mathcal{P}_L \to \mathbb{R}$ satisfies*

$$\arg\min_{\mathbf{z}} \mathbb{E}\{D(\mathbf{y}, \mathbf{z})|\mathbf{x}\} = \arg\min_{\mathbf{z}} \mathbb{E}\{D(\mathbf{p}, \mathbf{z})|\mathbf{x}\} \tag{2.5}$$

for any distribution $p(\mathbf{x}, \mathbf{y})$ if and only if D is a Bregman divergence for some entropy measure h.

The theorem shows that probability estimates minimizing the mean divergence can be found by minimizing $\mathbb{E}\{D(\mathbf{y}, \mathbf{z})\}$, which, in practice, can be estimated from

samples as in Eq. (2.3). Moreover, since $\arg\min_{\mathbf{z}} \mathbb{E}\{D(\mathbf{p},\mathbf{z})|\mathbf{x}\} = \mathbf{p}$, the posterior class probability vector is the minimizer of the expected divergence.

As a particular case, if $h(\mathbf{z}) = -\sum_{i=1}^{L} z_i \log(z_i)$ (i.e., the Shannon entropy), $D_h(\mathbf{p},\mathbf{z})$ is the Kullback-Leibler divergence, and $D_h(\mathbf{y},\mathbf{z})$ is the cross-entropy. A concise summary of many of the properties of Bregman divergences is given in Appendix A.

Theorem 2.2.1 is a reformulation of Th. 1 in [Cid-Sueiro et al., 1999] by using Bregman divergences (details of the proof can be found there), though the role of these divergences in the calibration of probabilities is well known in the area of subjective probability (see, for instance, a similar result in [Gneiting and Raftery, 2007]). A recent generalization can be found in [Banerjee et al., 2005a].

Our approach in this chapter is based on the idea (also explored in [Guerrero-Curieses et al., 2005]) of optimizing Bregman divergences which are very sensitive to deviations of $\mathbf{z}$ from values of $\mathbf{p}$ close to the decision boundaries. The strategy that we follow in the next section is to design specific divergence measures for each decision problem.

2.2.3 Sensitivity of a divergence measure

In general, posterior probability vector $\mathbf{p}$ is an unknown function of observation $\mathbf{x}$. If the final goal is to minimize a mean risk function, the accuracy of the probability estimates near the decision regions should be maximized. To do so, the Bregman divergence should have maximum *sensitivity* to changes at probability vectors near the decision regions. The sensitivity can be defined as follows:

Definition The *sensitivity* of a Bregman divergence at $\mathbf{p} \in \mathcal{P}_L$ in direction $\mathbf{a}$ (with $\|\mathbf{a}\| = 1$ and $\sum_i a_i = 0$) is

$$s(\mathbf{p},\mathbf{a}) = \left.\frac{\partial^2 D_h(\mathbf{p},\mathbf{p}+\alpha\mathbf{a})}{\partial\alpha^2}\right|_{\alpha=0} = -\mathbf{a}^T\mathbf{H}_{\mathbf{zz}}(\mathbf{p})\mathbf{a} \tag{2.6}$$

where $\mathbf{H}_{\mathbf{zz}}$ is the Hessian matrix of the corresponding entropy $h(\mathbf{z})$.

(note that condition $\sum_i a_i = 0$ is necessary for $\mathbf{p}+\alpha\mathbf{a}$ in Eq. (2.6) to be a probability vector). The sensitivity measures the velocity of change of the divergence around $\mathbf{p}$. It is always non negative, since $D_h(\mathbf{p}, \mathbf{z})$ is a convex function of $\mathbf{z}$ at $\mathbf{z} = \mathbf{p}$, for any $\mathbf{p} \in \mathcal{P}_L$.

2.3 Designing Bregman divergences

2.3.1 A parametric family of entropies

If decision rule in Eq. (2.1) is based on estimates of posterior probabilities p_i, small estimation errors near the decision boundaries may change decisions and reduce the overall performance. This is the motivation to search for Bregman divergences with the highest sensitivity at probability values close to the decision boundaries and in the direction orthogonal to the boundary.

Since, according to Def. 2.2.2, a Bregman divergence can be specified from an entropy function (Eq. (2.4)), we define the family of entropies given by

$$h_R(\mathbf{z}) = -\|\mathbf{s} - \mathbf{C}\mathbf{z}\|_R + \mathbf{b}^T\mathbf{z} \tag{2.7}$$

where $\mathbf{s} = \max_{\mathbf{z}} \left\{\mathbf{u_i^T C z}\right\} = \max_{\mathbf{j}} \left\{\mathbf{c_{ij}}\right\}$, $\|\cdot\|_R$ is the R-norm (i.e., for any $\mathbf{t} \in \mathbb{R}^L$, $\|\mathbf{t}\|_R = \left(\sum_{i=1}^L t_i^R\right)^{1/R}$), $\mathbf{C}$ is the cost matrix with components c_{ij} (the cost of deciding in favor of class i when the true class is j), and R is a smooth parameter.

Parameter vector $\mathbf{b}$ should be adjusted so that $h_R(\mathbf{u}_i) = 0$, for any $\mathbf{u}_i \in \mathcal{U}_L$, though, as we will see later, it has no influence on the Bregman divergence. It is easy to see that $b_i = \|\mathbf{s} - \mathbf{C}\mathbf{u}_i\|_R$.

The concavity of h_R arises from the fact that the R-norm is strictly convex for any finite R, and convexity is preserved after any affine transformation of the variables. Moreover, if $\mathbf{C}$ is invertible, h_R is strictly concave so that it satisfies Def. 2.2.2, and the divergence D_R emanated from h_R using Eq. (2.4) is actually a Bregman divergence.

2.3.2 Bregman Divergence

According to Eq. (2.4), and defining

$$\mathbf{t}(\mathbf{z}) = \mathbf{s} - \mathbf{C}\mathbf{z} \tag{2.8}$$

the Bregman divergence corresponding to h_R is

$$\begin{aligned} D_R(\mathbf{p},\mathbf{z}) &= \|\mathbf{t}(\mathbf{p})\|_R - \|\mathbf{t}(\mathbf{z})\|_R + \|\mathbf{t}(\mathbf{z})\|_R^{1-R}(\mathbf{t}^{R-1}(\mathbf{z}))^T\mathbf{C}(\mathbf{p}-\mathbf{z}) = \\ &= \|\mathbf{t}(\mathbf{p})\|_R - \|\mathbf{t}(\mathbf{z})\|_R + \|\mathbf{t}(\mathbf{z})\|_R^{1-R}(\mathbf{t}^{R-1}(\mathbf{z}))^T(\mathbf{t}(\mathbf{z})-\mathbf{t}(\mathbf{p})) = \\ &= \|\mathbf{t}(\mathbf{p})\|_R - \|\mathbf{t}(\mathbf{z})\|_R^{1-R}(\mathbf{t}^{R-1}(\mathbf{z}))^T\mathbf{t}(\mathbf{p}) \end{aligned} \tag{2.9}$$

Before analyzing the asymptotic behavior of the sample divergence, we show that, for large R, D_R has maximal sensitivity near the decision regions defined by $\min_i \left\{\sum_{j=1}^{L} c_{ij}p_j\right\}$.

2.3.3 Sensitivity analysis

This section highlights the relevance of sensitivity when designing Bregman divergences. See Figures 2.4 and 2.3 to check how the difference between two points $\mathbf{p}$ and $\mathbf{z}$ is the same in both examples but the divergence turns out to be completely different. This example evince that the value of the divergence drastically depends on the curvature of the entropy at the point of interest, justifying the use of the second derivative as sensitivity measure. According to Eq. (2.6), the sensitivity is a function of the Hessian matrix of the divergence. Since the gradient vector of h_R has components

$$\frac{\partial h_R}{\partial z_i} = \|\mathbf{t}\|_R^{1-R}\sum_{n=1}^{L}|t_n|^{R-1}c_{ni} + b_i \tag{2.10}$$

the second-order derivatives are

$$\begin{aligned} \frac{\partial^2 h_R}{\partial z_i \partial z_j} = (R-1)\|\mathbf{t}\|_R^{1-2R}\sum_{m=1}^{L}|t_m|^{R-1}c_{mi}\sum_{n=1}^{L}|t_n|^{R-1}c_{nj} - \\ -(R-1)\|\mathbf{t}\|_R^{1-R}\sum_{n=1}^{L}|t_n|^{R-2}c_{ni}c_{nj} \end{aligned} \tag{2.11}$$

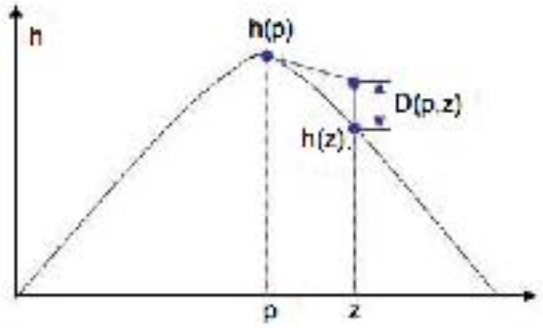


Figure 2.3: Bregman divergence, for large R, close to the high-sensitivity region.

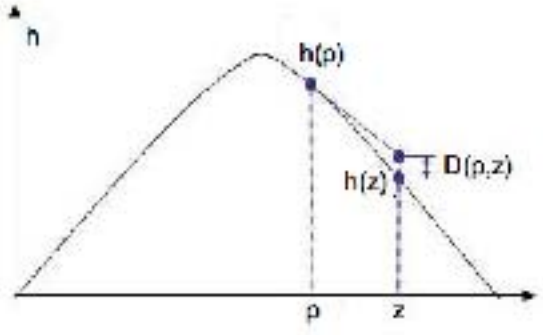


Figure 2.4: Bregman divergence, for large R, far from the high-sensitivity region.

Thus, the Hessian matrix of h_R can be expressed as

$$\mathbf{H_{zz}} = (R-1)\|\mathbf{t}\|_R^{1-2R}\left(\mathbf{C}^T\mathbf{t}^{R-1}(\mathbf{t}^{R-1})^T\mathbf{C} - \|\mathbf{t}\|_R^R\mathbf{C}^T\mathbf{D}_\mathbf{t}^{R-2}\mathbf{C}\right) \tag{2.12}$$

where $\mathbf{D_t}$ is a diagonal matrix with $\mathrm{diag}(\mathbf{D_t}) = \mathbf{t}$ and $\mathbf{t}^{R-1}$ denotes a vector whose i-th component is t_i^{R-1}, for $i = 1, \ldots, L$.

Using Eq. (2.12) the sensitivity defined in Eq. (2.6) is

$$\begin{aligned} s(\mathbf{z},\mathbf{a}) &= -(R-1)\|\mathbf{t}\|_R^{1-2R}\left(\mathbf{a}^T\mathbf{C}^T\mathbf{t}^{R-1}(\mathbf{t}^{R-1})^T\mathbf{C}\mathbf{a} - \|\mathbf{t}\|_R^R\mathbf{a}^T\mathbf{C}^T\mathbf{D}_\mathbf{t}^{R-2}\mathbf{C}\mathbf{a}\right) \\ &= -(R-1)\|\mathbf{t}\|_R^{1-2R}\left(\left((\mathbf{t}^{R-1})^T\mathbf{C}\mathbf{a}\right)^2 - \|\mathbf{t}\|_R^R\mathbf{a}^T\mathbf{C}^T\mathbf{D}_\mathbf{t}^{R-2}\mathbf{C}\mathbf{a}\right) \end{aligned} \tag{2.13}$$

For any decision problem given by cost matrix $\mathbf{C}$ and posterior probability vector $\mathbf{z}$,

any class m satisfying

$$\sum_j c_{mj} z_j = \min_n \left\{ \sum_j c_{nj} z_j \right\} \tag{2.14}$$

is optimal (because it minimizes the expected cost). Let k be the number of optimal classes for some $\mathbf{z}$. Note that, if $k = 1$, $\mathbf{z}$ is an interior point of a decision region. If $k > 1$, $\mathbf{z}$ is a point in the boundary between k decision regions. For large R, the powers of t_i for any non-optimal class i can be neglected, and we can approximate

$$\begin{aligned} s(\mathbf{z}, \mathbf{a}) &\approx -(R-1)k^{\frac{1-2R}{R}} t_m^{1-2R} \left(\left(k t_m^{R-1} \mathbf{u}^T \mathbf{C} \mathbf{a} \right)^2 - k t_m^R t_m^{R-2} \mathbf{a}^T \mathbf{C}^T \mathbf{D_u} \mathbf{C} \mathbf{a} \right) \\ &\approx -(R-1)k^{\frac{1-R}{R}} t_m^{-1} \left(k \left(\mathbf{u}^T \mathbf{C} \mathbf{a} \right)^2 - \mathbf{a}^T \mathbf{C}^T \mathbf{D_u} \mathbf{C} \mathbf{a} \right) \end{aligned} \tag{2.15}$$

where $\mathbf{u}$ is a vector with components equal to 1 at the optimal classes, and zero otherwise, and $\mathbf{D_u}$ is a diagonal matrix with $\mathbf{u}$ in the diagonal.

Analyzing the value of Eq. (2.15), it is not difficult to see that:

1. Far from the boundary: when $R \to \infty$, then $\|t\|_R \to \max_i\{t_i\}$ and

$$s(\mathbf{z}, \mathbf{a}) \to 0 \tag{2.16}$$

2. At the boundary between two or more decision regions, the sensitivity goes to infinity for any direction $\mathbf{a}$, (because of the factor $R - 1$ in Eq. (2.15)), unless some other factor is zero: it is not difficult to see that, for any vector $\mathbf{a}$ along the boundary decision, the right hand side of Eq. (2.15) is zero. Thus, at each point $\mathbf{z}$ in the boundary between several decision regions, the sensitivity to directions along the boundary tend to zero, while it tends to ∞ for any orthogonal direction.

Check Figure 2.5 for a graphical interpretation of the role of R. When the value of R is small, the corresponding entropy is smooth and its curvature varies slowly as we get closer to the interest region (the boundaries). On the contrary, large values of R result in a very high curvature around the decision boundaries, and the sensitivity decreases drastically as we move away from the boundaries.

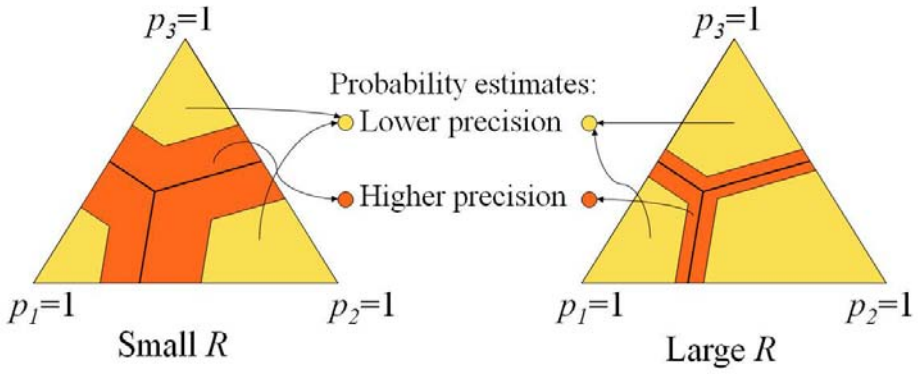


Figure 2.5: Sensitivity analysis for large values of R (right) and low values of R (left).

2.4 Asymptotic analysis

Replacing probability vector $\mathbf{p}$ by the label vector, $\mathbf{y}$, we obtain the *Bregman loss*

$$D_R(\mathbf{y},\mathbf{z}) \quad = \quad \|\mathbf{t}(\mathbf{y})\|_R - \|\mathbf{t}(\mathbf{z})\|_R^{1-R}(\mathbf{t}^{R-1}(\mathbf{z}))^T\mathbf{t}(\mathbf{y}) \tag{2.17}$$

The sum of the above expression computed over a set of training samples (as in Eq. (2.3)) is the objective function that should be minimized.

In order to analyze the behavior of D_R for large values of R, we will use an alternative expression. Let m be the index of the true class (i.e., $\mathbf{y} = \mathbf{u}_m$) and $\hat{m}$ the index of the classifier decision given $\mathbf{z}$, i.e.,

$$\hat{m} = \arg\min_i \left\{ \sum_{j=1}^{L} c_{ij} z_j \right\} = \arg\max_i t_i(\mathbf{z}) \tag{2.18}$$

Then, D_R can be written as

$$D_R(\mathbf{y},\mathbf{z}) \quad = \quad \|\mathbf{t}(\mathbf{u}_m)\|_R - \|\mathbf{t}(\mathbf{z})\|_R \frac{\sum_{i=1}^{L} t_i^{R-1}(\mathbf{z}) t_i(\mathbf{u}_m)}{\sum_{i=1}^{L} t_i^R(\mathbf{z})} \tag{2.19}$$

2.4.1 Non-separable data

For large R, Eq. (2.19) becomes

$$\lim_{R\to\infty} D_R(\mathbf{y},\mathbf{z}) \quad = \quad \max_i t_i(\mathbf{u}_m) - t_{\hat{m}}(\mathbf{z}) \frac{t_{\hat{m}}(\mathbf{u}_m)}{t_{\hat{m}(\mathbf{z})}} = c_{\hat{m}m} - \min_i c_{im} \tag{2.20}$$

(Usually, $\min_i c_{im} = c_{mm} = 0$ and the above limit is $c_{\hat{m}m}$). Thus, the divergence converges to the difference between the cost of the classifier decision and the cost of the correct decision. If the classifier makes the correct decision (i.e., the one minimizing c_{im}), the divergence is zero. Thus, in the limit, the objective function given by Eq. (2.3) converges to

$$\lim_{R\to\infty} R_R(\mathbf{w}) = \sum_{k=1}^{K} \left(c_{\hat{m}^k m^k} - \min_i c_{im^k} \right) \tag{2.21}$$

where m^k and $\hat{m}^k$ represent the index of the true class and the assigned class for sample $\mathbf{x}^k$, respectively. That is, the divergence converges to the difference in the total classification cost and the minimum achievable cost. In the MAP case, this equals the number of decision errors.

2.4.2 Separable data

If data are separable, then the limit in Eq. (2.21) is zero for any separating boundary. In this section we analyze which zero-error boundary is obtained when the loss in Eq. (2.3) is minimized.

It is interesting to analyze the behavior of this classifier for large R, when the sample is correctly classified. Though we will restrict our analysis to the MAP case, we provide a formula for the asymptotic divergence for an arbitrary cost matrix $\mathbf{C}$. Using Eq. (2.19), we can write

$$D_R(\mathbf{y}, \mathbf{z}) \;=\; \|\mathbf{t}(\mathbf{u}_m)\|_R - \left(\sum_{j=1}^{L} t_j^R(\mathbf{z}) \right)^{\frac{1}{R}} \sum_{i=1}^{L} \frac{t_i^R(\mathbf{z})}{\sum_{j=1}^{L} t_j^R(\mathbf{z})} \frac{t_i(\mathbf{u}_m)}{t_i(\mathbf{z})} \tag{2.22}$$

Consider an arbitrary sample, $\mathbf{x}$, from class m, that is out of any decision boundary. If decision $\hat{m}$ in Eq. (2.18) is correct, then $\max_i\{t_i(\mathbf{z})\} = t_m(\mathbf{z})$, and we can make first order approximations

$$\frac{t_i^R(\mathbf{z})}{\sum_{j=1}^{L} t_j^R(\mathbf{z})} = \frac{t_i^R(\mathbf{z})}{t_m^R(\mathbf{z})} \frac{1}{1 + \sum_{j\neq m}^{L} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})}} \approx \frac{t_i^R(\mathbf{z})}{t_m^R(\mathbf{z})} \left(1 - \sum_{j\neq m}^{L} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})} \right) \tag{2.23}$$

and,

$$\left(\sum_{j=1}^{L} t_j^R(\mathbf{z})\right)^{\frac{1}{R}} = t_m(\mathbf{z}) \left(1 + \sum_{j \neq m} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})}\right)^{\frac{1}{R}} \approx$$
$$\approx t_m(\mathbf{z}) \left(1 + \frac{1}{R} \sum_{j \neq m} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})}\right) \tag{2.24}$$

Using Eqs. (2.23) and (2.24) in Eq. (2.22), we get

$$\begin{aligned}
D_R(\mathbf{y}, \mathbf{z}) &\approx \|\mathbf{t}(\mathbf{u}_m)\|_R \\
&- t_m(\mathbf{z}) \left(1 + \frac{1}{R} \sum_{j \neq m} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})}\right) \sum_{i=1}^{L} \frac{t_i^R(\mathbf{z})}{t_m^R(\mathbf{z})} \left(1 - \sum_{j \neq m}^{L} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})}\right) \frac{t_i(\mathbf{u}_m)}{t_i(\mathbf{z})} \\
&= \|\mathbf{t}(\mathbf{u}_m)\|_R \\
&- \left(1 + \frac{1}{R} \sum_{j \neq m} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})}\right) \left(1 - \sum_{j \neq m} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})}\right) \sum_{i=1}^{L} \frac{t_i^{R-1}(\mathbf{z})}{t_m^{R-1}(\mathbf{z})} t_i(\mathbf{u}_m) \\
&\approx \|\mathbf{t}(\mathbf{u}_m)\|_R - \left(1 - \frac{R-1}{R} \sum_{j \neq m} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})}\right) \sum_{i=1}^{L} \frac{t_i^{R-1}(\mathbf{z})}{t_m^{R-1}(\mathbf{z})} t_i(\mathbf{u}_m) \\
&\approx \|\mathbf{t}(\mathbf{u}_m)\|_R - t_m(\mathbf{u}_m) + \frac{R-1}{R} t_m(\mathbf{u}_m) \sum_{j \neq m} \frac{t_j^R(\mathbf{z})}{t_m^R(\mathbf{z})} \\
&- \sum_{i \neq m} \frac{t_i^{R-1}(\mathbf{z})}{t_m^{R-1}(\mathbf{z})} t_i(\mathbf{u}_m)
\end{aligned} \tag{2.25}$$

A further approximation can be made if we note that, as R grows, only the terms with the highest values of t_j/t_m are relevant. Let n be the index of a "2nd-best" class, such that $n = \arg\max_{i \neq m} t_i(\mathbf{z})$, and q the number of classes satisfying this condition, and Q the set of indices of such classes. For large R, we can further approximate

$$\begin{aligned}
D_R(\mathbf{y}, \mathbf{z}) &\approx \|\mathbf{t}(\mathbf{u}_m)\|_R - t_m(\mathbf{u}_m) + \frac{R-1}{R} q t_m(\mathbf{u}_m) \frac{t_n^R(\mathbf{z})}{t_m^R(\mathbf{z})} \\
&- \frac{t_n^{R-1}(\mathbf{z})}{t_m^{R-1}(\mathbf{z})} \sum_{i \in Q} t_i(\mathbf{u}_m)
\end{aligned} \tag{2.26}$$

2.4.3 Maximum margin as a limit classifier

Starting from Eq. (2.26), we will show that, in the Maximum A Posteriori (MAP) case and using an exponential probability map, the classifier minimizing the asymptotic divergence tends to behave like a maximum margin classifier. To do so, let us assume that $\mathbf{C} = \mathbf{1}\mathbf{1}^T - \mathbf{I}$ (the MAP case), so that $\mathbf{t}(\mathbf{z}) = \mathbf{z}$, and Eq. (2.26) becomes

$$D_R(\mathbf{y}, \mathbf{z}) \approx \frac{R-1}{R} q \frac{z_n^R}{z_m^R} \tag{2.27}$$

Consider the exponential posterior probability estimate given by

$$\mathbf{z} = \mathbf{f}_{\mathbf{W}}(\mathbf{x}) = \frac{\exp(\mathbf{y}^T(\mathbf{W}\phi(\mathbf{x}) + \mathbf{b}))}{\sum_i \exp(\mathbf{u}_i^T(\mathbf{W}\phi(\mathbf{x}) + \mathbf{b}))} \tag{2.28}$$

where $\mathbf{W}$ is a parameter matrix, $\mathbf{b}$ is a parameter vector and $\phi : \mathcal{X} \rightarrow \mathbb{R}^{N'}$ is a nonlinear feature map. In such case, Eq. (2.27) reduces to

$$D_R(\mathbf{y}, \mathbf{z}) \approx q \exp(R(\mathbf{w}_n - \mathbf{w}_m)\phi(\mathbf{x}) + b_n - b_m) \tag{2.29}$$

where $\mathbf{w}_n$ is the n-th row in $\mathbf{W}$. If $P_{n,m}$ is the hyperplane defined by the equation $(\mathbf{w}_n - \mathbf{w}_m)\phi(\mathbf{x}) + b_n - b_m = 0$, and $d(\mathbf{x}, P_{n,m})$ is the euclidean distance (in the feature space) from $\phi(\mathbf{x})$ to $P_{n,m}$, we can write

$$D_R(\mathbf{y}, \mathbf{z}) \approx q \exp(R\|\mathbf{w}_n - \mathbf{w}_m\|_2 d(\mathbf{x}, P_{n,m})) \tag{2.30}$$

For the whole training set, we get

$$\begin{aligned} R_R(\mathbf{W}) &\approx \sum_{k=1}^{K} q^k \exp(-R\|\mathbf{w}_{n^k} - \mathbf{w}_{m^k}\|_2 d(\mathbf{x}^k, P_{n^k,m^k})) \\ &\approx q^\ell \exp(-R\|\mathbf{w}_{n^\ell} - \mathbf{w}_{m^\ell}\|_2 d(\mathbf{x}^\ell, P_{n^\ell,m^\ell})) \end{aligned} \tag{2.31}$$

where ℓ is the index of the sample in the training set that minimizes the negative of the exponent,

$$\ell = \arg\max_k \left\{ \|\mathbf{w}_{n^k} - \mathbf{w}_{m^k}\|_2 d(\mathbf{x}^k, P_{n^\ell,m^k}) \right\} \tag{2.32}$$

(if several samples attain this minimum, q_ℓ must be replaced by its sum over all that samples). This expression can be maximized by making $\|\mathbf{w}_{n^k} - \mathbf{w}_{m^k}\|_2$ large (which is easy to do by multiplying matrix $\mathbf{W}$ and $\mathbf{b}$ by a constant factor, which does not modify the decision boundaries). However, imposing some constraints on the size of $\mathbf{W}$, the minimum of $R_R(\mathbf{W})$ is obtained by maximizing the distances from samples to decision boundaries. Thus, for large R, the classifier optimizing $R_R(\mathbf{W})$ tends to behave as a maximum margin classifier.

