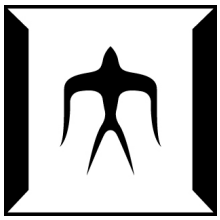
論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Learning under Class-Balance Change: Distribution Matching via Direct Divergence Estimation
著者(和文)	MARTHINUSC.DUPLES
Author(English)	Marthinus Christoffel du Plessis
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9558号, 授与年月日:2014年3月26日, 学位の種別:課程博士, 審査員:杉山　将,佐藤　泰介,徳永　健伸,村田　剛志,瀬々　潤
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9558号, Conferred date:2014/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Learning under Class-Balance Change: Distribution Matching via Direct Divergence Estimation

Marthinus Christoffel du Plessis

January 2014



Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

Thesis Committee:
Masashi Sugiyama, Chair
Taisuke Sato
Takenobu Tokunaga
Tsuyoshi Murata
Jun Sese

Submitted in partial fulfillment of
the requirements for the degree of

Doctor of Engineering

Keywords: Distribution matching; Pearson divergence; class-prior change; Learning from positive and unlabeled data

For my parents.

Abstract

In most machine learning algorithms, it is assumed that the training and target environment are the same and that the supervisor (teacher) assigns labels to all training samples. In many real-world datasets, these assumptions are however violated due to a changing environment or imperfect supervision.

Many of these situations in the classification scenario can be characterized as a change in class balance. In this thesis, three such problems are considered: classification under class-balance change, labeling of unsupervised datasets differing by class balance, and classification of partially labeled data.

Training a classifier on labeled training data and then applying it on target data with a different class prior may cause an excess misclassification rate. This can, however, be corrected for by reweighting training samples with the class prior of the target environment. In practice, however, the target class priors are unknown. We show that these class priors can be estimated in a semi-supervised setup by matching distributions. Moreover, this distribution matching can be performed efficiently and directly without resorting to density estimation. Empirically we show that the proposed method obtains an accurate estimate of the class priors, leading to a lower misclassification rate.

Secondly, the problem of labeling of a dataset without any supervisory information is considered. We show that, if two unlabeled datasets differing by class prior are available, labeling can be performed. This labeling is performed by directly estimating the sign of the density difference between the two datasets.

Finally, the problem of classification using positive only labeled data is considered. In this problem only some samples from a single class is labeled. It has previously been shown that a classifier can be trained from positive and unlabeled data if the class prior is known. We show that this class prior can be estimated via

partially matching distributions.

We conclude that many different learning problems characterized by a change in class balance can be formulated as distribution matching problems. Furthermore, by selecting an appropriate divergence and directly estimating it, practical algorithms can be obtained that yield excellent experimental results.

Acknowledgments

First of all, I am deeply indebted to my academic supervisor, Prof. Masashi Sugiyama, for his supervision the past three years. He has provided a rich environment for research and study. I would also like to express my gratitude to Prof. Taisuke Sato, Prof. Takenobu Tokunaga, Prof. Tsuyoshi Murata, and Prof. Jun Sese for reviewing and evaluating my thesis.

Furthermore, I am grateful for the secretaries of the Sugiyama laboratory, Ms. Yasuyo Obana and Mrs. Ayako Tamai, for helping with numerous administration tasks and ensuring that the laboratory is a pleasant environment. Dr. Hirotaka Hachiya and Dr. Makoto Yamada, post-doctoral research fellows in the Sugiyama laboratory, also helped me significantly.

I am also grateful to for my friends and fellow students whom shared the trenches with me: Tingting Zhao, Wittawat (Nuke) Jitkrittum, Akihiro Yamashita, Dr. Ning Xie, Dr. Gang Niu, Dr. Toby Dylan Hocking, Voot Tangkaratt, Tomoya Sakai, Duong Tuan Nguyen and Ikko Yamane and Hyunha Nam. My friends outside lab, Sanja Joka, Vida Macikenaite, and Mitsunori Nakamura made Tokyo a much more interesting place.

I am grateful that my research were supported by the Japanese government MEXT scholarship. Without this financial assistance, it would have been impossible for me to pursue a PhD.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the standard machine learning setup, it is assumed that there is perfect supervision and that the environment is unchanging. In practice however, these assumptions are often violated. This thesis is devoted to developing machine learning methods that can be applied to such non-standard settings when the deviation is characterized by a change in class prior.

1.1 Standard machine learning model

The goal of machine learning is to construct algorithms that can automatically learn from data. Data exists in the form of labeled samples originating from the environment. The labels are assigned by a supervisor (teacher) in order to facilitate learning. The goal of the learner is then to infer a rule from the labeled dataset provided by the teacher in order to emulate the teacher on unseen data.

We illustrate the supervised learning problem in Figure 1.1. Consider the problem of learning to classify whether a document is unsolicited marketing (i.e. spam) or not. According to the illustration, the supervisor assigns a class label "spam"/"not spam" to a training set of documents from the environment. This labeled dataset is in turn used by the machine learning algorithm to infer a rule. The learned rule can then be applied to unseen observations in order to determine the class label.

The training dataset can be modeled as samples drawn independently and iden-

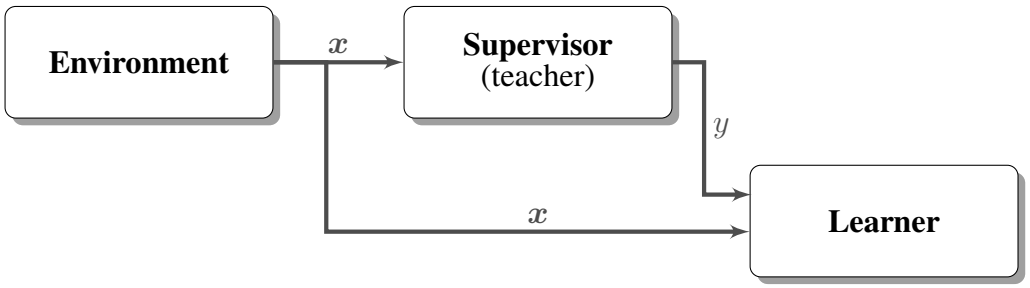


Figure 1.1: Illustration of supervised learning (i.e. learning with a teacher).

tically distributed from an unknown density $p_{\text{tr}}(\boldsymbol{x}, y)$,

$$\mathcal{X}_{\text{train}} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p_{\text{tr}}(\boldsymbol{x}, y).$$

In the above $\boldsymbol{x}$ is a feature vector (or covariate) and y is the associated class label (in the case of classification) or real value (in case of regression).

The overall goal of the learning algorithm is to construct a rule that can generalize on the basis of the training dataset to unseen test data. Assume that we have an unlabeled test dataset drawn as

$$\mathcal{X}_{\text{test}} := \{\boldsymbol{x}'_i\}_{i=1}^{n'} \overset{\text{i.i.d.}}{\sim} p_{\text{te}}(\boldsymbol{x}).$$

The standard assumption in machine learning is that the environment remains unchanged between the training and test phase. In other words, the training and test distributions are the same:

$$p(\boldsymbol{x}, y) = p_{\text{tr}}(\boldsymbol{x}, y) = p_{\text{te}}(\boldsymbol{x}, y),$$

where $p(\boldsymbol{x}, y)$ is shorthand for an unchanged training distribution. In many real-world machine learning scenarios, however, this implicit assumption is violated. Firstly, the environment may change between when the training data is collected and when the learning rule is applied to unseen data. This change is often referred to as *non-stationarity* between the test and training datasets (Quiñonero-Candela et al., 2009).

Furthermore, we may have imperfect supervision: the supervisor may selec-

tively label samples leading to a partially labeled training dataset (Niu, 2012).

In the subsequent sections we discuss several situations where the standard model breaks down, either due to inadequate supervision, or due to non-stationarity between training and test datasets. We will show that a wide variety of these problems can be modeled as a change in class balance between the test and training datasets.

1.2 Machine learning and supervision

We may distinguish between three learning settings, based on the presence or absence of labeled training data: unsupervised learning, supervised learning and semi-supervised learning. In *unsupervised learning*, we only have an unlabeled dataset:

$$\mathcal{X}_{\text{unlabeled}} := \{\boldsymbol{x}_i\} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}).$$

The goal of an unsupervised is to find an interesting or useful structure in the data (Chapelle et al., 2006). The unsupervised task of *clustering* attempts to assign each sample to a cluster (see Fig 1.2(a)). A tacit assumption in many clustering methods is that the underlying labels coincide with the *cluster structure* of the data. *Novelty detection* is another unsupervised learning method that makes the assumption that novel samples lie in low-density areas of the data (see Fig. 1.2(b)). Methods such as one-class support vector machines estimate the support for the high-density regions (Schölkopf et al., 1999) enabling the identification of such novel samples.

In *supervised learning*, the learning algorithm has access to a set of samples that are labeled:

$$\mathcal{X}_{\text{labeled}} := \{\boldsymbol{x}_i, y_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y).$$

The goal is to learn an input-output relation between the features and the outputs (Chapelle et al., 2006). Examples of such supervised learning tasks are regression and classification. The performance of such supervised learning methods is

measured in terms of the *risk* (or *expected loss*),

$$\mathbb{E}_p \left[L(f(\boldsymbol{x}), y) \right], \tag{1.1}$$

where f is the learned mapping from $\mathcal{X}$ to $\mathcal{Y}$, and $L(y', y)$ is a *loss function* between the estimated value y' and the true value y. $\mathbb{E}_p$ denotes the expectation with respect to $p(\boldsymbol{x}, y)$, the distribution from which the test samples were drawn. The preferred learner is therefore the one that minimizes the above expected risk. Since in practice the generating distribution $p(\boldsymbol{x}, y)$ is unknown, the above is empirically approximated as

$$\widehat{f} = \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{i=1}^{n} L(f(\boldsymbol{x}_i), y_i),$$

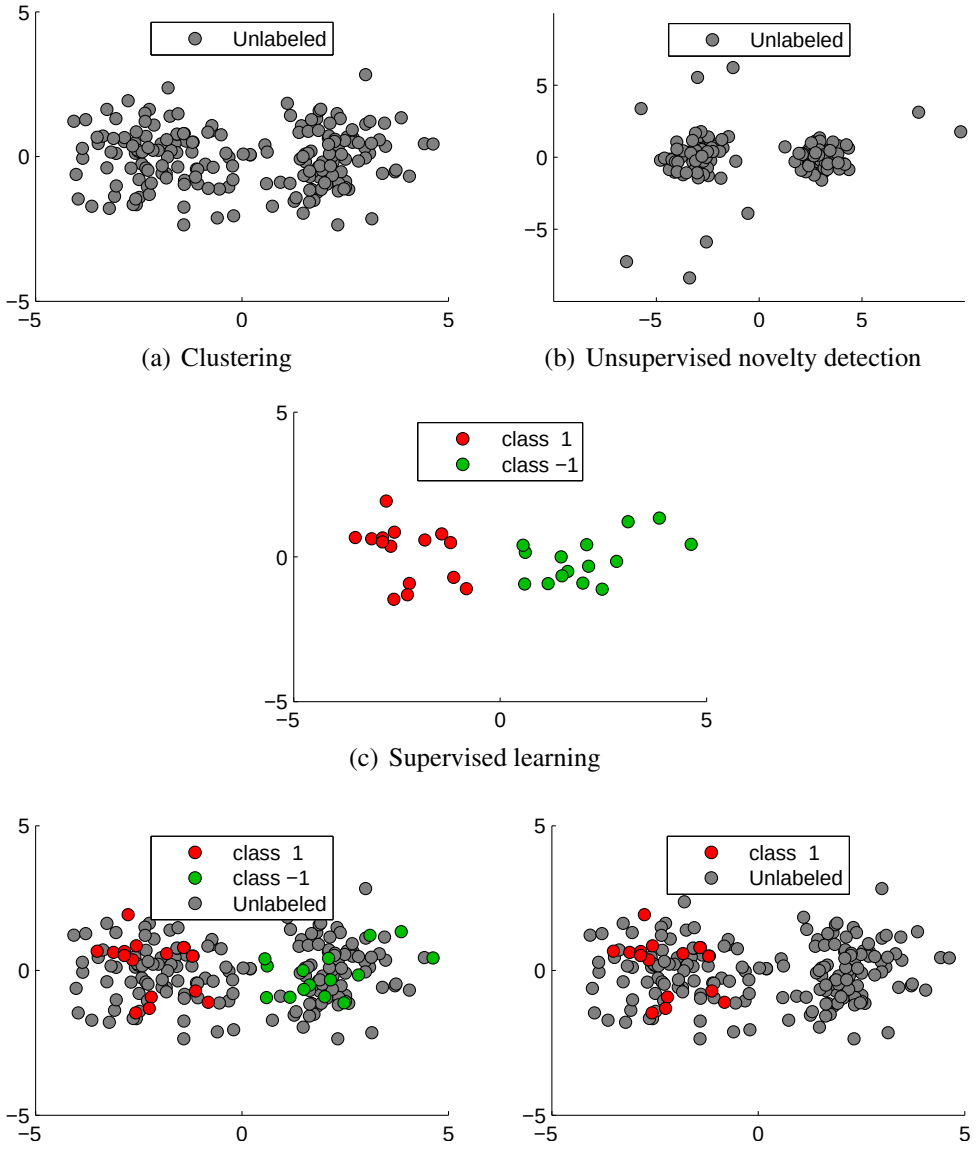
where f is searched from a set of functions $\mathcal{F}$. This learning framework is known as *empirical risk minimization* (Vapnik, 1998).

In *semi-supervised learning*, unlabeled samples are available in addition to labeled samples:

$$\mathcal{X}_{\text{labeled}} := \{\boldsymbol{x}_i, y_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y), \text{ and } \mathcal{X}_{\text{unlabeled}} := \{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}).$$

The goal here is to learn an input-output relation. However, it is hoped that by using the additional unlabeled samples, the learning task can be better accomplished.

In many practical problems, we may have a *partially labeled* dataset. Partially labeled means that the supervisor may selectively assign labels to samples. An instance of such a problem, called *learning from positive and unlabeled data*, is illustrated in Fig. 1.2(e) (Elkan and Noto, 2008). In this problem, the supervisor only gives labels to some positive samples. Such a partial labeling often occurs in land-cover identification problems: the user may label some samples of a specific land-cover type that he wants to identify in a larger dataset. The dataset therefore consists of labeled samples of a single class of interest and unlabeled samples, which is either of the class of interest or not of the class of interest (Li et al., 2011). Just as in the supervised learning setting, the goal is to learn an input-output relation from the partially labeled dataset.



(a) Clustering (b) Unsupervised novelty detection

(c) Supervised learning

(d) Semi-supervised learning (e) Learning from positive and unlabeled data

Figure 1.2: Illustration of different problem settings according to the level of supervision.

1.3 Non-stationarity in machine learning problems

Labeled training data is drawn from the source domain $p_{\text{tr}}(\boldsymbol{x}, y)$. We wish to apply the learned algorithm on unlabeled samples from a target domain distributed as $p_{\text{te}}(\boldsymbol{x}, y)$. Traditional machine learning settings discussed above assume that the source and target domain are similarly distributed, i.e.,

$$p(\boldsymbol{x}, y) = p_{\text{tr}}(\boldsymbol{x}, y) = p_{\text{te}}(\boldsymbol{x}, y).$$

In real-world problems however, the data often exhibit *non-stationarity*, causing the training and test dataset to differ. This non-stationarity may be caused by factors such as biased sampling or distribution shift over time (Zadrozny, 2004; Heckman, 1979; Sugiyama and Kawanabe, 2012).

If the test and training distributions are arbitrarily different, learning may not be possible. However, it is possible to make reasonable assumptions on the nature of the difference between the training and test distributions. These assumptions are often empirically validated in practical datasets and may be justified by reasonable sample-selection bias models.

One common assumption is *covariate shift*. In covariate shift, the assumption is that the training and test points follow different distributions, but the class posteriors are unchanged (Sugiyama and Kawanabe, 2012)

$$p_{\text{tr}}(y|\boldsymbol{x}) = p_{\text{te}}(y|\boldsymbol{x}).$$

In other words, only the input densities (or *covariates*) differ between the training and test distributions

$$p_{\text{tr}}(\boldsymbol{x}) \neq p_{\text{te}}(\boldsymbol{x}).$$

Learning on the labeled training data with a misspecified model may cause a large bias (Shimodaira, 2000). Due to the assumption of an unchanging class posterior, the expected test error may be rewritten as

$$\mathbb{E}_{p_{\text{te}}}\left[L(f(\boldsymbol{x}), y)\right] = \mathbb{E}_{p_{\text{tr}}}\left[L(f(\boldsymbol{x}), y)\frac{p_{\text{te}}(\boldsymbol{x})}{p_{\text{tr}}(\boldsymbol{x})}\right].$$

The risk can therefore be corrected by weighting each point with the *density ratio*

$$r(\boldsymbol{x}) = \frac{p_{\text{te}}(\boldsymbol{x})}{p_{\text{tr}}(\boldsymbol{x})}.$$

In the semi-supervised setup, unlabeled samples drawn according to $p_{\text{te}}(\boldsymbol{x})$ are available, enabling the estimation of this density ratio. This estimation can be performed by selecting $r(\boldsymbol{x})$ so that the induced distribution $r(\boldsymbol{x})p_{\text{tr}}(\boldsymbol{x})$ is as similar to $p_{\text{te}}(\boldsymbol{x})$ as possible (Huang et al., 2007; Sugiyama et al., 2008a). This may also be estimated via direct density-ratio estimation (Kanamori et al., 2009).

In the *class-prior change* setting, it is assumed that the class priors differ between the test and training distribution (Saerens et al., 2001),

$$p_{\text{tr}}(y) \neq p_{\text{te}}(y),$$

but the class-conditional densities remain unchanged:

$$p_{\text{tr}}(\boldsymbol{x}|y) = p_{\text{te}}(\boldsymbol{x}|y).$$

Due to this assumption, the expected test error can be written as

$$\mathbb{E}_{p_{\text{te}}}\left[L(f(\boldsymbol{x}), y)\right] = \mathbb{E}_{p_{\text{tr}}}\left[L(f(\boldsymbol{x}), y)\frac{p_{\text{te}}(y)}{p_{\text{tr}}(y)}\right].$$

In a semi-supervised setup, unlabeled samples from $p_{\text{te}}(\boldsymbol{x})$ are available. We show in Chapter 4 that, in this setup, class priors can be estimated via divergence estimation.

1.4 Divergences

As we see in the next section, divergences are central to solving many problems that arise in non-stationary datasets. Broadly speaking, a divergence defines the difference between two distributions. In other words, let a divergence between two probability densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ be denoted as $\mathrm{D}\left(p\|q\right)$. Then divergence

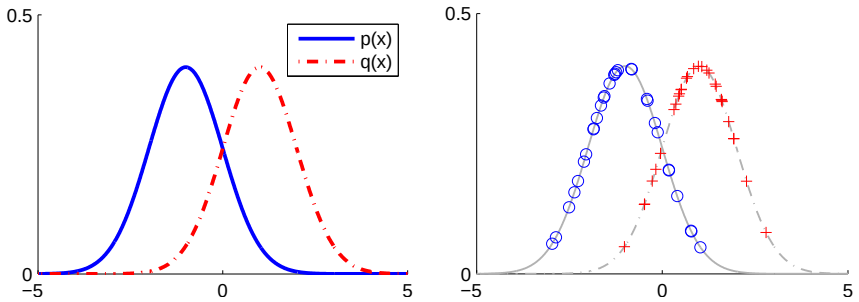


Figure 1.3: Figure 1.3(a) shows the densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$. Figure 1.3(b) shows samples drawn from these densities.

is always positive and

$$\mathrm{D}\left(p\|q\right) = 0 \text{ if and only if } p(\boldsymbol{x}) = q(\boldsymbol{x}).$$

The well-known Kullback-Leibler divergence is defined as (Kullback and Leibler, 1951),

$$\mathrm{KL}\left(p\|q\right) = \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}. \tag{1.2}$$

Divergences, such as the Kullback-Leibler divergence given above can be defined in terms of probability densities (e.g. Figure 1.3(a)). In practice, however, we only have samples drawn from these densities (Figure 1.3(b)). A possible approach is to first estimate the densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ from the samples and then plug the estimated densities into the definition of the divergence. However, we are not necessarily interested in the densities, but in the value of the divergence. Therefore, estimating the divergence in this way violates Vapnik's principle (Vapnik, 2000):

> If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

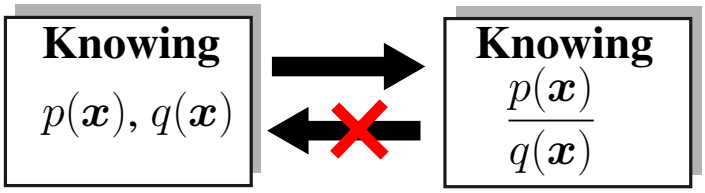


Figure 1.4: Illustration of density estimation and density-ratio estimation. The densities can not be calculated from the density ratio. Therefore, following Vapnik's principle, density estimation is a more general and more difficult problem than directly estimating the density ratio.

An alternative approach is to first estimate the density ratio $p(\boldsymbol{x})/q(\boldsymbol{x})$. The density ratio may be more accurately estimated than the densities, since density estimation is a more general problem (see Figure 1.4 for an illustration). Several efficient methods for estimating the density ratio has been introduced (Sugiyama et al., 2008a; Kanamori et al., 2009; Que and Belkin, 2013; Vapnik et al., 2013). It is also possible to directly estimate several types of divergences in terms of the density ratio (Keziou, 2003; Nguyen et al., 2010b; Sugiyama et al., 2013b). Interpreting existing problems in terms of divergences and using these divergence estimators, can lead to new and efficient algorithms.

1.5 Contributions

The overview of the three major contributions of the thesis is presented in this section.

1.5.1 Estimation of class priors in the semi-supervised setup

Assume that we have a training dataset

$$\{\boldsymbol{x}_i, y_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p_{\text{tr}}(\boldsymbol{x}, y),$$

but there is a change in class priors between the test and training distributions,

$$p_{\text{tr}}(\boldsymbol{x}|y) = p_{\text{te}}(\boldsymbol{x}|y), \text{ but } p_{\text{tr}}(y) \neq p_{\text{te}}(y).$$

In addition, we assume a semi-supervised setup with unlabeled samples

$$\mathcal{X}_{\text{unlabeled}} := \{\boldsymbol{x}'_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p_{\text{te}}(\boldsymbol{x}).$$

Note that the labeled and unlabeled data are not similarly distributed: the labeled data follows the training distribution, but the unlabeled data follows the test distribution. Due to the prior-change assumption, the test input distribution can be expressed as

$$q_{\text{te}}(\boldsymbol{x}) = \sum_{y=1}^{c} \theta_y p_{\text{tr}}(\boldsymbol{x}|y),$$

where θ_y models the unknown test class prior $p_{\text{te}}(y)$. Matching the model $q_{\text{te}}(\boldsymbol{x})$ to $p_{\text{te}}(\boldsymbol{x})$ under some divergence leads to a framework for class-prior estimation.

This allows for the existing method of Saerens et al. (2001) to be interpreted as distribution matching under the Kullback-Leibler divergence. Analysis using this framework shows that the method estimates the Kullback-Leibler divergence in an indirect manner, which may lead to inferior performance.

Furthermore, careful analysis of the optimization problem introduced in that paper shows that the problem was convex and the optimization can be interpreted as fixed point iteration. This fixed point iteration may however terminate at spurious fixed points and therefore not reach the (unique) optimal value.

To overcome these weaknesses, we proposed a method to directly estimate the class prior via a lower bound of any f-divergence. The f-divergence family of divergences includes the Kullback-Leibler divergence (Kullback and Leibler, 1951) and the Pearson divergence (Pearson, 1900). Using the Pearson divergence, we showed that a simple estimator with an analytic result may be obtained. Furthermore, the superior performance of this method was illustrated on numerous benchmark datasets. The proposed method was further extended to estimate the class-prior by matching the L_2-distance between probability densities.

This work is discussed in Chapter 4.

1.5.2 Labeling of data differing by class balance

As discussed previously, clustering is a hard problem due to the absence of class labels. The tacit assumption is that the underlying class labels coincide with the cluster structure of the data. When this assumption is violated, cluster separation may not coincide with class separation and the resulting clustering may be of dubious practical utility.

We assume that we have two unlabeled datasets,

$$\mathcal{X}_a := \{\boldsymbol{x}_i^a\} \overset{\text{i.i.d.}}{\sim} p_a(\boldsymbol{x}), \text{ and } \mathcal{X}_b := \left\{\boldsymbol{x}_i^b\right\} \overset{\text{i.i.d.}}{\sim} p_b(\boldsymbol{x}),$$

where the two datasets differ by class prior:

$$p_a(\boldsymbol{x}|y) = p_b(\boldsymbol{x}|y), \text{ but } p_a(y) \neq p_b(y).$$

Using these datasets, we show that we can obtain a labeling for unlabeled samples. If the class priors of the two datasets are unknown, the exact class labels can not be determined, but the data is partitioned into two disjoint sets. Moreover, this labeling does not depend on the cluster structure of the data. Therefore, this method would work even in multimodal datasets, where traditional clustering methods may fail.