The analysis of the non-MAP case is more complex. However, Eq. (2.26) shows that, for large R, the asymptotical divergence depends critically on the factor $\frac{t_n^R(\mathbf{y})}{t_m^R(\mathbf{z})}$. Using an exponential model $\mathbf{t}(\mathbf{z}) \propto \exp(\mathbf{y}^T(\mathbf{W}\phi(\mathbf{x}) + \mathbf{b}))$, it is easy to see that the divergence sum is similar to Eq. (2.32) and the boundary decision of the optimal classifier (when data are separable) does not depend on the cost matrix. Though this may seem surprising, it is in accordance with the boundary decision provided by other maximum margin classifiers, such as cost-sensitive support vector machines, which usually include the costs parameters in the slack variables, without apparent influence when dealing with separable data. In Chapter 4 we will show that this result is true for a general class of Bregman divergences.

2.5 Examples

In this section we show the results of the experiments carried out to test our approach.

2.5.1 Synthetic data

This example tries to illustrate the difference between minimizing a cost-sensitive divergence (given by our parametric family Eq. (2.7), BD) and a cost-insensitive divergence (cross-entropy, CE).

Consider the two-class problem with classes "0" and "1" and the probability map given by

$$P(1|\mathbf{x}) = \frac{1}{3}(\Phi(\mathbf{w_0}^T\mathbf{x} + 2) + \Phi(\mathbf{w_1}^T\mathbf{x}) + \Phi(\mathbf{w_2}^T\mathbf{x} - 2)) \tag{2.33}$$

where $\mathbf{x} \in \mathbb{R}$, $\mathbf{w_0} = (4, 0)$, $\mathbf{w_1} = (2, 2)$ and $\mathbf{w_2} = (0, 4)$. The setting of this example is a replica of experiment 5.C in [Cid-Sueiro and Figueiras-Vidal, 2001]. Function Φ is the Logistic function given by

$$\Phi(\xi) = \frac{1}{1 + \exp(\xi)} \tag{2.34}$$

Obviously, $P(0|\mathbf{x}) = 1 - P(1|\mathbf{x})$. An example of the contour-plot of this probabilistic map is represented in Figures 2.6. Colder colours correspond to higher values of the posterior probability of class "0". We generated 8000 training samples uniformly distributed in the square $[-1, 1] \times [-1, 1]$. The label of every sample was assigned stochastically according to the previous probability map. A single layer perceptron (SLP) with soft decisions given by

$$z_0 = \Phi(\mathbf{w}^T \mathbf{x}) \tag{2.35}$$

and $z_1 = 1 - z_0$, was used to estimate this map. Since the SLP has not capacity enough to do it exactly, different Bregman loss functions provide different approximations.

Learning consists in estimating parameters $\mathbf{w}$ by means of the stochastic gradient minimization of BD and CE. For instance, the stochastic gradient learning rule to minimize the divergences with a probabilistic model with parameters $\mathbf{w}$ is given by

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \rho \nabla_{\mathbf{w}} L(\mathbf{z}, \mathbf{y}) = \mathbf{w}(k) - \rho (\mathbf{y} - \mathbf{z})^T \mathbf{H_{zz}} \nabla_{\mathbf{w}} \mathbf{z} \tag{2.36}$$

where ρ is the step size, L is the loss function and $\mathbf{H_{zz}}$ is the Hessian matrix (given by Eq. (2.12) for BD). This shows the key role of $\mathbf{H_{zz}}$: the Hessian matrix modulates the error correcting term in the learning rule.

Figures 2.6, 2.7, 2.8 show the probability map and the decision boundaries for $R = \{2, 8, 16\}$ respectively (solid blue line, BD, and solid red line, CE), with a cost matrix $\mathbf{C_1} = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$. Figures 2.9, 2.10, 2.11 show the probability map and the decision boundaries for $R = \{2, 8, 16\}$ respectively (solid blue line, BD, and solid red line, CE), with a cost-matrix $\mathbf{C_2} = \left(\begin{smallmatrix} 0 & 1 \\ 2 & 0 \end{smallmatrix}\right)$. Two conclusions are clear: BD becomes a

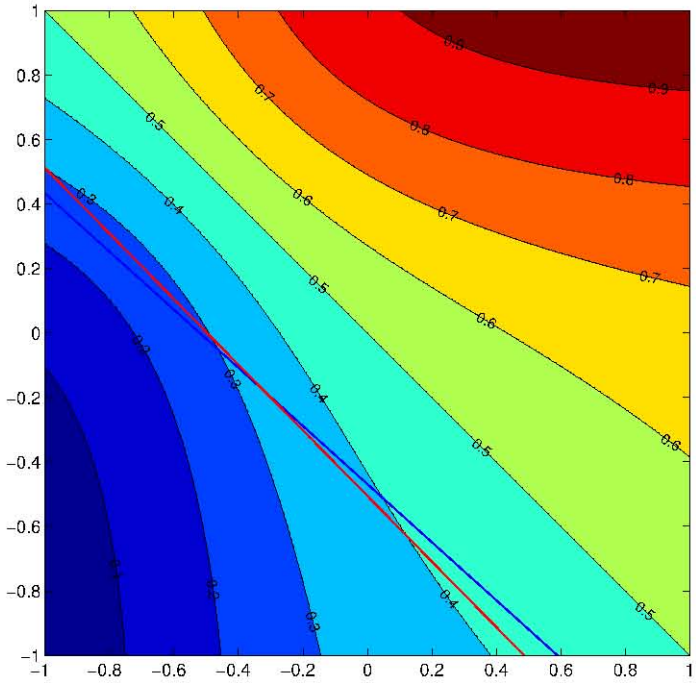


Figure 2.6: Probability map as defined in Eq. (2.33), $R = 2$, $\mathbf{C_1}$.

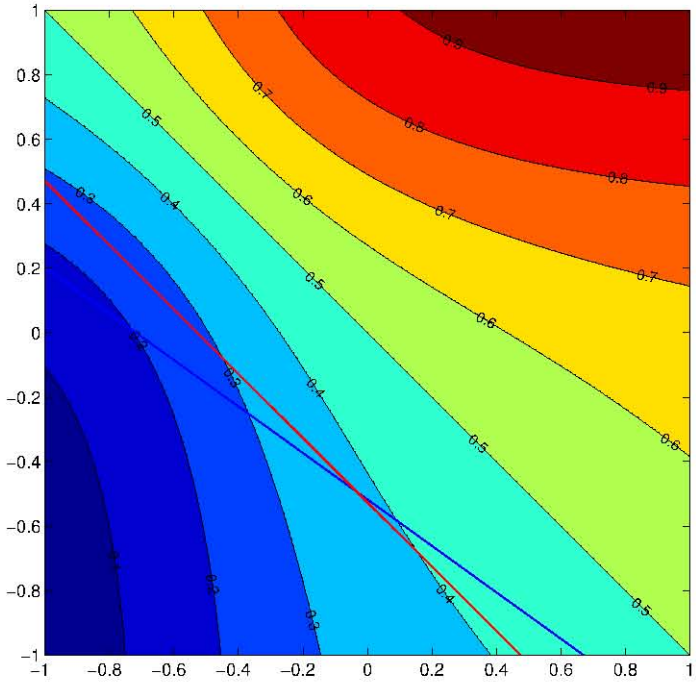


Figure 2.7: Probability map as defined in Eq. (2.33), $R = 8$, $\mathbf{C_1}$.

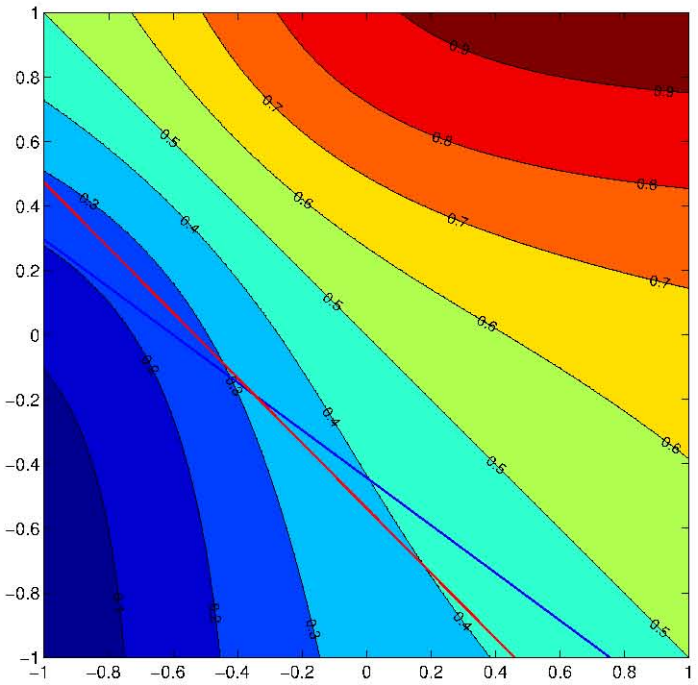


Figure 2.8: Probability map as defined in Eq. (2.33), $R = 16$, $\mathbf{C_1}$.

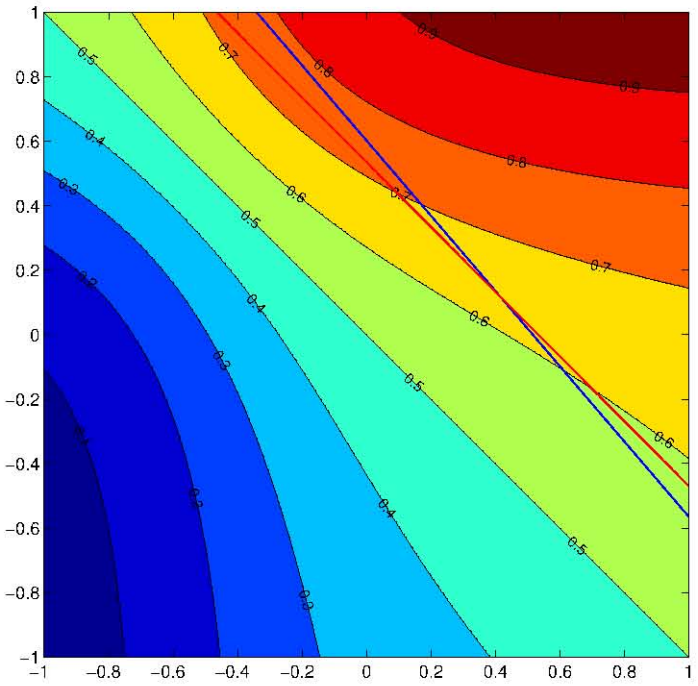


Figure 2.9: Probability map as defined in Eq. (2.33), $R = 2$, $\mathbf{C_2}$.

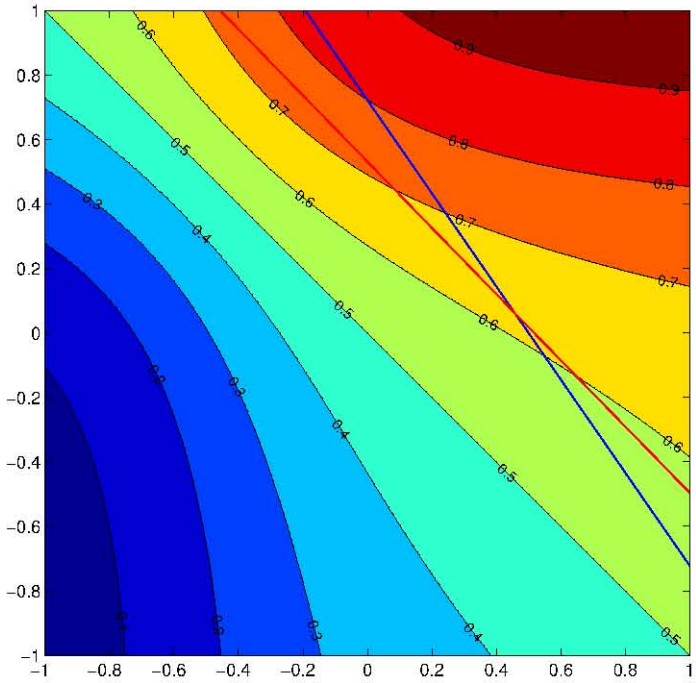


Figure 2.10: Probability map as defined in Eq. (2.33), $R = 8$, $\mathbf{C_2}$.

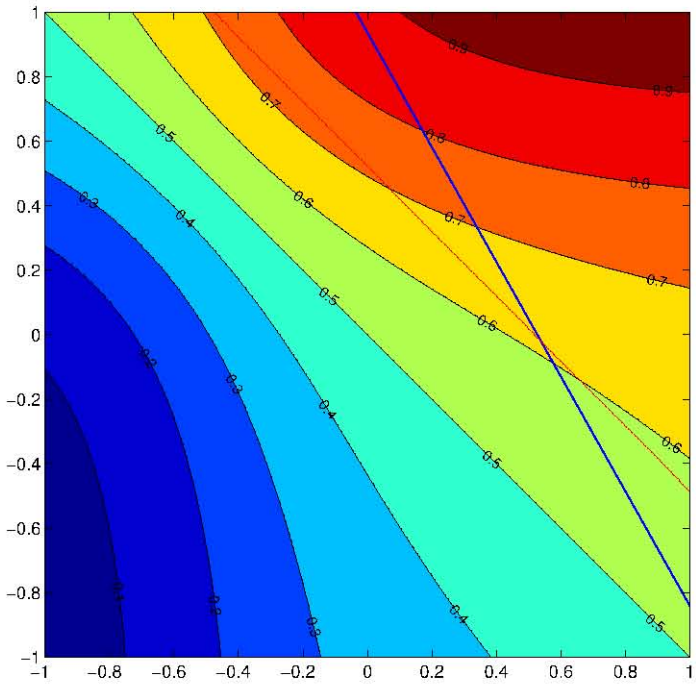


Figure 2.11: Probability map as defined in Eq. (2.33), $R = 16$, $\mathbf{C_2}$.

better approximation when the capacity of the selected architecture is limited and does not include the optimal boundary. Moreover, as R increases, the boundary obtained from BD varies its direction towards the optimal boundary.

2.5.2 UCI datasets

We conducted systematic experiments to compare the performance of the proposed method with a number of existing algorithms: an cost-insensitive architecture based on the classical CE objective function; oversampling and threshold-moving to train cost-sensitive architectures (we refer the reader to [Liu and Zhou, 2006] for the detailed description of the comparison methods we use); using datasets from the UCI repository.

We deal with two different objective functions, CE versus BD, in both cases using a probability map which computes the probability model given by

$$z_i = \sum_j z_{ij} \tag{2.37}$$

being

$$z_{ij} = \frac{\exp(\mathbf{w}_{ij}^T \mathbf{x})}{\sum_l \sum_m \exp(\mathbf{w}_{lm}^T \mathbf{x})} \tag{2.38}$$

Both, oversampling and threshold moving algorithms were coupled with the CE scheme described above. They were selected due to its simplicity and the fact that in two-class tasks were shown to be effective in cost-sensitive learning, reducing the misclassification costs. In this case, the results are obtained using a network architecture with $m = 2$ in Eq. (2.38); it is the configuration chosen to be trained with both CE and BD loss functions.

Two datasets from the UCI Machine Learning Repository are used to evaluate the algorithms: Heart Disease and German Credit data. The description of each dataset is shown in Table 2.1.

In the same way as [Liu and Zhou, 2006], three types of cost matrices are suitable with the selected UCI databases, defined as:

Table 2.1: UCI datasets description (C: continuous).

Data set	Size	Attribute	Class distribution
German	1000	24C	700/300
Heart	303	13C	164/139

1. $1.0 < c_{ij} \leq 10.0$ only for a single value $j = v$ and $c_{ij \neq v} = 1 \; \forall j \neq i$.

2. $1.0 \leq c_{ij} = V_i \leq 10.0$ for each $j \neq i$. At least one $V_i = 1$.

3. $1.0 \leq c_{ij} \leq 10.0$ for each $j \neq i$. At least one $c_{ij} = 1$.

The three conditions are the same in case we work with a binary classification task. As an example, cost matrices ($\mathbf{C}$) used in the experiments are chosen similar to the next one:

$$\mathbf{C} = \begin{pmatrix} 0 & 5 \\ 1 & 0 \end{pmatrix} \tag{2.39}$$

The experiments are carried out in the following way: first of all, we generate ten random cost matrices to estimate the average misclassification cost. Then, a 10-fold cross validation scheme is implemented: each dataset is partitioned into ten subsets with similar sizes and distributions, using nine of them as the training set and the remaining subset as the test set. This procedure is repeated ten times to use each set as test set at least once. The whole process is then performed for ten random permutations of the dataset and the average results are recorded as the final results.

Table 2.2 and Table 2.3 summarize the results of our experiments, giving the average test set error, misclassification cost and standard error for each of the datasets, and for each of four methods considered. Table 2.2 compares the average error of all comparison methods. The column corresponding to CI-CE contains the results of a cost-insensitive algorithm using CE as objective function for comparison. From

Table 2.2: Average error rate for different classification procedures (BD, CI-CE, Oversampling and Th. Moving) and different datasets.

Dataset	**Error rate** (Test)			
	BD	$CI-CE$	$Oversampling$	$Th.Moving$
German	0.232 ± 0.032	0.247 ± 0.031	0.244 ± 0.061	0.253 ± 0.041
Heart	0.184 ± 0.049	0.187 ± 0.053	0.193 ± 0.044	0.226 ± 0.060

Table 2.3: Average cost and standard error for different classification procedures (BD, CI-CE, Oversampling and Th. Moving) and different datasets.

Dataset	**Cost** (Test)			
	BD	$CI-CE$	$Oversampling$	$Th.Moving$
German	43.2 ± 1.7	57.9 ± 3.3	45.9 ± 4.1	47.7 ± 1.5
Heart	9.1 ± 0.9	12.1 ± 1.3	11.7 ± 2.1	12.3 ± 1.4

these results, it appears convincing that the designed Bregman divergences family performs better or equals all the comparison methods we have considered. Table 2.3 compares the performance, in misclassification cost, of the algorithms for both datasets, which is the main point of interest of our approach. It is confirmed that using BD in cost-sensitive learning, for high values of R, seems to be a good alternative to be further developed, which coincides with what was expected by our previous motivation.

The main conclusion of the performed experiments is that the improvement in the obtained error rate results is not statistically noteworthy but we can highlight the behavior in average cost.

Another aspect to be stressed is the difficulty of finding out the optimum value of R, which is a crucial and decisive factor to get adequate results, as well as a high sensitive parameter. This problem, together with the drawbacks of the stochastic

gradient learning rule used to minimize the loss function (in general the algorithm converges only to a local optimum), points at the necessity of exploring alternative optimization algorithms.

2.6 Summary

In this chapter we propose a general procedure to train multiclass classifiers for particular cost-sensitive decision problems, which is based on estimating posterior probabilities using Bregman divergences. We have proposed a parametric family of Bregman divergences that can be tuned to a specific cost matrix. Our asymptotic analysis shows that the optimization of the Bregman divergence for large values of parameter R becomes equivalent to minimizing the overall cost regret in non-separable problems, and to maximizing a margin in separable problems. We show that using the learning algorithm based on Bregman divergences with a simple classifier, the error/cost results obtained are lower than those given by the cross-entropy solely or combined with some well-known cost-sensitive algorithms.

Chapter 3

Cost-sensitive semi-supervised learning

The one where we change the learning paradigm from supervised to semi-supervised classification within cost-sensitive learning. Semi-supervised learning is a special form of learning. Traditional classifiers use only labeled data (feature/label pairs) to train. However, labeled instances are often difficult, expensive, or time consuming to obtain, while unlabeled data may be relatively easy to collect. We aim to make the most of this extra data inside the framework of Chapter 2.

3.1 Introduction

Frequently, even though a large database may be available, only a small fraction of the data can be labeled for supervised learning. Labeled instances are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced human annotators. Meanwhile unlabeled data may be relatively easy to collect, but few ways to exploit them are successful. Semi-supervised learning addresses this problem by using large amounts of unlabeled data, together with the labeled data, to build better classifiers. Moreover, the sampling process is not always com-

pletely random and the labeled data do not preserve the statistical features of the complete dataset. In these applications, extracting information from the unlabeled data becomes interesting. Because semi-supervised learning requires less human effort and gives higher accuracy, it is of great interest both in theory and in practice [Zhu, 2005a, Chapelle et al., 2006].

In general, the discriminative value of the unlabeled data during the learning process is highly determined by how much statistical information about the sample space is known a priori. This is the reason why unlabeled date not always helps the performance. Notice $p(\mathbf{x})$ is usually all we can get from unlabeled data. It is believed that if $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$ do not share parameters, semi-supervised learning can not help. We refer the reader to [Seeger, 2001] for further detail.

Very recently, [Li et al., 2010] proposed the first attempt of a cost-sensitive approach to binary semi-supervised learning tasks: the cost-sensitive semi-supervised vector machine (CS4VM) which considers a SVM-like objective function with unequal misclassification costs and the utilization of unlabeled data simultaneously. In the following sections we describe a more general solution to deal with multiclass cost-sensitive semi-supervised learning problems.

Our proposal is based on the *Entropy Minimization principle* (i.e. [Cid-Sueiro and Figueiras-Vidal, 2001]), which is a widely extended method in semi-supervised algorithms. For instance, the hyperparameter learning method in Section 7.2 of [Zhu, 2005b] uses entropy minimization. [Lee et al., 2006] apply the principle of entropy minimization for semi-supervised learning on 2-D conditional random fields for image pixel classification. In particular, the training objective is to maximize the standard conditional log likelihood, and at the same time minimize the conditional entropy of label predictions on unlabeled image pixels. Furthermore, we try to establish links with different well-know strategies that avoid changes in dense regions, such as Transductive SVM (TSVM) [Vapnik, 1998] and Entropy Regularization [Grandvalet and Bengio, 2004, Chapelle et al., 2006]. On the one hand, TSVM propose to broaden the margin definition to unlabeled examples, by taking

the smallest Euclidean distance between any (labeled and unlabeled) training point to the classification boundary. On the other hand, Grandvalet and Bengio use the label entropy on unlabeled data as a regularizer.

3.2 Problem formulation

The scenario can be described as follows: consider the set $\mathcal{S} = \mathcal{S}_L \cup \mathcal{S}_\mathcal{U}$ that consists of the labeled dataset $\mathcal{S}_L = \{\mathbf{x}_k, \mathbf{y}_k\}_{k=1}^{K_L}$ and the unlabeled dataset $\mathcal{S}_U = \{\mathbf{x}_k\}_{k=K_L+1}^{K}$, with K_L and $K_U = (K - K_L)$ i.i.d. samples respectively. Let us define a missing label indicator M^k (equal to 1 when the label of the k-th sample is available and 0 otherwise). Again, $\mathbf{C}$ is the cost matrix with components c_{ij} (the cost of deciding in favor of class i when the true class is j).

Let us take up again the definition of the (cost-sensitive) Bregman loss between labels and posterior probability estimates

$$L_h(\mathbf{z}, \mathbf{y}) = D_h(\mathbf{y}, \mathbf{z}) = h(\mathbf{z}) + (\mathbf{y} - \mathbf{z})^T \nabla_\mathbf{z} h(\mathbf{z}) \tag{3.1}$$

Our objective again is to minimize the expectation of the loss function. Note that can be expressed as

$$\arg\min_\mathbf{w} \mathbb{E}\{L_h(\mathbf{z}, \mathbf{y})\} = \arg\min_\mathbf{w} \mathbb{E}\{h(\mathbf{z}) + (\mathbf{y} - \mathbf{z})^T \nabla_\mathbf{z} h(\mathbf{z})\} \tag{3.2}$$

where the expectation is to be estimated making use of all the available information in $\mathcal{S}$. Different empirical risks can be defined to solve the problem in Eq. (3.2) depending on the role we assign the labeled and unlabeled data.

3.3 The Entropy Minimization Principle: derivation of the criterion

We follow a similar approach as the Entropy Minimization (HM) presented in [Cid-Sueiro and Figueiras-Vidal, 2001]. By minimizing the entropy on unlabeled

data, the method is based in the assumption that placing the decision boundaries in low density regions will improve the generalization (minimal class overlap is desired). The HM principle stems from the fact that only the second term in the left of Eq. (3.2) depends on the class labels, and can be simply stated as use unlabeled data to minimize the entropy. An empirical risk functional based on this is

$$R_1(\mathbf{w}) = \frac{1}{K}\sum_{k=1}^{K} h(\mathbf{z}_k) + \frac{1}{K_L}\sum_{k=1}^{K_L}(\mathbf{y}_k - \mathbf{z}_k)^T \nabla_{\mathbf{z}} h(\mathbf{z}_k) \tag{3.3}$$

The mean risk for this expression is

$$\mathbb{E}\{R_1\} = \mathbb{E}\{h(\mathbf{z})\} + \mathbb{E}\{(\mathbf{p}_m - \mathbf{z})^T \nabla_{\mathbf{z}} h(\mathbf{z}) | M = 1\} \tag{3.4}$$

where $\mathbf{p}_m = p(\mathbf{y}|\mathbf{x}, M = 1)$.