This work is discussed in Chapter 5.

1.5.3 Class-prior estimation in learning from positive and unlabeled data

In many situations, we may have a dataset that is partially labeled, with only some positive samples having a label. This occurs, for example, in land-cover classification problems (Li et al., 2011): we may label only a subset of the landcover type that we wish to identify as $y = 1$. The remainder of the dataset remains unlabeled and consists of landcover types of interest ($y = 1$) and landcover types not of interest ($y = -1$).

It was shown in Elkan and Noto (2008) that if the class prior $p(y = 1)$ is known, classification can be performed. We interpret the problem of estimating the class prior as distribution matching under a divergence. This leads to the

development of an efficient method that gives an analytical solution to the estimate of the class prior.

Furthermore, we show that the existing method can also be interpreted as distribution matching under a divergence. However, the existing method does not directly estimate the divergence, resulting in inferior empirical results.

This work is discussed in Chapter 6.

1.6 Organization

The dissertation is organized into seven chapters. The organization of the dissertation is given in Figure 1.5.

Chapter 2 discusses divergences and other measures of similarity between distributions in detail. The family of f-divergences is discussed. A new class of divergences, defined based on the density difference is also introduced. The direct estimation of these divergences is reviewed in Chapter 3.

These two chapters provide us with the tools necessary to view machine learning problems from the divergence estimation vantage point. In Chapter 4 we show that the class-prior change problem can be interpreted as distribution matching.

In Chapter 5, a method is introduced to perform clustering on when we have datasets differing by a class prior.

We show in Chapter 6 that learning from positive and unlabeled data is a special case of learning from data differing by a class prior. In order to achieve accurate classification in this setting, the class prior of the unlabeled dataset must be known. We introduce a method in this chapter to estimate the class prior in this setting.

Finally, we present a conclusion and discuss future work in Chapter 7.

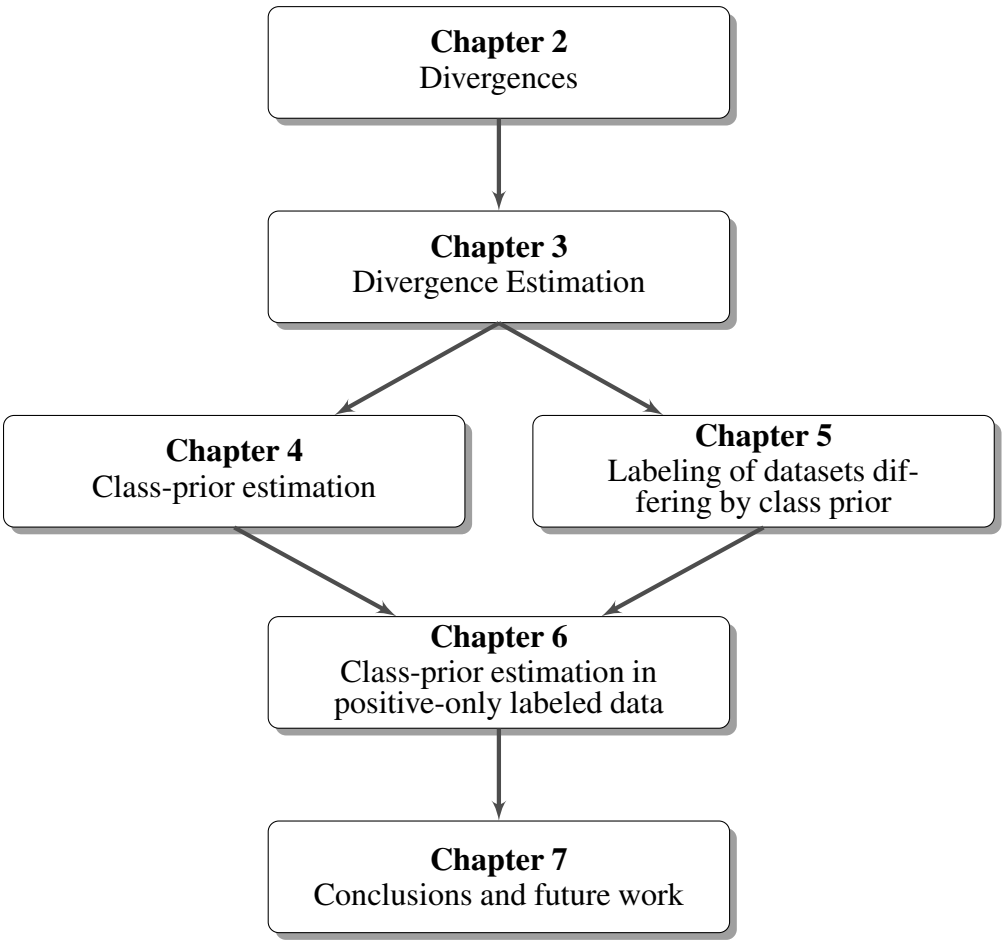


Figure 1.5: Organization of thesis.

Chapter 2

Divergences

This chapter discusses several divergences between probability distributions. Discussion on the estimation of these divergences is postponed until the next chapter.

2.1 Introduction and motivation

As discussed in the preceding chapter, a divergence measures the difference between two probability distributions.

Divergences and their estimation from samples have several applications such as determining whether two sets of samples originated from the same distribution or not (referred to as *homogeneity* testing) (Sugiyama et al., 2011a; Gretton et al., 2012a). Divergence estimation is also used to detect change-points in non-stationary data (Kawahara and Sugiyama, 2012; Liu et al., 2013). Comparing distributions using samples is also the basis for transfer learning (Huang et al., 2007; Sugiyama and Kawanabe, 2012).

Furthermore, several information-theoretic quantities, such as *mutual information*, can be recast as divergences. This can, in turn, be used to test whether two covariates are independent or not. Mutual information can also be used for clustering (Sugiyama et al., 2011b), dimensionality reduction (Suzuki and Sugiyama, 2013), canonical dependency analysis (Karasuyama and Sugiyama, 2012), and independent component analysis (Suzuki and Sugiyama, 2011).

From this perspective of machine learning, we see that divergence measures and their estimation is of cardinal importance and central to the task of machine

learning. In the next three sections we discuss three measures of similarity between distributions: f-divergences, density-difference divergences and the maximum mean discrepancy.

2.2 f-divergences

An f-divergence between two probability distributions P and Q, with densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, is defined as (Ali and Silvey, 1966; Csiszár, 1967)

$$\mathrm{D}_f(p\|q) := \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}, \tag{2.1}$$

where f is a convex function such that $f(1) = 0$. The above defines a class of divergences with different properties for different choices of f. It is obvious from the above definition that $\mathrm{D}_f(p\|q) = 0$ when $p(\boldsymbol{x}) = q(\boldsymbol{x})$. An f-divergence is also not necessarily symmetric. It is symmetric, however, when $f(t) = tf(\frac{1}{t})$. Since the function $f(t)$ is convex, all divergences are convex in their first argument. Arguably the most well-known f-divergence is the Kullback-Leibler (KL) divergence(Kullback and Leibler, 1951), with $f(t) = t\log(t)$,

$$\mathrm{KL}(p\|q) = \int \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Mutual information can be defined as the Kullback-Leibler divergence from $p(\boldsymbol{x}, \boldsymbol{y})$ to $p(\boldsymbol{x})p(\boldsymbol{y})$,

$$\begin{aligned}\mathrm{MI} &:= \mathrm{KL}\left(p(\boldsymbol{x}, \boldsymbol{y})\|p(\boldsymbol{x})p(\boldsymbol{y})\right) \\ &= \iint \log\left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right) p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.\end{aligned}$$

The *Pearson* or χ^2 divergence (Pearson, 1900) is defined with $f(t) = \frac{1}{2}(t-1)^2$ or equivalently $f(t) = \frac{1}{2}t^2 - \frac{1}{2}$,

$$\mathrm{PE}(p\|q) = \frac{1}{2}\int \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1\right)^2 q(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{2.2}$$

$$= \frac{1}{2}\int \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^2 q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{2}. \tag{2.3}$$

An advantage of the above divergence over the KL divergence is that often the quadratic term leads to simpler expressions that simplifies analysis and enables efficient estimation.

Similarly to ordinary mutual information, the *squared-loss mutual information* (SMI) can be defined as (Sugiyama, 2013)

$$\begin{aligned}\text{SMI} &:= \text{PE}\left(p(\boldsymbol{x},\boldsymbol{y})\|p(\boldsymbol{x})p(\boldsymbol{y})\right), \\ &= \frac{1}{2}\iint \left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1\right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.\end{aligned}$$

An application for SMI is information-maximization clustering, where using this leads to an efficient solution to clustering problems (Sugiyama et al., 2011b).

The f-divergence reduces to the *total variation* distance when $f(t) = |t-1|$,

$$\begin{aligned}\mathrm{TV}(p\|q) &= \int \left|\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1\right| q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \\ &= \int |p(\boldsymbol{x}) - q(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

This is also known as the L_1 distance between probability densities. Note that the total variational distance has the additional property of being symmetric. The above and some additional f-divergences are summarized in Table 2.1.

2.3 Density-difference divergences

The basic idea for f-divergence approximation is to derive a lower bound via *convex duality* (Boyd and Vandenberghe, 2004). Optimizing with respect to this lower bound results in a divergence estimator that estimates the divergence with respect to the density ratio $p(\boldsymbol{x})/q(\boldsymbol{x})$.

However, a potential weakness of f-divergences is that the density ratio can diverge to infinity even for simple setups such as the ratio of two Gaussian densities (Cortes et al., 2010). This may make f-divergence approximation unreliable

Table 2.1: Summary of commonly encountered f-divergences. Simple algebra confirms that $\mathrm{KL}(p\|q) = \mathrm{rKL}(q\|p)$ and $\mathrm{PE}(p\|q) = \mathrm{rPE}(q\|p)$.

Name	$f(t)$	**Definition**
Kullback-Leibler (KL)	$t\log(t)$	$\mathrm{KL}(p\|q) = \int \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$
Pearson (PE)	$\frac{1}{2}(t-1)^2, \frac{1}{2}t^2 - \frac{1}{2}$	$\mathrm{PE}(p\|q) = \frac{1}{2}\int \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1\right)^2 q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$
Total variation	$\lvert t-1\rvert$	$\mathrm{TV}(p\|q) = \int \left\lvert\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1\right\rvert q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$
reverse KL	$-\log(t)$	$\mathrm{rKL}(p\|q) = \int \log(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})})q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$
reverse PE	$\frac{1}{2}\frac{1}{t} - \frac{1}{2}$	$\mathrm{rPE}(p\|q) = \frac{1}{2}\int \left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{2}$

in practice. To overcome this problem, we define a class of divergences based on the *density difference*. Since the density difference is always bounded when the densities are bounded, these divergences do not suffer from the problem described above.

The class of density-difference divergences is defined as

$$\mathrm{DD}_\psi(p, q) = \int \psi(p(\boldsymbol{x}) - q(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}, \tag{2.4}$$

where the function $\psi(t)$ is convex with a minimum at $\psi(0) = 0$. A useful property of these divergences is that they are always symmetric.

An obvious choice for is the squared function $\psi(t) = \frac{1}{2}t^2$, resulting in the L_2 distance between distributions,

$$L_2(p, q) = \frac{1}{2}\int (p(\boldsymbol{x}) - q(\boldsymbol{x}))^2 \,\mathrm{d}\boldsymbol{x}.$$

Another possible choice is the $\psi(t) = |t|$, which results in the L_1 distance between

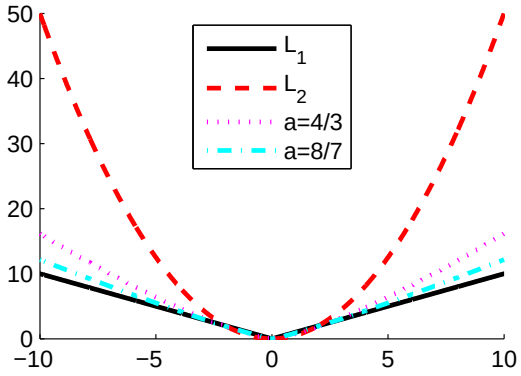


Figure 2.1: Illustration of different $\psi(t)$ functions for the density-difference divergence.

distributions,

$$L_1(p,q) = \int |p(\boldsymbol{x}) - q(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x}.$$

The L_1 distance is an example of a function that is both a density-difference divergence and an f-divergence. To generalize the above, we can consider a *power divergence*

$$\psi(t) = \tfrac{b-1}{b} \, |t|^{\frac{b}{b-1}} \, ,$$

where b is an even number. This gives a divergence as

$$L_P(p,q) = \frac{b-1}{b} \int |p(\boldsymbol{x}) - q(\boldsymbol{x})|^{\frac{b}{b-1}} \, \mathrm{d}\boldsymbol{x}.$$

By varying b, we can obtain divergences "between" the L_1 and L_2 distances. These are illustrated in Figure 2.1

2.4 Maximum mean discrepancy

The *maximum mean discrepancy* (MMD) is a computationally efficient method to compare distributions. The MMD between two distributions p and q is defined as (Gretton et al., 2007, 2012a)

$$\mathrm{MMD}\left[\mathcal{F}, p, q\right] = \sup_{f \in \mathcal{F}} \quad \mathbb{E}_p\left[f(\boldsymbol{x})\right] - \mathbb{E}_q\left[f(\boldsymbol{x})\right], \tag{2.5}$$

where $f \in \mathcal{F}$ is a function $f : \mathcal{X} \to \mathbb{R}$ from the feature space to the real line. The above definition is also known as the integral probability metric (Müller, 1997).

To obtain a practical estimator, the set of possible functions, $\mathcal{F}$ should be chosen and the optimization performed. The discussion on this aspect is deferred until Section 3.4.

MMD as a measure of similarity between distributions has been used to test if two samples are homogeneous (Gretton et al., 2007; Borgwardt et al., 2006) and to test whether covariates are independent (Gretton et al., 2008). By comparing distributions, it has also been used to adapt for covariate shift (Huang et al., 2007).

Chapter 3

Estimation of divergences

This chapter discusses the estimation of divergences from samples.

3.1 Introduction

The divergences discussed in the preceding chapter are defined in terms of probability distributions. In practical problems however, these divergences must be estimated from samples in order to measure the similarity of the (unknown) underlying distributions. We assume that the following two sets of samples are available for estimating the divergence:

$$\mathcal{X}_p := \{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) \quad \text{and} \quad \mathcal{X}_q := \{\boldsymbol{x}'_i\}_{i=1}^{n'} \overset{\text{i.i.d.}}{\sim} q(\boldsymbol{x}).$$

A naïve approach to estimate the divergences is to first estimate the densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ separately from $\mathcal{X}_p$ and $\mathcal{X}_q$. These estimates can then be plugged into the definitions to obtain estimates for the divergences. However, as discussed in Section 1.4, this violates Vapnik's principle: density estimation is a more general and difficult problem than divergence estimation. Therefore, following this principle, we wish to estimate the divergence directly in a single shot approach.

We review how, using the squared function, it is possible to directly estimate the Pearson divergence in terms of the density ratio. This is done by obtaining an inequality that is a lower bound for the squared function. By applying this inequality to the Pearson divergence a lower bound, which is linear in the unknown

densities, is obtained. This in turn enables estimation via sample averages and the subsequent estimation of the Pearson divergence. Using the same inequality, the L_2 distance between probability densities may be lower bounded allowing for direct estimation in terms of the *density difference*.

A similar inequality can be obtained for general convex functions via *Fenchel duality*. Applying this inequality to general f-divergences leads to practical estimators (Keziou, 2003; Nguyen et al., 2010b). We adapt this approach to estimate density-difference divergences introduced in Chapter 2.

3.2 Squared-error bound

The least-squares method is used to bound the Pearson divergence and the L_2-distance (Kanamori et al., 2009; Sugiyama et al., 2013c). Consider the following simple lower bound, which follows directly from expanding the square term,

$$\begin{aligned} \frac{1}{2}(t-r)^2 &\geq 0, \\ \frac{1}{2}t^2 - tr + \frac{1}{2}r^2 &\geq 0, \\ \frac{1}{2}t^2 &\geq tr - \frac{1}{2}r^2. \end{aligned} \tag{3.1}$$

We show below that this simple inequality can be used to create a single-shot estimator for the Pearson divergence and L_2 distance between densities.

3.2.1 Squared-error bound of the Pearson divergence

By using the above inequality, we can obtain a pointwise bound for the expression in Eq. (2.3), where $r(\boldsymbol{x})$ fulfils the role of r as

$$\begin{aligned} \frac{1}{2}\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^2 &\geq \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) r(\boldsymbol{x}) - \frac{1}{2}r(\boldsymbol{x})^2, \\ \frac{1}{2}\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^2 q(\boldsymbol{x}) &\geq p(\boldsymbol{x})r(\boldsymbol{x}) - \frac{1}{2}r(\boldsymbol{x})^2 q(\boldsymbol{x}) \end{aligned}$$

Integrating the above and selecting the tightest bound gives,

$$\frac{1}{2}\int\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^2 q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{2} \geq \sup_r \int r(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{2}\int r(\boldsymbol{x})^2 q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{2}.$$

The function $r(\boldsymbol{x})$ in the above inequality can be approximated as linear-in-parameter model,

$$\widehat{r}(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}),$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_b)^\top$ are the parameters and $\boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \varphi_2(\boldsymbol{x}), \ldots, \varphi_b(\boldsymbol{x}))^\top$ are the basis functions. In practice, we use Gaussian basis functions,

$$\varphi_i(\boldsymbol{x}) = \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{x} - \boldsymbol{x}_i\right\|^2\right),$$

centered around the training points. This model allows us to write the lower bound as

$$\max_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}^\top \boldsymbol{h} - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha} - \frac{1}{2},$$

where

$$\boldsymbol{H} := \int \boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad \text{and} \quad \boldsymbol{h} := \int \boldsymbol{\varphi}(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Estimating the integrals with sample averages gives

$$\widehat{\boldsymbol{H}} = \frac{1}{n'}\sum_{i=1}^{n'} \boldsymbol{\varphi}(\boldsymbol{x}'_i)\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top,$$
$$\widehat{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i).$$

An ℓ_2 regularizer can be added, which results in

$$\widehat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{h}} - \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \frac{1}{2}\lambda\boldsymbol{\alpha}^\top\boldsymbol{\alpha}.$$

This leads to the following analytic estimate of the function $r(\boldsymbol{x})$

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}\right)^{-1} \widehat{\boldsymbol{h}}.$$

By using this estimate for the function $r(\boldsymbol{x})$, the following estimate for the Pearson divergence can be obtained:

$$\widehat{\mathrm{PE}} = \widehat{\boldsymbol{h}}^\top \left(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}\right)^{-1} \widehat{\boldsymbol{h}} - \frac{1}{2} \widehat{\boldsymbol{h}}^\top \left(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}\right)^{-1} \widehat{\boldsymbol{H}} \left(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}\right)^{-1} \widehat{\boldsymbol{h}} - \frac{1}{2}.$$

3.2.2 Squared-error bound of the L_2 distance

The same inequality in Eq. (3.1), can be applied to bound the L_2 distance between probability densities (Sugiyama et al., 2012c, 2013c). The pointwise bound is

$$(p(\boldsymbol{x}) - q(\boldsymbol{x}))^2 \geq [p(\boldsymbol{x}) - q(\boldsymbol{x})]\, g(\boldsymbol{x}) - \frac{1}{2} g(\boldsymbol{x})^2$$

By integrating both sides and selecting the tightest bound via maximization, we get

$$\int (p(\boldsymbol{x}) - q(\boldsymbol{x}))^2 \, \mathrm{d}\boldsymbol{x} \geq \sup_g \int g(\boldsymbol{x}) \, [p(\boldsymbol{x}) - q(\boldsymbol{x})] \, \mathrm{d}\boldsymbol{x} - \frac{1}{2} \int g(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}. \quad (3.2)$$

As in the Pearson divergence case, the function $g(\boldsymbol{x})$ can again be modeled using a linear-in-parameter model,

$$\widehat{g}(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}),$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_b)^\top$ are the parameters and $\boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \varphi_2(\boldsymbol{x}), \ldots, \varphi_b(\boldsymbol{x}))^\top$ are the basis functions. In practice, we can use Gaussian basis functions centered at the training points,

$$\varphi_i(\boldsymbol{x}) = \exp\left(-\frac{1}{2\sigma^2} \left\|\boldsymbol{x} - \boldsymbol{c}_i\right\|^2\right), \quad (3.3)$$

where $(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n, \boldsymbol{c}_{n+1}, \ldots, \boldsymbol{c}_{n+n'}) := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'})$. The optimal parameter choice can then be written as

$$\boldsymbol{\alpha} = \arg\max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top \boldsymbol{h} - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H} \boldsymbol{\alpha},$$

where

$$H_{\ell,\ell'} := \int \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_{\ell'}\|^2}{2\sigma^2}\right) = \left(\pi\sigma^2\right)^{d/2} \exp\left(-\frac{\|\boldsymbol{c}_\ell - \boldsymbol{c}_{\ell'}\|^2}{4\sigma^2}\right),$$
$$h_\ell := \int \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) q(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

where d is the dimensionality of $\boldsymbol{x}$. The expectation in $\boldsymbol{h}$ can be replaced by its empirical estimate. Adding an ℓ_2 regularizer, results in the following objective function:

$$\widehat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \ \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H} \boldsymbol{\alpha} - \frac{1}{2}\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha}.$$

The analytical solution of the above is then

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}.$$

Substituting the above analytical solution into the unregularized objective function results in the following estimate of the L_2 distance

$$\widehat{L}_2 = \widehat{\boldsymbol{h}}^\top (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}} - \frac{1}{2}\widehat{\boldsymbol{h}}^\top (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \boldsymbol{H} (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}.$$

3.3 Fenchel duality bound

The Fenchel duality bounding technique was first introduced in Keziou (2003) and later popularized in Nguyen et al. (2010b). This bound uses the fact that $f(t)$ is convex to obtain a lower bound for any f-divergence. The bound can be

intuitively understood by considering the following self-evident inequality

$$\begin{aligned} tz - f(t) &\leq \sup_{t'} t'z - f(t') \\ f(t) &\geq tz - f^*(z), \end{aligned} \tag{3.4}$$

where

$$f^*(z) = \sup_{t'} t'z - f(t').$$

The function $f^*(z)$ is known as the *Fenchel dual* or *convex conjugate* (Boyd and Vandenberghe, 2004). This inequality provides a lower bound that is linear w.r.t. t for any function $f(t)$. Furthermore, if the function $f(t)$ is convex, this inequailty is tight (Boyd and Vandenberghe, 2004, p. 94). The above inequality is known as *Fenchel's inequality* (for an arbitrary f) or *Young's inequality* (when f is differentiable) (Boyd and Vandenberghe, 2004).

3.3.1 Fenchel lower-bound for f-divergences

We can obtain a lower-bound for the f-divergence in Eq. (2.1) by applying the bound in Eq. (3.4) in a pointwise manner. The first step is bounding the convex function $f(t)$,

$$f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \geq r(\boldsymbol{x})\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) - f^*(r(\boldsymbol{x})),$$

where $r(\boldsymbol{x})$ fulfils the role of z in Eq. (3.4). Multiplying both sides with $q(\boldsymbol{x})$ gives,

$$f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) q(\boldsymbol{x}) \geq r(\boldsymbol{x})p(\boldsymbol{x}) - f^*(r(\boldsymbol{x}))q(\boldsymbol{x}). \tag{3.5}$$

Integrating and then selecting the tightest bound gives the following estimator,

$$\mathrm{D}_f(p\|q) \geq \sup_r \int r(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int f^*(r(\boldsymbol{x}))q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

The above bound is already useful, since the expectations can be estimated via sample averages as

$$\widehat{\mathrm{D}}_f(p\|q) \geq \sup_r \frac{1}{n}\sum_{i=1}^{n} r(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{j=1}^{n'} f^*(r(\boldsymbol{x}'_i)).$$

Furthermore, the conjugate $f^*(z)$ of any function $f(t)$ is convex, so, if $r(\boldsymbol{x})$ is a linear model, the above is guaranteed to be a convex problem. Convex conjugates for functions defining several f-divergences are given in Table 3.1.

Table 3.1: Summary of estimators for commonly encountered divergences. Compare with Table 2.1

Name	$f(t)$	$f^*(v)$	$f^{*\prime}(v)$	**Estimates**
Kullback-Leibler	$t\log(t)$	e^{v-1}	e^{v-1}	$\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}+1$
Pearson (PE)	$\frac{1}{2}t^2-\frac{1}{2}$	$\frac{1}{2}v^2+\frac{1}{2}$	v	$\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$
Total variation	$\lvert t-1\rvert$	$\begin{cases}0 & 0\leq v\leq 1\\ \infty & \text{otherwise}\end{cases}$	$-$	$\operatorname{sign}\left[\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}-1\right]$
Reverse KL	$-\log(t)$	$-1-\log(-v)$	$-\frac{1}{v}$	$-\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$

From the table it is clear that the squared-error bound for the Pearson divergence is exactly the same as the Fenchel duality bound. For continuous $f(t)$, the following quantity is estimated

$$r(\boldsymbol{x}) = \left[f^{*\prime}\right]^{-1}\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right),$$

where $f^{*\prime}$ denotes the derivative of the conjugate f^*. This relation is obtained by solving the right-hand side of Eq. (3.5).

3.3.2 Fenchel lower bound for density-difference divergences

An estimator for the density-difference divergence can also be obtained by applying the Fenchel inequality. Applying this inequality pointwise gives,

$$\psi(p(\boldsymbol{x}) - q(\boldsymbol{x})) \geq g(\boldsymbol{x})\,[p(\boldsymbol{x}) - q(\boldsymbol{x})] - \psi^*(g(\boldsymbol{x})),$$
$$\int \psi(p(\boldsymbol{x}) - q(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \geq \int g(\boldsymbol{x})\,[p(\boldsymbol{x}) - q(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x} - \int \psi^*(g(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}.$$

Therefore, the tightest bound can be obtained as

$$\mathrm{DD}_\psi(p, q) \geq \sup_g \int g(\boldsymbol{x})\,[p(\boldsymbol{x}) - q(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x} - \int \psi^*(g(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}.$$

The expectations can be estimated by sample averages,

$$\widehat{\mathrm{DD}}_\psi(p, q) \geq \sup_g \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{j=1}^{n'} g(\boldsymbol{x}'_i) - \int \psi^*(g(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}. \tag{3.6}$$

The fact that the conjugate $\psi^*(v)$ of the function $\psi(t)$ is always convex ensures that the above optimization problem is always convex. A small difficulty with the above expression is the integration of the third term. For some problems, such as the L_2 distance with Gaussian basis functions, the last term can be calculated analytically.

A list of some possible functions are given in Table 3.2. From the table we confirm that the least-squares bound and the Fenchel duality bound are exactly the same.

By interpreting the conjugate as a constraint, we can rewrite the L_1 distance as,

$$L_1\,(p, q) = \sup_g \int g(\boldsymbol{x})\,[p(\boldsymbol{x}) - q(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x} \tag{3.7}$$
$$\text{subject to } |g(\boldsymbol{x})| \leq 1\,\forall\,\boldsymbol{x}.$$

For continuous functions, $g(\boldsymbol{x})$ estimates

$$g(\boldsymbol{x}) = \left[\psi^{*\prime}\right]^{-1}(p(\boldsymbol{x}) - q(\boldsymbol{x}))\,.$$

Consulting the table, we see that when estimating the L_2 distance, the density difference

$$g(\boldsymbol{x}) = p(\boldsymbol{x}) - q(\boldsymbol{x}),$$

is estimated. For the power density-difference divergence, the following quantity is estimated

$$g(\boldsymbol{x}) = \sqrt[b-1]{p(\boldsymbol{x}) - q(\boldsymbol{x})},$$

where b is an even number. This is plotted for separate values of b in Figure 3.1.

3.4 Maximum mean discrepancy estimation

In this section, we review the estimation of the maximum mean discrepancy (Gretton et al., 2012a). The relationship between the maximum mean discrepancy and the L_1 distance is also shown.

3.4.1 Estimation of MMD

In the maximum mean discrepancy, the set of functions $\mathcal{F}$ in

$$\text{MMD}\left[\mathcal{F}, p, q\right] = \sup_{f \in \mathcal{F}} \quad \mathbb{E}_p\left[f(\boldsymbol{x})\right] - \mathbb{E}_q\left[f(\boldsymbol{x})\right],$$

is selected as the unit ball in $\mathcal{H}$, a *reproducing kernel Hilbert space* (RKHS) (Gretton et al., 2012a). Let $k(x, x')$ be a reproducing kernel. The following property, known as the *reproducing property*, holds for any $f \in \mathcal{F}$, (Aronszajn, 1950; Schölkopf and Smola, 2001)

$$\langle f, k(x, \cdot)\rangle = f(x).$$

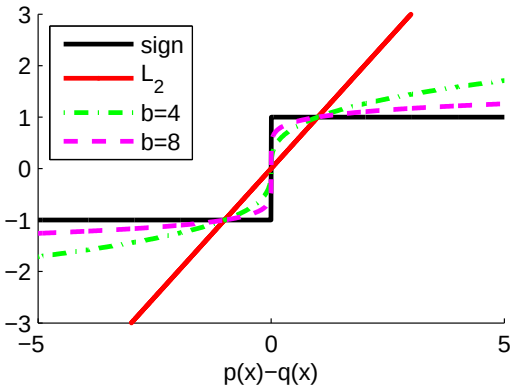


Figure 3.1: Values estimated using the L_1, L_2, and power density-difference divergences. As b increases, the estimated value becomes closer to the sign of the density difference.

Table 3.2: Summary of density-difference divergences

Name	$\psi(t)$	$\psi^*(v)$	$f^{*\prime}$
L_1 distance	$\lvert t\rvert$	$\begin{cases} 0 & -1 \le v \le 1, \\ \infty & \text{otherwise.} \end{cases}$	
L_2 distance	$\frac{1}{2}t^2$	$\frac{1}{2}v^2$	v
Power divergence	$\frac{b-1}{b}\lvert t\rvert^{\frac{b}{b-1}}$, b even	$\frac{1}{b}\lvert v\rvert^b$	v^{b-1}

MMD on the unit ball can then be expressed as

$$\begin{aligned}\text{MMD}\left[\mathcal{F}, p, q\right] &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_p\left[f(x)\right] - \mathbb{E}_q\left[f(x)\right], \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_p\left[\langle f, k(x, \cdot)\rangle_{\mathcal{H}}\right] - \mathbb{E}_q\left[\langle f, k(x, \cdot)\rangle_{\mathcal{H}}\right], \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mathbb{E}_p\left[k(x, \cdot)\right] - \mathbb{E}_q\left[k(x, \cdot)\right]\rangle_{\mathcal{H}}.\end{aligned}$$

Let $\mu_p, \mu_q \in \mathcal{H}$ be defined as

$$\mu_p = \mathbb{E}_p\left[k(x, \cdot)\right] \text{ and } \mu_q = \mathbb{E}_q\left[k(x, \cdot)\right].$$

The squared MMD can then be expressed as (Gretton et al., 2012a)

$$\begin{aligned}\text{MMD}\left[\mathcal{F}, p, q\right]^2 &= \left[\sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_p - \mu_q\rangle_{\mathcal{H}}\right]^2 \\ &= \left[\langle \frac{\mu_p - \mu_q}{\|\mu_p - \mu_q\|_{\mathcal{H}}}, \mu_p - \mu_q\rangle_{\mathcal{H}}\right]^2 \\ &= \left[\frac{1}{\|\mu_p - \mu_q\|_{\mathcal{H}}} \langle \mu_p - \mu_q, \mu_p - \mu_q\rangle_{\mathcal{H}}\right]^2 \\ &= \|\mu_p - \mu_q\|_{\mathcal{H}}^2.\end{aligned}$$

Expanding the above gives (Gretton et al., 2012a)

$$\begin{aligned}\text{MMD}^2\left[\mathcal{F}, p, q\right] &= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \\ &= \|\mathbb{E}_p\left[k(x, \cdot)\right] - \mathbb{E}_q\left[k(x, \cdot)\right]\|_{\mathcal{H}}^2 \\ &= \mathbb{E}_{x,x' \sim p}\left[k(x, x')\right] - 2\mathbb{E}_{x \sim p, x' \sim q}\left[k(x, x')\right] + \mathbb{E}_{x,x' \sim q}\left[k(x, x')\right].\end{aligned} \tag{3.8}$$

The above leads to the following unbiased estimator of MMD^2 (Borgwardt et al., 2006)

$$\widehat{\text{MMD}^2} = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(x_i, x_j) - \frac{2}{nn'} \sum_{i,j=1}^{n,n'} k(x_i, x'_j) + \frac{1}{n'(n'-1)} \sum_{i \neq j}^{n'} k(x_i, x_j).$$

The requirement to ensure that the MMD is a metric (i.e., a proper distance function) between distributions is that the RKHS is a universal RKHS (Gretton et al., 2012a). It was proven in Steinwart (2002) that the Gaussian RKHS is universal.

Using the Gaussian kernel however, still leaves the question of selecting the kernel width. A useful heuristic, that was used in Gretton et al. (2007), is setting the kernel width equal to the median distance between the samples. Recently, a method was introduced to automatically select the kernel width in the two-sample problem (Gretton et al., 2012b). It is still not clear how to perform kernel selection on other problems.

3.4.2 Relation to L_1 distance estimation

The function $f(\boldsymbol{x})$ in Eq. (2.5) is referred to as the "witness" function. If the witness function is chosen as the population version of the difference between kernel density estimators for $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$,

$$f(\boldsymbol{x}) = \int K(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} - \int K(\boldsymbol{x}, \boldsymbol{y}') q(\boldsymbol{y}') \mathrm{d}\boldsymbol{y}',$$

then normal MMD defined in Eq. (2.5) is equivalent to the squared MMD in Eq. (3.8). Furthermore, from the Cauchy-Schwartz inequality $|\langle f, g \rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|g\|_{\mathcal{H}}$, we have

$$f(\boldsymbol{x})^2 = \left(\langle f(\cdot), k(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}}\right)^2 \leq \|f\|_{\mathcal{H}}^2 \, k(\boldsymbol{x}, \boldsymbol{x}),$$

for all $\boldsymbol{x}$. When a Gaussian or Laplace kernel is used, $k(\boldsymbol{x}, \boldsymbol{x}) = 1$. Then, for any f such that $\|f\|_{\mathcal{H}} \leq 1$, we have $|f(\boldsymbol{x})| \leq 1$ for all $\boldsymbol{x}$. This implies that the squared MMD is equivalent to the L_1 distance in Eq. (3.7).

Chapter 4

Semi-supervised class-prior estimation

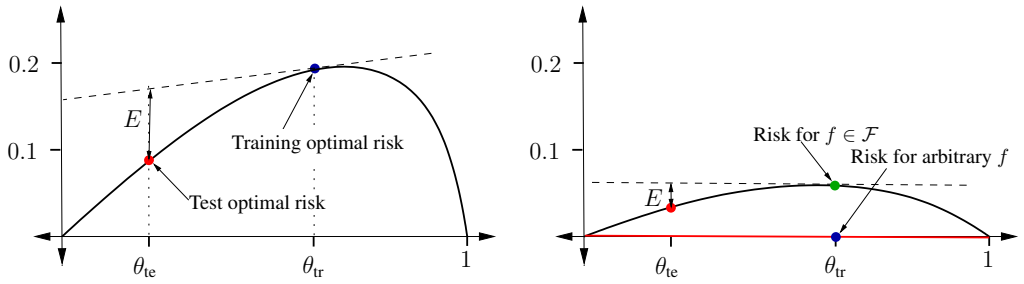
In this chapter, we discuss the problem of class-prior estimation in a semi-supervised setup.

4.1 Introduction

Most supervised learning algorithms assume that training and test data follow the same probability distribution (Vapnik, 1998; Hastie et al., 2001; Bishop, 2006). However, this de facto standard assumption is often violated in real-world problems, caused by intrinsic sample selection bias or inevitable non-stationarity (Heckman, 1979; Quiñonero-Candela et al., 2009; Sugiyama and Kawanabe, 2012).

In classification scenarios, changes in class balance are often observed – for example, the male-female ratio is almost fifty-fifty in the real-world (test set), whereas training samples collected in a research laboratory tends to be dominated by male data. Applying traditional classification methods in this *class-prior change* setting may lead to an excess misclassification rate.

We can gain an intuitive understanding of the adverse effect of class-prior change by considering the Bayes optimal risk. The risk for a binary experiment



(a) Overlapping class-conditional densities

(b) Non-overlapping class-conditional densities with a misspecified model.

Figure 4.1: Risk curves when classes are highly overlapping and when the classes are non-overlapping, but the model is misspecified. The x-axis shows the class prior and the y-axis the risk. The dashed line is the risk for the function f selected according to the training class prior.

with a class prior $p(y=1)=\theta$ can be expressed as

$$\mathbb{E}_p\left[L(f(\boldsymbol{x}),y)\right] = \min_f \quad \theta \int L(f(\boldsymbol{x}),1)p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} + (1-\theta)\int L(f(\boldsymbol{x}),-1)p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}, \tag{4.1}$$

where f is an arbitrary function. From the above, it is obvious that the function is concave with respect to θ and passes through zero when $t=0$ and $t=1$. An example of such a situation is given in Figure 4.1(a).

In the training phase, a function f is selected so as to minimize the risk according to the training class prior θ_{tr}. During the test phase, this function is applied to samples from a distribution with a prior θ_{te}. The difference between the optimal risk for the test distribution and the risk using the function f is denoted as E. We refer to this as the *excess risk*, since this is unnecessarily introduced due to the change in class prior.

It may appear that class-prior change is only a problem when the class-conditional densities significantly overlap. However, class-prior change is also a problem when the class-conditional densities do not overlap but a misspecified model is used. To model this, assume that f in Eq. (4.1) is selected from a set of functions $\mathcal{F}$. The result of varying the class prior is illustrated in Figure 4.1(b).

The bias caused by differing class balances can be systematically adjusted by instance reweighting or resampling if the class balance in the test dataset is known (Elkan, 2001; Lin et al., 2002).

However, the class ratio in the test dataset is often unknown in practice. A possible approach to mitigating this problem is to learn a classifier so that the performance for all possible class balances are improved, e.g., through maximization of the area under the ROC curve (Cortes and Mohri, 2004; Clémençon et al., 2009). Alternatively, in the minimax approach, a classifier is learned so as to minimize the worst-case performance for any change in the class prior (Duda et al., 2001; Van Trees, 1968). The disadvantage of the minimax approach is that it is often overly pessimistic. A more direct approach is to estimate the class ratio in the test dataset and use this estimate for instance reweighting or resampling. We focus on this scenario under a semi-supervised learning setup (Chapelle et al., 2006), where no labeled data is available from the test domain.

Saerens et al. (2001) is a seminal paper on this topic, which proposed to estimate the class ratio by the expectation-maximization (EM) algorithm (Dempster et al., 1977) – alternately updating the test class-prior and class-posterior probabilities from some initial estimates until convergence. This method has been successfully applied to various real-world problems such as word sense disambiguation (Chan and Ng, 2006) and remote sensing (Latinne et al., 2001).

In this chapter, we first reformulate the algorithm in Saerens et al. (2001), and show that this actually corresponds to approximating the test input distribution by a linear combination of class-wise training input distributions under the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). In this procedure, the class-wise input distributions are approximated via class-posterior estimation, for example, by kernel logistic regression (Hastie et al., 2001) or its squared-loss variant (Sugiyama, 2010b).

Since indirectly estimating the divergence by estimating the individual class-posterior distributions may not be the best scheme, the above reformulation motivates us to develop a more direct approach: matching the mixture of class-wise training input densities to the test input distribution. Historically, non-parametric estimation of the mixing proportions by matching the empirical distribution functions was investigated in Hall (1981), and its variant based on kernel density

estimation has been developed in Titterington (1983). However, these classical approaches do not perform well in high-dimensional problems (Sugiyama et al., 2013a). Recently, KL-divergence estimation based on *direct density-ratio estimation* has been shown to be promising (Nguyen et al., 2010b; Sugiyama et al., 2008b). Furthermore, a squared-loss variant of the KL divergence called the Pearson (PE) divergence (Pearson, 1900) can also be approximated in the same way, with an analytic solution that can be computed efficiently (Kanamori et al., 2009). Note that the PE divergence and the KL divergence both belong to the f-divergence class (Ali and Silvey, 1966; Csiszár, 1967), which share similar properties. In this chapter, with the aid of this density-ratio based PE-divergence estimator, we propose a new semi-supervised method for estimating the class ratio in the test dataset. Through experiments, we demonstrate the usefulness of the proposed method.

4.2 Problem formulation and existing method

In this section, we formulate the problem of semi-supervised class-prior estimation and review the existing method of Saerens et al. (2001).

4.2.1 Problem formulation

Let $\boldsymbol{x} \in \mathbb{R}^d$ be the d-dimensional input data, $y \in \{1, \ldots, c\}$ be the class label, and c be the number of classes. We consider class-prior change, i.e., the class-prior probability for training data $p_{\text{tr}}(y)$ and that for test data $p_{\text{te}}(y)$ are different. However, we assume that the class-conditional density for training data $p_{\text{tr}}(\boldsymbol{x}|y)$ and that for test data $p_{\text{te}}(\boldsymbol{x}|y)$ are the same:

$$p_{\text{tr}}(\boldsymbol{x}|y) = p_{\text{te}}(\boldsymbol{x}|y). \tag{4.2}$$

Note that training and test joint densities $p_{\text{tr}}(\boldsymbol{x}, y)$ and $p_{\text{te}}(\boldsymbol{x}, y)$ as well as training and test input densities $p_{\text{tr}}(\boldsymbol{x})$ and $p_{\text{te}}(\boldsymbol{x})$ are generally different under this setup.

For the purposes of classification, we are generally interested in selecting a classifier that minimizes the expected loss (or the risk) with respect to the test distribution. We can rewrite the test risk in terms of the training class conditional

density, $p_{\text{tr}}(\boldsymbol{x}|y)$, as

$$R = \mathbb{E}_{p_{\text{te}}}\left[L(f(\boldsymbol{x}), y)\right] \tag{4.3}$$

$$= \sum_{y} \int L(f(\boldsymbol{x}), y) p_{\text{te}}(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x}$$

$$= \sum_{y} \int L(f(\boldsymbol{x}), y) p_{\text{tr}}(\boldsymbol{x}|y) p_{\text{te}}(y) \mathrm{d}\boldsymbol{x}, \tag{4.4}$$

where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the loss function. Thus, if an estimate of the test class-priors is known, the expected loss can be calculated from the training class-conditional densities. The goal in this chapter is to estimate $p_{\text{te}}(y)$ from labeled training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ drawn independently from $p_{\text{tr}}(\boldsymbol{x}, y)$ and unlabeled test samples $\{\boldsymbol{x}'_i\}_{i=1}^{n'}$ drawn independently from $p_{\text{te}}(\boldsymbol{x})$[1]. Given test labels $\{y'_i\}_{i=1}^{n'}$, $p_{\text{te}}(y)$ can be naively estimated by n'_y/n', where n'_y is the number of test samples in class y. Here, however, we would like to estimate $p_{\text{te}}(y)$ *without* $\{y'_i\}_{i=1}^{n'}$.

4.2.2 Existing method

We give a brief overview of an existing method for semi-supervised class-prior estimation (Saerens et al., 2001), which is based on the expectation-maximization (EM) algorithm (Dempster et al., 1977).

In the algorithm, test class-prior and class-posterior estimates $\widehat{p}_{\text{te}}(y)$ and $\widehat{p}_{\text{te}}(y|\boldsymbol{x})$ are iteratively updated as follows:

1. Obtain an estimate of the training class-posterior probability, $\widehat{p}_{\text{tr}}(y|\boldsymbol{x})$, from training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, for example, by kernel logistic regression (Hastie et al., 2001) or its squared-loss variant (Sugiyama, 2010b).

2. Obtain an estimate of the training class-prior probability, $\widehat{p}_{\text{tr}}(y)$, from the labeled training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ as $\widehat{p}_{\text{tr}}(y) = n_y/n$, where n_y is the number of training samples in class y. Set the initial estimate of the test class-prior probability equal to it: $\widehat{p}^{0}_{\text{te}}(y) = \widehat{p}_{\text{tr}}(y)$.

[1] As we can confirm later, our proposed method does not actually require the independence assumption on $\{y_i\}_{i=1}^{n}$, but is valid for *deterministic* $\{y_i\}_{i=1}^{n}$ as long as $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ is drawn independently from $p_{\text{tr}}(\boldsymbol{x}|y = y_i)$. However, to remain consistent with other methods, we assume the independence condition here.

3. Repeat until convergence: $t = 1, 2, \ldots$

   (a) Compute a new test class-posterior estimate $\widehat{p}^{t}_{\text{te}}(y|\boldsymbol{x})$ based on the current test class-prior estimate $\widehat{p}^{t-1}_{\text{te}}(y)$ as

$$\widehat{p}^{t}_{\text{te}}(y|\boldsymbol{x}) = \frac{\widehat{p}^{t-1}_{\text{te}}(y)\widehat{p}_{\text{tr}}(y|\boldsymbol{x})/\widehat{p}_{\text{tr}}(y)}{\sum_{y'=1}^{c}\widehat{p}^{t-1}_{\text{te}}(y')\widehat{p}_{\text{tr}}(y'|\boldsymbol{x})/\widehat{p}_{\text{tr}}(y')}. \tag{4.5}$$

   (b) Compute a new test class-prior estimate $\widehat{p}^{t}_{\text{te}}(y)$ based on the current test class-posterior estimate $\widehat{p}^{t}_{\text{te}}(y|\boldsymbol{x})$ as

$$\widehat{p}^{t}_{\text{te}}(y) = \frac{1}{n'}\sum_{i=1}^{n'}\widehat{p}^{t}_{\text{te}}(y|\boldsymbol{x}'_i). \tag{4.6}$$

Note that Eq.(4.5) comes from the Bayes formulae,

$$p_{\text{tr}}(\boldsymbol{x}|y) = \frac{p_{\text{tr}}(y|\boldsymbol{x})p_{\text{tr}}(\boldsymbol{x})}{p_{\text{tr}}(y)} \text{ and } p_{\text{te}}(\boldsymbol{x}|y) = \frac{p_{\text{te}}(y|\boldsymbol{x})p_{\text{te}}(\boldsymbol{x})}{p_{\text{te}}(y)},$$

combined with Eq.(4.2):

$$p_{\text{te}}(y|\boldsymbol{x}) \propto \frac{p_{\text{te}}(y)}{p_{\text{tr}}(y)}p_{\text{tr}}(y|\boldsymbol{x}).$$

Eq.(4.6) comes from empirical marginalization of

$$p_{\text{te}}(y) = \int p_{\text{te}}(y|\boldsymbol{x})p_{\text{te}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

It was suggested that this procedure may converge to a local optimal solution (Saerens et al., 2001). In the following section, we will show that the objective function is actually convex, but that the method suggested in Saerens et al. (2001) may fail to converge to the unique optimal value.

4.3 Reformulation of the EM algorithm as distribution matching

In this section, we show that the class priors can be estimated by matching the test input density to a linear combination of class-wise training input distributions under the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). We show that the existing EM method performs this matching via an estimation of the class posterior. Furthermore, we show that this results in a convex problem, but that the existing EM method may not obtain the optimal result.

4.3.1 Class-Prior Estimation as Distribution Matching

Based on the assumption that the class-conditional densities for training and test data are unchanged (see Eq.(4.2)), let us model the test input density $p_{\text{te}}(\boldsymbol{x})$ by

$$q_{\text{te}}(\boldsymbol{x}) = \sum_{y=1}^{c} \theta_y p_{\text{tr}}(\boldsymbol{x}|y), \tag{4.7}$$

where θ_y is a coefficient corresponding to $p_{\text{te}}(y)$:

$$\sum_{y=1}^{c} \theta_y = 1. \tag{4.8}$$

We match the model $q_{\text{te}}(\boldsymbol{x})$ with the test input density $p_{\text{te}}(\boldsymbol{x})$ under the KL divergence:

$$\begin{aligned} \mathrm{KL}(p_{\text{te}} \| q_{\text{te}}) &:= \int p_{\text{te}}(\boldsymbol{x}) \log \frac{p_{\text{te}}(\boldsymbol{x})}{q_{\text{te}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}, \\ &= \int p_{\text{te}}(\boldsymbol{x}) \log p_{\text{te}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int p_{\text{te}}(\boldsymbol{x}) \log \left(\sum_{y=1}^{c} \theta_y p_{\text{tr}}(\boldsymbol{x}|y) \right) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{4.9}$$

We wish to select the class prior, under the constraint Eq.(4.8), that minimizes this KL divergence.