An alternative empirical risk based on the HM principle is given by

$$\begin{aligned} R_2(\mathbf{w}) &= \frac{1}{K}\left(\sum_{k=1}^{K} h(\mathbf{z}_k) + \sum_{k=1}^{K_L}(\mathbf{y}_k - \mathbf{z}_k)^T \nabla_{\mathbf{z}} h(\mathbf{z}_k)\right) \\ &= \frac{1}{K}\left(\sum_{k=1}^{K_L} L(\mathbf{y}_k, \mathbf{z}_k) + \sum_{k=K_L+1}^{K} L(\mathbf{z}_k, \mathbf{z}_k)\right) \end{aligned} \tag{3.5}$$

Note that the term $L(\mathbf{z}_k, \mathbf{z}_k) = h(\mathbf{z}_k)$ represents the entropy we want to minimize. It allows to interpret the minimization of R_2 as an imputation strategy: posterior probability estimate $\mathbf{z}_k$ is assigned to unlabeled data. The mean risk for R_2 is

$$\mathbb{E}\{R_2\} = \mathbb{E}\{h(\mathbf{z})\} + P(M = 1)\mathbb{E}\{(\mathbf{p}_m - \mathbf{z})^T \nabla_{\mathbf{z}} h(\mathbf{z}) | M = 1\} \tag{3.6}$$

Both approaches are particular cases of a more general parametric mean risk

$$\mathbb{E}\{R_\lambda\} = \mathbb{E}\{h(\mathbf{z})\} + \lambda\mathbb{E}\{(\mathbf{p}_m - \mathbf{z})^T \nabla_{\mathbf{z}} h(\mathbf{z}) | M = 1\} \tag{3.7}$$

where $\lambda \in \mathbb{R}^+$ regulates the trade-off between the labeled and unlabeled terms. It is easy to check that we get Eq. (3.6) when we consider $\lambda = P(M = 1) \approx \frac{K_L}{K}$. Obviously, for $\lambda = 1$, it becomes $\mathbb{E}\{R_1\}$. Therefore this expression provides a generalized HM principle.

Note that none of the expectations in Eqs. (3.4), (3.6), (3.7) coincide with the original Eq. (3.2). In the case of Eq. (3.4), if the data labeling process can be described as *Missing at Random* (that is, the missing process is independent of the values of the missing labels [Little and Rubin, 1987]), then we can disregard the condition $M = 1$ and the mean risk will be equivalent to Eq. (3.2). In general, the mean risks in Eq. (3.6) and Eq. (3.7) will be different from Eq. (3.2), but we consider them because they seem appropriate alternatives to achieve good performance in classification.

Example. Synthetic data: HM decision boundary Consider the following scenario with two classes. Two-dimensional data were generated according to

$$P\{y = 1\} = P\{y = 0\} = 1/2 \tag{3.8}$$

$$p(\mathbf{x}|y = 1) = N(\mathbf{x} - \mathbf{m_1}, \sigma^2) \tag{3.9}$$

$$p(\mathbf{x}|y = 0) = \frac{1}{2}N(\mathbf{x} - \mathbf{m_{0,1}}, \sigma^2) + \frac{1}{2}N(\mathbf{x} - \mathbf{m_{0,2}}, \sigma^2) \tag{3.10}$$

where $\mathbf{m_1} = (0, -2)$, $\mathbf{m_{0,1}} = (0, 2)$, $\mathbf{m_{0,2}} = (2, 1)$, $N(\mathbf{x}, \sigma^2)$ is the zero-mean Gaussian distribution with variance σ^2. We chose $\sigma^2 = 1$. 2000 independent samples were generated to minimize a HM empirical risk, using a cost-insensitive version of the entropy h_R (Section 2.3).

The behavior of the Entropy Minimization can be understood examining the boundary decisions in the sample space shown in Figure 3.1, where all data generated by the Gaussian with mean $\mathbf{m_{0,2}} = (2, 1)$ missed their labels. As expected, the HM boundary flows through the low density area that takes into account the unlabeled data.

This method assumes that the decision boundary should avoid regions with high $p(\mathbf{x})$. Nonetheless, imagine the data is generated from two heavily overlapping Gaussian, the decision boundary would go right through the densest region, and this method could perform badly.

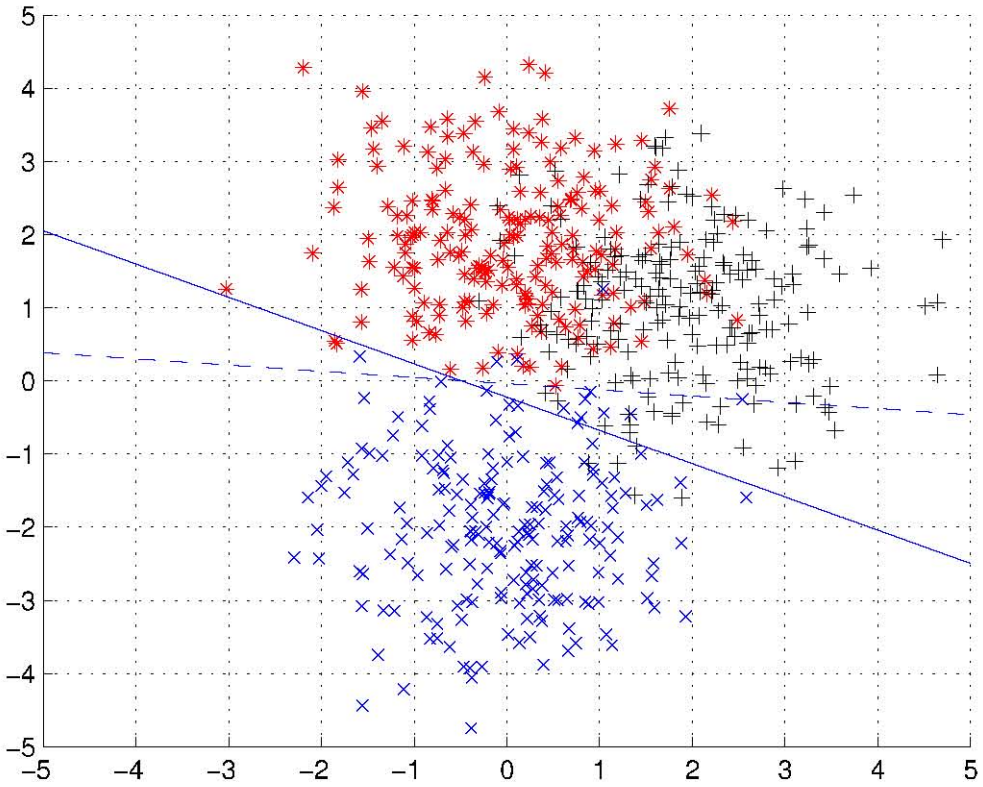


Figure 3.1: Entropy Minimization Principle. Sample distribution and boundary decisions. x: class $y = 1$ points in the training set; $*$: class $y = 0$ samples in the labeled set; $+$: unlabeled samples, belonging to class $y = 0$. The dashed line represents the boundary decision obtained using only labeled samples. The continuous line represents the boundary of the HM method.

3.4 Related Methods

3.4.1 Entropy Minimization and Entropy Regularization

In Chapter 9 of [Chapelle et al., 2006] and in [Grandvalet and Bengio, 2004], Grandvalet and Bengio propose an Entropy Regularization method through a *Maximum a Posteriori* estimation that enables to address the semi-supervised induction problem. They notice that the information content of unlabeled data decreases with class overlap, which can be measured by the conditional (Shannon) entropy of labels given patterns. Hence, the minimum entropy prior encodes a premise of semi-supervised induction, that is, the belief that unlabeled data may be useful. The strength of the prior is controlled by a tuning parameter, so that the contribution of unlabeled examples to the estimate may vanish. The proposed regularizer can be applied to local and global model of posterior probabilities. As a result, it can improve over local models when they suffer from the curse of dimensionality. Grandvalet and Bengio conclude that minimum entropy regularization may also be a serious contender to generative methods.

The (binary) MAP estimate is then defined as the maximizer of the posterior distribution, that is, the maximizer of

$$\begin{aligned} C(\mathbf{w}) &= L(\mathbf{w}) - \eta H_{emp}(\mathbf{y}|\mathbf{x}, M=1) \\ &= \sum_{k=1}^{K_L} \ln p(\mathbf{y}_k|\mathbf{x}_k) + \eta \sum_{k=K_L+1}^{K} \sum_{m=1}^{L} p(\mathbf{u}_m|\mathbf{x}_k) \ln p(\mathbf{u}_m|\mathbf{x}_k) \end{aligned} \tag{3.11}$$

which is closely linked to the generalized HM principle. In fact, for $\eta = 1$, an empirical risk based on Eq. (3.5), choosing h to be the Shannon entropy, $h(\mathbf{z}) = -\sum_{i=1}^{L} z_i \log(z_i)$, is exactly the optimization problem proposed by Grandvalet and Bengio.

3.4.2 Entropy Minimization and Transductive Support Vector Machines

Maximal margin separators are theoretically well founded models which have shown great success in supervised scenarios. Regarding the SVM, only the labeled data is exploited, and the goal is to find a maximum margin linear boundary in the Reproducing Kernel Hilbert Space. In the framework of transductive learning, the objective is to find a labeling of the unlabeled data, so that a linear boundary has the maximum margin on both the original labeled data and the (now labeled) unlabeled data. The decision boundary has the smallest generalization error bound on unlabeled data [Vapnik, 1998]. Intuitively, unlabeled data guides the linear boundary away from dense regions.

In order to compare our approach to the SVM, we can generalize the result of the asymptotical analysis (Section 2.4.3) in the supervised separable case with the Entropy Minimization criterion (assuming that the best boundary occurs in this situation). Consider the MAP case and the parametric entropy $h_R(\mathbf{z})$ given in Section 2.3. Using an exponential probability map for the posterior probability given by

$$\mathbf{z} = \mathbf{f}_{\mathbf{W}}(\mathbf{x}) = \frac{\exp(\mathbf{y}^T(\mathbf{W}\mathbf{x} + \mathbf{b}))}{\sum_i \exp(\mathbf{u}_i^T(\mathbf{W}\mathbf{x} + \mathbf{b}))}$$

leads to linear classifiers. If we apply the semi-supervised Entropy Minimization criterion form Eq. (3.5) and impose that the norm of the weights is arbitrarily large, the margin (distance from the closest samples to the decision boundaries) of that linear classifiers converges towards the maximum possible margin among all such linear classifiers as R goes to infinity.

This result can be obtained reconstructing the steps and approximations in Section 2.4.3. Asymptotically, for the whole training set, the risk converges to

$$R(\mathbf{W}) \approx \frac{1}{K}\left(\sum_{k=1}^{K_L} \frac{R-1}{R} q^k \frac{(z_n^k)^R}{(z_m^k)^R} + \sum_{k=K_L+1}^{K} z_m^k \left(1 + \frac{1}{R} q^k \frac{(z_n^k)^R}{(z_m^k)^R}\right)\right) \tag{3.12}$$

where m represents the true class and n is the index of a "2nd-best" class. Note that the left addend of the expression (labeled data) is equal to Eq. (2.27). The minimum

of the objective function is obtained by maximizing the distances from both labeled and unlabeled samples to decision boundaries.

Due to this result, the minimum entropy solution based on Eq. (3.5) can asymptotically approach the semi-supervised SVM [Vapnik, 1998, Bennett and Demiriz, 1999] because both methods converge to a solution maximizing the margin. It promotes classifiers with high confidence on the unlabeled examples. However, we recall that our objective function is non-convex, so that the convergence towards the global maximum cannot be guaranteed. Nevertheless this problem is shared by all inductive semi-supervised algorithms dealing with a large number of unlabeled data in reasonable time, such as mixture models or the Transductive SVM of Joachims [Joachims, 1999]. Explicitly or implicitly, most inductive semi-supervised algorithms impute labels which are somehow consistent with the decision rule returned by the learning algorithm. The enumeration of all possible configurations is only avoided thanks to heuristics which may fail.

3.5 Cost-sensitive learning and Entropy Minimization

One of the advantages of the proposed criterion is given by the freedom to choose the concave function h that will define the divergence to minimize. Different elections of h result in different properties of the objective function. In our case, the Entropy Minimization principle described above can be easily particularized for a cost-sensitive case by just plugging in a cost-sensitive entropy h_R (Section 2.3) in the empirical risk functionals proposed in this chapter.

The optimization is carried out using a quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) method instead of the stochastic gradient descent applied before. All presented semi-supervised learning strategies show some cases of convergence to wrong local minima. To avoid this, from now on, a simple solution is implemented: for each simulation, the training process was performed five times, and the case with the lowest value of the corresponding risk

functional on the training set was selected.

Example. UCI datasets In this example we provide an empirical evaluation of the cost-sensitive semi-supervised approach based on Bregman divergences. Using the same datasets as in Section 2.5, we follow the scenario proposed in[Li et al., 2010]. We compare our method (HM-BD) against the results reported in [Li et al., 2010] for the semi-supervised version of the cost-sensitive SVM (CS4VM), the supervised version of the cost-sensitive SVM (CSSVM, using only the labeled training examples), the supervised version of the cost-sensitive SVM (GT, CSSVM using the labeled training examples and the unlabeled examples with ground-truth labels), and a cost-sensitive version of the Transductive SVM (TSVM). GT provides an upper-bound of performance.

Each dataset is split into two equal subsets, containing the training and test sets. Each training set consists of ten labeled examples. Since there are too few labeled examples for a reliable model selection, all the algorithms use the linear kernel. We choose two versions of divergence-based methods: HM-BD1 with R_1 as objective function and HM-BD2 with R_2 as objective function. Regarding the costs, c_{10} is fixed at 1 while c_{01} is set to 2, 5, and 10 respectively. For each case, the experiment is repeated for 30 times and then the average results are reported. The objective of this setting is to be able to see how the performance of an approach changes as the cost varies.

Table 3.1: Average cost and standard error for different semi-supervised classification procedures and different datasets. Cost Ratio = 2.

Dataset	**Cost** (Test)					
	$HM-BD1$	$HM-BD2$	$CSSVM$	$TSVM$	$CS4VM$	GT
German	285.33 ± 32.74	269.95 ± 25.21	276.7 ± 27.39	268.2 ± 45.57	275.0 ± 28.82	196.6 ± 8.84
Heart	69.10 ± 12.25	68.11 ± 4.38	67.46 ± 12.58	48.69 ± 14.30	49.83 ± 12.08	34.60 ± 4.43

The results in Tables 3.1, 3.2, 3.3 show that the HM-BD2 performs better o worse than the state-of-the-art methods depending on the truthfulness of our assumption

Table 3.2: Average cost and standard error for different semi-supervised classification procedures and different datasets. Cost Ratio = 5.

Dataset	**Cost** (Test)					
	$HM-BD1$	$HM-BD2$	$CSSVM$	$TSVM$	$CS4VM$	GT
German	488.16 ± 63.93	348.83 ± 35.19	464.4 ± 76.69	462.9 ± 85.49	406.40 ± 63.30	294.4 ± 21.13
Heart	70.11 ± 12.73	73.10 ± 8.36	73.00 ± 10.77	92.24 ± 33.03	68.60 ± 7.61	54.43 ± 9.00

Table 3.3: Average cost and standard error for different semi-supervised classification procedures and different datasets. Cost Ratio = 10.

Dataset	**Cost** (Test)					
	$HM-BD1$	$HM-BD2$	$CSSVM$	$TSVM$	$CS4VM$	GT
German	797.16 ± 228.67	463.14 ± 68.11	777.2 ± 174.60	767.60 ± 157.40	600.70 ± 119.30	389.3 ± 56.62
Heart	89.33 ± 21.51	84.33 ± 9.25	79.00 ± 19.27	147.60 ± 52.63	75.53 ± 6.98	71.53 ± 8.52

in the different datasets. For instance, German Credit is known to be more separable than Heart Disease. Unsurprisingly, in Heart Disease, the boundary decision is pushed towards a low-density region so that it classifies all the training points as members of the lower-cost class. However, when the assumption holds, the advantages of our method become more prominent as the cost ratio increases. HM-BD1 struggles in the proposed scenario because of the lack of labeled examples in the training set (R_2 relies more on unlabeled data than R_1). In real-world problems the strategy would be based on minimizing a risk functional from Eq. (3.7), choosing λ by cross-validation to be adapted to the percentage of labeled examples available.

3.6 Extension: An alternative empirical risk estimation principle

The risk functionals described above are not the only options we have of using the available labels to minimize an empirical risk. For instance, imputations different

from the one in R_2 are also feasible.

One possible option is related to the Maximum Entropy Principle. The maximum entropy discrimination approach [Jaakkola et al., 1999] maximizes the margin, and is able to take into account unlabeled data, with SVM as a special case. For example, [Weston et al., 2006] learn with a *universum*, which is a set of unlabeled data that is known to come from neither of the two classes. The decision boundary is encouraged to pass through the universum. One interpretation is similar to the maximum entropy principle: the classifier should be confident on labeled examples, yet maximally ignorant on unrelated examples. An empirical risk based in this principle is:

$$R_{ME}(\mathbf{w}) = \frac{1}{K}\left(\sum_{k=1}^{K_L} L_h(\mathbf{y}_k, \mathbf{z}_k) + \sum_{k=K_L+1}^{K} L_h\left(\begin{pmatrix}0.5 & 0.5\end{pmatrix}^T, \mathbf{z}_k\right)\right) \quad (3.13)$$

This is an interesting line to explore in the future because all indications are that the method in [Li et al., 2010] could be explained as a cost-sensitive version of this kind of imputation: Li et al. propose the use of a cost-sensitive version of the standard hinge loss together with the imputation of the estimated label means of the unlabeled data.

3.7 Summary

In this chapter we propose a general procedure to train multiclass semi-supervised classifiers for particular cost-sensitive decision problems, which is based on estimating posterior probabilities using Bregman divergences. We establish an optimization problem relying on the empirical risk minimization of a Bregman loss together with what it is called Entropy Minimization principle. We link our work with two well-know semi-supervised approaches: Entropy regularization and Transductive SVM.

Under the assumption that inter-class separation is stronger than intra-class separation, the use of unlabeled data to minimize the average entropy is proposed as a multiclass cost-sensitive semi-supervised algorithm, with a performance comparable

with the state-of-the-art in binary classification tasks (when the assumption holds).

Part III

Cost-sensitive sequences of Bregman divergences

Chapter 4

Cost-sensitive sequences of Bregman losses

The one where Bregman divergences become sequences of Bregman divergences and some general properties are presented. We intend to extend the family of Bregman divergences that are suitable for cost-sensitive learning. The result is a set of conditions over what we call *sequences of weighted Bregman loss functions*. These sequences of Bregman loss functions can be constructed in such a way that their minimization guarantees, asymptotically, minimum number of errors in non-separable cases, and maximum margin classifiers in separable problems. Moreover, a wide family of Bregman sequences exists whose minimization provides, asymptotically, classifiers with minimum (cost-sensitive) risk in non-separable cases. Under very general conditions, these sequences converge to the same maximum margin classifiers in separable problems. The chapter subsumes the joint work with Jesus Cid-Sueiro and John Shawe-Taylor [Santos-Rodriguez et al., 2011a].

4.1 Introduction

Bregman divergences have been successfully translated into the cost-sensitive scenario, by defining Bregman divergences which are specially sensitive to changes near the decision boundaries (Chapter 2). In situations where the number of samples is large enough and the capacity of the learning machine is high enough, all proper losses achieve the same solution (namely, the Bayes classifier). However, this statement is not true when resources are limited. It is key to point out that the election of the loss function becomes especially relevant when the knowledge about the problem is restricted or the available train examples are somehow unsuitable (i.e., applications where the high-cost examples are scarce). In those cases, different loss functions lead to dramatically different posterior probabilities estimates and *tailoring* of the loss becomes very important [Buja et al., 2005]. In Chapter 2 a parametric family of cost-sensitive generalized entropy measures was defined in such a way that, asymptotically, the Bregman divergence associated with that family of functions, computed over a non-separable set of samples, minimizes to the number of errors. Moreover, if the dataset is separable, the classifiers minimizing the divergence converge to a maximum margin classifier.

When estimating a (posterior) probability, a parametric representation of the probability, $o : \mathcal{X} \rightarrow \mathbb{R}$, which has a natural scale not matching $[0, 1]$, can be used. This function o can be later converted to a probability estimate through a link function f^{-1}, leading to a probability estimate $f(o(\mathbf{x}))$. In the literature, f is traditionally referred to as the *inverse link*. Computationally, it is useful if the composite risk (between labels and probability estimate) is convex with respect to the parameters. Despite the nice properties, as we pointed before, a disadvantage of our parametric family is the fact that, together with an exponential probability map (inverse link), the resulting optimization problem is non-convex. That is inappropriate because convex optimization problems are much simpler and easier to be dealt with. The relationship between convexity and Bregman divergences have been studied recently

[Nock and Nielsen, 2009, Reid and Williamson, 2009a]. For instance, the optimization of other Bregman divergence, the cross-entropy, when combined with an exponential probability map, turns out to be a convex problem. So, if the underlying data distribution belongs to the model class, then minimizing the $0-1$ loss is equivalent (asymptotically, in the limit of training samples) to minimizing the cross-entropy and this divergence stands out as a sensible option. However, for finite samples or for a model not in the class, different loss functions provide different results and we saw that our parametric family with an exponential probability map, despite being non-convex, is still interesting for cost-sensitive learning. Thus, the first question we will try to answer in this chapter is the following: is it possible to find an inverse link to get a convex optimization problem using our parametric family?

In addition, we will intend to address another issue: if we regard our parametric family as a sequence (indexed by the free parameter), our second goal is to find out whether the properties of this sequence are common to other sequences or not. That is, we will try to discover if it is possible to find some general forms of sequences of Bregman losses whose minimization provides minimum (cost-sensitive) risk for non-separable problems and some type of maximum margin classifiers in separable cases. The connection between the minimization of Bregman divergence sequences and maximum margin classification is interesting, because even though all members of the sequence may be non-convex functions of the model parameters, the minimizer may converge to a maximum margin classifier, that can be efficiently obtained using convex optimization methods.

To keep the analysis simple we will describe again the scenario but for binary problems this time.

4.2 Cost-sensitive learning in binary experiments

In a similar way to Section 2.2, let $\mathcal{X}$ be an observation space with classes or labels $i \in \{0, 1\}$. In a classic classification problem, a pair $(\mathbf{x}, y) \in \mathcal{X} \times \{0, 1\}$ is generated

according to a probability model $p(\mathbf{x}, y)$. The goal is to predict the class y when only $\mathbf{x}$ is observed.

The probability model $p(\mathbf{x}, y)$ is unknown, and only a training set $\mathcal{S} = \left\{(\mathbf{x}^k, y^k), k = 1, \ldots, K\right\}$ of statistically independent samples (drawn from the given model) is available. The classical discriminative approach to the problem consists in estimating a posterior probability map $z = f_{\mathbf{w}}(\mathbf{x})$, where $f_{\mathbf{w}} : \mathcal{X} \to [0, 1]$ is a function with parameters $\mathbf{w}$, transforming every element of the observation space into an element of $[0, 1]$, and replace the true posterior probabilities (p) by their estimates (z).

Note that, although our analysis is restricted to linear decision boundaries, general non-linear boundaries can be easily considered by expanding the sample space with a non-linear transformation $t(\mathbf{x})$ mapping $\mathcal{X}$ into a higher dimensional space $t(\mathcal{X})$.

In a general setting, a cost $c_{iy}(\mathbf{x})$ can be associated with deciding in favor of class i when the true class is y and the observation is $\mathbf{x}$. The general decision problem consists in making decisions minimizing the mean risk

$$R = \mathbb{E}\{c_{iy}(\mathbf{x})\}$$

It is well-known that such minimum is reached by taking, for every sample $\mathbf{x}$, class i^* such that

$$i^* = \arg\min_i \{(1-p)\mathbb{E}\{c_{i0}(\mathbf{x})|\mathbf{x}\} + p\mathbb{E}\{c_{i1}(\mathbf{x})|\mathbf{x}\}\} \tag{4.1}$$

where $p = P\{y = 1 \mid \mathbf{x}\}$ is the posterior probability of class 1 given sample $\mathbf{x}$. Costs $c_{iy}(\mathbf{x})$ can be grouped in the cost matrix

$$\mathbf{C}(\mathbf{x}) = \begin{pmatrix} c_{00}(\mathbf{x}) & c_{01}(\mathbf{x}) \\ c_{10}(\mathbf{x}) & c_{11}(\mathbf{x}) \end{pmatrix}$$

In particular, taking $c_{ij} = (1 - \delta_{i-j})$, we get $i^* = \arg\max_i\{p_i\}$, which is the decision rule of the *Maximum A Posteriori* (MAP) classifier.

Therefore, the risk R is minimized, if $\mathbf{x}$ is assigned to class 1, if

$$p \geq \frac{c_{10}(\mathbf{x}) - c_{00}(\mathbf{x})}{c_{10}(\mathbf{x}) - c_{11}(\mathbf{x}) + c_{01}(\mathbf{x}) - c_{00}(\mathbf{x})} \tag{4.2}$$

holds, and to class 0 otherwise. In (4.2) we have assumed positive regrets $c_{10}(\mathbf{x}) - c_{11}(\mathbf{x})$ and $c_{01}(\mathbf{x}) - c_{00}(\mathbf{x})$. It follows that the classification of examples depends on the cost regrets, not on the absolute cost values. Therefore, without loss of generality, we will assume zero hit costs, $c_{00} = c_{11} = 0$ and $c_{01}, c_{10} > 0$.

Given a training set $\mathcal{S} = \left\{(\mathbf{x}^k, y^k, \mathbf{C}^k), k = 1, \ldots, K\right\}$ with $\mathbf{C}^k = \mathbf{C}(\mathbf{x}^k)$, the empirical risk is defined by

$$R_{emp} = \frac{1}{K} \sum_{k=1}^{K} c_{i^k, y^k}(\mathbf{x}^k) \tag{4.3}$$

where $c_{i^k, y^k}(\mathbf{x}^k)$ is the cost of sample k.

In this chapter we will address the cost-sensitive scenario where the cost matrix is deterministic and sample-independent, so that $\mathbf{C}(\mathbf{x}) = \mathbf{C}$ and q is the *normalized regret*

$$q = \frac{c_{10} - c_{00}}{c_{10} - c_{11} + c_{01} - c_{00}}.$$

4.3 Cost-sensitive learning with Bregman Divergences

Let us introduce the concept of Bregman divergence associated with a strictly *convex* function. Note that in previous chapters we define the concept of Bregman divergence in terms of a strictly concave function.