4.3.2 Equivalence of the EM method to divergence matching

When the KL divergence is minimized in Eq.(4.9), we can omit the term that is constant with respect to the class prior. This results in an optimization problem of

$$\begin{aligned}
\underset{\{\theta_y\}_{y=1}^c}{\arg\min}\ \mathrm{KL}(p_{\mathrm{te}}\|q_{\mathrm{te}}) &= \underset{\{\theta_y\}_{y=1}^c}{\arg\max} \int p_{\mathrm{te}}(\boldsymbol{x}) \log \left(\sum_{y=1}^{c} \theta_y p_{\mathrm{tr}}(\boldsymbol{x}|y) \right) \mathrm{d}\boldsymbol{x}, \\
&= \underset{\{\theta_y\}_{y=1}^c}{\arg\max} \int p_{\mathrm{te}}(\boldsymbol{x}) \log \left(p_{\mathrm{tr}}(\boldsymbol{x}) \sum_{y=1}^{c} \theta_y \frac{p_{\mathrm{tr}}(\boldsymbol{x}, y)}{p_{\mathrm{tr}}(\boldsymbol{x}) p_{\mathrm{tr}}(y)} \right) \mathrm{d}\boldsymbol{x}, \\
&= \underset{\{\theta_y\}_{y=1}^c}{\arg\max} \int p_{\mathrm{te}}(\boldsymbol{x}) \log p_{\mathrm{tr}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\qquad + \int p_{\mathrm{te}}(\boldsymbol{x}) \log \left(\sum_{y=1}^{c} \theta_y \frac{p_{\mathrm{tr}}(\boldsymbol{x}, y)}{p_{\mathrm{tr}}(\boldsymbol{x}) p(y)} \right) \mathrm{d}\boldsymbol{x}, \\
&= \underset{\{\theta_y\}_{y=1}^c}{\arg\max} \int p_{\mathrm{te}}(\boldsymbol{x}) \log \left(\sum_{y=1}^{c} \theta_y \frac{p_{\mathrm{tr}}(\boldsymbol{x}, y)}{p_{\mathrm{tr}}(\boldsymbol{x}) p_{\mathrm{tr}}(y)} \right) \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Approximating the expectation with its empirical average gives the following optimization problem:

$$\max_{\{\theta_y\}} \frac{1}{n'} \sum_{i=1}^{n'} \log \left(\sum_{y=1}^{c} \theta_y p_{\mathrm{tr}}(y|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y) \right), \tag{4.10}$$

subject to Eq.(4.8).

The above can be viewed as a convex problem since the concave $\log$ function is maximized and the constraints in Eq.(4.8) is linear. Therefore, the optimal points must satisfy the Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vandenberghe, 2004). The KKT conditions for the above problem is given by Eq.(4.8) and

$$\frac{1}{n'} \sum_{i=1}^{n'} \frac{p_{\mathrm{tr}}(y|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y)}{\sum_{y'=1}^{c} \theta_{y'} p_{\mathrm{tr}}(y'|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y')} = \nu, \quad \forall y = 1, \ldots, c, \tag{4.11}$$

where ν is a Lagrange multiplier. From these equations, we can determine ν as

$$\begin{aligned}\nu = 1 \cdot \nu &= \left(\sum_{y=1}^{c} \theta_y\right) \cdot \left(\frac{1}{n'}\sum_{i=1}^{n'} \frac{p_{\mathrm{tr}}(y|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y)}{\sum_{y'=1}^{c} \theta_{y'} p_{\mathrm{tr}}(y'|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y')}\right) \\ &= \frac{1}{n'}\sum_{i=1}^{n'} \frac{\sum_{y=1}^{c} \theta_y p_{\mathrm{tr}}(y|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y)}{\sum_{y'=1}^{c} \theta_{y'} p_{\mathrm{tr}}(y'|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y')} = 1.\end{aligned}$$

Then the solution $\{\theta_y\}_{y=1}^{c}$ can be calculated by fixed-point iteration as follows (McLachlan and Krishnan, 1997):

$$\theta_y \longleftarrow \theta_y \left(\frac{1}{n'}\sum_{i=1}^{n'} \frac{p_{\mathrm{tr}}(y|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y)}{\sum_{y'=1}^{c} \theta_{y'} p_{\mathrm{tr}}(y'|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y')}\right). \tag{4.12}$$

By using an estimator of the class-posterior, $\widehat{p}_{\mathrm{tr}}(y|\boldsymbol{x})$, in the above expression, we obtain an estimator for the test class-prior $\widehat{p}_{\mathrm{te}}(y)$. The above is actually the same as Eq.(4.6) with Eq.(4.5) substituted.

4.3.3 Fixed-point iteration

The unknown class-priors can therefore be obtained as the solution to the non-linear equation given by Eq.(4.11). A simple way to construct a solution to a non-linear equation is via a fixed-point iteration (as in Eq.(4.12)). For conciseness, we rewrite the fixed-point iteration as a mapping $T : \mathbb{R}^c \to \mathbb{R}^c$:

$$[T(\boldsymbol{\theta})]_y = \frac{1}{n'}\sum_{i=1}^{n'} \frac{\theta_y p_{\mathrm{tr}}(y|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y)}{\sum_{y'=1}^{c} \theta_{y'} p_{\mathrm{tr}}(y'|\boldsymbol{x}'_i)/p_{\mathrm{tr}}(y')}, \tag{4.13}$$

where $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \cdots \ \theta_c]^\top$ and $[\ \ ]_y$ denotes the yth component of a vector. The solution is then iteratively calculated as

$$\boldsymbol{\theta} \leftarrow T(\boldsymbol{\theta}),$$

until a fixed point $\boldsymbol{\theta} = T(\boldsymbol{\theta})$ is reached. Since the problem is convex, we would expect that there is a single unique fixed point. The *Banach fixed-point theorem* (also known as the *contraction mapping theorem*) (Hunter and Nachtergaele,

2001, p.62) guarantees a unique solution if T is a contraction mapping. T is a contraction mapping if

$$d\left(T(\boldsymbol{\theta}^j), T(\boldsymbol{\theta}^k)\right) < d\left(\boldsymbol{\theta}^j, \boldsymbol{\theta}^k\right), \quad \forall\, \boldsymbol{\theta}^j, \boldsymbol{\theta}^k \in \mathbb{R}^c, \tag{4.14}$$

where $d : \mathbb{R}^c \times \mathbb{R}^c \to \mathbb{R}$ is a metric.

However, we can actually show that Eq.(4.13) is *not* a contraction mapping. To explain this, we consider the counter example with vectors defined as

$$\left[\boldsymbol{\theta}^j\right]_i = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

where $1 \leq j \leq c$. By substituting this into Eq.(4.13), we obtain

$$T(\boldsymbol{\theta}^j) = \boldsymbol{\theta}^j, \qquad \forall j = 1, \ldots, c.$$

Therefore, for any two vectors $\boldsymbol{\theta}^j$ and $\boldsymbol{\theta}^k$, $j, k \in 1 \ldots c$, selected as above, $d\left(T(\boldsymbol{\theta}^j), T(\boldsymbol{\theta}^k)\right) = d\left(\boldsymbol{\theta}^j, \boldsymbol{\theta}^k\right)$. The condition in Eq.(4.14) is therefore violated, which means that T is not a contraction mapping. From this example, it is also immediately obvious that any $\boldsymbol{\theta}^j$ is a fixed point, but not necessarily the optimal value.

As shown above, the method of Saerens et al. (2001) can be regarded as solving a convex problem via fixed point iteration, but it may not result in the unique optimal value. These spurious optimal values is not a characteristic of the problem itself (which is convex), but due to solving the KKT conditions with a fixed-point iteration.

Spurious fixed points may be avoided by using several different initial values and then selecting the optimal value according to Eq.(4.10). Alternatively, the objective function Eq.(4.10) can be directly solved, e.g. through gradient descent and projection (Boyd and Vandenberghe, 2004). However, indirectly estimating the KL divergence via class-posterior estimation may not be the best scheme in practice.

4.4 Class-prior estimation by f-divergence matching

The analysis in the previous section motivates us to explore a more direct way to learn coefficients $\{\theta_y\}_{y=1}^c$. That is, given an estimator of a divergence from $p_{\text{te}}(\boldsymbol{x})$ to $q_{\text{te}}(\boldsymbol{x})$, coefficients $\{\theta_y\}_{y=1}^c$ are learned so that the divergence is minimized.

In this section, we first review a general framework for estimating the class prior via f-divergence matching (Ali and Silvey, 1966; Csiszár, 1967). We then review two specific methods of divergence estimation for the KL divergence and the Pearson (PE) divergence (Pearson, 1900). Finally, we propose to use the PE-divergence estimator for determining the coefficients $\{\theta_y\}_{y=1}^c$.

4.4.1 Framework for class-prior estimation

As discussed in the previous section, we wish to choose coefficients $\{\theta_y\}_{y=1}^c$, so that the model of the test input density,

$$q_{\text{te}}(\boldsymbol{x}) = \sum_{y=1}^{c} \theta_y p_{\text{tr}}(\boldsymbol{x}|y),$$

is the same as the input density $p_{\text{te}}(\boldsymbol{x})$. We wish to select $\{\theta_y\}_{y=1}^c$ so that the two distributions are the same under an f-divergence (Ali and Silvey, 1966; Csiszár, 1967)

$$\mathrm{D}_f(p_{\text{te}}\|q_{\text{te}}) := \int q_{\text{te}}(\boldsymbol{x}) f\left(\frac{p_{\text{te}}(\boldsymbol{x})}{q_{\text{te}}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x},$$

where the specific divergence is determined by the function $f(t)$. As discussed in Section 3.3.1, this f-divergence can be estimated as

$$D_f(p_{\text{te}}\|q_{\text{te}}) \geq \max_r \int p_{\text{te}}(\boldsymbol{x}) r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int q_{\text{te}}(\boldsymbol{x}) f^*(r(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}. \tag{4.15}$$

The above is a useful expression because the right-hand side only contains expectations of r and $f^*(r(\boldsymbol{x}))$, which can be approximated by sample averages. For a continuous f, the maximum is attained for a function r such that $p'(x)/q'(x) =$

$\partial f^*(r(x))$ (where ∂f^* is the derivative of f^*) (Nguyen et al., 2010a). Therefore, in contrast to the plug-in approach, the f-divergence is directly estimated in terms of the density ratio. This is intuitively advantageous since the estimation of densities is a more general problem than the estimation of a density ratio (Sugiyama et al., 2012b). Below, we show specific methods of divergence approximation for the KL and PE divergences under the model (4.7) and the following parametric expression of the density ratio $r(\boldsymbol{x})$:

$$r(\boldsymbol{x}) = \sum_{\ell=0}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}), \tag{4.16}$$

where $\{\alpha_\ell\}_{\ell=0}^b$ are parameters and $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=0}^b$ are basis functions. In practice, we use a constant basis and Gaussian kernels centered at the training data points, i.e., for $b = n$ and $\ell = 1, 2, \ldots, n$,

$$\varphi_0(\boldsymbol{x}) = 1 \quad \text{and} \quad \varphi_\ell(\boldsymbol{x}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_\ell\|^2}{2\sigma^2}\right).$$

The constant basis function is included since, if two distributions are equal, the density ratio would be $r(\boldsymbol{x}) = 1$. To prevent overfitting, we add a regularizer of the form $\lambda \boldsymbol{\alpha}^\top \boldsymbol{R} \boldsymbol{\alpha}$ to the objective function, where λ is a small constant, $\boldsymbol{R}$ is defined as

$$\boldsymbol{R} = \begin{bmatrix} 0 & \mathbf{0}_{1\times b} \\ \mathbf{0}_{b\times 1} & \boldsymbol{I}_{b\times b} \end{bmatrix}, \tag{4.17}$$

$\mathbf{0}_{a\times b}$ denotes the zero matrix of size $a \times b$, and $\boldsymbol{I}_{b\times b}$ denotes a $b \times b$ identity matrix. Since the regularizer should penalize non-smoothness, the constant basis function was not regularized. The model for the density ratio is then learned by the following regularized empirical maximization problem:

$$\max_{\{\alpha_\ell\}_{\ell=0}^b} \sum_{\ell=0}^{b} \frac{\alpha_\ell}{n'} \sum_{i=1}^{n'} \varphi_\ell(\boldsymbol{x}_i) - \sum_{y=1}^{c} \frac{\theta_y}{n_y} \sum_{i:y_i=y} f^*\left(\sum_{\ell=0}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i)\right) - \lambda \sum_{\ell=0}^{b}\sum_{\ell'=0}^{b} \alpha_\ell \alpha_{\ell'} R_{\ell,\ell'}. \tag{4.18}$$

The only remaining task in order to obtain an estimator of the class prior is to choose the function $f(t)$. The choice of $f(t)$ can be made by considering the sensitivity and robustness trade-off of f-divergences. Furthermore, the choice of $f(t)$ also affects the computational complexity of the resulting method.

4.4.2 KL-divergence approximation

For the reversed KL divergence, the function $f(t)$ and its conjugate $f^*(v)$ is

$$f(t) = -\log(t) \quad \text{and} \quad f^*(v) = -1 - \log(-v).$$

For the sake of convenience, we regard $-r(\boldsymbol{x})$ as $r(\boldsymbol{x})$ (Nguyen et al., 2010a). We can then write the empirical approximation of Eq.(4.15) under Eqs.(4.7) and (4.16) as follows (Nguyen et al., 2010a):

$$\mathrm{KL}\left(q_{\mathrm{te}} \| p_{\mathrm{te}}\right) \approx \max_{\{\alpha_\ell\}_{\ell=0}^{b}} -\frac{1}{n'}\sum_{i=1}^{n'}\sum_{\ell=0}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}'_i) + \sum_{y=1}^{c}\frac{\theta_y}{n_y}\sum_{i:y_i=y}\log\left(\sum_{\ell=0}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}_i)\right) + 1.$$

We note that when the above is used to estimate $\mathrm{KL}(p'\|q')$, the function $r(\boldsymbol{x})$ will be an estimate of the density ratio $q'(\boldsymbol{x})/p'(\boldsymbol{x})$. An alternative choice would be to estimate $\mathrm{KL}(q'\|p')$, which would lead to an estimate of $r(\boldsymbol{x}) = p'(\boldsymbol{x})/q'(\boldsymbol{x})$. The density $q'(\boldsymbol{x})$ may, however, have a small support for certain values of $\{\theta_y\}_{y=1}^{c}$, causing the density ratio $p'(\boldsymbol{x})/q'(\boldsymbol{x})$ to diverge. For this reason, the estimator $\mathrm{KL}(p'\|q')$, which estimates the density ratio $q'(\boldsymbol{x})/p'(\boldsymbol{x})$, is preferred.

The resulting regularized optimization problem,

$$\max_{\{\alpha_\ell\}_{\ell=0}^{b}} -\frac{1}{n'}\sum_{i=1}^{n'}\sum_{\ell=0}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}'_i) + \sum_{y=1}^{c}\frac{\theta_y}{n_y}\sum_{i:y_i=y}\log\left(\sum_{\ell=0}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}_i)\right) - \lambda\sum_{\ell=0}^{b}\sum_{\ell'=0}^{b}\alpha_\ell\alpha_{\ell'}R_{\ell,\ell'},$$

is convex and the solution can be obtained by naive optimization. The Gaussian width and regularization constant can be systematically optimized by cross-validation. The KL-divergence estimator obtained above was proved to possess

superior convergence properties both in parametric and non-parametric setups (Nguyen et al., 2010a; Sugiyama et al., 2008b).

However, in the current context of estimating the test class-priors, computing the KL-divergence estimator is rather time-consuming because optimization of $\{\alpha_\ell\}_{\ell=0}^{b}$ needs to be carried out for each $\{\theta_y\}_{y=1}^{c}$.

4.4.3 PE-divergence approximation

As an alternative to the KL divergence, let us consider the PE divergence defined by

$$\begin{aligned}\mathrm{PE}(q_{\mathrm{te}}\|p_{\mathrm{te}}) &:= \frac{1}{2}\int\left(\frac{q'(\boldsymbol{x})}{p'(\boldsymbol{x})}-1\right)^2 p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \frac{1}{2}\int\left(\frac{q'(\boldsymbol{x})}{p'(\boldsymbol{x})}\right)^2 p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\frac{1}{2},\end{aligned}$$

which is an f-divergence with

$$f(u)=\frac{u^2}{2}-\frac{1}{2}.$$

For this f, the convex conjugate is given by

$$f^*(v)=\frac{v^2}{2}+\frac{1}{2}.$$

The function $r(\boldsymbol{x})$ will again be an estimate of the ratio $q'(\boldsymbol{x})/p'(\boldsymbol{x})$. The empirical approximation of Eq.(4.15) under Eqs.(4.7) and (4.16) is given as follows (Kanamori et al., 2009):

$$\mathrm{PE}(q_{\mathrm{te}}\|p_{\mathrm{te}}) \approx \max_{\boldsymbol{\alpha}}\left[-\frac{1}{2}\boldsymbol{\alpha}^\top\widehat{\boldsymbol{G}}\boldsymbol{\alpha}+\boldsymbol{\alpha}^\top\widehat{\boldsymbol{H}}\boldsymbol{\theta}-\frac{1}{2}\right],$$

where

$$\boldsymbol{\alpha} = [\alpha_0 \ \alpha_1 \ \cdots \ \alpha_b]^\top, \quad \widehat{\boldsymbol{G}} = \frac{1}{n'}\sum_{i=1}^{n'} \boldsymbol{\varphi}(\boldsymbol{x}'_i)\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top,$$
$$\boldsymbol{\varphi}(\boldsymbol{x}) = [\varphi_0(\boldsymbol{x}) \ \varphi_1(\boldsymbol{x}) \ \cdots \ \varphi_b(\boldsymbol{x})], \ \widehat{\boldsymbol{H}} = \left[\widehat{\boldsymbol{h}}_1 \ \cdots \ \widehat{\boldsymbol{h}}_c\right],$$
$$\widehat{\boldsymbol{h}}_y = \frac{1}{n_y}\sum_{i:y_i=y} \boldsymbol{\varphi}(\boldsymbol{x}_i), \quad \boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \cdots \ \theta_c]^\top.$$

A regularized solution to the above maximization problem can be obtained analytically as

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{G}} + \lambda \boldsymbol{R}\right)^{-1} \widehat{\boldsymbol{H}}\boldsymbol{\theta}, \tag{4.19}$$

where the regularization matrix is defined in Eq.(4.17). The PE-divergence estimator obtained above was proved to have superior convergence properties both in parametric and non-parametric setups (Kanamori et al., 2009, 2012a). The kernel width and regularization parameter can be systematically optimized by cross-validation (Kanamori et al., 2009, 2012a).

4.4.4 Learning class ratios by PE divergence matching

As shown above, the KL and PE divergences can be systematically estimated without density estimation via Legendre-Fenchel convex duality. Among them, the PE-divergence estimator is more useful for our purpose of learning class ratios, because of the following reasons: The PE divergence was shown to be more robust against outliers than the KL divergence, based on power divergence analysis (Basu et al., 1998; Sugiyama et al., 2012a). This is a useful property in practical data analysis suffering high noise and outliers. Furthermore, the above PE-divergence estimator was shown to possess the minimum condition number among a general class of estimators, meaning that it is the most stable estimator (Kanamori et al., 2012b).

Another practically more important advantage of the PE-divergence estimator is that it can be computed efficiently and analytically. This analytical solution

allows us to express the PE divergence directly in terms of the class priors:

$$\widehat{\mathrm{PE}}(\boldsymbol{\theta}) := -\frac{1}{2}\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}^\top \left(\widehat{\boldsymbol{G}} + \lambda \boldsymbol{R}\right)^{-1} \widehat{\boldsymbol{G}} \left(\widehat{\boldsymbol{G}} + \lambda \boldsymbol{R}\right)^{-1} \widehat{\boldsymbol{H}}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}^\top \left(\widehat{\boldsymbol{G}} + \lambda \boldsymbol{R}\right)^{-1} \widehat{\boldsymbol{H}}\boldsymbol{\theta} - \frac{1}{2}.$$

The solution can then be obtained by minimizing the above expression with respect to $\boldsymbol{\theta}$.

4.4.5 Experiments

In this section, we report experimental results.

Benchmark datasets

The following five methods are compared:

- **EM-KLR**: The method of Saerens et al. (2001) (see Section 4.2.2). The class-posterior probability of the training dataset is estimated using ℓ_2-penalized kernel logistic regression with Gaussian kernels. The L-BFGS quasi-Newton implementation included in the 'minFunc' package is used for logistic regression training (Schmidt, 2005).

- **KL-KDE**: The estimator of the KL divergence $\mathrm{KL}(p'\|q')$ using kernel density estimation (KDE). The class-wise input densities are estimated by KDE with Gaussian kernels. The kernel widths are estimated using likelihood cross-validation (Silverman, 1986).

- **PE-KDE**: The estimate of the Pearson divergence $\mathrm{PE}(q'\|p')$ using KDE. The class-wise input densities are estimated by KDE with Gaussian kernels. The kernel widths are estimated using least-squares cross-validation (Silverman, 1986).

- **KL-DR**: The proposed method (see Section 4.4.2) using a KL-divergence estimator based on the density ratio (DR). For the optimization, the L-BFGS implementation 'minFunc' is used (Schmidt, 2005).

Table 4.1: Datasets used in the experiments. The SAHeart dataset was taken from Hastie et al. (2001). All other datasets were taken from the *LIBSVM* webpage (Chang and Lin, 2011).

Dataset	d	# samples	# positives	# negatives
Adult	123	32561	7841	24720
Australian	14	690	307	383
Diabetes	8	768	500	268
German	24	1000	300	700
Ionosphere	34	351	225	126
Ringnorm	20	7400	3664	3736
SAHeart	9	462	302	160
Statlog heart	13	270	120	150

- **PE-DR**: The proposed method (see Section 4.4.4) using the PE-divergence estimator based on DR.

Here, we use binary-classification benchmark datasets listed in Table 4.1. We select 10 samples from each of the two classes for the training dataset and 50 samples for the test dataset. The samples in the test set are selected with probability θ^* from the first class and with probability $(1-\theta^*)$ from the second class. The experiments are performed for several class-priors, selected as $\theta^* \in [0.1\, 0.2\, \ldots\, 0.8\, 0.9]$.

The squared error of the estimated class-priors averaged over 1000 runs are given in Figure 4.2. This shows that methods based on the KL and PE divergences overall outperform EM-KLR, implying that our reformulation of the EM algorithm as distribution matching (see Section 4.3) contributes to obtaining accurate class-ratio estimates. Among the divergence-based methods, PE-DR and KL-DR outperforms PE-KDE and KL-KDE, showing that directly estimating density ratios without density estimation is more promising as divergence estimators. Overall, PE-DR and KL-DR are shown to be the most accurate.

The average calculation time for the estimation of the class priors is given in Figure 4.3. From this, it can be seen that the speed of the PE-DR method is similar to the EM-KLR method and two orders of magnitude faster than the KL-DR method.

To illustrate how more accurate estimates of the class prior translate into higher classification accuracies, we train a classifier with the estimated prior. For the binary benchmark experiments, a weighted variant of the ℓ_2-regularized kernel

logistic regression classifier (Hastie et al., 2001) was used.

We minimize the prior-corrected expected loss of Eq.(4.4), where the expectation is approximated by its empirical average and the class priors are replaced by the estimated class-priors. Using the logistic loss, a classifier can be learned as,

$$\left(\widehat{\beta}_1, \ldots, \widehat{\beta}_n\right) := \arg\min \beta_1, \ldots, \beta_n \left[\sum_{y=1}^{2} \frac{\widehat{\theta}_y}{n_y} \sum_{i:y_i=y} L\left(z_i, \sum_{\ell=1}^{n} \beta_\ell K(\boldsymbol{x}_i, \boldsymbol{x}_\ell) \right) + \delta \sum_{\ell=1}^{n} \beta_\ell^2 \right],$$

where $L(z, f(\boldsymbol{x}))$ is the logistic loss defined as

$$L(z, f(\boldsymbol{x})) = \log\left(1 + \exp\left(-z f(\boldsymbol{x})\right)\right),$$

and the class labels $y \in \{1, 2\}$ are encoded as $z \in \{-1, 1\}$. The width of the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{x}')$ and the regularization parameter $\delta(\geq 0)$ are chosen by five-fold weighted cross-validation (Sugiyama et al., 2007) in terms of the misclassification error. The class label $\widehat{y}$ for the test input $\boldsymbol{x}$ is then estimated by

$$\widehat{y} = \begin{cases} 1 & \sum_{i=1}^{n} \widehat{\beta}_i K(\boldsymbol{x}, \boldsymbol{x}_\ell) < 0, \\ 2 & \text{otherwise.} \end{cases}$$

The results in Figure 4.4 show that, as expected, a more accurate estimate of the class prior tends to give a lower misclassification rate. Taking into account both the computation time and accuracy, the PE-DR method is overall the most promising method.