Definition (Bregman Divergence) Given a differentiable strictly convex function (Bregman generator) $\phi : \mathcal{A} \to \mathbb{R}$ defined in the convex set $\mathcal{A}$, and two points $p, z \in \mathcal{A}$, the Bregman divergence $D : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ relative to ϕ is defined as

$$D_\phi(p, z) = \phi(p) - \phi(z) - \langle \nabla\phi(z), p - z \rangle \tag{4.4}$$

The definition is slightly different from the one in Chapter 2: we saw how Bregman divergences can be alternatively defined in terms of a function $h = -\phi$, with the

additional constraints $h(i) = 0$. In such a case, as we mentioned earlier h can be interpreted as a generalized entropy (following [Kapur and Kesavan, 1993]).

Now we can simplify the previous definitions. It is easy to see the relationship between the definition of the Bregman divergence and the Taylor expansion [Reid and Williamson, 2009a].

Theorem 4.3.1 (Taylor's Theorem) *Let $[m, s]$ be a closed interval of $\mathbb{R}$ and let ϕ be a twice-differentiable real-valued function over $[m, s]$, then*

$$\phi(s) = \phi(m) + \phi'(m)(s - m) + \int_m^s (s - m)\phi''(m)dm \tag{4.5}$$

So, by comparing the equations Eq. (4.4) and Eq. (4.5), an integral representation of the Bregman divergence can be inferred from the Taylor theorem [Cid-Sueiro et al., 1999]. Given a differentiable real-valued strictly convex function ϕ and two points p, z, the Bregman divergence relative to ϕ is

$$D_\phi(p, z) = \int_z^p (p - \alpha)g(\alpha)d\alpha. \tag{4.6}$$

where $g = \phi''$ is a strictly positive function ($g(z) > 0$, for any $z \in (0, 1)$).

One of the main properties of Bregman divergences is that they correspond with the set of Fisher consistent or proper losses [Reid and Williamson, 2009a]. For any binary random variable $Y \in \{0, 1\}$ with $P\{Y = 1\} = p$, the surrogate loss $D_\phi(y, z)$ satisfies

$$\arg \min_{z \in [0,1]} \{\mathbb{E}_{y \sim p}\{D_\phi(y, z)\}\} = p \tag{4.7}$$

According to this, given a class $\mathcal{M}$ of functions $z : \mathcal{X} \to [0, 1]$, if the true posterior probability p is in $\mathcal{M}$, then, for any (strictly convex) ϕ,

$$\arg \min_{z \in \mathcal{M}} \{\mathbb{E}_{(\mathbf{x},y) \sim P}\{D_\phi(y, z)|\mathbf{x}\}\} = p \tag{4.8}$$

(note that p and z are scalar variables in Eq. (4.7) and functions in Eq. (4.8)). Thus, if the true posterior is in the function class, the choice of ϕ is not critical, and standard estimators, as the ML estimate (which is equivalent to $\phi(p) = p\log(p) + (1 - p\log(1 - p))$) can be efficient [Dmochowski et al., 2010].

However, if $p \notin \mathcal{M}$, Eq. (4.8) no longer holds, and different choices of ϕ provide different estimators.

The second derivative of $\phi(p)$ is an indicator of the sensitivity of D_ϕ to deviations of z from the true posterior, p (Chapter 2). The application to cost-sensitive learning becomes clear: since accurate posterior probability estimates are critical in the vicinity of the normalized regret, q, Bregman generators with higher sensitivity at q may be more efficient for classification than ML estimates.

Example Consider the binary version of the Bregman generator from Section 2.3, given by

$$\phi_n(z) = ((c_{01}z)^n + (c_{10}(1-z))^n)^{1/n} - c_{01}z - c_{10}(1-z) \tag{4.9}$$

where $n \in \mathbb{N}$ is a smoothness parameter. It naturally defines a cost-sensitive Bregman divergence $D_{\phi_n}(y, z)$. Figure 4.1 shows some plots of the generator and its associated divergence for different values of n. Note that the highest curvature region varies with $\mathbf{C}$, achieving greater sensitivity in areas close to q. As n grows larger, the sensitivity around the boundary increases, but the loss becomes less well-behaved from a numerical optimization point of view.

Other divergences show a similar sensitivity behavior.

Example The parametric family presented in [Guerrero-Curieses et al., 2005]:

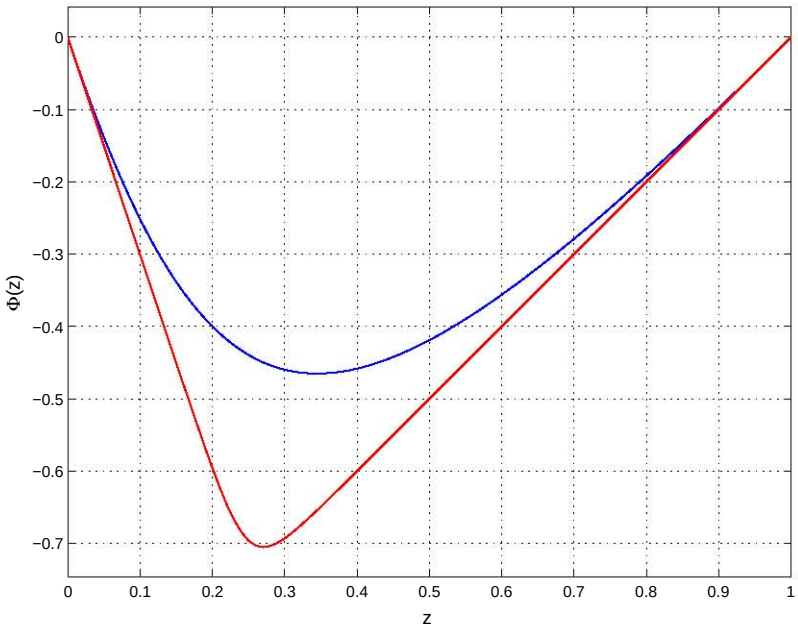
$$\phi_n(z) = -\frac{1}{n}\log\left(e^{-n(c_{01}z)} + e^{-n(c_{10}(1-z))}\right) - \frac{1}{n}\left(nc_{01}z + nc_{10}(1-z)\right) \tag{4.10}$$

where where $n \in \mathbb{N}$ is a smoothness parameter.

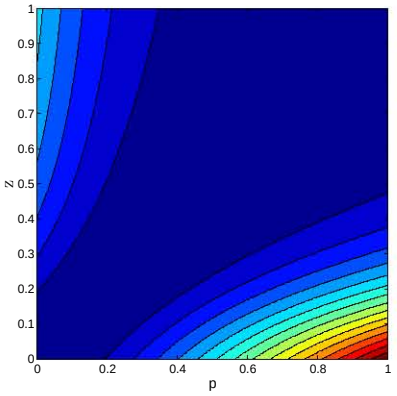
Example The Beta loss as defined in [Guerrero-Curieses et al., 2004]:

$$g_n(z) = a_n(z^{qn}(1-z)^{(1-q)n}) \tag{4.11}$$

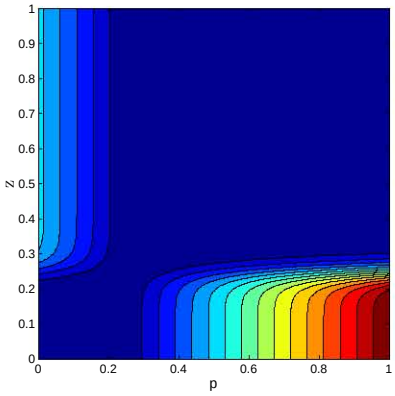
where $n \in \mathbb{R}$, $a_n = \beta((qn+1), (1-q)n+1)^{-1}$ (β is the beta function) is a normalizing constant, ensuring that g_n has unit area. The loss function associated with this g_n results in some relevant cases of interest. In particular, $L_{q=0.5,n=0}$ is the square error and $L_{q=0.5,n=-2}$ is the cross-entropy ($L(z, y) = -y\log z - (1-y)\log(1-z)$). It is easy to check that, when $n > 0$, the divergence has maximal sensitivity at $z = q$.



(a)



(b)



(c)

Figure 4.1: Cost-sensitive Bregman divergence ($c_{10} = 1, c_{01} = 3$). (a) Bregman generator (ϕ) for $n = 2$ (blue) and $n = 10$ (red), (b) Bregman divergence, $n = 2$, (c) Bregman divergence, $n = 10$.

4.4 Convexity and Activation Functions

In this section we deal with a question we left open in Chapter 2: is it possible to find an activation function in such a way that the resulting optimization problem is convex with respect to the model parameters?

4.4.1 The canonical link

In [Reid and Williamson, 2009a] the authors note that sometimes, when estimating a class probability p, a parametric representation of z, $o : \mathcal{X} \to \mathbb{R}$, which has a natural scale not matching $[0, 1]$, can be used [McCullagh and Nelder, 1989]. This function o can be later converted to a probability estimate through a link function ψ, leading to a probability estimate $z = \psi^{-1}(o(\mathbf{x}))$. Th. 4.4.1 is similar to Th. 5 in [Reid and Williamson, 2009a], and demonstrates that choosing $\psi = \phi'$, makes the resulting Bregman divergence convex with respect to o. In this case, ψ^{-1} is called *canonical link*.

Theorem 4.4.1 *Let $\psi = \phi'$. Then for $o : \mathcal{X} \to \mathbb{R}$ the Bregman divergence $D_\phi(p, \psi^{-1}(o))$ is convex in o.*

Details of the proof can be found in [Reid and Williamson, 2009a]. To establish a relationship between g and an activation function f, this condition can be reformulated in the following way for the binary case.

Corollary 4.4.2 *Let $g = -\phi''$. Then for $f = \psi^{-1}(o)$, $f_{\mathbf{w}} : im(o) \to \mathbb{R}$, linear in $\mathbf{w}$, if $g(f(o)) \cdot f'(o) = 1$ the Bregman divergence $D_\phi(p, f(o))$ is convex in $\mathbf{w}$.*

Proof Taking the first derivative of Eq. 4.6 with respect to o

$$\frac{\partial D_\phi(p, z)}{\partial o} = g(f(o))(f(o) - p)\frac{\partial f(o)}{\partial o} = g(f(o))(f(o) - p)f'(o) = (f(o) - p) \quad (4.12)$$

Taking the second derivative with respect to o

$$\frac{\partial^2 D_\phi(p,z)}{\partial o^2} = \frac{\partial f(o)}{\partial o} = f'(o) = \frac{1}{g(f(o))} \tag{4.13}$$

By definition $g(z) > 0$. So, we check that

$$\frac{\partial^2 D_\phi(p,z)}{\partial o^2} > 0. \tag{4.14}$$

and, if $o = o_{\mathbf{w}}$ is parametrized linearly in parameter vector $\mathbf{w}$, the Bregman divergence is also convex with respect to the vector $\mathbf{w}$. ∎

This condition may be useful in the case in which the generator of the Bregman divergence is given (e.g. the Bregman generator/entropy is chosen to have some desired properties). Using the corollary, an activation function $f_{\mathbf{w}}(o)$ adapted to the entropy may be chosen. This selection makes the whole minimization problem convex with respect to the parameters, which is convenient computationally. $f_{\mathbf{w}}(o)$ can be determined by solving the following differential equation:

$$f'_{\mathbf{w}}(o) = \frac{1}{g(f_{\mathbf{w}}(o))} \tag{4.15}$$

In case it is feasible, this procedure provides a tool to find functions that are suitable to estimate probabilities. Unfortunately, in some cases it is difficult to find a solution of Eq. 4.15 and most of the times the adapted activation function is not adequate for classification purposes. The following examples illustrate the relationship between the Bregman generator ϕ, its second derivative g and the activation function f.

Example When $-\phi$ is set to be the Shannon entropy:

$$\phi_1(z) = z \log z + (1-z) \log (1-z)$$

Taking two derivatives

$$g_1(z) = \frac{1}{z(1-z)}$$

It is well-know that the corresponding loss is the cross-entropy (particular case of the Beta loss in Eq. 4.11). Using Eq. 4.15 is easy to check that the *Logistic* function

$$f_1(o) = \frac{1}{1+e^{-o}}$$

is the canonical link for the cross-entropy:

$$f'(o) = f(o)(1 - f(o)) = \frac{1}{g(f_o)}$$

Example Let us obtain the canonical link associated with a given generator function

$$\phi_2(z) = -(\sqrt{z(1-z)})$$

Again, taking the second derivative

$$g_2(z) = \frac{1}{4z^{3/2}(1-z)^{3/2}}$$

This g_n is also a particular case of Eq. 4.11, with $q = 0.5$ and $n = 3$. The canonical link follows this expression

$$f_2(o) = -\frac{1}{2}\left(1 - \text{sign}(o)\frac{\sqrt{(o^2(o^2+1))}}{(o^2+1)}\right)$$

Example Analogously, consider the generator function given by

$$\phi_3(z) = -\frac{\exp(-\text{erf}^{-1}(2z-1)^2)}{2\sqrt{\pi}}$$

where erf(x) is the error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

In this case, the second derivative of the generator is

$$g_3(z) = \sqrt{\pi}\exp(\text{erf}^{-1}(2z-1)^2)$$

and the adapted activation function is given by

$$f_3(o) = \frac{\text{erf}(o)+1}{2}$$

Example An interesting case is the function that generates the squared loss

$$\phi_4(z) = -z(1-z)$$

The second derivative of this generator is

$$g_4(z) = 2$$

In order to satisfy Eq. 4.15, the activation function should be

$$f_4(o) = \frac{1+o}{2}$$

Example Lastly, consider a symmetric case of the example we used to motivate the beginning of this chapter, Eq. 4.9, with $n = 2$:

$$\phi_5(z) = (z^2 + (1-z)^2)^{1/2} - 1$$

In this case, the second derivative is

$$g_5(z) = \frac{1}{(2z^2 - 2z + 1)^{3/2}}$$

The canonical link for this function would be

$$f_5(o) = -\frac{1}{2} - \text{sign}(o)\frac{\sqrt{-o^2(o^2-2)}}{2(o^2-2)}$$

Figures 4.2, 4.3, 4.4 represent the different ϕ, g and f functions. Note that the first three examples are cases where the activation function is adapted to classification tasks while in the last example the domain of activation function is not entirely satisfactory. The fourth example represents the limit case of the desirable activations functions.

4.4.2 Convexity and potential functions

In this section we briefly discuss the relationship between activation functions and convexity from the point of view of *potential functions*, as described in [Mora-Jimenez and Cid-Sueiro, 2005]. In this case, we consider the activation function to be the derivative of a potential function, so that

$$f = \nabla P$$

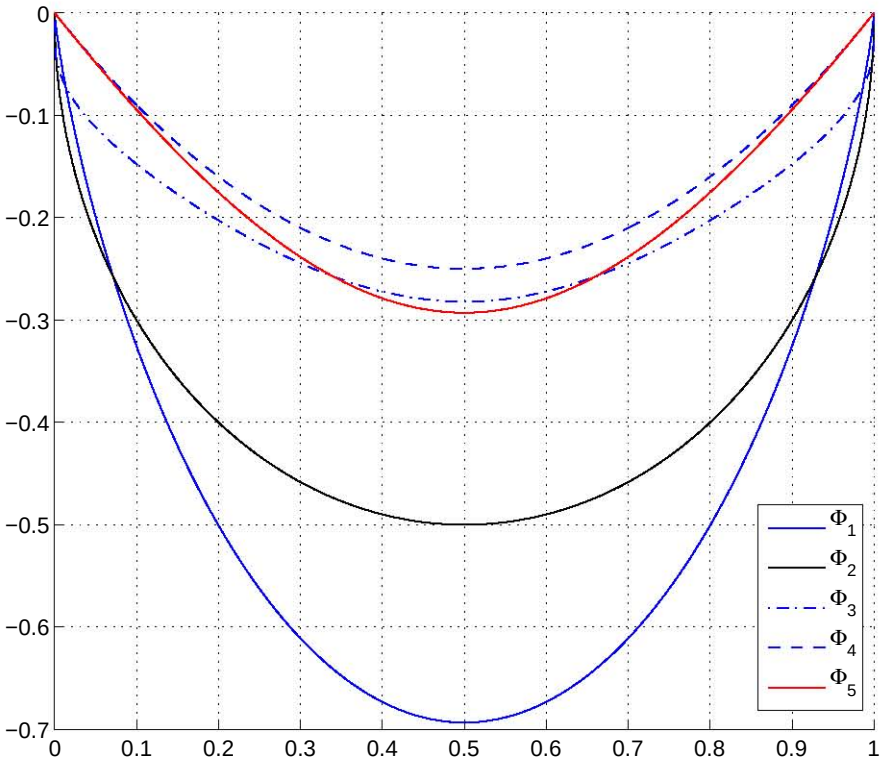


Figure 4.2: Examples of Bregman generator functions

This would allow us to apply all the results from potential theory.

Let us denote by θ^* the Legendre conjugate of θ defined by

$$\theta^*(u) := \sup_z [v^T u - \theta(u)] \tag{4.16}$$

The next theorem reformulates the same ideas of Th. 4.4.1 in terms of potential functions.

Theorem 4.4.3 *$D_\phi(p, z)$ is convex with respect to $z = f(o)$ if*

$$P = \phi^* \tag{4.17}$$

where ϕ^* is the Legendre conjugate of the Bregman generator ϕ.

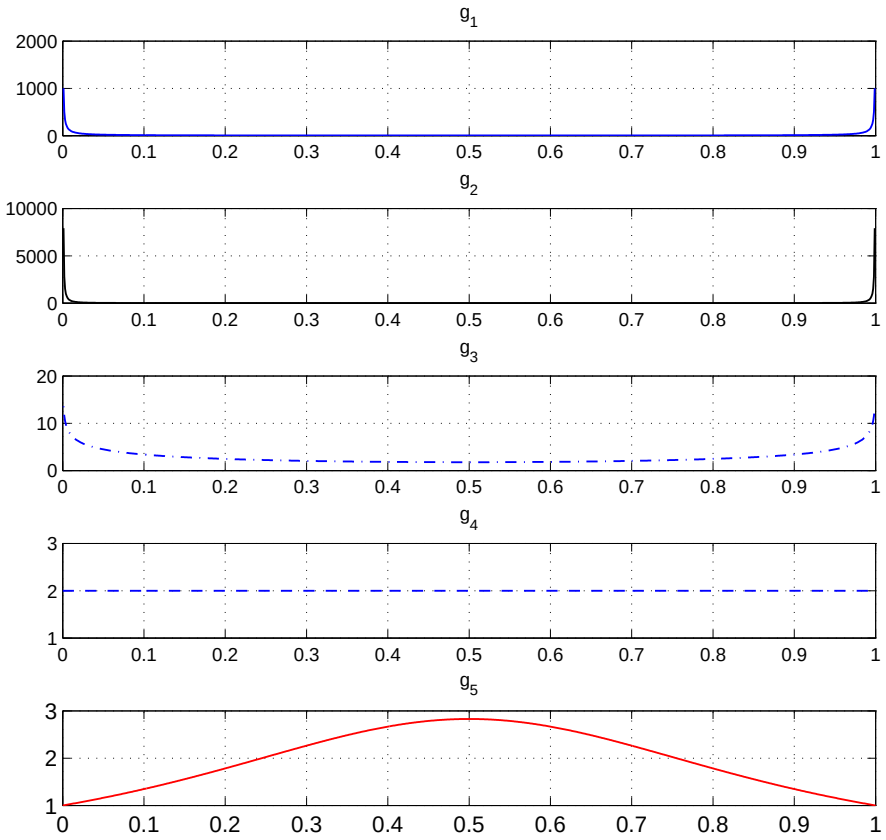


Figure 4.3: Examples of the second derivative of Bregman generator functions

Proof From Th. 4.4.1, if $f(o) = \psi^{-1}(o)$, then $D_\phi(p, z)$ is convex with respect to $z = f(o)$. As we know, $\psi = \nabla\phi$. Moreover, taking into account that the Legendre transform satisfies that $\nabla\phi = \nabla^{-1}\phi^*$ (also $\nabla\phi^* = \nabla^{-1}\phi$), it is straightforward to state that if $P = \phi^*$, the divergence is convex in $f(o)$. ∎

We can sum up this result in the following relationship

$$\nabla P = \nabla^{-1}\phi \tag{4.18}$$

[Nock and Nielsen, 2009] makes use of an equivalent result to define a family of functions in such a way that minimizing affine transformations of the Legendre conjugate

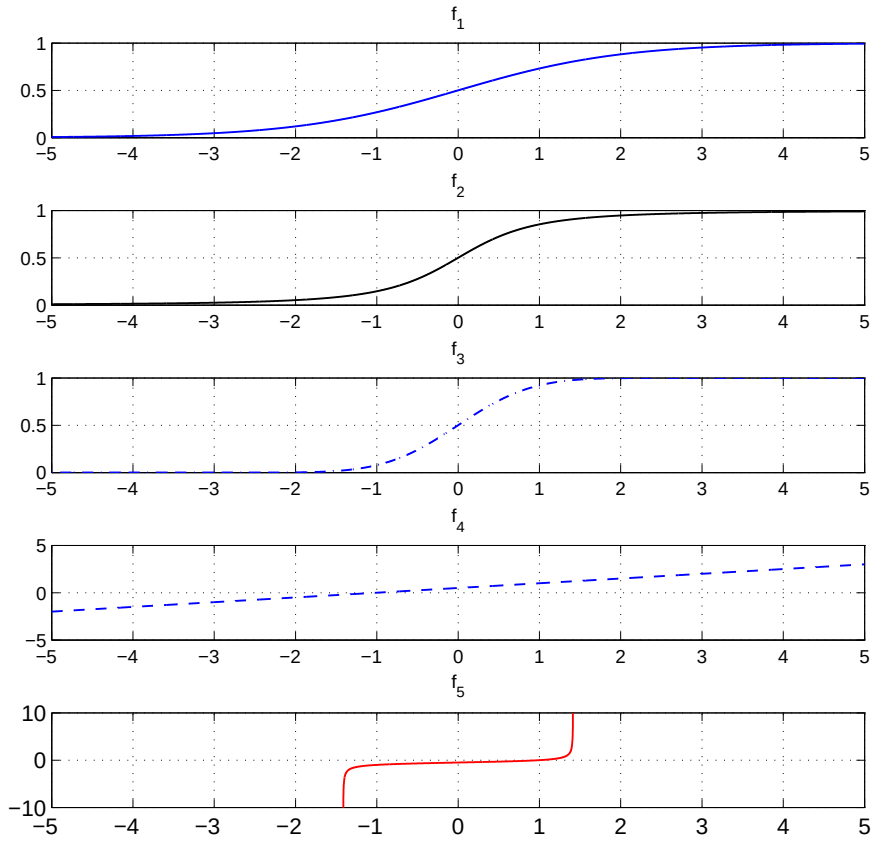


Figure 4.4: Examples of canonical links

of a reflected version of Bregman generators turns out to be equivalent to minimizing convex surrogates of the empirical risk. This family spans a subclass of the convex surrogates called *permissible convex surrogates*. This subset does not include multiclass nor asymmetric losses. Using the multiclass setting described in Chapter 2 is immediate to extend the set of permissible convex surrogates to multiclass scenarios using potential functions and therefore generalize the concept of canonical link.

4.4.3 An example

This example illustrates the use of the canonical link in a parametric family as those described in Section 4.3.

For different values of the free parameter n, the canonical link determines different activation functions. As we know, these activation functions should be defined as $f(o) : \mathbb{R} \to [0, 1]$. However as we have seen in previous examples, it is easy to find cases where the activation function coming from a canonical link is just defined in a finite interval. In practice, this forces us to the undesirable situation of having to normalize the data and encompass the possible value range of the weights of the classifier. Otherwise, the activation functions are similar to a sigmoid function, which are easier to deal with.

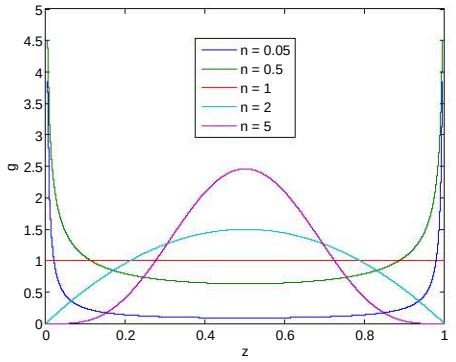
We present the results for the Beta family in Eq. (4.11). Figures 4.5(a), 4.5(b), 4.5(c) display the functions g, ϕ' and the canonical link f for different values of n, $n = \{0.05, 0.5, 1, 2, 5\}$. Check that for values of $n < 1$ the activation functions are defined in a infinite interval while for values of $n > 1$ the activation functions are defined just in a finite interval. Alternatively, studying the values of n that make g convex also separates both cases.

We generate 50 1-dimensional data points in such a way that we draw samples from classes "0" and "1", where class "0" is characterized as a Gaussian distribution with 0 mean and standard deviation equal to 3, while class "1" is characterized as mixture of two Gaussian distributions with means 1 and 2 and standard deviation equal to 1. The data is then normalized data to lie in a predefined interval, $[-0.2, 0.2]$. In this setting, we use an extremely simple classifier of the form $f(x + w)$, with only one free parameter, but useful to visualize 1-dimensional plots. The optimization is carried out via the well-known *Regula falsi* method.

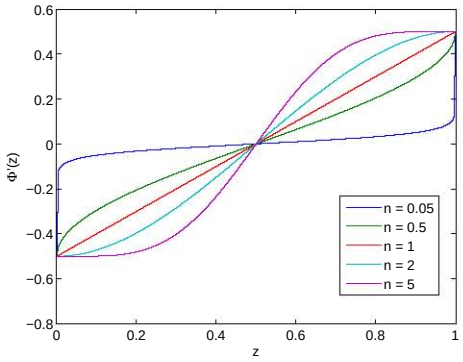
First of all, as expected, we can see in Figure 4.5(d) that, in all cases, the loss function is indeed convex with respect to the weights of the classifier. Figure 4.5(e) displays the points belonging to both classes together with the cost. For different values of n, it also illustrates the input/output relationship of the corresponding

classifier. The second conclusion of the experiment, although unclear, let us say that, large values of n move the boundary to a position with less number of errors. The differences are very small but this is the motivation to start studying the behavior of this kind of *sequences* for large values of n, even though the canonical links associated with them are less appropriate.