4.4.6 Real-world application

Finally, we demonstrate the usefulness of the proposed approach in a real-world problem of military vehicle classification from geophone recordings (Duarte and Hu, 2004). This is a three-class problem: two vehicle classes and a class of recorded noise. The features are 50-dimensional. In this vehicle classification task, class-prior change is inevitable because the types of vehicles passing through differ depending on time (e.g., day and night).

n samples are drawn from each of the classes for the training set, whereas 100

samples are drawn with probabilities $p = [0.6\ 0.1\ 0.3]$ from each of the classes for the test set. Due to the prohibitive computational cost, KL-DR was not included in this experiment.

In Figure 4.5, we plot the ℓ_2-distance between the true and estimated class-priors and the misclassification rate based on instance-weighted kernel logistic regression (Hastie et al., 2001) averaged over 1000 runs as functions of the number of training samples. As can be seen from the graphs, the performance of all methods improves as the number of training samples increases. Among the compared methods, PE-DR provides the most accurate estimates of the class prior and thus yields the lowest classification error.

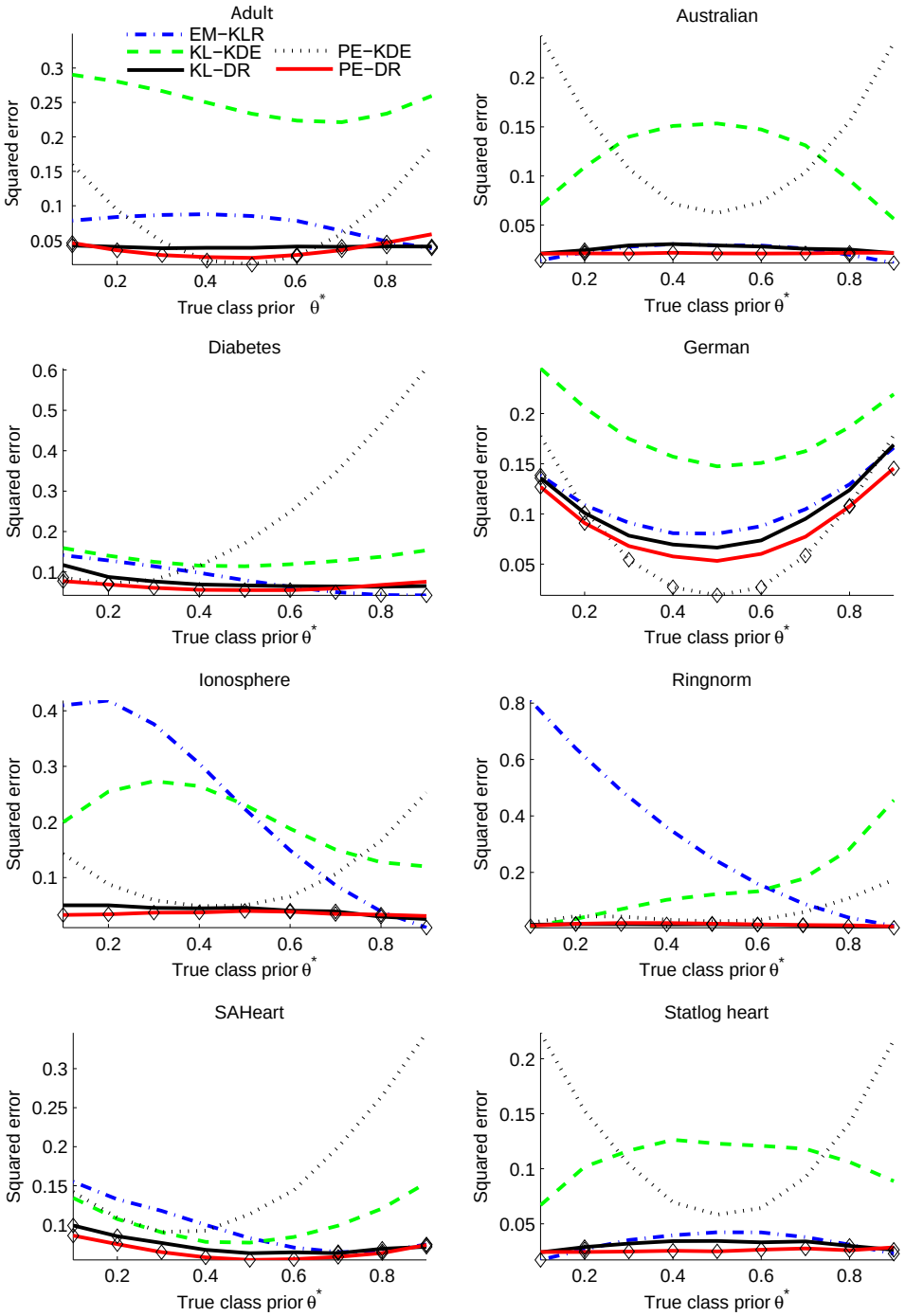


Figure 4.2: Average squared error between the true class-prior θ^* and estimated class-prior $\widehat{\theta}$ for the benchmark datasets listed in Table 4.1. The true class prior is indicated on the x-axis and the accuracy is indicated on the y-axis. The best method and comparable methods according to the t-test at significance level of 5% are indicated with '$\diamond$'

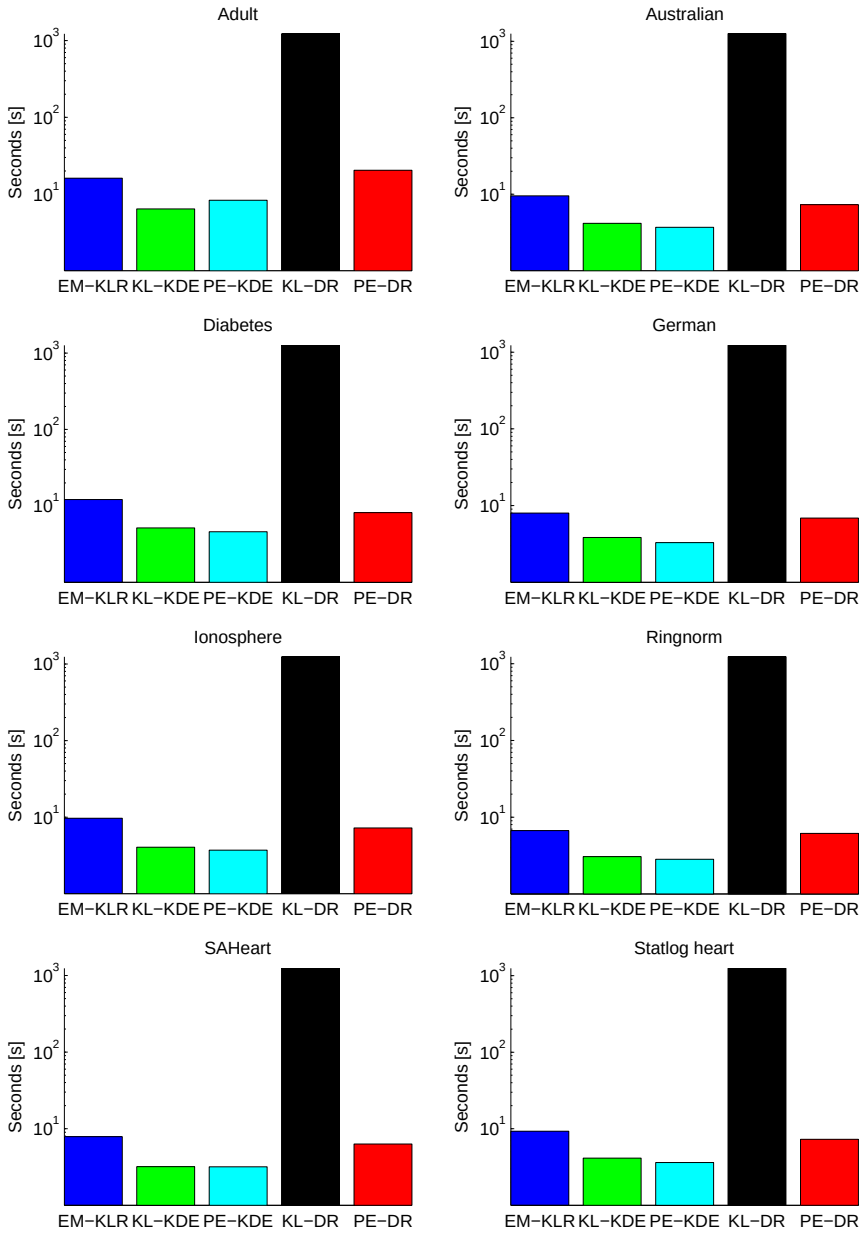


Figure 4.3: Average calculation time for the estimation of the class priors for the datasets listed in Table 4.1.

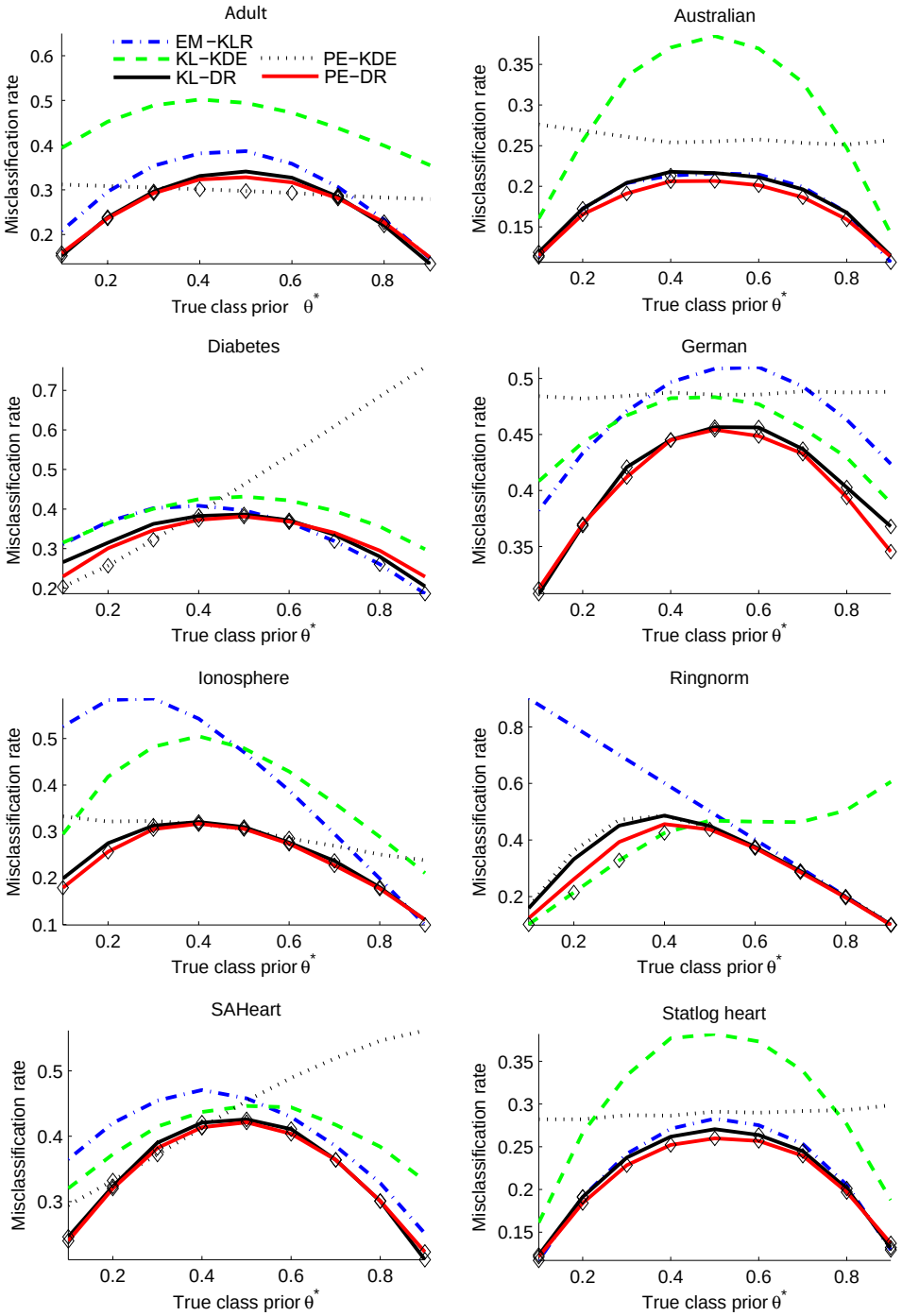
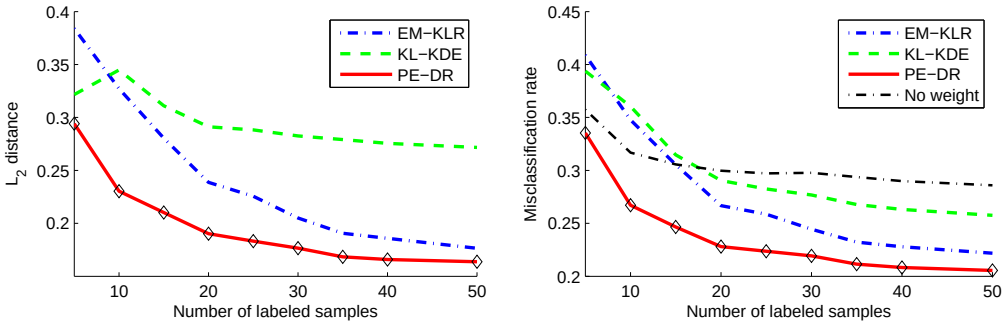


Figure 4.4: Average misclassification rates for the datasets listed in Table 4.1. Classification is performed using a regularized kernel logistic regression classifier with instance weighting. The true class prior is indicated on the x-axis and the resulting misclassification rate is indicated on the y-axis. The best method and comparable methods according to the t-test at significance level of 5% are indicated with '$\diamond$'.



(a) ℓ_2-distance between true and estimated class-priors.

(b) Misclassification rate with instance-weighted kernel logistic regression.

Figure 4.5: Experimental results for the vehicle classification problem. The best method and comparable methods according to the t-test at significance level of 5% are indicated with '$\diamond$'.

4.5 Class-prior estimation via L_2 distance matching

The main idea in class-prior estimation is to match the two densities $q_{\text{te}}(\boldsymbol{x})$ and $p_{\text{te}}(\boldsymbol{x})$ under some divergence. In the previous section, we used f-divergences to perform the matching and showed that it may be directly estimated in terms of the density ratio. A potential weakness of f-divergence estimation is that the density ratio may diverge.

In this section, we show that the class prior may be estimated by matching the distributions under the L_2 distance between probabilty densities. Furthermore, this is directly estimated in terms of the *density difference*, which is always bounded if the densities are bounded.

The L_2 distance between the two probability densities is then

$$\begin{aligned} L_2(p_{\text{te}}, q_{\text{te}}) &= \frac{1}{2}\int \left(p_{\text{te}}(\boldsymbol{x}) - q_{\text{te}}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}, \\ &= \frac{1}{2}\int \left(p_{\text{te}}(\boldsymbol{x}) - \sum_{y=1}^{c} \theta_y p_{\text{tr}}(\boldsymbol{x}|y)\right) \mathrm{d}\boldsymbol{x}. \end{aligned}$$

Then, using the squared-loss lower-bound discussed in Section 3.2 we obtain the following expression

$$L_2(p_{\text{te}}(\boldsymbol{x}), p_{\text{tr}}(\boldsymbol{x})) \geq \sup_{g} \int g(\boldsymbol{x}) \left[p_{\text{te}}(\boldsymbol{x}) - \sum_{y=1}^{c} \theta_y p_{\text{tr}}(\boldsymbol{x}|y)\right] \mathrm{d}\boldsymbol{x} - \frac{1}{2}\int g(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}.$$

The density difference is modeled with the following linear-in-parameter model,

$$g(\boldsymbol{x}) := \boldsymbol{\alpha}^\top \varphi(\boldsymbol{x}),$$

where the basis functions are the same as in Eq. (3.3). Using this model, the above expression can be written as

$$\begin{aligned} L_2(p_{\text{te}}(\boldsymbol{x}), p_{\text{tr}}(\boldsymbol{x})) \geq \sup_{\boldsymbol{\alpha}} \; & \boldsymbol{\alpha}^\top \int \varphi(\boldsymbol{x}) p_{\text{te}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \sum_{y=1}^{c} \theta_y \boldsymbol{\alpha}^\top \int \boldsymbol{\varphi}(\boldsymbol{x}) p_{\text{tr}}(\boldsymbol{x}|y) \mathrm{d}\boldsymbol{x} \\ & - \frac{1}{2}\boldsymbol{\alpha}^\top \left(\int \boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x}\right) \boldsymbol{\alpha} \end{aligned}$$

The following definitions can be used to simplify the above expression:

$$\boldsymbol{h}_{\text{te}} = \int \boldsymbol{\varphi}(\boldsymbol{x}) p_{\text{te}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \qquad \boldsymbol{h}_y = \int \boldsymbol{\varphi}(\boldsymbol{x}) p_{\text{tr}}(\boldsymbol{x}|y) \mathrm{d}\boldsymbol{x},$$
$$\boldsymbol{G} = [\boldsymbol{h}_1\, \boldsymbol{h}_2\, \ldots\, \boldsymbol{h}_c], \qquad \boldsymbol{H} = \int \boldsymbol{\varphi}(\boldsymbol{x}) \boldsymbol{\varphi}(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x}.$$

This leads to

$$L_2(p_{\text{te}}(\boldsymbol{x}), p_{\text{tr}}(\boldsymbol{x})) \geq \sup_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top \boldsymbol{h}_{\text{te}} - \boldsymbol{\alpha}^\top \boldsymbol{G}\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha}.$$

Empirical estimates can be used to approximate $\boldsymbol{h}_{\text{te}}$ and $\boldsymbol{h}_y$:

$$\widehat{\boldsymbol{h}}_{\text{te}} = \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \boldsymbol{\varphi}(\boldsymbol{x}'_i), \quad \text{and} \quad \widehat{\boldsymbol{h}}_y = \frac{1}{n_y} \sum_{i, y_i = y}^{n_{\text{tr}}} \boldsymbol{\varphi}(\boldsymbol{x}_i).$$

This results in a solution of:

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \left[\widehat{\boldsymbol{h}}_{\text{te}} - \widehat{\boldsymbol{G}}\boldsymbol{\theta} \right],$$

where the $\lambda \boldsymbol{I}$ term is due to the addition of an ℓ_2 regularization term. At this point, the above estimator is already useful. The estimate of the L_2 distance can be explicitly expressed as

$$\begin{aligned}\widehat{L}_2 (p_{\text{te}}, q_{\text{te}}) = {} & \widehat{\boldsymbol{h}}_{\text{te}}^\top (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}_{\text{te}} - 2 \widehat{\boldsymbol{h}}_{\text{te}}^\top (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{G}}\boldsymbol{\theta} \\ & + \boldsymbol{\theta}^\top \boldsymbol{G}^\top (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{G}}\boldsymbol{\theta} - \boldsymbol{\theta}^\top \boldsymbol{G}^\top (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \boldsymbol{H} (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \boldsymbol{G}\boldsymbol{\theta}.\end{aligned}$$

Minimizing this above expression with respect to $\boldsymbol{\theta}$ under the constraints $\boldsymbol{\theta} \succeq 0$ and $\boldsymbol{\theta}^\top \mathbf{1} = 1$ results in an estimate of the class prior.

4.5.1 Experiments

Four UCI benchmark datasets were used, where we randomly chose 10 labeled training samples from each class and 50 unlabeled test samples following true class prior $\theta^* = 0.1, 0.2, \ldots, 0.9$. The L_2 method was compared with the following methods:

- **KDEi:** Kernel density estimation (KDE) is used to approximate $p_{\mathrm{te}}(\boldsymbol{x})$ and $q_{\mathrm{te}}(\boldsymbol{x})$ from the data and then the L^2-distance is computed (Silverman, 1986). The Gaussian widths are independently chosen based on fivefold least-squares cross-validation (Titterington, 1983).

- **KDEj:** In the KDE-based method, two Gaussian widths are *jointly* chosen based on fivefold cross-validation int terms of the least-squares criterion (Hall and Wand, 1988). That is, the cross-validated least-squares criterion is computed as a function of two Gaussian widths and the best pair that minimizes the criterion is selected.

- **EM:** The class-prior estimation method based on the expectation maximization algorithm (Saerens et al., 2001).

The mean and standard error of the squared-error between the true and estimated class-balances over 1000 runs is plotted in Figure 4.6. From these graphs, it is clear that the L_2 method tends to provide better estimates of the class-balance than other approaches.

The classification accuracy was tested using a weighted ℓ_2 regularized least-squares classifier (Rifkin et al., 2003). The label $\widehat{y}$ for a test input $\boldsymbol{x}$ is estimated by

$$\widehat{y} = \mathrm{sign}\left(\sum_{\ell=1}^{n} \widehat{\beta}_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell)\right),$$

where $K(\boldsymbol{x}, \boldsymbol{x}')$ is the Gaussian kernel function with a kernel width κ. $\left\{\widehat{\beta}_\ell\right\}_{\ell=1}^{n}$ are the learnt parameters given by

$$\left(\widehat{\beta}_1, \ldots, \widehat{\beta}_n\right) := \underset{\beta_1,\ldots,\beta_n}{\arg\min} \left[\sum_{i=1}^{n} \frac{\theta_{y_i}}{n_{y_i}/n}\left(\sum_{\ell=1}^{n} \beta_\ell K(\boldsymbol{x}_i, \boldsymbol{x}_\ell) - y_i\right)^2 + \delta \sum_{\ell=1}^{n} \beta_\ell^2\right].$$

The Gaussian width κ and regularization parameter δ are chosen by weighted cross-validation.

The misclassification rate over 1000 runs are plotted on the right-hand side of Figure 4.6. The results show the LSDD-based method provides a lower misclas-

sification rate, which would be brought by good estimates of the test class prior.

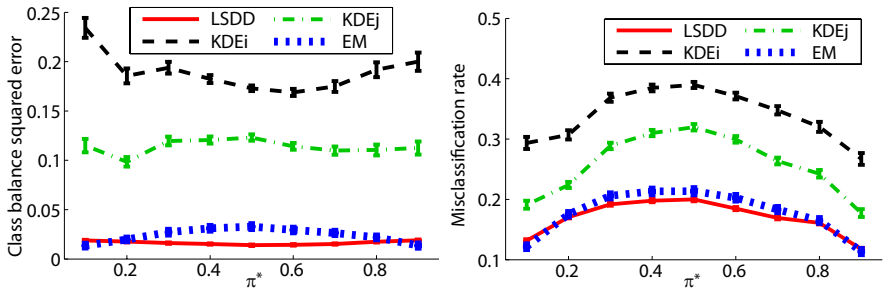
4.6 Conclusion

Class-prior change is a problem that is conceivable in many real-world datasets, and it can be systematically corrected for if the class prior of the test dataset is known. In this chapter, we discussed the problem of estimating the test class-priors under a semi-supervised learning setup.

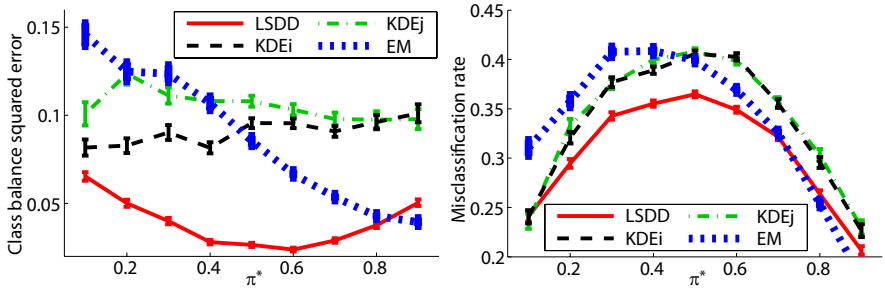
We first showed that the EM-based estimator introduced in Saerens et al. (2001) can be regarded as indirectly approximating the test input distribution by a linear combination of class-wise input distributions. From this viewpoint, there are two criticisms of the EM method: the fixed-point iteration does not necessarily lead to the unique global optimal value of a convex problem. More importantly, the problem estimates the divergence in an indirect two-step approach: the KL divergence is estimated based on the posterior. However, this posterior is not estimated with respect to the divergence.

Based on this view, we proposed to use explicit f-divergence estimators for learning test class-priors. Through experiments, we showed that the class ratios estimated by the proposed method are more accurate than competing methods, which can be translated into better classification accuracy.