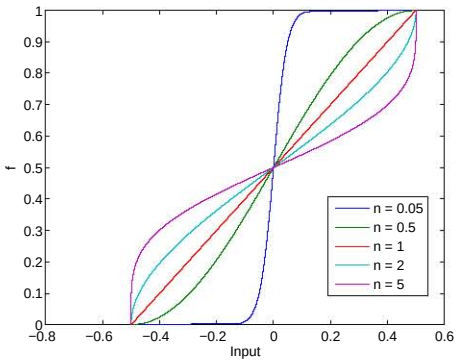
We conjecture that minimizing divergences with large values of n will lead us to minimize the total cost. Of course, there is no free lunch. If we decide to go for high values of n, we must give up using canonical links due to their increasing inconsistency in those cases (classification-wise). Remember that given up canonical links means sacrificing convexity. Note that minimizing the total cost is a non-convex problem and if we aim to get closer to it maybe is reasonable not to impose convexity.



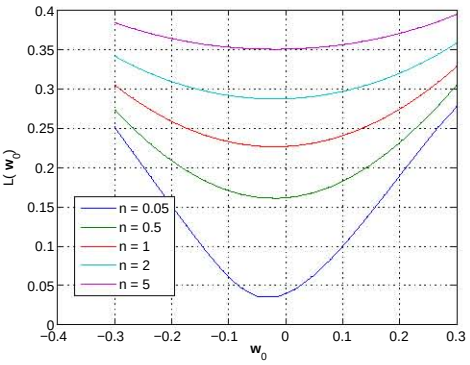
(a)



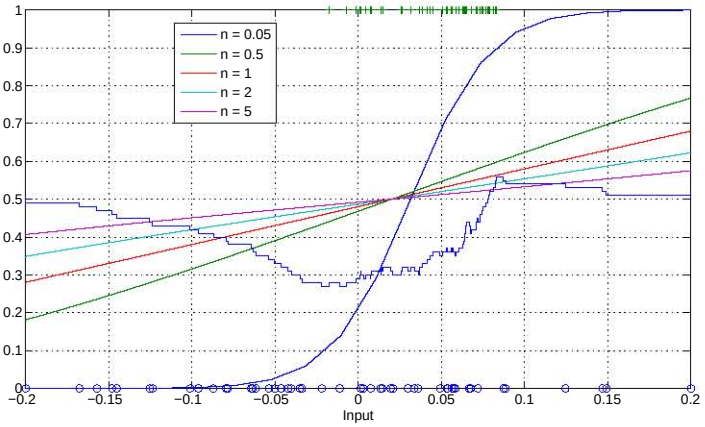
(b)



(c)



(d)



(e)

Figure 4.5: Beta family. Results for different values of n, (a) function g, (b) function ϕ', (c) canonical link f, (d) Loss function, (e) Average cost.

4.5 Cost-sensitive sequences of Bregman divergences

In Chapter 2 we derived a relationship between a particular multiclass Bregman divergence and maximum margin. In this section we forget about achieving convexity through the canonical link and we focus on studying the conditions we need to impose on Bregman divergences to obtain maximum margin as a limit classifier in general. This analysis will allow us not to rule out convex optimization methods and determine the type of margin we can get in an asymmetric scenario as well.

We devote the rest of the chapter to provide some properties to characterize parametric cost-sensitive Bregman divergences that maintain the behavior of parametric family given by Eq. 2.9. We will refer to them as *sequences of weighted Bregman loss functions*.

4.5.1 Sequences of weighted Bregman loss functions

We are interested here in sequences of weighted Bregman loss functions $L_n(z, y)$. Given a sequence of *weighting* functions $g_n(z)$ depending on a parameter $n \in \mathbb{N}$, then $L_n(z, y)$ is a sequence of weighted Bregman loss functions iff it can be written in the form

$$L_n(z, y) = \int_y^z g_n(\alpha)(\alpha - y)d\alpha \tag{4.19}$$

where $g_n = \phi_n''$ is a strictly positive function ($g_n(z) > 0$, for any $z \in (0, 1)$).

In general, we are interested in both symmetric and asymmetric loss functions. Symmetric losses are those verifying $L_n(z, y) = L_n(1-z, 1-y)$ for any z and y. It can be shown that a symmetric loss function is a Bregman loss iff g_n is also symmetric, i.e.: $g_n(z) = g_n(1 - z)$. We will consider the symmetric case as a particular initial case of our study.

4.5.2 Sequences minimizing the total cost

In this section we present and analyze the asymptotic behavior of the sequence L_n for large values of n in the case where we deal with non-separable data.

Consider the cost-sensitive loss given by

$$\hat{L}_q(z, y) = q(1 - y)\hat{y} + (1 - q)y(1 - \hat{y}) \tag{4.20}$$

where $\hat{y} = u(z - q)$ (and u is the step function). Note that, for $q = 1/2$, $\hat{L}_q$ becomes half the zero-one loss. For any other q, the loss is a cost indicator: for any cost-sensitive problem with normalized costs q and $1-q$, $\hat{L}_q$ computes the cost of decision rule $\hat{y} = u(z-q)$ (which is optimal if z is the posterior probability of class 1). Given function $f_{\mathbf{w}}(\mathbf{x})$, the total cost computed over sample set $\mathcal{S} = \{(\mathbf{x}^k, y^k), k = 1, \ldots, K\}$ is

$$\hat{R}_q(\mathbf{w}) = \sum_{k=1}^{K} \hat{L}_q(f(\mathbf{x}^k), y^k) = q \cdot n_{e_0} + (1 - q) \cdot n_{e_1} \tag{4.21}$$

where n_{e_0} is the number of errors (NOE) of deciding in favor of class 1 (0) when the true class is 0 (1).

We will say that f is a separating function for the sample set $\mathcal{S}$ if $\hat{R}_q(\mathbf{w}) = 0$. Also, we will say that $\mathcal{S}$ is separable by a function class $\mathcal{F}$ if there exists a separating function $f \in \mathcal{F}$.

First we show a general condition on a sequence of Bregman divergences to converge to the total cost.

Theorem 4.5.1 *Consider the sequence of Bregman losses $\{L_n(z, y), n = 0, 1, 2, ...\}$ given by weighting functions $\{g_n(z), n = 0, 1, 2, \ldots\}$, and let $R_n(\mathbf{w})$ be the corresponding sequence of empirical risks given by*

$$R_n(\mathbf{w}) = \sum_{k=1}^{K} L_n(z^k, y^k) \tag{4.22}$$

where $y^k \in \{0, 1\}$ and $z^k \in [0, 1]$. If g_n converges to a delta distribution shifted to some $q \in (0, 1)$ (i.e.,

$$\lim_{n\to\infty} \int_0^1 f(z) g_n(z) dz = f(q) \tag{4.23}$$

for any continuous function f), then

$$\lim_{n\to\infty} R_n(\mathbf{w}) = \hat{R}_q(\mathbf{w}) \tag{4.24}$$

Proof The proof is straightforward. Since L_n is a Bregman Loss, for any n, we can express Eq. 4.22 as

$$R_n(\mathbf{w}) = \sum_{k=1}^{K} \int_{y^k}^{z^k} g_n(\alpha)(\alpha - y^k)d\alpha \tag{4.25}$$

Taking the limit,

$$\begin{aligned}
\lim_{n\to\infty} R_n(\mathbf{w}) =& \sum_{k=1}^{K} \lim_{n\to\infty} \int_{y^k}^{z^k} g_n(\alpha)(\alpha - y^k)d\alpha \\
=& \sum_{k=1}^{K} \lim_{n\to\infty} \int_{0}^{1} g_n(\alpha)(\alpha - y^k)\cdot \\
& \qquad \cdot \left(u\left(z^k - \alpha\right) - u\left(y^k - \alpha\right)\right) d\alpha \\
=& \sum_{k=1}^{K} (q - y^k)\left(u\left(z^k - q\right) - u\left(y^k - q\right)\right) \\
=& \sum_{k=1}^{K} (q - y^k)(\hat{y}^k - y^k) \\
=& q \cdot n_{e_0} + (1 - q) \cdot n_{e_1}
\end{aligned} \tag{4.26}$$

∎

In summary, if g_n behaves asymptotically as a delta function centered in q, as n goes to ∞, R_n is minimum for a function minimizing the total cost.

4.5.3 Sequences and maximum margin

If data are separable, the previous analysis shows that, for a large n, R_n is minimum for a separating function (such that $n_{e_0} = n_{e_1} = 0$). But the number of such functions

in the function class may be infinity. In this section we establish some connections between the minimizers of a sequence of weighted Bregman losses and some kind of large margin classifiers. To do so, we first state some conditions for which the empirical risk is dominated by the cost on a reduced subset of samples, providing a sparse representation of the problem.

To make the analysis independent of a particular function class, we consider sample *multisets*[1] in the form $\mathcal{S} = \{(z^k, y^k) |\, z^k \in [0,1], y^k \in \{0,1\}, k = 1, \ldots, K\}$. Eventually, scalars z^k will be the outputs of a particular predictor f for some input $\mathbf{x}^k$. To analyze the asymptotic behavior of Bregman loss sequences, we will focus in zero-error multisets, whose elements (z^k, y^k) satisfy $y^k = u(z^k - q)$, where q is the decision threshold.

In both the following theorem and its proof, we use the bar symbol " $\bar{}$ " over an arbitrary variable z to denote the operation

$$\bar{z} = \min\{z, 1 - z\} \tag{4.27}$$

Symmetric losses

The following theorem describes the separable case for losses with decision threshold $q = 1/2$ (cost-insensitive).

Theorem 4.5.2 *Consider the sequence of symmetric Bregman losses $\{L_n\}$ given by their respective symmetric weighting functions g_n (such that $g_n(z) = g_n(1-z)$, for any $z \in [0,1]$). For any multiset $\mathcal{S} = \{(z^k, y^k) |\, z^k \in [0,1], y^k \in \{0,1\}, k = 1, \ldots, K\}$, let $R_n(\mathcal{S})$ be the empirical loss given by*

$$R_n(\mathcal{S}) = \sum_{k=1}^{K} L_n(z^k, y^k) \tag{4.28}$$

and let $\mathcal{M}_\mathcal{S} = \arg\max_k \{\bar{z}^k\}$ and let $|\mathcal{M}_\mathcal{S}|$ be the cardinality of $\mathcal{M}_\mathcal{S}$.

The following conditions are equivalent

[1] We use multisets instead to sets to allow $\mathcal{S}$ to have repeated elements.

1. *For any zero-error multiset $\mathcal{S}$ (for decision threshold 1/2) and any $m \in \mathcal{M}_{\mathcal{S}}$,*

$$\lim_{n\to\infty} \frac{R_n(\mathcal{S})}{L_n(z^m, y^m)} = |\mathcal{M}_{\mathcal{S}}| \tag{4.29}$$

2. *For almost any pair $(z, z') \in [0,1]^2$ such that $\bar{z} < \bar{z'}$,*

$$\lim_{n\to\infty} \frac{g_n(z)}{g_n(z')} = 0 \tag{4.30}$$

Proof First we prove that Condition 1 implies Condition 2. Note that the empirical risk satisfies

$$\frac{R_n(\mathcal{S})}{L_n(z^m, y^m)} = |\mathcal{M}_{\mathcal{S}}| + \sum_{k\notin\mathcal{M}_{\mathcal{S}}} \frac{L_n(z^k, y^k)}{L_n(z^m, y^m)} \tag{4.31}$$

Since L_n is symmetric and $\mathcal{S}$ is a zero-error multiset, $L_n(z^k, y^k) = L_n(\bar{z}^k, 0)$, for any k, and we can write

$$\frac{R_n(\mathcal{S})}{L_n(z^m, y^m)} = |\mathcal{M}_{\mathcal{S}}| + \sum_{k\notin\mathcal{M}_{\mathcal{S}}} \frac{L_n(\bar{z}^k, 0)}{L_n(\bar{z}^m, 0)} \tag{4.32}$$

If (4.29) is true for any $\mathcal{S}$, then, using (4.32) we get

$$\lim_{n\to\infty} \frac{L_n(\bar{z}^k, 0)}{L_n(\bar{z}^m, 0)} = 0 \tag{4.33}$$

for any $\bar{z}^k$, $\bar{z}^m$ such that $0 \le \bar{z}^k < \bar{z}^m \le 1$. This is equivalent to

$$\lim_{n\to\infty} \frac{L_n(\bar{z}^m, 0) - L_n(\bar{z}^k, 0)}{(\bar{z}^m - \bar{z}^k) L_n(\bar{z}^m, 0)} = \frac{1}{\bar{z}^m - \bar{z}^k} \tag{4.34}$$

Since (4.34) is true for any $\bar{z}^k < \bar{z}^m$ and L_n is a differentiable cost, we can take the limit

$$\lim_{n\to\infty} \lim_{\bar{z}^k\to\bar{z}^m} \frac{L_n(\bar{z}^m, 0) - L_n(\bar{z}^k, 0)}{(\bar{z}^m - \bar{z}^k) L_n(\bar{z}^m, 0)} = \infty \tag{4.35}$$

Defining L_n' as the first order derivative of $L_n(z, 0)$ with respect to z, and taking into account that L_n is a Bregman loss given by weighting function g_n,

$$\begin{aligned}
\lim_{\bar{z}^k\to\bar{z}^m} \frac{L_n(\bar{z}^m, 0) - L_n(\bar{z}^k, 0)}{(\bar{z}^m - \bar{z}^k) L_n(\bar{z}^m, 0)} &= \frac{L_n'(\bar{z}^m, 0)}{L_n(\bar{z}^m, 0)} \\
&= \frac{\bar{z}^m g_n(\bar{z}^m)}{\int_0^{\bar{z}^m} \alpha g_n(\alpha) d\alpha} \\
&= \left(\int_0^{\bar{z}^m} \frac{\alpha g_n(\alpha)}{\bar{z}^m g_n(\bar{z}^m)} d\alpha \right)^{-1}
\end{aligned} \tag{4.36}$$

Combining (4.35) and (4.36) we get

$$\lim_{n\to\infty} \frac{\alpha g_n(\alpha)}{\bar{z}^m g_n(\bar{z}^m)} = 0 \tag{4.37}$$

for all but an enumerable set of values $\alpha \in [0, \bar{z}^m]$. This is true for any $\bar{z}^m \in [0, 1/2]$. Because of the symmetry of g we get Condition 2.

Now we prove that Condition 2 implies Condition 1. If (4.30) is true for almost any pair $(z, z') \in [0,1]^2$ with $\bar{z} < \bar{z'}$, then

$$\lim_{n\to\infty} \frac{z g_n(z)}{\bar{z'} g_n(\bar{z'})} = 0 \tag{4.38}$$

so that

$$\lim_{n\to\infty} \frac{\int_0^{\bar{z'}} \alpha g_n(\alpha) d\alpha}{\bar{z'} g_n(\bar{z'})} = 0 \tag{4.39}$$

and, using the equality relations in (4.36), we get

$$\lim_{n\to\infty} \frac{L'_n(z,0)}{L_n(z,0)} = \infty \tag{4.40}$$

Since L_n is convex, for any $\bar{z} < \bar{z'}$

$$L_n(\bar{z'}, 0) > L_n(\bar{z}, 0) + L'_n(\bar{z}, 0)(\bar{z'} - \bar{z}) \tag{4.41}$$

Combining (4.40) and (4.41), we have

$$\lim_{n\to\infty} \frac{L_n(\bar{z}, 0)}{L_n(\bar{z'}, 0)} < \lim_{n\to\infty} \frac{L_n(\bar{z}, 0)}{L_n(\bar{z}, 0) + L'_n(\bar{z}, 0)(\bar{z'} - \bar{z})} = 0 \tag{4.42}$$

Thus, using (4.31) and (4.32) we get

$$\begin{aligned} \lim_{n\to\infty} \frac{R_n(\mathcal{S})}{L_n(z^m, y^m)} &= |\mathcal{M}_\mathcal{S}| + \sum_{k \notin \mathcal{M}_\mathcal{S}} \lim_{n\to\infty} \frac{L_n(\bar{z^k}, 0)}{L_n(\bar{z^m}, 0)} \\ &= |\mathcal{M}_\mathcal{S}| \end{aligned} \tag{4.43}$$

∎

Th. 4.5.2 shows that, for large n, the empirical risk can be approximated as

$$R_n(\mathcal{S}) \approx |\mathcal{M}_\mathcal{S}| L_n(z^m, y^m) \tag{4.44}$$

provided that Condition 2 is satisfied. This in turn implies that the risk depends primarily on the loss of the critical samples $L_n(z^m, y^m)$. In order to minimize the risk, $L_n(z^m, y^m)$ should be minimized. Since L_n is convex, this means that $\hat{z}^m$ should be maximized. If $z^m = f(\mathbf{w}^T\mathbf{x})$ with f a sigmoid-type function, this is equivalent to maximizing the usual margin, defined as the distance from samples to the decision boundary.

Asymmetric losses

This section describes the asymmetric case (cost-sensitive). First of all let us introduce the definition of *order-preserving* sequence.

Definition (Order-preserving sequence)

Let $\mathcal{L} = \{L_n, n = 0, 1, ...)\}$ be a sequence of loss functions such that $\lim_{n\to\infty} L_n(1-y, y) = c_y > 0$, for $y \in \{0, 1\}$. Sequence $\mathcal{L}$ is *order-preserving* with parameter $q \in [0, 1]$ if there exists some $n_0 > 0$ such that, for any $z, z' \in [0, 1] \setminus \{q\}$, if $y = u(z - q)$ and $y' = u(z' - q)$, either $L_n(z, y) \leq L_n(z', y')$ for all $n > n_0$ or $L_n(z, y) \geq L_n(z', y')$ for all $n > n_0$.

Note that, if $\mathcal{L}$ is order-preserving, we can define the relation $z \prec z'$ iff $L_n(z, u(z-q)) \leq L_n(z', u(z'-q))$ for n large enough. This relation has the following properties:

- It is a *preorder*: reflexive, antisymmetric and transitive.
- It is complete: for any $z, z' \in [0, 1]$, $z \prec z'$ or $z' \prec z$.

If $\mathcal{L}$ is a sequence of Bregman losses, the following properties are also satisfied:

- If $z, z' \in [0, q]$ with $z \neq z'$ and $z \prec z'$, then $z < z'$. This is because $L_n(z, 0)$ is strictly increasing for any Bregman loss L_n.
- If $z, z' \in [q, 1]$ with $z \neq z'$ and $z \prec z'$, then $z' > z$ (because $L_n(z, 1)$ is strictly decreasing for all n)

Lemma 4.5.3 *(Existence of a pairing function)*

If the sequence of Bregman Losses $\mathcal{L} = \{L_n, n = 0, 1, ...)\}$ is order-preserving with parameter $q \in [0, 1]$, there exists a map $v : [0, 1] \to [0, 1]$ with $v([0, q]) \subset [q, 1]$ and $v([q, 1]) \subset [0, q]$ such that, for any $z, z' \in [0, 1]$ with $u(z - q) \neq u(z' - q)$,

1. *If $z \prec z'$, then $v(z) \prec z'$.*
2. *If $z' \prec z$, then $z' \prec v(z)$.*

Proof We complete the proof for $z \in (q, 1]$ (and, thus, $y = 1$). Due to the symmetry of the problem, the proof for $z \in [0, q]$ follows similar steps. For any $z \in [q, 1]$ we define the sets

$$\mathcal{V}_-(z) = \{z' \in [0, q] |\ z' \prec z\} \tag{4.45}$$

$$\mathcal{V}_+(z) = \{z' \in [0, q] |\ z \prec z'\} \tag{4.46}$$

and we define the extreme values

$$z_- = \sup_z \mathcal{V}_-(z) \tag{4.47}$$

(which is well defined because, since $0 \in \mathcal{V}_-(z)$, $\mathcal{V}_-(z)$ is never empty), and

$$z_+ = \begin{cases} \inf_z \mathcal{V}_+(z), & \text{if } \mathcal{V}_+(z) \neq \emptyset \\ q, & \text{if } \mathcal{V}_+(z) = \emptyset \end{cases} \tag{4.48}$$

Note that, for any $z' \in [0, q]$,

- If $z' \prec z$, then $z' \in \mathcal{V}_-(z)$ and, thus, $z' \leq z_-$,
- If $z \prec z'$, then $z' \in \mathcal{V}_+(z)$ and, thus, $z' \geq z_+$,

Since the preorder is complete, for any $z' \in [0, q]$ we have $z' \in \mathcal{V}_-(z)$ or $z' \in \mathcal{V}_+(z)$ and, thus $z_- \geq z_+$.

Note also, that, for any $z_0 \in \mathcal{V}_-(z)$ and $z_1 \in \mathcal{V}_+(z)$, we have $z_0 \prec z \prec z_1$ and, thus, $z_0 \leq z_1$. Thus, $z_- \leq z_+$

Thus, $z_- = z_+$. The function $v(z) = z_- = z_+$ satisfies the desired properties.

- If $z \prec z'$, then $z' \in \mathcal{V}_+(z)$, thus $v(z) \leq z'$, thus $v(z) \prec z'$.
- If $z' \prec z$, then $z' \in \mathcal{V}_-(z)$, thus $z' \leq v(z)$, thus $z' \prec v(z)$.

In both the following theorem and its proof, we use the bar symbol " $\bar{}$ " over an arbitrary variable z to denote the operation

$$\bar{z} = \min\{z, v(z)\} \tag{4.49}$$

Theorem 4.5.4 *Consider the order-preserving sequence (with parameter q) of Bregman losses $\{L_n\}$ given by their respective weighting functions g_n. For any multiset $\mathcal{S} = \{(z^k, y^k) |\, z^k \in [0,1], y^k \in \{0,1\}, k = 1, \ldots, K\}$, let $R_n(\mathcal{S})$ be the empirical loss given by*

$$R_n(\mathcal{S}) = \sum_{k=1}^{K} L_n(z^k, y^k) \tag{4.50}$$

and let $\mathcal{M}_\mathcal{S} = \arg\max_k\{\bar{z^k}\}$ and let $|\mathcal{M}_\mathcal{S}|$ be the cardinality of $\mathcal{M}_\mathcal{S}$.

The following conditions are equivalent

1. *For any multiset $\mathcal{S}$ with zero errors with decision threshold q, and any $m \in \mathcal{M}_\mathcal{S}$,*

$$\lim_{n\to\infty} \frac{R_n(\mathcal{S})}{L_n(z^m, y^m)} = |\mathcal{M}_\mathcal{S}| \tag{4.51}$$

2. *For almost any pair $(z, z') \in [0,1]^2$ such that $\bar{z} < \bar{z'}$,*

$$\lim_{n\to\infty} \frac{g_n(z)}{g_n(z')} = 0 \tag{4.52}$$

Proof First we prove that Condition 1 implies Condition 2. Note that the empirical risk satisfies (4.31). If (4.51) is true for any $\mathcal{S}$, then, using (4.32) we get

$$\lim_{n\to\infty} \frac{L_n(z^k, y^k)}{L_n(z^m, y^m)} = 0 \tag{4.53}$$

Since (4.53) must be true for any (z^k, y^k), (z^m, y^m) with no errors ($y^k = u(z^k - q)$ and $y^m = u(z^m - q)$), we can take them such that $y^k = y^m = y$, so that

$$\lim_{n\to\infty} \frac{L_n(z^k, y)}{L_n(z^m, y)} = 0 \tag{4.54}$$

which is true for any z^k, z^m such that $0 \leq z^k < z^m \leq q$ or $q \leq z^m < z^k \leq 1$. This is equivalent to

$$\lim_{n\to\infty} \frac{L_n(z^m, y) - L_n(z^k, y)}{(z^m - z^k)L_n(z^m, y)} = \frac{1}{z^m - z^k} \tag{4.55}$$

Taking the limit

$$\lim_{n\to\infty} \lim_{z^k \to z^m} \frac{L_n(z^m, y) - L_n(z^k, y)}{(z^m - z^k)L_n(z^m, y)} = \infty \tag{4.56}$$

Defining L'_n as the first order derivative of $L_n(z, y)$ with respect to z, and taking into account that L_n is a Bregman loss given by weighting function g_n,

$$\begin{aligned} \lim_{z^k \to z^m} \frac{L_n(z^m, y) - L_n(z^k, y)}{(z^m - z^k)L_n(z^m, y)} &= \frac{L'_n(z^m, y)}{L_n(z^m, y)} \\ &= \left(\int_y^{\bar{z^m}} \frac{(y-\alpha)g_n(\alpha)}{(y - z^m)g_n(z^m)} d\alpha \right)^{-1} \end{aligned} \tag{4.57}$$

Combining (4.56) and (4.57) we get

$$\lim_{n\to\infty} \frac{(y-\alpha)g_n(\alpha)}{(y - z^m)g_n(z^m)} = 0 \tag{4.58}$$

for all but a set of values α with zero-Lebesgue measure. Since this is true for any $z^m \in [0, 1]$, we get Condition 2.