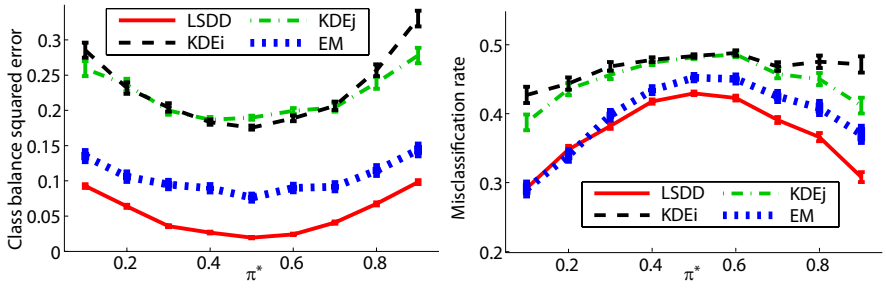
Alternatively to matching under f-divergences, the L_2-distance was used. This resulted in a simple estimator which gave better results than the EM methods and plug-in based approaches.



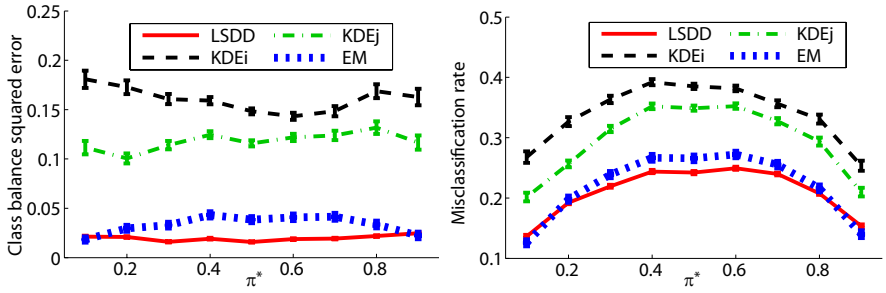
(a) Australian



(b) Diabetes



(c) German



(d) Statlogheart

Figure 4.6: Results of class-prior estimation via L_2 distance minimization. Left: Squared error of class-balance estimation. Right: Misclassification error by a weighted ℓ_2-regularized least-squares classifier with weighted cross-validation.

Chapter 5

Labeling data differing by class balance

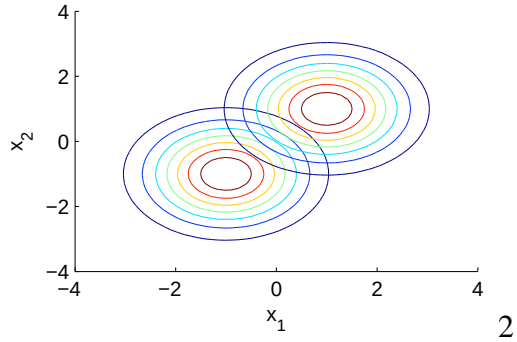
This chapter discusses the problem of labeling unlabeled samples from datasets with a differing class prior.

5.1 Introduction

Gathering labeled data is expensive and time-consuming in many practical machine learning problems, and therefore class labels are often absent. In this chapter, we consider the problem of *labeling*, which is aimed at giving a label to each unlabeled sample. Labeling is similar to, but slightly simpler than classification, because classes do not have to be specified. That is, labeling just attempts to split unlabeled samples into disjoint subsets, and class labels such as male/female or positive/negative are not assigned to samples.

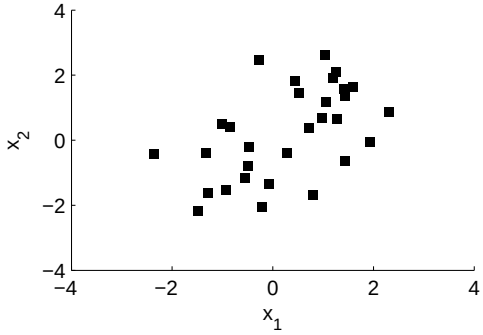
A naive approach to the labeling problem is to use a clustering technique, which is aimed at assigning a label to each sample of the dataset to divide the dataset into disjoint clusters. The tacit assumption in clustering is that the clusters correspond to the underlying classes. However, this assumption is often violated in practical datasets, for example, when clusters are not well separated or a dataset exhibits within-class multimodality.

An example of the labeling problem is illustrated in Figure 5.1. Figure 5.1(a) denotes the densities of the two classes. Figure 5.1(b) shows samples drawn from

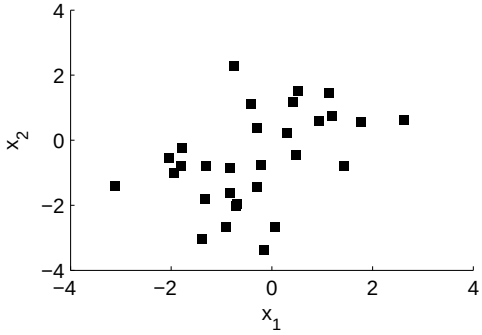


2

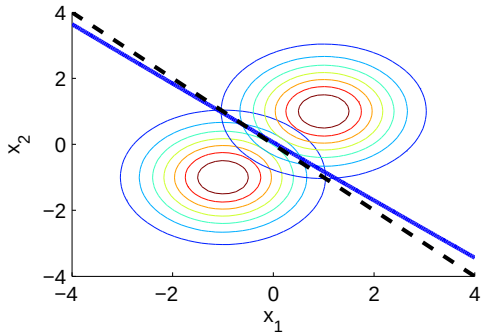
(a) Original class-conditional densities



(b) Samples of the first dataset



(c) Samples of the second dataset



(d) Optimal (dashed black) and estimated (solid blue) separation lines for the equal class-prior

Figure 5.1: Illustrative example of labeling. (a) Original class-conditional densities $p(\boldsymbol{x}|y)$. The bottom-left and top-right Gaussians correspond to class $y = 1$ and $y = -1$, respectively. (b) Samples of the first dataset $\mathcal{X}_p$ with $p(y = 1) = 0.3$. (c) Samples of the second dataset $\mathcal{X}_{p'}$ with $p'(y = 1) = 0.7$. (d) Optimal discriminant under the equal class-prior (dashed black) and the discriminant estimated by our proposed method (solid blue).

a mixture of the two original densities. Because the two clusters are highly overlapping, it may not be possible to properly label them via a clustering method.

In this chapter we show that if one more dataset with a different class balance is available (Figure 5.1(c)), we can obtain a discriminant for the equal class priors (Figure 5.1(d)). More specifically, we show that this labeling of the samples can be obtained by estimating *the sign of the difference between probability densities of two unlabeled datasets*. Thus, the challenge is to estimate the sign of the density difference as accurately as possible.

A naive way to estimate the sign of the density difference is to first separately estimate two densities from two sets of samples and then take the sign of their difference to obtain a labeling. However, this naive procedure violates *Vapnik's principle* (Vapnik, 2000):

> If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

This principle was successfully used in the development of *support vector machines* (SVMs): Rather than modeling two classes of samples, SVM directly learns a decision boundary that is sufficient for performing pattern recognition.

In the current context, estimating two densities is more general than labeling samples. Thus, the above naive scheme may be improved by estimating the density difference directly and then taking its sign to obtain the class labels. Recently, a method was introduced to directly estimate the density difference, called the *least-squares density difference (LSDD)* estimator (Sugiyama et al., 2013c). Thus, the use of LSDD for labeling is expected to improve performance.

However, the LSDD-based procedure is still indirect; directly estimating the sign of the density difference would be the most suitable approach to labeling. In this chapter, we show that the sign of the density difference can be directly estimated by lower-bounding the L_1-distance between probability densities.

Based on this, we give a practical algorithm for labeling and illustrate its usefulness through experiments on various real-world datasets.

The remainder of this chapter is structured as follows. In Section 5.2, we formulate the problem of labeling, give our fundamental strategy, and consider two naive approaches. In Section 5.3, we describe the detail of our proposed method. In Section 5.4, we experimentally investigate the behavior of the proposed method. Finally, in Section 5.5, we offer concluding remarks.

5.2 Problem formulation and fundamental approaches

In this section, we formulate the problem of labeling, give our fundamental strategy, and consider two naive approaches.

5.2.1 Problem formulation

Suppose that there are two joint probability distributions on $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \{1, -1\}$ with densities $p(\boldsymbol{x}, y)$ and $p'(\boldsymbol{x}, y)$, which are different only in class balances:

$$p(y) \neq p'(y) \qquad \text{but} \qquad p(\boldsymbol{x}|y) = p'(\boldsymbol{x}|y). \tag{5.1}$$

Here $p(y)$ and $p'(y)$ denote the marginal probabilities of y and $p(\boldsymbol{x}|y)$ and $p'(\boldsymbol{x}|y)$ denote the conditional densities of $\boldsymbol{x}$ given y, respectively. Figure 5.1(a) shows an example of class-conditional densities $p(\boldsymbol{x}|y)$ as two Gaussians with different means.

From these distributions, we are given two sets of unlabeled samples:

$$\mathcal{X}_p = \{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) \text{ and } \mathcal{X}_{p'} = \{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{\text{i.i.d.}}{\sim} p'(\boldsymbol{x}),$$

where $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ denote the marginal densities of $\boldsymbol{x}$. Figures 5.1(b) and 5.1(c) show examples of $\mathcal{X}_p$ and $\mathcal{X}_{p'}$ with class priors $p(y = 1) = 0.3$ and $p'(y = 1) = 0.7$, respectively.

The goal of labeling is to find the optimal discriminant for the equal class-prior (see Figure 5.1(d)). That is, for

$$q(y = 1) = q(y = -1) = \frac{1}{2},$$

our goal is to obtain the decision boundary such that

$$q(y = 1|\boldsymbol{x}) = q(y = -1|\boldsymbol{x}),$$

where

$$q(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)q(y)}{q(\boldsymbol{x})},$$
$$q(\boldsymbol{x}) = p(\boldsymbol{x}|y = 1)q(y = 1) + p(\boldsymbol{x}|y = -1)q(y = -1).$$

We note that, different from classification, labeling does not require correct class labels, but only correct class *separation* up to label commutation.

5.2.2 Fundamental strategy

To find the optimal decision boundary for the equal class-prior, let us consider a classifier

$$d(\boldsymbol{x}) = \operatorname{sign}\left[q(y = 1|\boldsymbol{x}) - q(y = -1|\boldsymbol{x})\right]. \tag{5.2}$$

First, we show the following lemma:

Lemma 5.1. *The classifier* (5.2) *can be rewritten as*

$$d(\boldsymbol{x}) = A \cdot \operatorname{sign}\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right], \tag{5.3}$$

where

$$A = \operatorname{sign}\left[p(y = 1) - p'(y = 1)\right].$$

Proof: We can write the difference between class-posteriors as

$$\begin{aligned} q(y = 1|\boldsymbol{x}) - q(y = -1|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|y = 1)\frac{1}{2}}{q(\boldsymbol{x})} - \frac{p(\boldsymbol{x}|y = -1)\frac{1}{2}}{q(\boldsymbol{x})} \\ &= \frac{1}{2q(\boldsymbol{x})}\left(p(\boldsymbol{x}|y = 1) - p(\boldsymbol{x}|y = -1)\right). \end{aligned}$$

Since $1/(2q(\boldsymbol{x}))$ is always positive, the criterion becomes

$$d(\boldsymbol{x}) = \operatorname{sign}\left[p(\boldsymbol{x}|y=1) - p(\boldsymbol{x}|y=-1)\right].$$

We do not have any labeled samples to calculate $p(\boldsymbol{x}|y=1) - p(\boldsymbol{x}|y=-1)$, but we can rewrite it in terms of marginal distributions. Indeed, we have

$$\begin{aligned} A\left[p(\boldsymbol{x}|y=1) - p(\boldsymbol{x}|y=-1)\right] &\propto \left[p(y=1) - p'(y=1)\right]\left[p(\boldsymbol{x}|y=1) - p(\boldsymbol{x}|y=-1)\right] \\ &= p(\boldsymbol{x}, y=1) - p'(\boldsymbol{x}, y=1) \\ &\quad - p(y=1)p(\boldsymbol{x}|y=-1) + p'(y=1)p(\boldsymbol{x}|y=-1). \end{aligned}$$

To write the third and fourth term as a joint distribution, we add and subtract $p(\boldsymbol{x}|y=-1)$, giving

$$\begin{aligned} A\left[p(\boldsymbol{x}|y=1) - p(\boldsymbol{x}|y=-1)\right] \propto\ & p(\boldsymbol{x}, y=1) - p'(\boldsymbol{x}, y=1) + \left[1 - p(y=1)\right] p(\boldsymbol{x}|y=-1) \\ & - \left[1 - p'(y=1)\right] p(\boldsymbol{x}|y=-1). \end{aligned}$$

Since $p(y=-1) = 1 - p(y=1)$ and $p'(y=-1) = 1 - p'(y=1)$, we can express the above as

$$q(y=1|\boldsymbol{x}) - q(y=-1|\boldsymbol{x}) \propto A\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right].$$

Substituting this into Eq.(5.2), we obtain Eq.(5.3). **(Q. E. D.)**

The expression (5.3) means that, if we know the class proportions in $\mathcal{X}_p$ and $\mathcal{X}_{p'}$, we can compute A and thus class labels can be obtained only from unlabeled samples. In practice, however, we may not know the class proportions and thus we can only split unlabeled samples into disjoint subsets that correspond to the original class labels.

Thus, now our challenge is to obtain a good estimator of the sign of density difference,

$$\operatorname{sign}\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right].$$

5.2.3 Kernel density estimation

A naive approach to estimating the sign of density-difference is to use *kernel density estimation* (KDE) (Silverman, 1986). For Gaussian kernels, the KDE solutions are given by

$$\widehat{p}(\boldsymbol{x}) \propto \sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma^2}\right),$$
$$\widehat{p}'(\boldsymbol{x}) \propto \sum_{j=1}^{n'} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'_j\|^2}{2\sigma'^2}\right).$$

The Gaussian widths σ and σ' may be determined based on least-squares cross-validation (Härdle et al., 2004). Finally, a labeling is obtained as

$$y = \operatorname{sign}[\widehat{p}(\boldsymbol{x}) - \widehat{p}'(\boldsymbol{x})].$$

5.2.4 Direct estimation of the density difference

KDE is a good estimator for the density, but it is not necessarily suitable for density-difference estimation, because small estimation errors incurred in each density estimate can cause a big error in the final density-difference estimate. More intuitively, good density estimators tend to be smooth and thus a density-difference estimator obtained from such smooth density estimators tends to be over-smoothed (Hall and Wand, 1988; Anderson et al., 1994).

The density difference can be estimated in a single shot using the *least-squares density difference* (LSDD) approach (Sugiyama et al., 2013c). In this approach, a model $g(\boldsymbol{x})$ is directly fitted to the density difference under the square loss:

$$\widehat{g} = \arg\min_{g} \frac{1}{2} \int \Big(g(\boldsymbol{x}) - [p(\boldsymbol{x}) - p'(\boldsymbol{x})]\Big)^2 \mathrm{d}\boldsymbol{x}.$$

The above square-loss for estimating the density difference is equivalent to the lower-bound that is used to estimate the L_2 distance. Therefore, the details and analytical form for estimating the density-difference is discussed in Section 3.2.2. When an estimate of the density difference $\widehat{g}(\boldsymbol{x})$ is obtained, a label can be as-

signed as

$$\widehat{y} = \widehat{g}(\boldsymbol{x}).$$

5.3 Direct estimation of the sign of the density difference

We expect that an improved solution can be obtained by LSDD over KDEs due to the more direct nature of LSDD. However, LSDD is still indirect because the sign of density difference is inspected after the density difference is estimated. In this section, we show how to directly estimate the sign of the density difference.

5.3.1 Derivation of the objective function

By lower-bounding the L_1-distance between probability densities, defined as

$$\int |p(\boldsymbol{x}) - p'(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x}, \tag{5.4}$$

we can obtain the sign of the density difference.

We begin by considering the following self-evident relation:

$$|t| \geq tz, \text{ if } |z| \leq 1.$$

We can apply this relation at each point $\boldsymbol{x}$, to obtain

$$|p(\boldsymbol{x}) - p'(\boldsymbol{x})| \geq g(\boldsymbol{x}) \, [p(\boldsymbol{x}) - p'(\boldsymbol{x})] \text{ if } |g(\boldsymbol{x})| \leq 1, \ \forall \boldsymbol{x}.$$

By applying the above inequality to Eq.(5.4) and maximizing with respect to $g(\boldsymbol{x})$, we can obtain the tightest lower bound as

$$\int |p(\boldsymbol{x}) - p'(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x} \geq \sup_{g} \int g(\boldsymbol{x}) \, [p(\boldsymbol{x}) - p'(\boldsymbol{x})] \, \mathrm{d}\boldsymbol{x} \tag{5.5}$$
$$\text{s.t. } |g(\boldsymbol{x})| \leq 1, \ \forall \boldsymbol{x}.$$

It is straightforward to verify that the above relation will be met with equality when

$$g(\boldsymbol{x}) = \mathrm{sign}\left(p(\boldsymbol{x}) - p'(\boldsymbol{x})\right).$$

What makes the right-hand side of the expression in Eq.(5.5) especially useful is that the probability densities occur linearly in the integral. By replacing the integrals (i.e., the expectations) with sample averages and searching $g(\boldsymbol{x})$ from a parametric family (denoted as $g_{\boldsymbol{\alpha}}(\boldsymbol{x})$), we can write the above as

$$\begin{aligned} \widehat{\boldsymbol{\alpha}} = \quad & \underset{\boldsymbol{\alpha}}{\arg\min} \quad \frac{1}{n'}\sum_{i=1}^{n'} g_{\boldsymbol{\alpha}}(\boldsymbol{x}'_i) - \frac{1}{n}\sum_{j=1}^{n} g_{\boldsymbol{\alpha}}(\boldsymbol{x}_j) \\ & \text{s.t.} \qquad |g_{\boldsymbol{\alpha}}(\boldsymbol{x})| \leq 1, \ \ \forall \boldsymbol{x}. \end{aligned} \tag{5.6}$$

5.3.2 Optimization

Here we briefly discuss how to solve the optimization problem in Eq.(5.6). The function in Eq.(5.6) should satisfy the constraint $|g(\boldsymbol{x})| \leq 1, \forall \boldsymbol{x}$. We can consider a clipped version of the function that always satisfies the constraint:

$$\widetilde{g}(\boldsymbol{x}) = R(g(\boldsymbol{x})), \ \ \text{where} \ \ R(z) = \begin{cases} 1 & z > 1, \\ -1 & z < -1, \\ z & \text{otherwise.} \end{cases}$$

We use a linear-in-parameter model,

$$g(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}), \tag{5.7}$$

where $\varphi_\ell(\boldsymbol{x})$ are the basis functions. Using the above definitions and including a regularizer, we arrive at the following objective function to be minimized:

$$J(\boldsymbol{\alpha}) = \frac{1}{n'}\sum_{i=1}^{n'} R\left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_i)\right) - \frac{1}{n}\sum_{j=1}^{n} R\left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j)\right) + \frac{\lambda}{2}\sum_{\ell=1}^{b} \alpha_\ell^2. \tag{5.8}$$

Although the above objective function is non-convex, a local minimizer can be efficiently found using the *convex-concave procedure* (CCCP) (Yuille and Rangarajan, 2002). We note that the above objective can be reformulated as the ramp loss. A ramp-loss version of the support vector machine (SVM) has been previously solved using the CCCP algorithm (Collobert et al., 2006). We use a set of linear basis functions (Eq.(5.7)) as our model instead of the kernel embedding of the ramp-loss SVM. This leads to a slightly different optimization problem that can be directly solved in the primal (instead of the ramp-loss SVM formulation, which is solved in the dual due to the kernel embedding).

CCCP requires the objective function to be split into convex and concave parts:

$$J(\boldsymbol{\alpha}) = J_{\mathrm{vex}}(\boldsymbol{\alpha}) + J_{\mathrm{cave}}(\boldsymbol{\alpha}).$$

This is done by expressing $R(z)$ as

$$R(z) = C_{-1}(z) - C_1(z) - 1,$$

where $C_\epsilon(z) = \max(0, z - \epsilon)$. This results in the following convex and concave functions:

$$J_{\mathrm{vex}}(\boldsymbol{\alpha}) = \frac{1}{n'}\sum_{i=1}^{n'} C_{-1}\left(\sum_{\ell=1}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}'_i)\right) + \frac{1}{n}\sum_{j=1}^{n} C_1\left(\sum_{\ell=1}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}_j)\right) + \frac{\lambda}{2}\sum_{\ell=1}^{b}\alpha_\ell^2,$$

$$J_{\mathrm{cave}}(\boldsymbol{\alpha}) = -\frac{1}{n'}\sum_{i=1}^{n'} C_1\left(\sum_{\ell=1}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}'_i)\right) - \frac{1}{n}\sum_{j=1}^{n} C_{-1}\left(\sum_{\ell=1}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}_j)\right).$$

We use the following *Fenchel inequality* (discussed in Section 3.3)

$$f(z) \geq tz - f^*(t),$$

where the *Fenchel dual* of the function $f(z)$ is defined as

$$f^*(t) = \sup_{y\in\mathbb{R}} yt - f(y).$$

Applying the above inequality to $C_\epsilon(z)$, we can obtain a bound as

$$C_\epsilon(z) \geq zt - C_\epsilon^*(t),$$

where $C_\epsilon^*(t)$ is the *Fenchel dual* of $C_\epsilon(z)$,

$$C_\epsilon^*(t) = \begin{cases} \epsilon t & 0 \leq t \leq 1, \\ \infty & \text{otherwise.} \end{cases}$$

Applying this to the concave part gives

$$J_{\mathrm{cave}}(\boldsymbol{\alpha}) \leq \bar{J}_{\mathrm{cave}}(\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{c}),$$

where the bound is specified by $\boldsymbol{b}$ and $\boldsymbol{c}$:

$$\begin{aligned} \bar{J}_{\mathrm{cave}}(\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{c}) =& \frac{1}{n'} \sum_{i=1}^{n'} \left(C_1^*(b_i) - b_i \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_i) \right) \\ &+ \frac{1}{n} \sum_{j=1}^{n} \left(C_{-1}^*(c_j) - c_j \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j) \right). \end{aligned}$$

This bound is convex w.r.t. $\boldsymbol{b}$ and $\boldsymbol{c}$ if $\boldsymbol{\alpha}$ is fixed. Using this bound, we have

$$J(\boldsymbol{\alpha}) \leq J_{\mathrm{vex}}(\boldsymbol{\alpha}) + \bar{J}_{\mathrm{cave}}(\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{c}).$$

The strategy to minimize $J(\boldsymbol{\alpha})$ is then to alternately minimize the right-hand side by minimizing w.r.t. $\boldsymbol{\alpha}$ (keeping $\boldsymbol{b}$ and $\boldsymbol{c}$ fixed) and minimize w.r.t. $\boldsymbol{b}$ and $\boldsymbol{c}$ (keeping $\boldsymbol{\alpha}$ fixed). Minimization w.r.t. $\boldsymbol{\alpha}$ minimizes the current upper bound and minimization w.r.t. $\boldsymbol{b}$ and $\boldsymbol{c}$ corresponds to tightening the bound at the current point.

Our final optimization algorithm is summarized below:

1. *Initialize the starting value:*

$$\boldsymbol{\alpha}^1 \leftarrow \arg\min_{\boldsymbol{\alpha}} J_{\mathrm{vex}}(\boldsymbol{\alpha}).$$

2. *For $t = 1, \ldots,$ until convergence:*

   (a) *Tighten the upper-bound:* Obtain $\boldsymbol{b}$ and $\boldsymbol{c}$ as

$$\boldsymbol{b}^t, \boldsymbol{c}^t \leftarrow \operatorname*{arg\,min}_{\boldsymbol{b},\boldsymbol{c}} \bar{J}_{\text{cave}}(\boldsymbol{\alpha}^t, \boldsymbol{b}, \boldsymbol{c}),$$

   which can be analytically performed as

$$b_i^t \leftarrow \begin{cases} 0 & \text{if } \sum_{\ell=1}^{b} \alpha_\ell^t \varphi_\ell(\boldsymbol{x}'_i) < 1, \\ 1 & \text{otherwise,} \end{cases}$$

$$c_j^t \leftarrow \begin{cases} 0 & \text{if } \sum_{\ell=1}^{b} \alpha_\ell^t \varphi(\boldsymbol{x}_j) < -1, \\ 1 & \text{otherwise.} \end{cases}$$

   (b) *Minimize the upper bound:* Set

$$\boldsymbol{\alpha}^{t+1} \leftarrow \operatorname*{arg\,min}_{\boldsymbol{\alpha}} J_{\text{vex}}(\boldsymbol{\alpha}) + \bar{J}_{\text{cave}}(\boldsymbol{\alpha}, \boldsymbol{b}^t, \boldsymbol{c}^t),$$

   which can be performed by solving the following convex quadratic problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & -\sum_{\ell=1}^{b} \alpha_\ell \left(\frac{1}{n'} \sum_{i=1}^{n'} b_i^t \varphi_\ell(\boldsymbol{x}'_i) + \frac{1}{n} \sum_{j=1}^{n} c_j^t \varphi_\ell(\boldsymbol{x}_j) \right) \\ & \qquad + \frac{1}{n'} \sum_{i=1}^{n'} \xi'_i + \frac{1}{n} \sum_{j=1}^{n} \xi_j + \frac{\lambda}{2} \sum_{\ell=1}^{b} \alpha_\ell^2 \\ \text{s.t.} \quad & \xi'_i \geq 0,\ \xi'_i \geq \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_i) + 1, \ \ \forall i = 1, \ldots, n' \\ & \xi_j \geq 0,\ \xi_j \geq \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j) - 1 \ \ \forall j = 1, \ldots, n. \end{aligned}$$

In practice, Gaussian kernels centered at the sample points in $\mathcal{X}_p$ and $\mathcal{X}_{p'}$ are chosen as the basis functions. All hyper-parameters are set by cross-validation. We call this proposed method *direct sign density difference (DSDD)* estimation.

5.3.3 Finite-sample error bounds

The conditional density $p(\boldsymbol{x}|y)$ is shared by the two training distributions $p(\boldsymbol{x}, y)$ and $p'(\boldsymbol{x}, y)$. Assume that $p(\boldsymbol{x}|y)$ is also shared by the test distribution $p_{\text{te}}(\boldsymbol{x}, y)$ and

$$p_{\text{te}}(\boldsymbol{x}, y) = \theta p(\boldsymbol{x}, y) + (1 - \theta)p'(\boldsymbol{x}, y), \tag{5.9}$$

where $0 \leq \theta \leq 1$, that is, $p_{\text{te}}(\boldsymbol{x}, y)$ is the convex combination of $p(\boldsymbol{x}, y)$ and $p'(\boldsymbol{x}, y)$. Integrating Eq.(5.9) w.r.t. $\boldsymbol{x}$ gives

$$p_{\text{te}}(y) = \theta p(y) + (1 - \theta)p'(y).$$

For the sake of conciseness, we use the following shorthand:

$$\begin{aligned}\pi_{\text{te}} &:= p_{\text{te}}(y = 1),\\ \pi &:= p(y = 1),\\ \pi' &:= p'(y = 1).\end{aligned}$$

If π_{te}, π, and π' are available for evaluation of the algorithm (not training), where π_{te} must be between π and π', θ can be computed as

$$\theta = \frac{\pi_{\text{te}} - \pi'}{\pi - \pi'}.$$

We consider a decision function of the form

$$g(\boldsymbol{x}) = \sum_{i=1}^{n+n'} \alpha_i k(\boldsymbol{x}, \boldsymbol{c}_i), \tag{5.10}$$

where k is a kernel function, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n+n'})$, and $\boldsymbol{c}_i = \boldsymbol{x}_i$ for $1 \leq i \leq n$ and $\boldsymbol{c}_i = \boldsymbol{x}'_{i-n}$ for $n + 1 \leq i \leq n + n'$. Let $\mathbb{E}[\cdot]$ and $\hat{\mathbb{E}}[\cdot]$ stand for the true expectation and the empirical expectation, $\ell(z)$ be the *indicator loss* such that

$$\ell(z) = \begin{cases} 0 & \text{if } z > 0, \\ 1 & \text{if } z \leq 0, \end{cases}$$

and $\ell_\eta(z)$ be the *surrogate loss* (Bartlett and Mendelson, 2002) such that

$$\ell_\eta(z) = \begin{cases} 0 & \text{if } z > \eta, \\ 1 - z/\eta & \text{if } 0 < z \leq \eta, \\ 1 & \text{if } z \leq 0. \end{cases}$$

Note that, for any $\eta > 0$, $\ell_\eta(z)$ is lower bounded by $\ell(z)$ and approaches $\ell(z)$ as η approaches zero.

We then have the following theorem and corollary:

Theorem 5.2. *Assume that*

$$\exists B_k > 0, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d, k(\boldsymbol{x}, \boldsymbol{x}') \leq B_k^2.$$

Let $\boldsymbol{\alpha}^*$ be an optimal solution to DSDD, $g(\boldsymbol{x})$ be the decision function defined in Eq.(5.10) *with parameter $\boldsymbol{\alpha}^*$, and*

$$B_{\mathcal{F}} = \sqrt{\boldsymbol{\alpha}^{*\top} \boldsymbol{K} \boldsymbol{\alpha}^*}, \quad B'_{\mathcal{F}} = \|\boldsymbol{\alpha}^*\|_1,$$

where $\boldsymbol{K}$ is the kernel matrix. Assume that the ground truth class labels $y_1, \ldots, y_n, y'_1, \ldots, y'_{n'}$ are available for evaluation. With probability at least $1 - \delta$, we have

$$\begin{aligned} \mathbb{E}_{p_{te}}[\ell(yg(\boldsymbol{x}))] \leq \frac{\theta}{n} \sum_{i=1}^{n} \ell_\eta(y_i g(\boldsymbol{x}_i)) + \frac{1-\theta}{n'} \sum_{i=1}^{n'} \ell_\eta(y'_i g(\boldsymbol{x}'_i)) \\ + \left(\frac{\theta}{\sqrt{n}} + \frac{1-\theta}{\sqrt{n'}} \right) \frac{2 B_k B_{\mathcal{F}}}{\eta} \\ + \left(\frac{\theta}{\sqrt{n}} + \frac{1-\theta}{\sqrt{n'}} \right) \min \left(3, 1 + \frac{4 B_k^2 B'_{\mathcal{F}}}{\eta} \right) \sqrt{\ln(2/\delta)/2}, \end{aligned} \tag{5.11}$$

where the expectation $\mathbb{E}_{p_{te}}[\ell(yg(\boldsymbol{x}))]$ follows the test distribution $p_{te}(\boldsymbol{x}, y)$.

Corollary 5.3. *Under the assumptions of Theorem 5.2, further assume that the ground truth class prior π_{te}, π, and π' are available for evaluation, where π_{te} is*

between π and π'. With probability at least $1-\delta$, we have

$$\begin{aligned}\mathbb{E}_{p_{te}}[\ell(yg(\boldsymbol{x}))] \leq & \frac{\pi_{\mathrm{te}}-\pi'}{(\pi-\pi')n}\sum_{i=1}^{n}\ell_\eta(y_i g(\boldsymbol{x}_i)) + \frac{\pi-\pi_{\mathrm{te}}}{(\pi-\pi')n'}\sum_{i=1}^{n'}\ell_\eta(y'_i g(\boldsymbol{x}'_i)) \\ & + \frac{(\pi-\pi_{\mathrm{te}})\sqrt{n}+(\pi_{\mathrm{te}}-\pi')\sqrt{n'}}{(\pi-\pi')\sqrt{nn'}}\cdot\frac{2B_k B_{\mathcal{F}}}{\eta} \\ & + \frac{(\pi-\pi_{\mathrm{te}})\sqrt{n}+(\pi_{\mathrm{te}}-\pi')\sqrt{n'}}{(\pi-\pi')\sqrt{nn'}}\min\left(3, 1+\frac{4B_k^2 B'_{\mathcal{F}}}{\eta}\right)\sqrt{\ln(2/\delta)/2}.\end{aligned} \tag{5.12}$$

Note that these bounds for unsupervised classification using DSDD can also be applied to unsupervised classification using LSDD.

From the above, we see that the order of the bounds is $O\left(1/\sqrt{n}+1/\sqrt{n'}\right)$. Compared to supervised classification from i.i.d. data such as support vector machines (Bartlett and Mendelson, 2002), which has an order of $O(1/\sqrt{n+n'})$, our bounds converge slower. However, we *do not require class labels for training* in our problem setting.

5.4 Experiments

We first illustrate the operation of our method and characterize the failures of other methods on various toy problems. Then we use real-world benchmark data to show the superiority of our algorithm.

5.4.1 Numerical illustration

We first illustrate the problem of labeling and our method with a simple example. Suppose that the class-conditional densities for the two classes are given as

$$\begin{aligned}p(\boldsymbol{x}|y=1) &= \mathcal{N}_{\boldsymbol{x}}\left(-\mathbf{1}_2, \boldsymbol{I}_{2\times 2}\right), \\ p(\boldsymbol{x}|y=-1) &= \mathcal{N}_{\boldsymbol{x}}\left(\mathbf{1}_2, \boldsymbol{I}_{2\times 2}\right),\end{aligned}$$

where $\mathcal{N}_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ w.r.t. $\boldsymbol{x}$. $\mathbf{1}_2$ is a 2×1 vector of ones and $\boldsymbol{I}_{2\times 2}$ is a 2×2 identity matrix. We generate 2 sets of 30 samples with class-priors $p(y=1)=0.3$ and $p'(y=1)=0.7$,

respectively. The result is illustrated in Figure 5.1. As can be seen from this example, we are able to obtain a discriminant that roughly corresponds to the true (unknown) one.

Another possible way to obtain a labeling of unlabeled samples is to use clustering. The tacit assumption in clustering is that samples in the same cluster belong to the same class. This assumption, however, is not always be true, for example, when the class conditional densities are multimodal. Next, we consider a problem with the following class conditional densities:

$$p(\boldsymbol{x}|y=1) = \frac{1}{2}\mathcal{N}_{\boldsymbol{x}}([3\ 0]^\top, \boldsymbol{I}_{2\times 2}) + \frac{1}{2}\mathcal{N}_{\boldsymbol{x}}([-3\ 0]^\top, \boldsymbol{I}_{2\times 2})$$
$$p(\boldsymbol{x}|y=-1) = \frac{1}{2}\mathcal{N}_{\boldsymbol{x}}([0\ 3]^\top, \boldsymbol{I}_{2\times 2}) + \frac{1}{2}\mathcal{N}_{\boldsymbol{x}}([0\ -3]^\top, \boldsymbol{I}_{2\times 2}).$$

The two distributions are plotted in Figure 5.2(a). We can try to obtain a class label by performing clustering on $\mathcal{X}_p \cup \mathcal{X}_{p'}$ [1]. The results for k-means (MacQueen, 1967) and spectral clustering (Shi and Malik, 2000), given in Figures 5.2(d) and 5.2(e), show that these methods fail to reveal the true labeling. On the other hand, the proposed method still gives a reasonable result (Figure 5.2(f)).

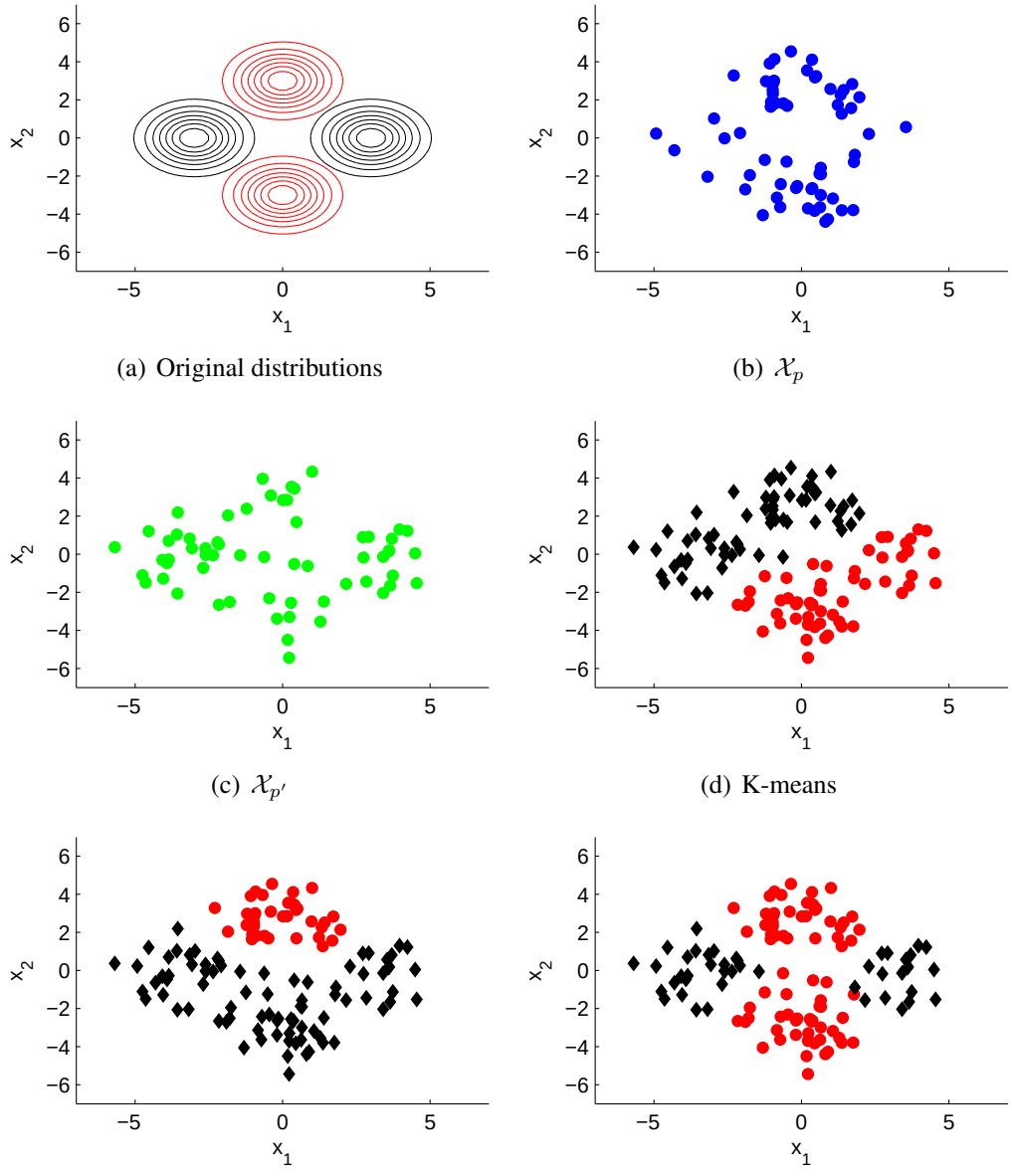
5.4.2 Benchmark datasets

We compared our method against several competing methods on benchmark datasets.

For each experiment, we constructed the datasets $\mathcal{X}_p$ and $\mathcal{X}_{p'}$ by drawing n and n' samples from the positive and negative classes of the binary classification datasets according to $p(y=1)$ and $p'(y=1)$. The labeling was then performed using these two datasets. A label was assigned to each sample according to the sign of the density difference.

Since we can obtain a labeling, but cannot determine the original class labels, we cannot measure the performance using the misclassification rate directly. As-

[1] If clustering is performed separately on $\mathcal{X}_p$ and $\mathcal{X}_{p'}$, we do not know which clusters in each dataset correspond to the clusters in the other dataset. We can also not perform clustering on one dataset and apply it to the other dataset, since most clustering methods do not give out-of-sample labeling. For these reasons, it makes most sense to perform clustering on the combined dataset.



(a) Original distributions (b) $\mathcal{X}_p$

(c) $\mathcal{X}_{p'}$ (d) K-means

(e) Spectral clustering (f) Direct estimation

Figure 5.2: Illustration of within-class multimodality and clustering.

sume that the label l_i is assigned to $\boldsymbol{x}_i$ and l'_j is assigned to $\boldsymbol{x}'_j$ as

$$l_i = \begin{cases} -1 & p(\boldsymbol{x}_i) - q(\boldsymbol{x}_i) < 0, \\ 1 & \text{otherwise.} \end{cases}$$

The misclassification rate (MCR) assuming that the current labels are correct is

$$\text{MCR} := \frac{1}{n} \sum_{i:l_i \neq y_i} 1 + \frac{1}{n'} \sum_{j:l'_j \neq y'_j} 1.$$

The misclassification rate assuming that the labels are the opposite is $1 - \text{MCR}$. We define the *labeling error rate* (LER) as

$$\text{LER} := \min\left(\text{MCR}, 1 - \text{MCR}\right).$$

We compared the following methods:

- **Direct Sign Density Difference (DSDD) estimation** (proposed): Directly estimate $\operatorname{sign}\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right]$ using the method described in Section 5.3. Hyperparameters are selected via cross-validation.

- **Least-Squares Density Difference (LSDD) estimation**: Estimate $\operatorname{sign}\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right]$ by estimating $p(\boldsymbol{x}) - p'(\boldsymbol{x})$ using the least squares fitting method (Sugiyama et al., 2013c). Hyperparameters are selected via cross validation.

- **Kernel Density Estimation (KDE):** Estimate $\operatorname{sign}\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right]$ by estimating the densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ with KDE. Hyperparameters are selected using least-squares cross validation.

- **K-Means++ (KM)**: Cluster the data into two clusters using the K-means algorithm (MacQueen, 1967). The algorithm was seeded according to Arthur and Vassilvitskii (2007).

- **Spectral Clustering (SC)**: Cluster the data into two clusters using the spectral clustering algorithm (Shi and Malik, 2000). The affinity matrix was constructed with 7 nearest neighbors.

- **Squared-loss Mutual Information based Clustering (SMIC)**: Cluster the data according to the SMIC method (Sugiyama et al., 2013a). SMIC was chosen since it provides model selection, avoiding the need for subjective parameter tuning.

We compare the performance of the methods by varying the class balance. Two class balances were selected: one with a large difference between the classes ($p(y = 1) = 0.2$ and $p'(y = 1) = 0.8$) and one with a small difference between the classes ($p(y = 1) = 0.35$ and $p'(y = 1) = 0.65$). The average and standard deviation of the labeling error rate for the two experiments, with $|\mathcal{X}_p| = |\mathcal{X}_{p'}| = 40$ are given in Tables 5.1 and 5.2.

Table 5.1: Labeling error rate for experiments with a class prior of $p(y = 1) = 0.2$ and $p'(y = 1) = 0.8$. The size of each dataset was $|\mathcal{X}_p| = 40$ and $|\mathcal{X}_{p'}| = 40$. The best method in terms of the mean error and comparable methods according to the two-sided paired t-test at the significance level of 5% are specified by bold face. The standard deviation of the labeling error rate is given in brackets.

Dataset	DSDD	LSDD	KDE	KM	SC	SMIC
australian	**.142**(.046)	.175(.109)	.211(.155)	.257(.145)	.379(.121)	.304(.109)
banana	.180(.094)	**.171**(.071)	.240(.152)	.432(.068)	.428(.141)	.425(.149)
diabetes	.246(.134)	**.223**(.080)	**.227**(.052)	.376(.088)	.381(.092)	.371(.115)
german	.268(.076)	.285(.135)	**.210**(.051)	.438(.153)	.447(.133)	.439(.062)
heart	**.175**(.050)	**.173**(.047)	.209(.046)	.257(.113)	.310(.038)	.324(.118)
image	**.197**(.078)	.206(.047)	.201(.123)	.387(.093)	.352(.121)	.382(.133)
ionosphere	**.157**(.059)	.182(.137)	.193(.128)	.339(.145)	.321(.061)	.312(.146)
saheart	.310(.109)	**.205**(.049)	.238(.116)	.425(.126)	.394(.137)	.384(.064)
thyroid	**.102**(.051)	.122(.113)	.206(.068)	.330(.114)	.327(.111)	.306(.094)
twonorm	.043(.086)	.051(.067)	.200(.029)	**.035**(.048)	.042(.071)	.048(.071)

From the results we see that methods which follow the approach proposed in Section 4.2 of estimating the sign of the density difference (i.e., DSDD, LSDD, and KDE) generally work better than methods using the cluster structure of the data (i.e., KM, SC, and SMIC). The thyroid dataset lends itself to interpretation of why these methods work better. The labels in the thyroid dataset correspond to healthy and diseased. The diseased label is caused by either a hyper-functioning

Table 5.2: Labeling error rate for experiments with a class prior of $p(y = 1) = 0.35$ and $p'(y = 1) = 0.65$. The size of each dataset was $|\mathcal{X}_p| = 40$ and $|\mathcal{X}_{p'}| = 40$. The best method in terms of the mean error and comparable methods according to the two-sided paired t-test at the significance level of 5% are specified by bold face. The standard deviation of the labeling error rate is given in brackets.

Dataset	**DSDD**	**LSDD**	**KDE**	**KM**	**SC**	**SMIC**
australian	**.245**(.115)	.260(.113)	.355(.086)	.256(.062)	.374(.083)	.306(.139)
banana	**.338**(.093)	**.336**(.100)	.367(.097)	.431(.057)	.428(.070)	.425(.074)
diabetes	**.339**(.076)	.359(.112)	.345(.033)	.374(.055)	.380(.038)	.371(.111)
german	.375(.045)	.381(.097)	**.354**(.057)	.438(.030)	.445(.062)	.437(.048)
heart	.271(.097)	**.248**(.085)	.353(.095)	.256(.062)	.315(.092)	.327(.116)
image	**.332**(.079)	.352(.066)	.350(.039)	.386(.031)	.353(.072)	.385(.036)
ionosphere	**.290**(.098)	.356(.070)	.345(.059)	.341(.070)	.322(.080)	.315(.089)
saheart	.377(.094)	**.353**(.057)	.362(.051)	.422(.058)	.395(.023)	.385(.039)
thyroid	**.225**(.099)	.251(.116)	.302(.042)	.331(.053)	.329(.038)	.307(.079)
twonorm	.160(.186)	.151(.119)	.352(.096)	**.033**(.043)	.041(.122)	.048(.120)

or hypo-functioning thyroid. These two underlying causes induce within-class multimodality (Sugiyama, 2007), which may cause clustering-based methods to fail.

Among the methods that estimate the sign of the density difference, we see that DSDD generally performs better than LSDD and LSDD, in turn, performs better than KDE. This is as expected since KDE solves a more general problem than LSDD, and LSDD solves a more general problem than DSDD. This pattern is even more pronounced on the more difficult case where the class balances are close to each other (Table 5.2).

5.5 Discussion and conclusion

In this chapter, the problem of unsupervised labeling of two unbalanced datasets was considered.

Since an estimate of the sign of the density difference is needed for solving the labeling, we introduced a method to directly estimate the sign of the density difference and avoid density estimation. We derive finite-sample error bounds which theoretically guarantee the converge of our solution to the optimal one with a reasonable rate. The method was shown on various datasets to outperform competing methods that either estimate the density difference or use the cluster structure of the data.

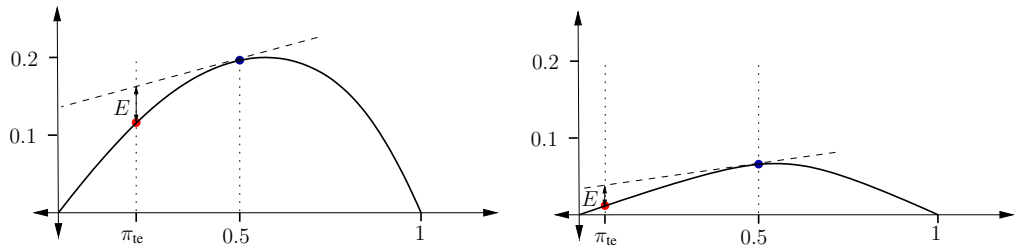
Because the sign of density difference corresponds to the Bayes optimal classifier under the equal class-prior, it may be estimated by any classifier that separates $\mathcal{X}_p$ and $\mathcal{X}_{p'}$. Following this idea, we tested the *support vector machine* (SVM) (Vapnik, 2000) for estimating the sign of density difference. However, this did not work well due to the high overlap of $\mathcal{X}_p$ and $\mathcal{X}_{p'}$—both the datasets are mixtures of two classes, only with different mixing ratios. From this classification point of view, we can actually see that our objective function (5.8) corresponds to the *robust SVM* (Shawe-Taylor and Cristianini, 2004) that minimizes the ramp loss (a clipped hinge loss). Thanks to the robustness brought by the ramp loss, the overlapped datasets $\mathcal{X}_p$ and $\mathcal{X}_{p'}$ can be separated more reliably, and thus we obtained a good estimation of the sign of density difference. Furthermore, this view conversely shows that the robust SVM is actually a suitable classification method

because it directly estimates the Bayes optimal classifier, the sign of density difference. Labeling and classification are different problems, but one can actually give insight into the other.