Now we prove that Condition 2 implies Condition 1. If (4.52) is true for almost any pair $(z, z') \in [0, 1]^2$ with $\bar{z} < \bar{z'}$, then

$$\lim_{n\to\infty} \frac{z g_n(z)}{\bar{z'} g_n(\bar{z'})} = 0 \tag{4.59}$$

so that

$$\lim_{n\to\infty} \frac{\int_0^{\bar{z'}} \alpha g_n(\alpha) d\alpha}{\bar{z'} g_n(\bar{z'})} = 0 \tag{4.60}$$

and, using the equality relations in (4.36), we get

$$\lim_{n\to\infty} \frac{L'_n(z, 0)}{L_n(z, 0)} = \infty \tag{4.61}$$

Since L_n is convex, for any $\bar{z} < \bar{z'}$

$$L_n(\bar{z'}, 0) > L_n(\bar{z}, 0) + L'_n(\bar{z}, 0)(\bar{z'} - \bar{z}) \tag{4.62}$$

Combining (4.61) and (4.62), we have

$$\lim_{n\to\infty} \frac{L_n(\bar{z},0)}{L_n(\bar{z'},0)} < \lim_{n\to\infty} \frac{L_n(\bar{z},0)}{L_n(\bar{z},0) + L'_n(\bar{z},0)(\bar{z'}-\bar{z})} = 0 \tag{4.63}$$

Thus, using (4.31) and (4.32) we get

$$\begin{aligned} \lim_{n\to\infty} \frac{R_n(\mathcal{S})}{L_n(z^m, y^m)} &= |\mathcal{M}_\mathcal{S}| + \sum_{k\notin\mathcal{M}_\mathcal{S}} \lim_{n\to\infty} \frac{L_n(\bar{z^k},0)}{L_n(\bar{z^m},0)} \\ &= |\mathcal{M}_\mathcal{S}| \end{aligned} \tag{4.64}$$

∎

It is easy to check that both weighting functions given by the Bregman generators in Eq. (4.9) and Eq. (4.10) satisfy the conditions of the theorems. Also note that any sequence $g_n(z)$ in the form

$$g_n(z) = a_n b(z)^n \tag{4.65}$$

where a_n is a normalizing constant (to ensure unit area) and $b(z)$ is bell-shaped (decreasing around q), satisfies the requirements of the theorem. The next examples show different functions that can be deduced from Eq. (4.65):

Example Gaussian weighting function.

$$g_n(z) = \frac{1}{\sqrt{2\pi\frac{1}{n}}} \exp\left(-n\left(z-\frac{1}{2}\right)^2\right)$$

where the corresponding function $b(z)$ is given by

$$b(z) = exp\left(-(z-1/2)^2\right)$$

Example Beta loss as defined in Eq. (4.11) but with $n \in \mathbb{N}$.

$$g_n(z) = a_n(z^{qn}(1-z)^{(1-q)n})$$

where $a_n = \beta((qn+1),(1-q)n+1)^{-1}$ (β is the beta function) is a normalizing constant, ensuring that g_n has unit area. The corresponding function $b(z)$ is given by

$$b(z) = z^q(1-z)^{(1-q)}$$

The theorem focuses on the study of the behavior of samples $\mathbf{x}^m$, where m is the index of any sample maximizing the value of function g_n, that is, $g_n(z^m) = \max_k(g_n(z^k))$. This condition can be thought as a definition of a *generalized margin*, in the sense that $\mathbf{x}^m$ must be a sample that makes z as close as possible to q (where the value of $g_n(z)$ is maximum). This means that $g_n(z)$ will be close to $g_n(p)$ and, therefore, z will be close to p (being p the true posterior probability). This is equivalent to saying that the sample $\mathbf{x}^m$ is close to the boundary according to the metric defined by the function f.

We can now move towards an analysis based on the standard concept of margin. Given a dataset $\mathcal{S}$ and a linear classifier with separating boundary $\mathbf{w}$, the margin is defined as

$$\text{margin}(\mathbf{w}, \mathcal{S}) = \min_{\mathbf{x} \in \mathcal{S}} \text{dist}(\mathbf{x}, \mathbf{w}) \tag{4.66}$$

where $\text{dist}(\mathbf{x}, \mathbf{w})$ is the Euclidean distance from sample $\mathbf{x}$ to the boundary hyperplane determined by $\mathbf{w}$.

In the particular case in which we consider a posterior probability map $z = f(\mathbf{w}^T\mathbf{x})$, for an arbitrary increasing f which is linear in some feature space, and a function $g_n(z)$ locally symmetric around q (e.g., if $g_n(z) = a_n b(z)^n$ and $b(z)$ is a Gaussian function centered in q), then we can state that the sample $\mathbf{x}^m$ maximizing the $\text{margin}(\mathbf{w}, \mathcal{S})$ is the same as the sample maximizing $g_n(z^k)$. Therefore, if $g_n(z)$ is locally symmetric around q, the minimizer of the Bregman sequence converges to a maximum margin classifier which does not depend on q.

This is in agreement with some cost-sensitive learning methods based on weighting samples, that do not modify the decision boundary in separable problems [Bach et al., 2006, Wu and Srihari, 2003, Davenport et al., 2006]. This fact makes a difference with respect to threshold shifting methods, such as those based on shifting decision boundaries depending on the logarithm of cost ratios [Dmochowski et al., 2010]. We will further explore this issue in Chapter 5.

4.6 Summary

In this chapter we study some properties of Bregman divergences in binary experiments. First of all, we explore the relationship between Bregman divergences and convexity through the canonical link. We find that some of the families that are particularly interesting to us lead to canonical links that are not appropriate for classification purposes.

This chapter also describes a procedure to approximate cost-sensitive losses using Bregman divergences. It provides a characterization of the construction of sequences of Bregman divergences in order to achieve some nice properties in a cost-sensitive setting. It is desired that their minimization to guarantee, asymptotically, minimum number of errors in non-separable cases, and maximum margin classifiers in separable problems.

The asymptotic results in the non-separable case confirm that the described sequences of weighted Bregman losses accomplish the first goal: if the weighting function g_n behaves asymptotically as a delta function centered in q, the empirical risk is minimum for a function minimizing the total cost.

In the separable case, two different weighting functions are considered, starting with the simplest case (symmetric) and then extending the analysis to the general case (asymmetric). The main result provides that, under very general conditions, it is proven that the minimization of sequences of weighted Bregman losses is equivalent to the maximization of a generalized margin.

One of the goals behind this work was to derive a convex loss from the asymptotic analysis of the sequence of Bregman divergences. In this respect, a positive result is achieved in separable problems for a large class of Bregman divergences sequences (the classifiers converge to a maximum margin classifier which is independent on the cost), with a negative side (the limit classifier does not depend on the cost parameters).

These results are coherent with intuition that we have to give up convexity in

order to approach the non-convex loss in Eq. (4.20) and achieve the minimum total cost.

Chapter 5

Learning with example-dependent costs

The one where we explore the advantages and disadvantages of cost-sensitive sequences of Bregman divergences in example-dependent cost scenarios. In this chapter we analyze the problem of designing cost-sensitive classifiers under example-dependent costs, when complete cost matrices for each sample are available during training and test. Our proposal consists in estimating the posterior probability map using a combination of surrogate losses based on Bregman divergences. The contribution of each sample to the empirical risk is given by a loss that depends on the cost parameters for that sample. Our experiments show that the appropriate choice of these sample-dependent losses can outperform conventional cost-independent posterior probability estimators, at least in terms of classification performance. Moreover, we show that probability estimators make a more efficient use of cost information during training and test with respect to other discriminative approaches, like cost-sensitive support vector machines. This chapter is a summary of the work with Dario Garcia-Garcia and Jesus Cid-Sueiro in [Santos-Rodriguez et al., 2011b].

5.1 Introduction

Most decision problems are intrinsically related to example-dependent costs. By this we mean that the cost associated with a deciding a label for a given sample depends not only on the actual and estimated labels, but on the sample itself. Consider for instance the problem of credit risk classification. The task consist in classifying borrowers as good or bad clients (more or less likely to return the whole credit). The cost of each decision can be influenced by a number factors that vary from customer to customer, including the amount in demand, the duration of the credit or the previous links between client and entity.

As we mentioned before, the literature in machine classification is mainly devoted to the cost-insensitive or *minimum error probability* scenario. However, in a case such as the one defined above, minimizing the error probability is not the natural choice. A first approach to go beyond cost-insensitive classifiers consists in introducing deterministic label-dependent costs to asymmetrically penalize incorrect decisions and to reward correct ones [Elkan, 2001a], yielding the standard cost-sensitive learning scenario. In the credit risk example, this amounts to considering, for example, that the cost of classifying a risky borrower as good can be much higher that the cost of classifying a potentially good customer as bad. This is just a special case of the example-dependent cost learning framework and, as such, considers only a part of the whole picture.

The example-dependent cost framework has been hardly treated in the literature. It can be traced back to [Lenarcik and Piasta, 1998] and is also mentioned by Provost and Fawcett in [Provost and Fawcett, 2001]. Our approach follows mainly from [Zadrozny and Elkan, 2001a], where misclassification costs are different for different examples in the same way that class membership probabilities are example-dependent. Zadrozny and Elkan describe domains where both costs and probabilities are unknown for test examples, so both cost estimators and probability estimators must be learned. Few other works have dealt specifically with cost-sensitive learning

with example-dependent costs. In [Brefeld et al., 2003, Zadrozny et al., 2003] the classical support vector machine (SVM) was extended to handle example-dependent costs. Very recently [Scott, 2011] studies surrogate example-dependent losses to provide conditions for the existence of surrogate regret bounds.

In this chapter we consider the application of Bregman divergences as loss functions to generate finely tuned posterior probability estimates in example-dependent costs scenarios. In our proposal, each sample contributes to the objective function depending not only on the distance to the boundary but also through its cost. This way we are able to obtain loss functions which try to accurately approximate the posterior probability in the areas of interest from a classification perspective, that is to say, near the optimal classification boundary of the example-dependent cost problem. This implies that, if the capacity of the learning machine (or, equivalently, the number of training examples) is limited, our approach automatically optimizes the classification performance, in contrast with standard losses for probability estimation. At the same time, estimating posterior probabilities near the boundary also allows for a more robust handling of imprecisions in the cost definitions, in contrast with methods which directly optimize a measure of classification accuracy, such as the SVM and its cost-sensitive extensions. In this sense, our approach can be seen as enjoying benefits of both purely classification-oriented methods and standard posterior probability estimation.

5.2 Cost-sensitive learning with example-dependent costs

We slightly change the notation in order to make it more suitable for example-dependent cost scenarios. Let $\mathcal{X}$ be a sample space and $\mathcal{C}$ the space of 2×2 cost matrices. Let (X, Y, C) be a triple of random variables taking values on $\mathcal{X} \times \{0, 1\} \times \mathcal{C}$, according to a joint probability distribution $P(X, Y, C)$. The component c_{iy} of matrix $\mathbf{C} \in \mathcal{C}$ represents the cost of deciding in favor of class i when the correct class is y.

The goal of the cost-sensitive classification problem discussed in this chapter is to predict the correct label y when both sample $\mathbf{x}$ and cost matrix $\mathbf{C}$ are observed, by minimizing the expected cost or *risk*

$$R = \mathbb{E}_{(\mathbf{x},y,\mathbf{C})\sim P}\{c_{iy}\}. \tag{5.1}$$

For every sample $\mathbf{x}$, the optimal decision maker selects class i^* such that

$$i^*(\mathbf{x},\mathbf{C}) \in \arg\min_i \{(1-p(\mathbf{x}))c_{i0} + p(\mathbf{x})c_{i1}\}, \tag{5.2}$$

where $p(\mathbf{x}) = P(Y = 1 \mid X = \mathbf{x})$ is the *posterior probability function*. Assuming positive regrets $c_{10} - c_{11}$ and $c_{01} - c_{00}$, it is easy to see that the risk R is minimized by the assignment

$$i^*(\mathbf{x},\mathbf{C}) = I_{p(\mathbf{x})\geq q(\mathbf{C})}, \tag{5.3}$$

where I denotes the indicator function and $q(\mathbf{C})$ is the *normalized regret*

$$q(\mathbf{C}) = \frac{c_{10} - c_{00}}{c_{10} - c_{11} + c_{01} - c_{00}}. \tag{5.4}$$

It follows that the classification depends just on those regrets, not on the absolute cost values themselves.

In a standard setting, $P(X,Y,C)$ is unknown and only a training set $\mathcal{S} = \left\{(\mathbf{x}^k, y^k, \mathbf{C}^k), k = 1,\ldots,K\right\}$ of statistically independent pairs drawn from P and their corresponding cost matrices $\mathbf{C}^k = \begin{pmatrix} c^k_{00} & c^k_{01} \\ c^k_{10} & c^k_{11} \end{pmatrix}$ (or, equivalently, the corresponding regrets) is available. The classical discriminative approach to the problem consists on estimating a posterior probability map $z(\mathbf{x})$ using sample $\mathcal{S}$. That estimation can be carried out through the Empirical Risk Minimization. The empirical risk given the samples in $\mathcal{S}$ is defined by

$$R_{emp} = \frac{1}{K}\sum_{k=1}^{K} c^k_{i^k,y^k} \tag{5.5}$$

where $i^k = I_{z(\mathbf{x}^k)\geq q(\mathbf{x}^k)}$ is the decision for the k-th sample according to z.

5.3 Example-dependent costs and Bregman divergences

The motivation follows the same reasoning as the ideas presented in Section 4.3. Since the loss c_{iy} in Eq. (5.1) is neither convex nor differentiable it is then convenient from a practical perspective to explore the use of other *surrogate* losses. As we have shown previously, Bregman divergences are a natural choice.

In the general scenario discussed in this chapter, the threshold q is sample dependent and, thus, there is no single choice of the Bregman generator that is appropriate for any sample. Our approach in this chapter is based on using a different generator for each sample. Noting that

$$\mathbb{E}_{(\mathbf{x},y,C)\sim P}\{D_{\phi_C}(y,z)\} = \mathbb{E}_{\mathbf{C}\sim P}\{\mathbb{E}_{(\mathbf{x},y)\sim P}\{D_{\phi_\mathbf{C}}(y,z)|\mathbf{C}\}\} \tag{5.6}$$

Since the inner expectation is minimized by $z = p$ for any $\mathbf{C}$, the whole expectation is also minimized by p.

Given the above discussion, we propose to optimize the empirical risk based on Eq. (5.6), given by

$$R(\mathbf{w}) = \frac{1}{K}\sum_{k=1}^{K} D^k(y^k, z^k) \tag{5.7}$$

where $z^k = z(\mathbf{x}^k)$, D^k is the Bregman divergence given by generator $\phi_{\mathbf{C}^k}$, and $\mathbf{w}$ is the parameter vector specifying z. Note that each sample is associated with its very own divergence. $R(\mathbf{w})$ represents the average of these divergences evaluated in their corresponding samples.

5.3.1 An example

This synthetic example tries to illustrate the difference between minimizing the empirical risk from Eq. (5.7) (from now on, EDBD) and example-independent divergences in a scenario where the capacity of the learning machine is limited. Consider

the two-class problem with classes "0" and "1" and the probability map given by

$$p(\mathbf{x}) = \frac{1}{3}\left(\Phi(\mathbf{w_0}^T\mathbf{x}) + \Phi(\mathbf{w_1}^T\mathbf{x}) + \Phi(\mathbf{w_2}^T\mathbf{x})\right) \tag{5.8}$$

where $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{w_0} = (4, 0)$, $\mathbf{w_1} = (0, 4)$ and $\mathbf{w_2} = (1, 1)$. The inverse link function $\Phi : \mathbb{R} \to [0, 1]$ is the logistic function given by $\Phi(z) = 1/\left(1 + \exp(z)\right)$ The contour-plot of this probabilistic map is represented in Figure 5.2. Colder colours correspond to higher values of $1-p(\mathbf{x})$, the posterior probability of class "0". We generated 8000 training samples uniformly distributed in the square $[-4, 4] \times [-4, 4]]$. The label of every sample was assigned stochastically according to the previous probability map. A single layer perceptron (SLP) with soft decisions given by

$$z(\mathbf{x}) = \Phi(\mathbf{w}^T\mathbf{x}) \tag{5.9}$$

was used to estimate this map. Since the SLP has not enough capacity to do it exactly, different Bregman loss functions provide different approximations.

Learning consists of estimating parameters $\mathbf{w}$ by means of the minimization of the Bregman divergence using a quasi-Newton method. In this case, the selected one was the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. In quasi-Newton methods the Hessian matrix of second derivatives of the function to be minimized does not need to be computed at any stage. The Hessian is updated by analyzing successive gradient vectors instead.

Figure 5.1 shows the two cost policies are assigned to the samples depending on their position in the input space: for points satisfying that their ordinate is greater or equal than their abscissa, $\mathbf{C} = \begin{pmatrix} 0 & 3 \\ 7 & 0 \end{pmatrix}$; in any other case, $\mathbf{C} = \begin{pmatrix} 0 & 7 \\ 3 & 0 \end{pmatrix}$. Here, for the purpose of illustrating the advantages of using the proposed approach, D_{ϕ_n} is the divergence given by the convex function in Eq. (4.9) (but it could have been replaced with any other cost-sensitive Bregman divergence). The example-independent divergence is also based on in Eq. (4.9) but makes use of the mean costs to estimate the posterior probability.

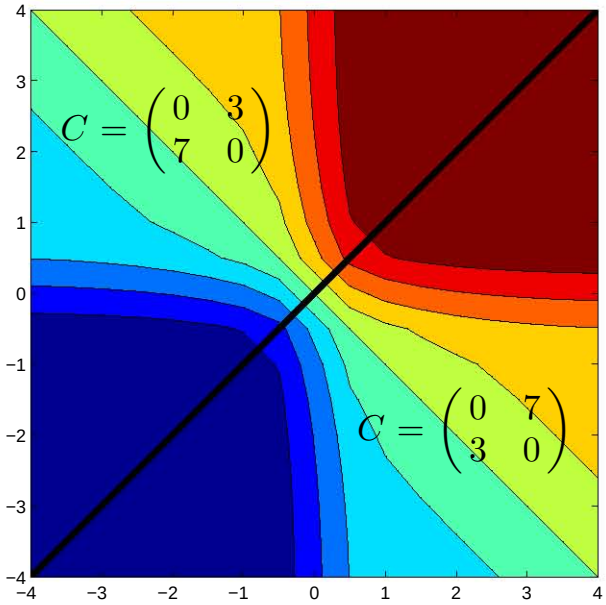
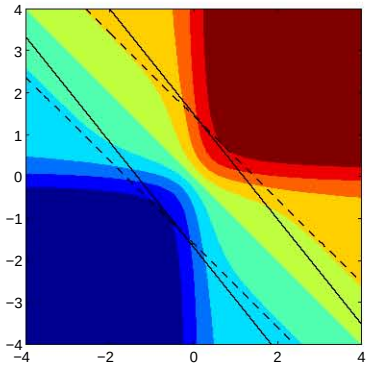
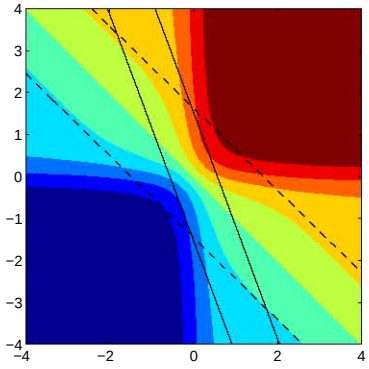


Figure 5.1: Cost policies assigned in Ex. 5.3.1.

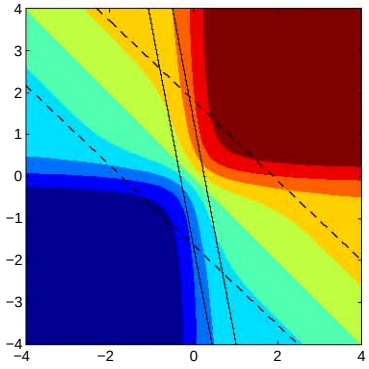
The result of any comparison between Bregman losses depends on how we measure the quality of a probability estimate. Figures 5.2(a), 5.2(b) and 5.2(c) show the probability map and the decision boundaries for both cost matrices and $n = \{2, 4, 8\}$ respectively. It becomes clear that, as n increases, the boundary obtained from EDBD varies its direction towards the Bayes solution, while the boundary corresponding to example-independent divergence remains practically unchanged. In this scenario, our method clearly improves locally the accuracy of the probability estimates. The importance of n is highlighted in the next section.



(a)



(b)



(c)

Figure 5.2: Probability map as defined in Eq. (5.8). Dashed line: boundary minimizing the divergence given by the mean costs; solid line: boundary minimizing EDBD, (a) $n = 2$, (b) $n = 4$, (c) $n = 8$. 100

5.4 Convergence to the minimum total cost

In this section we show that the minimization of a empirical risk of the form of Eq. (5.7) with an adequately chosen Bregman generator leads, asymptotically, to the minimization of the overall cost regret (or the minimum number of errors in a cost-insensitive scenario). Note that the following results are asymptotical with respect to a parameter of the Bregman generator and do not depend on the number of samples.

Consider the cost-weighted classification loss given by

$$\hat{L}_c(z, y) = c(1 - y)\hat{y} + (1 - c)y(1 - \hat{y}) \tag{5.10}$$

where $\hat{y} = I_{z \geq c}$. For $c = 1/2$ this becomes half the zero-one loss. For any other constant, the loss is a cost indicator: for any cost-sensitive problem with normalized costs c and $1 - c$, $\hat{L}_c$ computes the cost of decision rule $\hat{y} = I_{z \geq c}$ (which is optimal if $z = p$). Given a sample set $\mathcal{S} = \{(\mathbf{x}^k, y^k, c^k), k = 1, \ldots, K\}$, the corresponding risk is

$$\hat{R}_c(\mathbf{w}) = \sum_{k=1}^{K} \hat{L}_c^k(z^k, y^k)) = \sum_{k=1}^{K} (c^k - y^k)(\hat{y}^k - y^k) \tag{5.11}$$

We show a general condition on our sequence of Bregman divergences to converge to the optimal risk.

Theorem 5.4.1 *Consider the sequence of Bregman losses* $\{L_n^k, n = 0, 1, 2, ...\}$ *with corresponding weighting functions* $\{g_n^k(z) = \frac{\partial^2 \phi_n^k}{\partial z^2}, n = 0, 1, 2, \ldots\}$, *and let* $R_n(\mathbf{w})$ *be the corresponding sequence of empirical risks given by* $R_n(\mathbf{w}) = \sum_{k=1}^{K} L_n^k(z^k, y^k)$, *where* $y^k \in \{0, 1\}$ *and* $z^k \in [0, 1]$. *If* g_n^k *converges to a delta distribution shifted to* c^k *, then*

$$\lim_{n \to \infty} R_n(\mathbf{w}) = \hat{R}_c(\mathbf{w}) \tag{5.12}$$

Proof The proof is straightforward. Since L_n^k is a Bregman loss, we can express the empirical risk as $R_n(\mathbf{w}) = \sum_{k=1}^{K} \int_{y^k}^{z^k} g_n^k(\alpha)(\alpha - y^k) d\alpha$ (see e.g. [Miller et al., 1991]).

Taking the limit,

$$\begin{aligned}\lim_{n\to\infty} R_n(\mathbf{w}) &= \sum_{k=1}^{K} \lim_{n\to\infty} \int_{y^k}^{z^k} g_n^k(\alpha)(\alpha - y^k)d\alpha \\ &= \sum_{k=1}^{K} \lim_{n\to\infty} \int_{0}^{1} g_n^k(\alpha)(\alpha - y^k) \cdot \left(I_{z^k \geq \alpha} - I_{y^k \geq \alpha}\right) d\alpha \\ &= \sum_{k=1}^{K} (z^k - y^k) \left(I_{z^k \geq c^k} - I_{y^k \geq c^k}\right) \\ &= \sum_{k=1}^{K} (c^k - y^k)(\hat{y}^k - y^k),\end{aligned} \tag{5.13}$$

The third equality follows from the condition on the convergence to a delta distribution, which implies $\lim_{n\to\infty} \int_0^1 f(z) g_n^k(z) dz = f(c^k)$ for any continuous f. ∎

In summary, if the individual generators g_n^k behave asymptotically as delta distribution centered in c^k, as n goes to ∞, R_n converges to the minimum total cost. The Bregman divergence defined by (4.9) was shown in Chapter 2 to satisfy the conditions of the above theorem. Thus, for large n, the empirical risk in (5.7) converges to the total empirical cost. On the other hand, smaller values of n provide smoother approximations to the total cost. This way, n provides a "knob" to adjust the behaviour of the empirical risk between a generalized 0-1 loss and numerically and analytically better-behaved versions.

5.5 An application: Credit risk classification

In this section we show the results of experiments carried out to test our approach. To demonstrate the effects of the example dependent costs , we have conducted experiments in a real-world dataset of credit risk classification. Throughout the section we keep the architecture and optimization procedure of Section 5.3.1: a hypothesis class based on Eq. (5.9) and the BFGS method to minimize the objective function. We

utilize the example-dependent cost SVM (EDSVM) proposed in [Brefeld et al., 2003] as reference for comparison.

EDSVM extends the standard SVM formulation by adding example-dependent coefficients for the slack variables:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2_{\mathcal{H}} + \lambda \left(\sum_{k:y^k=1} c_{01}(\mathbf{x}^k)\xi_k^m + \sum_{k:y^k=0} c_{10}(\mathbf{x}^k)\xi_k^m \right) \tag{5.14}$$

$$\text{s.t. } y^k \left(\left\langle \mathbf{w}, \psi(\mathbf{x}^k) \right\rangle + b \right) \geq 1 - \xi_k \text{ and } \xi_k \geq 0 \tag{5.15}$$

For samples with a margin smaller than 1, ξ_k represents how much the example fails to fulfill the margin requirement, and it is then weighted with the cost value of the sample. For $m = 2$ we recover the 2-nom EDSVM. In this case, it is shown in [Brefeld et al., 2003] that the EDSVM converges to the Bayes rule for large training sets. Note that the SVM optimizes the *hinge loss*, which is widely known not to be a proper loss. This implies that SVMs bypass the probability estimation and directly aims at optimizing the classification performance [Buja et al., 2005].

5.5.1 German credit

The German credit dataset consists of 1000 instances, 700 of which correspond to creditworthy applicants and the other remaining to applicants to whom credit should not be extended. Each applicant is described by 24 attributes describing the status of existing accounts, credit history records, loan amount and purpose, employment status and an assortment of personal information. We use the purely numerical version of the dataset, provided by Strathclyde University. We will examine both EDSVM and EDBD on the numerical version and we will use some information from the symbolic attributes to computed the costs (exact quantity and period of the loan). The task consists of classifying customers in bad or good clients. Let us consider a simplified cost policy. Classifying a good credit customer as bad incurs a loss of $c_{10}^k = Q^k \cdot ((1+i)^{t^k} - 1)$. where i represent a quantity related to the interest rate, Q^k

Table 5.1: Total average loss for each method on the German Credit dataset (K$).

EDBD	**EDSVM**
113.67 ± 8.04	114.13 ± 7.36

is the amount of the loan and t^k is the period of the load for sample k. Equivalently, classifying a bad credit customer as good incurs a loss of $c_{01}^k = Q^k$. We chose $i = 0.05$ but any other value behaves similarly. To compare the performance of EDSVM and EDBD, a linear kernel is selected for the first and a linear model is chosen for the second. The training set and the test set contain the same number of samples. The slack penalty λ is fixed using cross validation. Regarding the election of n, we start with a small value and then we increase it progressively, following the philosophy of continuation methods [Allgower and Georg, 1990]. The results displayed in Table 5.1 show that the two algorithms perform very similarly.

However, there are several advantages in using EDBD instead of EDSVM (and related methods). Firstly, EDSVM suffers in scenarios where costs consist of a example-dependent part and a random term (i.e. the costs are noisy). Let us model the additive random term as a Rayleigh distribution ($p(x;\sigma) = x/\sigma^2 \exp(-x^2/2\sigma^2), x \geq 0$) while the example-dependent term remains the same. Table 5.2 shows the effect of σ in the performance of both methods. As σ increases, the performance of the EDSVM degrades since it can not handle test costs which differs from the training ones. Even inserting the costs as features for learning (ED^{2}SVM) does not help. However, EDBD behaves in a more robust manner since it can naturally take into account the test costs in the Bayes decision framework. Note that the mean cost increases with σ because of the non-zero mean of the chosen Rayleigh distribution. In our formulation, we assume that the cost information is available in test (at least partially). If this is not the case then we should also estimate the costs [Zadrozny and Elkan, 2001a].

Secondly, the technique of biased penalties included in the EDSVM optimization

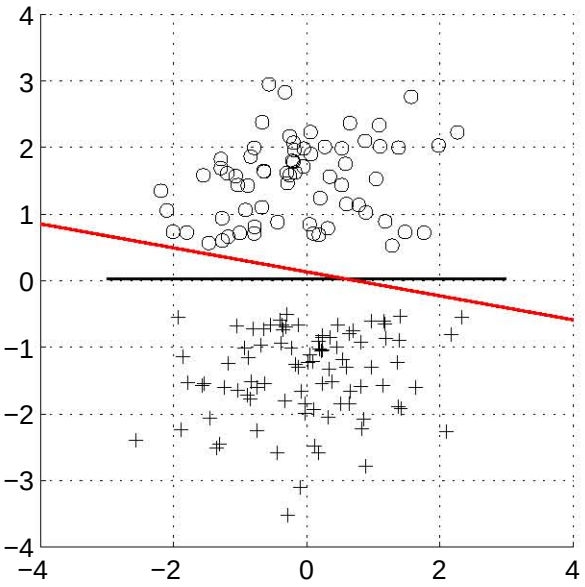


Figure 5.3: Separable training set: EDBD (red line), EDSVM (black line).

leads to an obvious problem. It has a limited ability to enforce cost-sensitive behavior when the training data is separable. Even for large slack penalty λ, the slack variables ξ_k are zero-valued and the optimization degenerates into that of the standard SVM, where the decision boundary is placed just in between the two classes, instead of moving the boundary far from the high-cost examples. This is solved when using a method based on Bayes decision theory. For instance, Figure 5.3 shows an example of a training set where the data are separable. The costs are defined as

$$c^k_{o+} = 1/(1 + e^{-x^k_1})$$

and

$$c^k_{+o} = 1/(1 + e^{x^k_1})$$

where x^k_1 is the abscissa of the sample k.

Note that SVM-based methods for example-independent cost-sensitive learning have been studied in [Karakoulas and Shawe-Taylor, 1999,

Table 5.2: Total average loss for each method on the German Credit dataset with noisty costs (K$).

	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$
EDBD	208.11 ± 10.21	311.25 ± 13.26	427.44 ± 42.74	673.29 ± 75.32	919.51 ± 92.88
EDSVM	219.09 ± 9.65	325.07 ± 10.28	532.69 ± 31.34	789.14 ± 101.98	1144.36 ± 88.97
ED^{2}SVM	231.15 ± 9.20	336.11 ± 10.01	542.93 ± 28.42	769.82 ± 91.26	1141.17 ± 88.31

Masnadi-Shirazi and Vasconcelos, 2010], and their ideas could potentially be extended to a example-dependent cost scenario, even though the boundary movement does not arise in such a natural way as in the Bayes decision framework.

5.6 Summary

Example-dependent cost scenarios are pervasive and generalize standard minimum-error and (class-dependent) cost-sensitive classification. We have explored the application of Bregman divergences to learning in example-dependent cost scenarios. The key idea is to associate each sample in the training set with its very own divergence, which intrinsically reflects the cost structure of the corresponding sample. Then, a global divergence is constructed by averaging the individual divergences. Optimizing this global divergence leads to posterior probability estimates which are specially accurate near the optimal boundary of the example-dependent cost-sensitive classification problem.

This approach naturally exploits the capacity of the learning machine for classification purposes, by focusing that capacity on most interesting area of the posterior probability map. This way, we can enjoy benefits of both classification-based methods (SVMs) and posterior probability estimation methods. As an example of this, experimental results show that the performance of our proposal is similar to that

of an example-dependent cost sensitive SVM when the costs are perfectly specified. However, when the costs are noisy the divergence-based method clearly outperforms the SVM, due to its decision-theoretic roots. The extension of the proposed approach to multiclass problems is straightforward, since Bayes decision theory handles those scenarios gracefully.

Part IV

Conclusions

Chapter 6

Conclusions and future research lines

The last one. This chapter compiles the most relevant conclusions of this work and outlines the future research lines and application fields.

6.1 Conclusions

This dissertation was devoted to the study of cost-sensitive learning in general and cost-sensitive classification in particular. Contrary to most of the algorithms that pursue to minimize the error rate, we focused on minimizing the cost of our decisions, which turns out to be a much more realistic situation. The background and motivation of the Thesis was briefly depicted in Chapter 1.

The main goal of the Thesis was to provide with a wide framework for cost-sensitive classification based on Bayes decision theory. Throughout previous chapters, we addressed the application to cost-sensitive learning of a well-known family of measures, the so-called Bregman divergences. The flexibility that this formulation achieves was proven to be vast: we dealt with a large range of scenarios, covering supervised and semi-supervised learning, binary and multiclass problems, all together with example-dependent and class-dependent costs. The key idea behind this work can be condensed as the establishment of a link between each individual sample or set

of samples in a training set and their very own specific divergence, perfectly adapted to their costs. Thereby, the loss function associated with this divergence intrinsically reflects the structure of the cost information given in the problem. Optimizing the resulting loss function leads to posterior probability estimates that are particularly sensible and accurate near the optimal decision boundaries. This approach makes the most of the natural capacity of learning machines in classification tasks, emphasizing the most relevant areas of the posterior probability map.

The main conclusions we can extract from the Thesis, according to the different contributions are listed below.

- In Chapter 2 we proposed a general procedure to train multiclass classifiers for particular cost-sensitive decision problems, which was based on estimating posterior probabilities using Bregman divergences. We designed a parametric family of Bregman divergences that can be tuned to a specific cost matrix. We showed that the highest curvature region of these divergences varies with $\mathbf{C}$, achieving greater sensitivity in areas close to the decision boundaries. As R grows larger, the sensitivity around the boundary increases, but the loss becomes less well-behaved from a numerical optimization point of view.

- Our asymptotic analysis demonstrated that the optimization problem based on the parametric family of Bregman divergences became equivalent to minimizing the overall cost regret in non-separable problems, and to maximizing a margin in separable problems.

- Additionally, we showed that using the learning algorithm based on Bregman divergences with a simple linear classifier, the error/cost results obtained are comparable to (or better than) those given by the cross-entropy solely or combined with some well-known cost-sensitive algorithms. A major drawback was the optimization stage, as the problem is non-convex in general.

- In Chapter 3 we proposed a general procedure to train multiclass semi-supervised classifiers for particular cost-sensitive decision problems, which was

also based on estimating posterior probabilities using Bregman divergences. We established an optimization problem relying on the empirical risk minimization of a Bregman loss together with what is called Entropy Minimization principle. We linked our work with two well-know semi-supervised approaches: Entropy regularization and Transductive SVM. Under the assumption that inter-class separation is stronger than intra-class separation, the use of unlabeled data to minimize the average entropy is proposed as a multiclass cost-sensitive semi-supervised algorithm (the first one up to our knowledge), with a performance comparable with the state-of-the-art in binary classification tasks.

- Due to the results in supervised and semi-supervised learning for our parametric family of Bregman divergences, in Chapter 4 we decided to broaden the approach in several directions: can we find an inverse link (activation function) in order to obtain a convex optimization problem? To answer the question we presented a some results concerning the canonical link and analyzed some examples. We also established some links with potential functions. The answer to this first question was not entirely positive but allowed us to motivate the need for a further the study of the sequences of Bregman divergences.

- Then, we wanted to address the following two additional questions: is it possible to define other cost-sensitive Bregman divergences that also minimize the total cost in non-separable problems? is there a connection between maximum margin classifiers and Bregman divergences under more general conditions? We derived some results about the identification and characterization of sequences of Bregman divergences that are suitable in the cost-sensitive context. In particular, we found some very general conditions to define sequences whose minimization provides minimum (cost-sensitive) risk for non-separable problems and some type of maximum margin classifiers in separable cases.

- The final generalization involved substituting the once-deterministic cost matrices with example-dependent cost matrices. While previous results had been

based on the assumption that misclassification costs were equal for all samples, in Chapter 5 we extended the cost-sensitive sequences of Bregman divergences to tackle non-deterministic cost matrices and studied their performance in both synthetic and real data. In our proposal, each sample contributed to the objective function depending not only on the distance to the boundary but also through its cost. This chapter was particularly interesting because the example-dependent cost framework has been hardly treated in the literature.

Let us highlight that the proposed method benefits from advantages belonging to classical discriminative classification methods, such as SVMs, as well as those from methods that rely on probability estimation. Along the chapters we showed several examples, namely, the relationship with maximum margin classifiers, the asymptotic results on the minimization of the total cost or the possibility of using the costs in test (when available).

6.2 Future research lines

This very last section of the Thesis compiles the on-going and future research lines.

1. **Improve the cost information:** Throughout the Thesis we realized the importance of what we mentioned right in the beginning of Chapter 1: in non-synthetic problems, the cost is often a non-homogeneous measure, consisting of a mixture of factors. Therefore, evaluating the costs is a non-trivial task and is the main reason why cost information is not widely available in benchmark datasets. For this reason our experiments were restricted to synthetic data and just a few examples of real-world data (UCI). Therefore, we aim to conduct further experiments in real-world datasets with real costs and that means start looking for new datasets and new applications of cost-sensitive learning where this information is available.

   Additionally, given the different costs defined in Chapter 1, we should consider how different types costs, other than, misclassification costs can be taken into

account: we are missing the cost of acquiring data, the cost of labeling data and many others. Make our model complex enough to deal with some of them is a great challenge.

2. **Comparative study of the performance of different sequences of Bregman divergences:** We defined different sequences of Bregman divergences based on [Santos-Rodriguez et al., 2009c, Guerrero-Curieses et al., 2005, Guerrero-Curieses et al., 2004]. Even though they share a common asymptotic behavior, providing minimum cost in non-separable problems a maximum margin in separable data, the study of other properties would be interesting. In cost-insensitive learning, some divergences guarantee a faster convergence rate or might preserve sparsity better than others. We are looking for further properties to decide which divergence we should apply to different scenarios.

3. **Extension to clustering:** In Chapter 3 we proposed the following expectation to be minimized

$$\mathbb{E}\{R_\lambda\} = \mathbb{E}\{h(\mathbf{z})\} + \lambda\mathbb{E}\{(\mathbf{p}_m - \mathbf{z})^T \nabla_{\mathbf{z}} h(\mathbf{z}) | M = 1\}$$

Remember that the role of λ is to adjust the trade-off between labeled and unlabeled data. This parameter poses the possibility of learning from partially labeled data, with clustering as the limit scenario. Clustering deals with finding a structure in a collection of unlabeled data and organize samples into groups whose members are similar in some way [Hastie et al., 2003, Garcia-Garcia and Santos-Rodriguez, 2011, Garcia-Garcia and Santos-Rodriguez, 2009].

A different approach based on the concept of *Bregman information* (Section B.5) was successfully explored for clustering in [Banerjee et al., 2005b]. Mixing cost-sensitive learning and clustering is somehow related to the well-know weighted K-means and similar algorithms [Dhillon et al., 2004], where a natural extension of the K-means problem allows us to include some information,

namely, a set of weights associated with the data points. These might represent a measure of the cost. The intent is that a point with a weight of 5.0 is twice as important as a point with a weight of 2.5, for instance.

4. **Extension to active learning:** The key idea behind active learning [Settles, 2009] is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns. An active learner may pose queries, usually in the form of unlabeled data instances to be labeled by an oracle (e.g., a human annotator). Active learning is well-motivated in many modern machine learning problems where unlabeled data may be abundant or easily obtained, but labels are difficult, time-consuming, or expensive to obtain. There is a clear cost information associated with the value of obtaining labels. The cost of each new query could potentially be introduced in our model.

5. **Connections between cost-sensitive learning, reinforcement learning and on-line learning:** This three topics share some points in common. Reinforcement learning ([Sutton and Barto, 1998, Kaelbling et al., 1996]) is concerned with how an agent ought to take actions in an environment so as to maximize some notion of cumulative reward. Reinforcement learning differs from standard supervised learning in that correct input/output pairs are never presented, nor sub-optimal actions explicitly corrected. Further, there is a focus on on-line performance, which involves finding a balance between exploration (of uncharted territory) and exploitation (of current knowledge). In our case, we deal with costs instead of rewards. Also the example-dependent cost framework presented in Chapter 5 seems to fit well with on-line learning. Our problem could then be formulated as taking decisions so as to minimize a cumulative cost.

6. **Learn costs and posterior probabilities jointly in a high-dimensional space:** Other promising approach consists in unifying the example-dependent

framework in Chapter 5 and [Zadrozny and Elkan, 2001a], where both costs and probabilities in test are unknown. In our work we assume that the cost matrix for each test sample is given. In many problems that could not be the case and we would need estimate the example-specific misclassifications costs as well as the example-specific class probabilities. The task of estimating the costs could be even more important than the probability estimation itself. In a first approach to the problem we can use a vanilla linear regression to estimate the costs as in [Zadrozny and Elkan, 2001a]. An alternative more sophisticated approach would involve Multiple Kernel Learning [Bach et al., 2004] to estimate both costs and probabilities in one step.

7. **Explore the connections with related families of divergences:** [Reid and Williamson, 2011] clarifies the relationship between Bregman divergences and other families of divergences, such as the f-divergences [Ali and Silvey, 1966] and the (f, l)-divergences [Garcia-Garcia et al., 2011] presented in Appendix B. This link opens an unexplored line to make the most of the properties, tools and bounds defined for f-divergences and utilize them for Bregman divergences and cost-sensitive scenarios.

8. **Definition of an extended learning paradigm:** Apart from supervised, semi-supervised and unsupervised learning, regarding the relationship between training and test we can distinguish between induction and transduction. Roughly speaking, inductive inference is concerned with the estimation of a model based on data from the whole problem space and using this model to predict output values for a new input vector, which can be any point in this space. In contrast to the inductive inference, transductive inference methods estimate the value of a potential model only for a specific set of points of the space utilizing additional information related to this set.

   We are interested in a learning setting where three types of data are given: labeled data, unlabeled data and objective data (half-way between semi-

supervised learning and transductive learning). Consider the set $\mathcal{S} = \mathcal{S}_L \cup \mathcal{S}_\mathcal{U} \cup \mathcal{S}_\mathcal{T}$ that consists of the labeled dataset $\mathcal{S}_L = \{\mathbf{x}^k, y^k\}_{k=1}^{K_L}$, the unlabeled dataset $\mathcal{S}_U = \{\mathbf{x}^k\}_{k=K_L+1}^{K_L+K_U}$ and the objective set $\mathcal{S}_T = \{\mathbf{x}^k\}_{k=K_L+K_U+1}^{K}$, with K_L, K_U, $K_T = (K - K_L - K_U)$ samples respectively. Note that we expect the samples in $\mathcal{S}_U$ to be i.i.d but we do not impose the same constraint on the samples in $\mathcal{S}_T$. For instance, based on the Entropy Minimization principle, consider the following empirical risk functional

$$R_{TS}(\mathbf{w}) = \sum_{k=1}^{K_L} \alpha(\mathbf{x}^k, \mathcal{S}_T) L_\phi(y^k, z^k) + \lambda \sum_{k=K_L+1}^{K_L+K_U} \alpha(\mathbf{x}^k, \mathcal{S}_T) L_\phi(z^k, z^k)$$

where α is a function measuring the distance from input examples $\mathbf{x}^k$ to the objective set $\mathcal{S}_T$ and $\lambda \in \mathbb{R}^+$ regulates the trade-off between the labeled and unlabeled terms. Note that the term $L_\phi(z^k, z^k) = -\phi(z^k)$ represents a generalized entropy. Therefore, the unlabeled set $\mathcal{S}_\mathcal{U}$ acts as a regularizer, pushing the solution to low-density regions. The unlabeled objective set $\mathcal{S}_\mathcal{U}$ reinforces the weight of the samples that are close to the targets. Note that the targets are not taken into account in the estimation because they are not forced to be i.i.d.

Part V

Appendix

Appendix A

Some properties of Bregman divergences

In this appendix we present some interesting properties of Bregman divergences. We refer the reader to Appendix A in [Banerjee et al., 2005b] for additional properties.

Let ϕ be a strictly convex and differentiable function, then Bregman divergence between $x, y \in dom\, \phi$ is

$$D_\phi(y, x) = \phi(y) - \phi(x) - (y - x)^T \nabla\phi(x) \tag{A.1}$$

Remember that the Bregman divergence is the vertical distance at y between the graph of ϕ and the tangent to the graph of ϕ in x.

Some useful properties can be easily derived from the definition of Bregman divergence.

Non-negativity

$D_\phi \geq 0$ (the tangent to the epigraph is always below the graph)

Convexity

The divergence is convex in y (due to the convexity of ϕ)

Linearity

Linear in ϕ (by definition)

Invariant to the addition of affine functions

$D_{\phi+b^T x+c} = D_\phi$ (by definition)

Linear separation

$\{x|D_\phi(x,u) = D_\phi(x,v)\}$ is a hyperplane.

Proof

$$D_\phi(x,u) = D_\phi(x,v)$$

$$\phi(x) - \phi(u) - (x-u)^T\nabla\phi(u) = \phi(x) - \phi(v) - (x-v)^T\nabla\phi(v)$$

$$x^T(\nabla\phi(u) - \nabla\phi(v)) - [u^T\nabla\phi(u) - v^T\nabla\phi(v) - \phi(u) - \phi(v)] = 0$$

This last equation defines a hyperplane.

Expected value of a Bregman divergence

$\arg\min_u \mathbb{E}_p[D_\phi(X,u)] = \mathbb{E}_p[X] \equiv \mu$ for any probability distribution p over X.

Proof Denote $J(u) = \mathbb{E}_p[D_\phi(X,u)]$, then,

$$\begin{aligned}
J(u) - J(v) &= \sum_x p(x)D_\phi(x,u) - \sum_x p(x)D_\phi(x,\mu) \\
&= \sum_x p(x)(\phi(x) - \phi(u) - (x-u)^T\nabla\phi(u) - \phi(x) - \phi(v) - (x-v)^T\nabla\phi(v)) \\
&= \phi(\mu) - \phi(u) - \left(\sum_x p(x)x - u\right)^T \nabla\phi(u) - \left(\sum_x p(x)x - \mu\right)^T \nabla\phi(\mu) \\
&= \phi(\mu) - \phi(u) - (\mu-u)^T\nabla\phi(u) \\
&= D_\phi(u,\mu)
\end{aligned}$$

Convex duality (1)

Let $\phi^*(\theta) = sup_u \left[\theta \cdot u - \phi(u)\right]$ be the Legendre conjugate of $\phi(u)$. Then $D_\phi(u_1, u_2) = D_{\phi^*}(\theta_1, \theta_2)$

Proof

$$\begin{aligned} D_\phi(u_1, u_2) &= \phi(u_1) - \phi(u_2) - (u_1 - u_2)^T \nabla\phi(u_2) \\ &= \phi(u_1) - \phi(u_2) - (u_1 - u_2)^T \theta_2 + (\mu_1^T \theta_1 - \mu_1^T \theta_1) \\ &= [-\mu_1^T \theta_1 + \phi(\mu_1)] + [\mu_2^T \theta_2 - \phi(\mu_2)] - \mu_1^T \theta_2 + \mu_1^T \theta_1 \\ &= -\phi^*(\theta_1) + \phi^*(\theta_2) - \mu_1^T(\theta_2 - \theta_1) \\ &= D_{\phi^*}(\theta_2, \theta_1) \end{aligned}$$

Convex duality (2)

$D_\phi(u, v) = D_{\phi^*}(\nabla\phi(v), \nabla\phi(u))$ (consequence of convex duality (1) and $\nabla\phi^* = (\nabla\phi)^{-1}$)

Generalized Pythagorean Theorem

$D_\phi(u, v) \geq D_\phi(u, z) + D_\phi(z, v) + (u - z)^T(\nabla\phi(z) - \nabla\phi(v))$ (by definition)

Appendix B

f and (f, l)-divergences

In this appendix we describe other well-know families of divergences that are closely related to Bregman divergences, such as f-divergences [Csiszár, 1967] and the (f, l)-divergences we derived in [Garcia-Garcia et al., 2011]. To do so we first introduce the standard notation for these divergences, some definitions and properties.

B.1 Notation and definitions

Let P, Q be a pair of probability distributions, and M their convex combination $M := \pi P + (1 - \pi)Q$ for $\pi \in [0, 1]$. Given a classification task (π, P, Q) whose goal is to assign labels $Y = 1$ to points coming from P and $Y = 0$ to points from Q, we denote by $\eta = P(Y = 1|X = x)$ and $\hat{\eta}$ the posterior class probability and its estimate, respectively. The representations (π, P, Q) and (η, M) are interchangeable.

We write $\mathbb{E}_P[f]$ for the expectation of a function $f(x)$ of a random variable $x \sim P$. Let l be a loss function $l : \{0, 1\} \times [0, 1] \to \mathbb{R}$. The point-wise risk L_l associated to l is given by $L_l(\eta(x), \hat{\eta}(x)) = \eta(x)l(1, \hat{\eta}(x)) + (1 - \eta(x))l(0, \hat{\eta}(x))$, and the (expected) risk $\mathbb{L}_l$ is thus $\mathbb{L}_l(\eta, M) = \mathbb{E}_M[L_l(\eta(x), \hat{\eta}(x))]$. Optimal or *Bayes* risks are denoted by an underline, so $\underline{L}_l(\eta(x)) = \inf_{\hat{\eta}(x)} L_l(\eta(x), \hat{\eta}(x))$ and $\underline{\mathbb{L}}_l(\pi, P, Q) = \underline{\mathbb{L}}_l(\eta, M) = \mathbb{E}_M[\underline{L}_l(\eta(x))]$. The *prior* Bayes risk is the optimal risk when only the prior class

probability π is known $\underline{\mathbb{L}}_l(\pi) = \underline{L}_l(\pi)$.

B.2 f-divergences

In this section we recapitulate definitions and known facts about f-divergences. Given a convex function $f : (0, \infty) \to \mathbb{R}$, with $f(1) = 0$, the corresponding f-divergence [Ali and Silvey, 1966] between two probability distributions P, Q over an input space $\mathcal{X}$ is defined as

$$\mathbb{I}_f(P, Q) = \mathbb{E}_Q \left[f\left(\frac{dP}{dQ}\right)\right] = \int_{\mathcal{X}} dQ f\left(\frac{dP}{dQ}\right),$$

if P is absolutely continuous with respect to Q, and ∞ otherwise. Many well-known divergences can be cast into this framework by adequately choosing the generating function f. Some important examples include the variational, Kullback-Leibler (KL) and Pearson's χ^2 divergences.

Our discussion will be based mainly on a classical result (see e.g. [Österreicher and Vajda, 1993]) that shows how f-divergences can be represented by a weighted integral of *statistical informations* $\Delta\underline{\mathbb{L}}_{0-1}(\pi, P, Q)$ under the 0-1 loss. These informations can be intuitively interpreted as the risk reduction provided by the knowledge of the exact posterior probability η instead of just the prior probability π. They are defined as

$$\begin{aligned}\Delta\underline{\mathbb{L}}_{0-1}(\pi, P, Q) &= \underline{\mathbb{L}}_{0-1}(\pi) - \underline{\mathbb{L}}_{0-1}(\pi, P, Q) \\ &= \min(\pi, 1-\pi) - \underline{\mathbb{L}}_{0-1}(\pi, P, Q).\end{aligned}$$

The integral representation of f-divergences is given by

$$\mathbb{I}_f(P, Q) = \int_0^1 \Delta\underline{\mathbb{L}}_{0-1}(\pi, P, Q)\gamma_f(\pi)d\pi, \tag{B.1}$$

where the weight function $\gamma_f(\pi)$ is related to the curvature of the function f defining the divergence

$$\gamma_f(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right). \tag{B.2}$$

Since f is a convex function, the weights $\gamma_f(\pi)$ are non-negative. For a comprehensive list of well-known f-divergences and their associated f and weight functions please refer to [Reid and Williamson, 2011].

Symbol	$\gamma(\pi)$	$f(t)$	Name
$V(P,Q)$	$4\delta(\pi - \frac{1}{2})$	$\lvert t-1\rvert$	Variational Divergence
$\Delta(P,Q)$	8	$(t-1)^2/(t+1)$	Triangular Discrimination
$\mathrm{KL}(P,Q)$	$\frac{1}{\pi^2(1-\pi)}$	$t\ln t$	Kullback-Leibler Divergence
$I(P,Q)$	$\frac{1}{2\pi(1-\pi)}$	$\frac{t}{2}\ln t - \frac{t+1}{2}\ln(t+1) + \ln 2$	Jensen-Shannon Divergence
$J(P,Q)$	$\frac{1}{\pi^2(1-\pi)^2}$	$(t-1)\ln t$	Jeffreys Divergence
$\chi^2(P,Q)$	$\frac{2}{\pi^3}$	$(t-1)^2$	Pearson Chi Squared Divergence
$h^2(P,Q)$	$\frac{1}{2[\pi(1-\pi)]^{\frac{3}{2}}}$	$(\sqrt{t}-1)^2$	Hellinger Divergence

Table B.1: Some well-known f-divergences with their associated weights. Extracted from [Reid and Williamson, 2009b].

B.3 (f, l)-divergences

In [Garcia-Garcia et al., 2011] we propose a risk-based generalization of the family of f-divergences, based on the integral representation in Eq. B.1. The main idea is to substitute the 0-1 loss for an arbitrary loss function l. This way, we can express this new generalization as follows.

Definition For a convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, and a loss $l : \{0, 1\} \times [0, 1] \to \mathbb{R}_+$, we define the corresponding (f, l)-*divergence* $\mathbb{I}_{f,l}$ as

$$\mathbb{I}_{f,l} = \int_0^1 \Delta\underline{\mathbb{L}}_l(\pi, P, Q)\gamma_f(\pi)d\pi, \tag{B.3}$$

where $\gamma_f(\pi)$ is given by Eq. B.2 and

$$\Delta\underline{\mathbb{L}}_l(\pi, P, Q) = \underline{\mathbb{L}}_l(\pi) - \underline{\mathbb{L}}_l(\pi, P, Q). \tag{B.4}$$

Obviously, the original f-divergences can be obtained as a particular case of (f, l)-divergences by setting $l = l_{0-1}$. Note that the idea of substituting 0-1 for more general losses is at the core of almost every practical classifier. This is the idea of *surrogate losses* [Bartlett et al., 2006]: Since the 0-1 loss is not very well behaved and thus hard to handle, most learning algorithms use, explicitly or implicitly, other kind of losses that approximate the 0-1 loss while being much more amenable to theoretical analysis and numerical optimization. These surrogates are almost always[1] proper losses whose second term is mapped from $[0, 1]$ to $\mathbb{R}$. Thus, if the goal is to define divergences that can be nicely estimated using classification risks it is very natural to work with surrogate/proper losses, since they are what most practical classifiers optimize.

B.3.1 Some properties of (f, l)-divergences

In this section we will study how we can get interesting properties for (f, l)-divergences by adequately choosing the loss l. We will implicitly assume all losses to

[1] The most important exception being the hinge loss

be proper.

As we will show in Sec. B.4, (f, l) and f-divergences are deeply connected, so it is natural to recover most properties of standard f-divergences with a sensible selection of the loss function l. For an overview of the most important properties of f-divergences, please refer to [Österreicher, 2002]. We show in the form of short theorems a small representative selection of such properties, along with the conditions that the losses must satisfy in order for those properties to hold. We sketch the proofs, which are quite straight-forward.

Theorem B.3.1 (Non-negativity and identity of indiscernibles) *For any convex f and any proper loss l, $\mathbb{I}_{f,l}(P, Q) \geq 0$ for all P, Q. Moreover, if f is non-trivial ($\exists \pi \in (0, 1) \mid \gamma_f(\pi) > 0$) and l is such that $\underline{L}_l$ is strictly concave, then equality holds iff $P = Q$.*

This theorem can be easily proved by applying Jensen's inequality, noting that point-wise Bayes risks $\underline{L}_l$ induced by proper losses are always concave [Savage, 1971]. It is easy to check that most common proper losses, such as square or log-losses, induce strictly concave point-wise Bayes risks $\underline{L}_l$, so the condition is not very restrictive.

Theorem B.3.2 (Symmetry) *If l is a proper loss such that $l(0, \hat{\eta}) = l(1, 1 - \hat{\eta})$, then $\mathbb{I}_{f,l}(P, Q) = \mathbb{I}_{f,l}(Q, P)$ if $f(t) = f^*(t) + c(t-1)$, $c \in \mathbb{R}$, where f^* is the Csiszar's dual (or $*$-conjugate) of function f.*

This is analogous to the standard symmetry property of f-divergences. The proof uses the fact that the condition on f implies $\gamma_f(\pi) = \gamma_f(1 - \pi)$, and then it mainly involves showing that $\Delta\underline{\mathbb{L}}_l(\pi, P, Q) = \Delta\underline{\mathbb{L}}_l(1 - \pi, Q, P)$ for $\pi \in [0, 1]$. Once again, standard losses satisfy the simple and natural condition imposed on l for the symmetry property to hold.

Theorem B.3.3 (Information Processing) $\mathbb{I}_{f,l}(P, Q) \geq \mathbb{I}_{f,l}(\Phi(P), \Phi(Q))$, *where Φ is any transformation.*

This is also analogous to a standard f-divergences property. The proof relies on the non-decreasing property of Bayes risks under arbitrary transformations.

B.4 Connecting f and (f, l)-divergences

In this section we show how some (f, l)-divergences are equivalent to standard f-divergences via a transformation of the weight function depending on the loss l. This will provide insight into the effect of using a surrogate loss for divergence definition, as well as motivating surprising ways of estimating some well-known divergences.

The discussion is based on the one-to-one relationship between statistical informations and f-divergences, as stated in the following classical result

Theorem B.4.1 ([Österreicher and Vajda, 1993], Th. 2)
Given an arbitrary loss l, then defining

$$f_l^\pi(t) = \underline{L}_l(\pi) - (\pi t + 1 - \pi)\underline{L}_l\left(\frac{\pi t}{\pi t + 1 - \pi}\right) \tag{B.5}$$

for $\pi \in [0, 1]$ implies f_l^π is convex and $f_l^\pi(1) = 0$, and

$$\Delta\underline{\mathbb{L}}_l(\pi, P, Q) = \mathbb{I}_{f_l^\pi}(P, Q) \tag{B.6}$$

for all distributions P and Q.

This may seem at odds with the result in [Nguyen et al., 2009] which establish a many-to-one relationship between losses and f-divergences. However, note that in that work they are concerned with margin classification losses, while here we work with proper losses. The many link functions that can be coupled with a given proper loss to yield classification losses introduce that extra degree of freedom [Reid and Williamson, 2011].

Exploiting this representation of statistical information for arbitrary losses, Eq. B.3 can be rewritten as $\mathbb{I}_{f,l} = \int_0^1 \mathbb{I}_{f_l^\pi}(P, Q)\gamma_f(\pi)d\pi$. Now we can leverage the weighted integral representation of $\mathbb{I}_{f_l^\pi}$ as given by Eq. B.1, yielding

$$\begin{aligned}
\mathbb{I}_{f,l} &= \int_0^1 \left(\int_0^1 \Delta\underline{\mathbb{L}}_{0-1}(\pi', P, Q)\varphi_{l,\pi}(\pi')d\pi'\right)\gamma_f(\pi)d\pi \\
&= \int_0^1 \Delta\underline{\mathbb{L}}_{0-1}(\pi', P, Q)\left(\int_0^1 \varphi_{l,\pi}(\pi')\gamma_f(\pi)d\pi\right)d\pi' \\
&= \int_0^1 \Delta\underline{\mathbb{L}}_{0-1}(\pi, P, Q)\gamma_{f,l}(\pi)d\pi,
\end{aligned} \tag{B.7}$$

where $\varphi_{l,\pi}(\pi')$ is the weight function corresponding to f_l^π, as given by Eq. B.2

$$\varphi_{l,\pi}(\pi') = \frac{1}{\pi^3} f_l^{\pi\prime\prime}\left(\frac{1-\pi}{\pi}\right). \tag{B.8}$$

So we get the following theorem.

Theorem B.4.2 *Assume a (f,l)-divergence with weight function $\gamma_f(\pi)$ and loss function l. Let $\varphi_{l,\pi}$ be given by Eq. B.8. Whenever*

$$\gamma_{f,l}(\pi) = (T_l\gamma_f)(\pi) = \int_0^1 \varphi_l(\pi,\pi')\gamma_f(\pi')d\pi'$$

converges, then that (f,l)-divergence is equivalent to a standard f-divergence with weight function $\gamma_{f,l}(\pi)$.

In this case, both divergences are intrinsically the same one, but expressed on different bases. The relationships between the weight functions is given by a linear operator T_l with kernel $\varphi_l(\pi,\pi') \equiv \varphi_{l,\pi}(\pi')$. This connection has the important effect of allowing the estimation of standard f-divergences by using statistical informations under adequate proper/surrogate losses.

Note that [Reid and Williamson, 2011] connect losses and f-divergences by associating a loss l with a divergence with $f = f_l^{\frac{1}{2}}$ (see Th. B.4.1). That can be seen to be a particular case of (f,l)-divergences when f is chosen to represent the *variational divergence* V, since $\gamma_V \propto \delta(\pi - \frac{1}{2})$.

B.5 Link with Bregman divergences

We refer the reader to [Reid and Williamson, 2011], where Reid and Williamson unify f-divergences, Bregman divergences, surrogate regret bounds, proper scoring rules, cost curves, ROC-curves and statistical information.

[Banerjee et al., 2005a] introduced the notion of the *Bregman information* $\mathbb{B}_\phi(\mathcal{S})$ of a random variable $\mathcal{S}$ drawn according to some distribution σ over $\mathcal{S}$. It is the minimal σ-average Bregman divergence that can be achieved by an element $s^* \in \mathcal{S}$ (the Bregman representative).

Th. 10 in [Reid and Williamson, 2011] says that for each choice of π, the classes of f-divergences $\mathbb{I}_f$, statistical informations $\Delta\underline{\mathbb{L}}$ and (discriminative) Bregman informations $\mathbb{B}_\phi$ can all be defined in terms of the Jensen gap of some convex function. Additionally, there is a bijection between each of these classes due to the mapping $\lambda_\pi(c) := \frac{1-\pi}{\pi}\frac{c}{1-c}$, $c \in [0, 1)$ that identifies likelihood ratios with posterior probabilities.

Bibliography

[Abe et al., 2004] Abe, N., Zadrozny, B., and Langford, J. (2004). An iterative method for multi-class cost-sensitive learning. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 3–11, New York, NY, USA. ACM.

[Ali and Silvey, 1966] Ali, S. and Silvey, S. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society*, 28:131–142.

[Allgower and Georg, 1990] Allgower, E. and Georg, K. (1990). *Numerical continuation methods: an introduction*. Springer-Verlag New York, Inc., New York, NY, USA.

[Bach et al., 2006] Bach, F., Heckerman, D., and Horvitz, E. (2006). Considering cost asymmetry in learning classifiers. *Journal of Machine Learning Research*, 7:1713–1741.

[Bach et al., 2004] Bach, F. R., Lanckriet, G. R. G., and Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st international conference on Machine learning*, volume 69, pages 6+, New York, NY, USA. ACM.

[Banerjee et al., 2005a] Banerjee, A., Guo, X., and Wang, H. (2005a). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669.

[Banerjee et al., 2005b] Banerjee, A., Merugu, S., Dhillon, I., and Ghosh, J. (2005b). Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749.

[Bartlett et al., 2006] Bartlett, P., Jordan, M., and McAuliffe, J. (2006). Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101:138–156.

[Bennett and Demiriz, 1999] Bennett, K. and Demiriz, A. (1999). Semi-supervised support vector machines. In *Advances in neural information processing systems (NIPS) 12*. The MIT Press.

[Bradford et al., 1998] Bradford, J. P., Kunz, C., Kohavi, R., Brunk, C., and Brodley, C. E. (1998). Pruning decision trees with misclassification costs. In *Proceedings of the European Conference on Machine Learning*, pages 131–136. Springer Verlag.

[Brefeld et al., 2003] Brefeld, U., Geibel, P., and Wysotzki, F. (2003). Support vector machines with example dependent costs. In *Proceedings of the European Conference on Machine Learning*, pages 23–34. Springer.

[Bregman, 1967] Bregman, L. (1967). The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(10):200–217.

[Buja et al., 2005] Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, Department of of Statistics, University of Pennsylvania.

[Chapelle et al., 2006] Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

[Cid-Sueiro et al., 1999] Cid-Sueiro, J., Arribas, J., Urban-Muñoz, S., and Figueiras-Vidal, A. (1999). Cost functions to estimate a posteriori probabilities in multi-class problems. *IEEE Transactions on Neural Networks*, 10(3):645–656.

[Cid-Sueiro and Figueiras-Vidal, 2001] Cid-Sueiro, J. and Figueiras-Vidal, A. (2001). On the structure of strict sense bayesian cost functions and its applications. *IEEE Transactions on Neural Networks*, 12(3):445–455.

[Cristianini and Shawe-Taylor, 2000] Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines : and other kernel-based learning methods*. Cambridge University Press, 1st edition.

[Csiszár, 1967] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:29–318.

[Davenport et al., 2006] Davenport, M. A., Baraniuk, R. G., and Scott, C. D. (2006). Controlling false alarms with support vector machines. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 589–592, Toulouse, France.

[Devroye et al., 1996] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

[Dhillon et al., 2004] Dhillon, I. S., Guan, Y., and Kulis, B. (2004). Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 551–556. ACM.

[Dmochowski et al., 2010] Dmochowski, J., Sajda, P., and Parra, L. (2010). Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. *Journal of Machine Learning Research*, 11:3313–3332.

[Domingos, 1999] Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 155–164. ACM Press.

[Elkan, 2001a] Elkan, C. (2001a). The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978.

[Elkan, 2001b] Elkan, C. (2001b). Magical thinking in data mining: lessons from coil challenge 2000. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–431, New York, NY, USA. ACM.

[Fan et al., 1999] Fan, W., Stolfo, S. J., Zhang, J., and Chan, P. K. (1999). Adacost: misclassification cost-sensitive boosting. In *Proceedings of the 16th International Conference on Machine Learning*, pages 97–105. Morgan Kaufmann.

[Freund and Schapire, 1995] Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, pages 23–37, London, UK. Springer-Verlag.

[Garcia-Garcia and Santos-Rodriguez, 2009] Garcia-Garcia, D. and Santos-Rodriguez, R. (2009). Spectral clustering and feature selection for microarray data. In *Proceedings of the 2009 International Conference on Machine Learning and Applications*, ICMLA '09, pages 425–428, Washington, DC, USA. IEEE Computer Society.

[Garcia-Garcia and Santos-Rodriguez, 2011] Garcia-Garcia, D. and Santos-Rodriguez, R. (2011). Sphere packing for clustering sets of vectors in feature space. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP'11, pages 740–744. IEEE SP Society.

[Garcia-Garcia et al., 2011] Garcia-Garcia, D., von Luxburg, U., and Santos-Rodriguez, R. (2011). Risk-based generalizations of f-divergences. In *Proceedings of the 28th Internacional Conference on Machine Learning*.

[Gneiting and Raftery, 2007] Gneiting, T. and Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

[Grandvalet and Bengio, 2004] Grandvalet, Y. and Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *Advances in neural information processing systems (NIPS) 17*, 17:529–536.

[Guerrero-Curieses et al., 2005] Guerrero-Curieses, A., Alaiz-Rodriguez, R., and Cid-Sueiro, J. (2005). Loss function to combine learning and decision in multiclass problems. *Neurocomputing*, 69:3–17.

[Guerrero-Curieses et al., 2004] Guerrero-Curieses, A., Alaiz-Rodriguez, R., Cid-Sueiro, J., and Figueiras, A. (2004). Local estimation of posterior class probabilities to minimize classification errors. *IEEE Transactions on Neural Networks*, 15(2):309–317.

[Hastie et al., 2003] Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). *The Elements of Statistical Learning*. Springer, corrected edition.

[Jaakkola et al., 1999] Jaakkola, T., Meila, M., and Jebara, T. (1999). Maximum entropy discrimination. In *Advances in Neural Information Processing Systems (NIPS) 12*, pages 470–476. MIT Press.

[Ji and Carin, 2007] Ji, S. and Carin, L. (2007). Cost-sensitive feature acquisition and classification. *Pattern Recognition*, 40(5):1474–1485.

[Joachims, 1999] Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *16th Internacional Conference on Machine Learning*, pages 200–209. Morgan Kaufmann publishers.

[Kaelbling et al., 1996] Kaelbling, L., Littman, M., and Moore, A. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285.

[Kapur and Kesavan, 1993] Kapur, J. and Kesavan, H. (1993). *Entropy Optimization Principles with Applications*. Academic Press, San Diego, CA.

[Karakoulas and Shawe-Taylor, 1999] Karakoulas, G. and Shawe-Taylor, J. (1999). Optimizing classifers for imbalanced training sets. In *NIPS*, pages 253–259.

[Kukar and Kononenko, 1998] Kukar, M. Z. and Kononenko, I. (1998). Cost-sensitive learning with neural networks. In *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI-98)*, pages 445–449. John Wiley and Sons.

[Lee et al., 2006] Lee, C., Jiao, F., Wang, S., Schuurmans, D., and Greiner, R. (2006). Learning to model spatial dependency: Semi-supervised discriminative random fields. In *Advances in Neural Information Processing Systems (NIPS) 19*.

[Lenarcik and Piasta, 1998] Lenarcik, A. and Piasta, Z. (1998). Rough classifiers sensitive to costs varying from object to object. In *Proceedings of the First International Conference on Rough Sets and Current Trends in Computing*, pages 222–230. Springer-Verlag.

[Li et al., 2010] Li, Y., Kwok, J., and Zhou, Z. (2010). Cost-sensitive semi-supervised support vector machine. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 500–505.

[Ling and Sheng, 2008] Ling, C. and Sheng, V. (2008). Cost-sensitive learning and the class imbalance problem. *Encyclopedia of Machine Learning*.

[Little and Rubin, 1987] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, New York, 1st edition.

[Liu and Zhou, 2006] Liu, X. and Zhou, Z. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans. on Knowl. and Data Eng.*, 18(1):63–77.

[Lozano and Abe, 2008] Lozano, A. C. and Abe, N. (2008). Multi-class cost-sensitive boosting with p-norm loss functions. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 506–514, New York, NY, USA. ACM.

[Marrocco and Tortorella, 2004] Marrocco, C. and Tortorella, F. (2004). A cost-sensitive paradigm for multiclass to binary decomposition schemes. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 753–761.

[Masnadi-Shirazi and Vasconcelos, 2010] Masnadi-Shirazi, H. and Vasconcelos, N. (2010). Risk minimization, probability elicitation, and cost-sensitive SVMs. In Fürnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning*, pages 759–766. Omnipress.

[McCullagh and Nelder, 1989] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, London.

[Miller et al., 1991] Miller, J., Goodman, R., and Smyth, P. (1991). Objective functions for probability estimation. In *Proceedings of the International Conference on Neural Networks*, volume 1, pages 881–886.

[Miller et al., 1993] Miller, J., Goodman, R., and Smyth, P. (1993). On loss functions which minimize to conditional expected values andposterior probabilities. *IEEE Transactions on Information Theory*, 39(4):1404–1408.

[Mora-Jimenez and Cid-Sueiro, 2005] Mora-Jimenez, I. and Cid-Sueiro, J. (2005). A universal learning rule that minimizes well-formed cost functions. *IEEE Transactions on Neural Networks*, 16(4):810–820.

[Nguyen et al., 2009] Nguyen, X., Wainwright, M., and Jordan, M. (2009). On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904.

[Nock and Nielsen, 2009] Nock, R. and Nielsen, F. (2009). Bregman divergences and surrogates for learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:2048–2059.

[O'Brien and Gray, 2005] O'Brien, D. and Gray, R. M. (2005). Improving classification performance by exploring the role of cost matrices in partitioning the estimated class probability space. In *Proceedings of the ICML Workshop on ROC Analysis*, pages 79–86.

[O'Brien et al., 2008] O'Brien, D. B., Gupta, M. R., and Gray, R. M. (2008). Cost-sensitive multi-class classification from probability estimates. In *Proceedings of the 25th international conference on Machine learning*, pages 712–719, New York, NY, USA. ACM.

[Platt, 1999] Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

[Provost and Fawcett, 2001] Provost, F. and Fawcett, T. (2001). Robust classification systems for imprecise environments. *Machine Learning*, 42(3):203–231.

[Qin et al., 2010] Qin, Z., Zhang, C., Wang, T., and Zhang, S. (2010). Cost sensitive classification in data mining. In *Proceedings of the 6th international conference on Advanced data mining and applications: Part I*, ADMA'10, pages 1–11, Berlin, Heidelberg. Springer-Verlag.

[Reid and Williamson, 2009a] Reid, M. and Williamson, R. (2009a). Surrogate regret bounds for proper losses. In *Proceedings of the 26th International Conference on Machine Learning*, pages 897–904. ACM.

[Reid and Williamson, 2011] Reid, M. and Williamson, R. C. (2011). Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817.

[Reid and Williamson, 2009b] Reid, M. D. and Williamson, R. C. (2009b). Information, divergence and risk for binary experiments. arXiv:0901.0356v1 [stat.ML].

[Saitta and Lavrač, 2001] Saitta, L. and Lavrač, N., editors (2001). *Machine learning: a technological roadmap*. University of Amsterdam.

[Santos-Rodriguez et al., 2011a] Santos-Rodriguez, R., Cid-Sueiro, J., and Shawe-Taylor, J. (2011a). Cost-sensitive sequences of bregman losses. *submitted to IEEE Transactions on Neural Networks*.

[Santos-Rodriguez and Garcia-Garcia, 2010] Santos-Rodriguez, R. and Garcia-Garcia, D. (2010). Cost-sensitive feature selection based on the set covering machine. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pages 740–746. IEEE Computer Society.

[Santos-Rodriguez et al., 2009a] Santos-Rodriguez, R., Garcia-Garcia, D., and Cid-Sueiro, J. (2009a). Cost-sensitive classification based on bregman divergences for medical diagnosis. In *Proceedings of the 2009 International Conference on Machine Learning and Applications*, ICMLA '09, pages 551–556, Washington, DC, USA. IEEE Computer Society.

[Santos-Rodriguez et al., 2011b] Santos-Rodriguez, R., Garcia-Garcia, D., and Cid-Sueiro, J. (2011b). Classification focused probability estimation for example-dependent cost scenarios. *submitted to Advances in neural information processing systems (NIPS) 25*.

[Santos-Rodriguez et al., 2009b] Santos-Rodriguez, R., Guerrero-Curieses, A., Alaiz-Rodriguez, R., and Cid-Sueiro, J. (2009b). Cost-sensitive learning based on bregman divergences. In *Proceedings of the European Conference on Machine*

Learning and Knowledge Discovery in Databases: Part I, ECML PKDD '09, pages 12–12. Springer-Verlag.

[Santos-Rodriguez et al., 2009c] Santos-Rodriguez, R., Guerrero-Curieses, A., Alaiz-Rodriguez, R., and Cid-Sueiro, J. (2009c). Cost-sensitive learning based on bregman divergences. *Machine Learning*, 76:271–285.

[Savage, 1971] Savage, L. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

[Scott, 2011] Scott, C. (2011). Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In *Proceedings of the 28th International Conference on Machine Learning*.

[Seeger, 2001] Seeger, M. (2001). Learning with labeled and unlabeled data. Technical report.

[Settles, 2009] Settles, B. (2009). Active Learning Literature Survey. Technical Report 1648, University of Wisconsin–Madison.

[Settles et al., 2008] Settles, B., Craven, M., and Friedland, L. (2008). Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, pages 1069–1078.

[Shirazi and Vasconcelos, 2008] Shirazi, H. M. and Vasconcelos, N. (2008). On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost. In *Advances in neural information processing systems (NIPS) 17*, pages 1049–1056.

[Sutton and Barto, 1998] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press.

[Turney, 1995] Turney, P. (1995). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *J. Artif. Intell. Res. (JAIR)*, 2:369–409.

[Turney, 2000] Turney, P. (2000). Types of cost in inductive concept learning. In *Proceedings of the Cost-Sensitive Learning Workshop at the 17th ICML-2000 Conference*.

[Vapnik, 1998] Vapnik, V. (1998). *Statistical learning theory*. Wiley-Interscience.

[Weston et al., 2006] Weston, J., Collobert, R., Bottou, L., and Vapnik, V. (2006). Inference with the universum. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 1009–1016. ACM Press.

[Wu and Srihari, 2003] Wu, X. and Srihari, R. K. (2003). New ν-support vector machines and their sequential minimal optimization. In *Proceedings of the 20th International Conference on Machine Learning*, pages 824–831.

[Yang and Wu, 2006] Yang, Q. and Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(4):597–604.

[Zadrozny and Elkan, 2001a] Zadrozny, B. and Elkan, C. (2001a). Learning and making decisions when costs and probabilities are both unknown. In *KDD '01: Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–213. ACM.

[Zadrozny and Elkan, 2001b] Zadrozny, B. and Elkan, C. (2001b). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 609–616, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Zadrozny and Elkan, 2002] Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *KDD '02: Proceedings*

of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 694–699, New York, NY, USA. ACM.

[Zadrozny et al., 2003] Zadrozny, B., Langford, J., and Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *ICDM '03: Proc. of the 3rd IEEE Int. Conf. on Data Mining*, page 435, Washington, DC, USA. IEEE Computer Society.

[Zhang, 2003] Zhang, T. (2003). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–134.

[Zhou and Liu, 2010] Zhou, Z. and Liu, X. (2010). On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257.

[Zhu, 2005a] Zhu, X. (2005a). Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

[Zhu, 2005b] Zhu, X. (2005b). *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA. AAI3179046.

[Österreicher, 2002] Österreicher, F. (2002). Csiszar's f-divergences-basic properties. Research report, Institute of Mathematics, University of Salzburg, Austria.

[Österreicher and Vajda, 1993] Österreicher, F. and Vajda, I. (1993). Statistical information and discrimination. *IEEE Transactions on Information Theory*, 39(3):1036–1039.