Since the priors of the dataset to be labeled is unknown, we decided to obtain a labeling for the equal class-prior. In this case, we showed that the discriminant can be obtained by estimating the sign of the density difference between the two datasets. The class priors may not be equal in many practical problems, so we wish to clarify what assumption on the data must be made to allow for an accurate labeling. The optimal expected loss (Bayes risk) according to a distribution $p(\boldsymbol{x}, y)$ with class prior π is

$$\mathbb{E}_p\left[\ell(g(\boldsymbol{x}))\right] = \min_g \pi \int \ell(yg(\boldsymbol{x}))p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} + (1-\pi)\int \ell(-g(\boldsymbol{x}))p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}.$$

where g can be an arbitrary function. From inspection it is obvious that the above function is concave with respect to π and passes through zero at $\pi = 0$ and $\pi = 1$ (see Figure 5.3). The expected loss of labeling using the optimal discriminant of the equal class-prior is given as the dashed line (which is tangent to the concave function at point $\pi = 0.5$). We define the "excess error" E as the difference between the risk with the equal class-prior and the optimal Bayes risk. This is the amount of extra error introduced to the Bayes error due to the assumption that the class priors are equal. We see that the excess error may be small when the true class-prior of the dataset is near 0.5 (Figure 5.3(a)) or when the different classes are well separated (Figure 5.3(b)). Datasets originating from distributions that satisfy these assumptions may be accurately labeled using the sign of the density difference.



(a) Test class-prior does not deviate far from the equal class-prior.

(b) Classes are separated well.

Figure 5.3: Risk curves for two hypothetical distributions. The discriminant is calculated according to the equal class-prior and applied to a test dataset with class prior π_{te}. E is the excess error due to classification with the equal class-prior.

Chapter 6

Prior estimation in positive-only labeled data

In this chapter, the problem of learning from positive-only labeled data is discussed.

6.1 Introduction

The standard assumption in supervised classification problems is that all training samples are labeled. In practical problems however, only partial labels may be available due to imperfect supervision. In this chapter, we consider the the problem of learning a classifier from positive-only labeled data. In this setting (illustrated in Figure 6.1) labels are only assigned to some (but not all) positive samples. Therefore, if a sample is not labeled, the underlying true label may be positive or negative. From the illustration we see that, unlike the fully-supervised classification setting, the class balance can not be directly estimated.

The goal is to obtain a classifier in order to assign the true class label to the underlying samples. This setting is often referred to as *semi-supervised novelty detection* (Blanchard et al., 2010), *inlier-based outlier detection* (Hido et al., 2008) or *learning from positive and unlabeled data* (Elkan and Noto, 2008).

This problem often occurs when a single class of interest must be separated from spurious or unwanted classes. For instance, in the land-cover identification task the user assigns labels only to samples of the class of interest (Li et al., 2011).

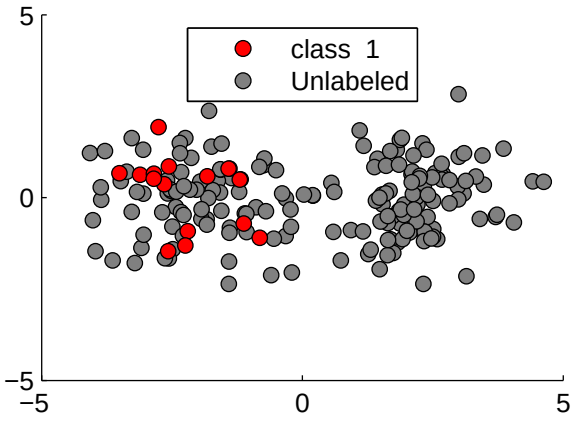


Figure 6.1: Illustration of the positive and unlabeled learning setting. Labels are only assigned to some examples from a single class. Unlabeled samples have a label of either $+1$ or -1.

A classifier should then be trained on the whole dataset in order to assign labels to the unlabeled samples.

In the next section, the problem is discussed in detail. We will see in that section that in order to train a classifier, an estimate for the class prior needs to be obtained.

6.2 Problem formulation

In this section the problem setting of learning from positive and unlabeled data is discussed. Following this, it is shown how, with a known class prior, classification in this setting can be performed.

6.2.1 Problem setting

We assume that data samples are drawn according to

$$(\boldsymbol{x}, y, s) \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y, s),$$

where $\boldsymbol{x}(\in \mathbb{R}^d)$ are the unlabeled features, $y(\in \{-1, 1\})$ are the (unknown) class labels, and $s(\in \{0, 1\})$ determines whether the sample is labeled or not. The assumption is that only positive samples are labeled (Elkan and Noto, 2008),

$$p(s = 1|\boldsymbol{x}, y = -1) = 0, \tag{6.1}$$

and that the probability that a sample is labeled depends only on the underlying label:

$$p(s = 1|\boldsymbol{x}, y = 1) = p(s = 1|y = 1). \tag{6.2}$$

Since we do not observe all class labels, the dataset would typically be

$$\mathcal{X} := \{(\boldsymbol{x}_i, s_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, s). \tag{6.3}$$

When $s_i = 1$, the sample $\boldsymbol{x}_i$ would have the label $y_i = 1$, according to the assumption in Eq. (6.1). When $s_i = 0$, the sample is unlabeled and the (unknown) underlying label may be $y_i = 1$ or $y_i = -1$.

We denote the subset of all labeled samples (i.e., $s_i = 1$) in $\mathcal{X}$ as

$$\mathcal{X}' = \{\boldsymbol{x}'_i\}_{i=1}^{n'}. \tag{6.4}$$

This dataset will therefore be drawn according to

$$p'(\boldsymbol{x}) = p(\boldsymbol{x}|y = 1)1 + p(\boldsymbol{x}|y = -1)0.$$

From this perspective, this is a special case of the problem in Chapter 5: We have two datasets $\mathcal{X}$ and $\mathcal{X}'$ that differ by class prior. However, here we wish to adjust the classifier to the class prior $p(y = 1)$ in the unlabeled test set $\mathcal{X}$.

Compared to a fully labeled dataset in the supervised learning case, this setting has an intrinsic problem: unlike the traditional classification setting, we can not trivially estimate the class prior $p(y = 1)$ from the dataset $\mathcal{X}$. The focus of this paper is to present a new method for estimating this class prior.

6.2.2 Classification

When the class prior $p(y = 1)$ is estimated or specified by the user, the classifier can be estimated from the dataset. The following lemma from Elkan and Noto (2008) holds:

Lemma 6.1. *The class posterior can be expressed as*

$$p(y = 1|\boldsymbol{x}) = \frac{1}{c} p(s = 1|\boldsymbol{x}), \tag{6.5}$$

where

$$c := p(s = 1|y = 1). \tag{6.6}$$

Proof: This lemma is proved as (Elkan and Noto, 2008) [1]

$$\begin{aligned} p(s = 1|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}, s = 1)}{p(\boldsymbol{x})} \\ &= \frac{1}{p(\boldsymbol{x})} \sum_{y} p(\boldsymbol{x}, s = 1, y) \\ &= \frac{1}{p(\boldsymbol{x})} \sum_{y} p(s = 1|\boldsymbol{x}, y) p(\boldsymbol{x}, y) \\ &= \frac{1}{p(\boldsymbol{x})} p(s = 1|\boldsymbol{x}, y = 1) p(\boldsymbol{x}, y = 1) \qquad (6.7) \\ &= p(y = 1|\boldsymbol{x}) p(s = 1|y = 1) \qquad (6.8) \end{aligned}$$

where (6.7) follows from the assumption in (6.1) and (6.8) follows from (6.2). **(Q. E. D.)**

The posterior $p(s = 1|\boldsymbol{x})$ is referred to in Elkan and Noto (2008) as a 'non-traditional' classifier. This can be estimated from the training set in (6.3) by a probabilistic classification method such as kernel logistic regression (Hastie et al., 2001) or its squared-loss variant (Sugiyama, 2010a).

The constant $c = p(s = 1|y = 1)$ that is used to reweight the non-traditional

[1] For the sake of clarity, the proof here is more verbose than in Elkan and Noto (2008, Lemma 1). The content however, is essentially the same.

classifier is unintuitive, but can be rewritten as,

$$\begin{aligned} p(s=1) &= \sum_y p(s=1,y) \\ &= \sum_y p(s=1|y)p(y) \\ &= p(s=1|y=1)p(y=1). \end{aligned} \tag{6.9}$$

Therefore, the constant weight is

$$\frac{1}{c} = \frac{p(y=1)}{p(s=1)}.$$

$p(s=1)$ can be directly estimated from (6.3), so the reweighting constant can be calculated if we can obtain an estimate of $p(y=1)$.

We propose a method in the next section to estimate this class prior from the training data. In Section 6.4 we show that the existing method of Elkan and Noto (2008) can be interpreted as indirectly estimating the same quantity as the proposed method. The superiority of our proposed method is illustrated on benchmark datasets in Section 6.5.

6.3 Prior estimation via partial matching

In this section we will propose a new method to estimate the class prior by partial matching.

6.3.1 Basic idea

From the assumptions, the set of labeled samples $\mathcal{X}'$ (defined in Eq. (6.4)) is drawn according to

$$p(\boldsymbol{x}|s=1) = p(\boldsymbol{x}|y=1). \tag{6.10}$$

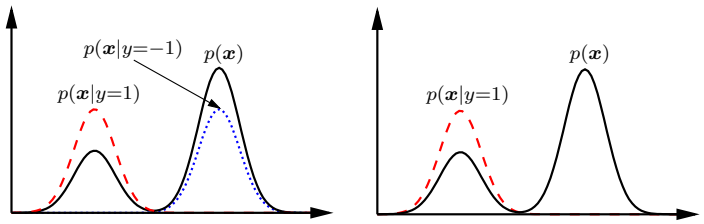


Figure 6.2: Estimating the class prior via *full matching* (left-hand side) and *partial matching* (right-hand side).

We model the input density as

$$q(\boldsymbol{x};\theta) := \theta p(\boldsymbol{x}|y=1) + (1-\theta)p(\boldsymbol{x}|y=-1),$$

where $\theta(\in [0,1])$ is a scalar value that represents the unknown class prior $p(y=1)$. The above model $q(\boldsymbol{x};\theta)$ would equal $p(\boldsymbol{x})$ if θ is the unknown class prior $p(y=1)$. Therefore, by selecting θ so that the two distributions are equal (illustrated in the left graph of Fig. 6.2), the class prior can be estimated as in Chapter 4. This setup will however not work in our current context, since we do not have samples drawn from $p(\boldsymbol{x}|y=-1)$ and consequently $q(\boldsymbol{x};\theta)$ can not be estimated.

Nevertheless, if the class-conditional densities $p(\boldsymbol{x}|y=1)$ and $p(\boldsymbol{x}|y=-1)$ are not strongly overlapping, we may estimate θ so that $\theta p(\boldsymbol{x}|y=-1)$ is as similar to $p(\boldsymbol{x})$ as possible (this is illustrated in the right graph of Fig. 6.2). Here we propose to use the Pearson (PE) divergence for matching $\theta p(\boldsymbol{x}|y=-1)$ to $p(\boldsymbol{x})$:

$$\theta^* = \arg\min_{\theta} \mathrm{PE}(\boldsymbol{\theta}),$$

where $\mathrm{PE}(\boldsymbol{\theta})$ denotes the PE divergence from $\theta p(\boldsymbol{x}|y=1)$ to $p(\boldsymbol{x})$[2]:

$$\begin{aligned}\mathrm{PE} &= \frac{1}{2}\int \left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})} - 1\right)^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\int \left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right)^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \theta + \frac{1}{2}. \end{aligned} \tag{6.11}$$

[2]Note that $\theta p(\boldsymbol{x})$ is not a density unless $\theta = 1$. This causes a small difference in the second line of the definition of the Pearson divergence (cf. Eq. (2.3))

The above PE divergence is defined in terms of unknown densities, and only samples drawn from these densities are available. A possible approach is to first estimate $p(\boldsymbol{x}|y=1)$ and $p(\boldsymbol{x})$ from the samples using, e.g., kernel density estimation and then plug these estimators into the above expression. This, however, does not work well since high-dimensional density estimation is a difficult problem (Vapnik, 2000). Furthermore, the division by an estimated density may exacerbate the estimation error.

6.3.2 Estimation algorithm

Here we show how we can avoid density estimation and directly minimize the PE divergence.

Our idea is to consider a lower bound which is linear in the unknown densities and can then be estimated from sample averages. Using the inequality $y^2/2 \geq ty - t^2/2$ we can lower bound (6.11) in a pointwise manner as follows[3]:

$$\frac{1}{2}\left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right)^2 \geq \left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right) r(\boldsymbol{x}) - \frac{1}{2} r(\boldsymbol{x})^2,$$

where $r(\boldsymbol{x})$ fulfills the role of t. This yields

$$\frac{1}{2}\left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right)^2 p(\boldsymbol{x}) \geq \theta p(\boldsymbol{x}|y=1) r(\boldsymbol{x}) - \frac{1}{2} r(\boldsymbol{x})^2 p(\boldsymbol{x}).$$

Therefore the PE divergence is lower bounded as

$$\mathrm{PE}(\theta) \geq \theta \int r(\boldsymbol{x}) p(\boldsymbol{x}|y=1) \mathrm{d}\boldsymbol{x} - \frac{1}{2} \int r(\boldsymbol{x})^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \theta + \frac{1}{2}.$$

The above lower bound can be turned into a practical estimator by using a parametric model for $r(\boldsymbol{x})$, replacing the integrals with sample averages, and selecting the tightest bound via maximization of the right-hand side.

We approximate the function $r(\boldsymbol{x})$ by a linear-in-parameter model $\widehat{r}(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$ are the parameters and $\boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \ldots, \varphi_n(\boldsymbol{x}))^\top$ are the basis functions. In practice, we use Gaussian basis functions centered at

[3]This inequality can be obtained via the squared-loss expansion (discussed in Section 3.2) or via Fenchel duality (discussed in Section 3.3).

the training points:

$$\varphi_i(\boldsymbol{x}) = \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{x}-\boldsymbol{x}_i\right\|^2\right), \quad i = 1, \ldots, n.$$

Using this model, we can rewrite the objective function as

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &:= \arg\max_{\boldsymbol{\alpha}} \theta\boldsymbol{\alpha}^\top \boldsymbol{h} - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha} - \theta + \frac{1}{2} \\ &= \arg\max_{\boldsymbol{\alpha}} \theta\boldsymbol{\alpha}^\top \boldsymbol{h} - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha},\end{aligned}$$

where

$$\boldsymbol{H} = \int \boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad \boldsymbol{h} = \int \boldsymbol{\varphi}(\boldsymbol{x}) p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}.$$

Estimating the integrals by their sample averages gives

$$\widehat{\boldsymbol{H}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i)\boldsymbol{\varphi}(\boldsymbol{x}_i)^\top, \quad \widehat{\boldsymbol{h}} = \frac{1}{n'}\sum_{i=1}^{n'} \boldsymbol{\varphi}(\boldsymbol{x}'_i).$$

Using these empirical estimates and adding an ℓ_2 regularizer leads to the following optimization problem:

$$\widehat{\boldsymbol{\alpha}} := \arg\max_{\boldsymbol{\alpha}} \theta\boldsymbol{\alpha}^\top \widehat{\boldsymbol{h}} - \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha},$$

where $\lambda(\geq 0)$ is the regularization parameter. This can be analytically solved as

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= \theta\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}}, \\ \widehat{\boldsymbol{G}} &= \widehat{\boldsymbol{H}} + \lambda\boldsymbol{I},\end{aligned}$$

where $\boldsymbol{I}$ denotes the identity matrix.

Substituting the analytical solution into the lower bound yields the following PE divergence estimator:

$$\widehat{\mathrm{PE}}(\theta) = \theta^2\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}} - \theta^2\frac{1}{2}\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{H}}\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}} - \theta + \frac{1}{2}.$$

This can be analytically minimized with respect to θ to yield the following estimator of the class prior:

$$\widehat{\theta} = \left[2\widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{H}}\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}}\right]^{-1}.$$

6.3.3 Theoretical analysis

Here, we theoretically investigate the bias of our algorithm when the assumption that class-conditional densities are non-overlapping is violated.

Assuming that the densities $p(\boldsymbol{x}|y=1)$ and $p(\boldsymbol{x})$ are known, we can analytically find the minimizer of (6.11) with respect to θ as

$$\theta = \left[\int \frac{p(\boldsymbol{x}|y=1)^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right]^{-1}. \tag{6.12}$$

Substituting the identity

$$p(\boldsymbol{x}|y=1) = \frac{p(\boldsymbol{x}) - (1-p(y=1))p(\boldsymbol{x}|y=-1)}{p(y=1)}$$

into the above gives

$$\theta = \frac{p(y=1)}{1-[1-p(y=1)]\int \frac{p(\boldsymbol{x}|y=1)p(\boldsymbol{x}|y=-1)}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}}.$$

If the class-conditional densities are completely non-overlapping, then $p(\boldsymbol{x}|y=1)p(\boldsymbol{x}|y=-1)=0$ and the estimator will be unbiased. Otherwise, we see that the value in the denominator is always smaller than 1, which means that the estimator will have a positive bias.

6.4 Analysis of existing method

In this section we analyze the method of estimating the class prior introduced in Elkan and Noto (2008). The paper proposed that a non-traditional classifier $g(\boldsymbol{x}) \approx p(s=1|\boldsymbol{x})$ is obtained from the training data. Using this classifier and a holdout set of positive samples P of size $|P|$, the constant c given by (6.6) is

estimated as

$$c \approx \frac{1}{|P|} \sum_{x \in P} g(\boldsymbol{x}). \tag{6.13}$$

Since samples in P are drawn from $p(\boldsymbol{x}|y = 1)$, (6.13) is essentially an estimate of

$$c = \int p(s = 1|\boldsymbol{x}) p(\boldsymbol{x}|y = 1) \mathrm{d}\boldsymbol{x},$$

where $p(s = 1|\boldsymbol{x})$ is estimated by a non-traditional classifier and the summation in (6.13) is due to estimation via an empirical average. Using (6.10) the above can be expressed as

$$c = \int \frac{p(\boldsymbol{x}|y = 1)p(s = 1)}{p(\boldsymbol{x})} p(\boldsymbol{x}|y = 1) \mathrm{d}\boldsymbol{x}.$$

Following from Eq. (6.6), the class prior is expressed as

$$p(y = 1) = \frac{p(s = 1)}{c} = \left[\int \frac{p(\boldsymbol{x}|y = 1)^2}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \right]^{-1},$$

which corresponds to Eq. (6.12).

Therefore, both methods can be viewed as estimating the class prior via PE divergence estimation. The important difference is how this estimation is performed. The existing method first learns a function g to estimate the posterior $p(s = 1|\boldsymbol{x})$ using a method such as kernel logistic regression. The PE divergence is then estimated using this function. This two-step approach may, however, not be optimal, since the best function estimated in the first step may not be the best for PE divergence estimation.

Our proposed method follows a single step approach: We directly learn a function based on how well the PE divergence is estimated. Therefore, our method is expected to perform better. We will experimentally investigate the superiority of our proposed approach in the next section.

6.5 Experiments

We compared the accuracy of the estimate of the class prior on several UCI benchmark datasets[4]. The following methods were compared:

- **PE** (proposed): The method described in Section 6.3 that directly estimates the PE divergence. All hyper-parameters was set using fivefold cross validation.

- **EN**: The method of Elkan and Noto (2008) discussed in Section 6.4. Data was split into 5 folds $\{\mathcal{X}_t\}_{t=1}^{5}$ and the posterior $p(s=1|\boldsymbol{x})$ was estimated from $\mathcal{X} \setminus \mathcal{X}_t$ (i.e., all samples except $\mathcal{X}_t$). The score in (6.13) was computed with $P = \mathcal{X}_t$. This was repeated for $t = 1, \ldots, 5$ and the average was used as the estimate of c.

The posterior $p(s=1|\boldsymbol{x})$ was estimated using kernel logistic regression (Hastie et al., 2001). The accuracy of the estimate of the class prior and the resulting classification accuracy are given in Fig. 6.3. The classification accuracy was computed by classifying an unlabeled hold-out dataset. As can be seen from the results, our proposed method gave a more accurate estimate of the class prior. Furthermore, the more accurate estimate of the class prior translated into a higher classification accuracy.

6.6 Conclusion

We proposed a new method to estimate the class prior from positive and unlabeled samples by partial matching under the PE divergence. By obtaining a lower bound for the PE divergence, we can directly get an analytical divergence estimator and estimate the class prior in a single step.

As was shown, the existing method of Elkan and Noto (2008) can also be interpreted as matching using the PE divergence. However, in that work, the estimation was indirectly performed using a two-step approach.

[4]The datasets can be obtained from '`http://archive.ics.uci.edu/ml/`'.

We experimentally illustrated on benchmark data that our single-step approach gave a more accurate estimate of the class prior, which, in turn, resulted in a higher classification accuracy.

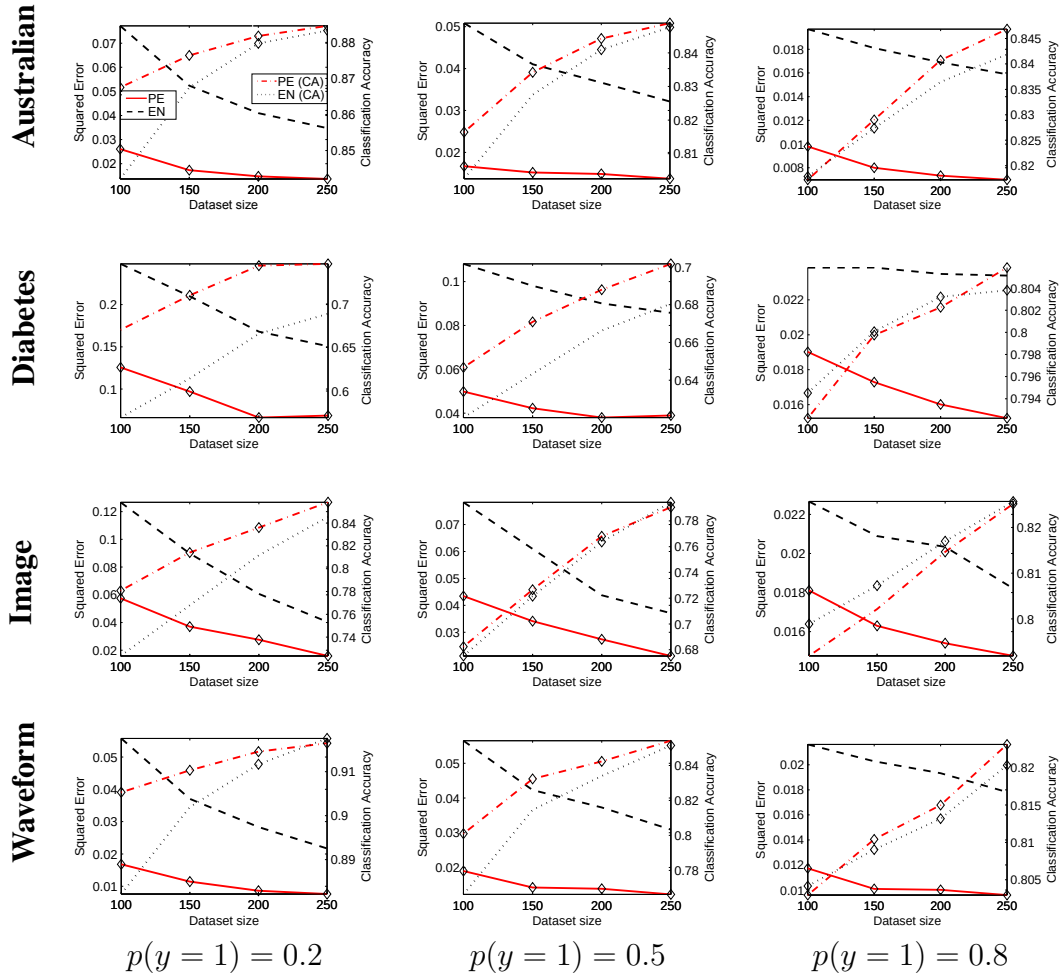


Figure 6.3: Experimental results on several UCI benchmark datasets. 'PE' and 'PE (CA)' indicates the squared-error and classification accuracy for class-prior estimation via direct PE divergence estimation. 'EN' and 'EN (CA)' indicates the squared error and classification accuracy for class-prior estimation using the method of (Elkan and Noto, 2008). The diamond symbol means that the method is the best or comparable in terms of the mean performance by t-test with significance level 5%.

Chapter 7

Conclusions and future work

In this chapter, a conclusion is presented and several future directions for research is discussed.

7.1 Conclusion

This thesis was devoted to situations where a class balance change occurs – either due to non-stationarity or imperfect supervision. Three such problems were investigated:

- Semi-supervised class-prior estimation in situations where class-prior change occurs.
- Unsupervised labeling of data using two unlabeled datasets differing by class prior.
- Class-prior estimation from positive-only labeled datasets.

These problems were solved using the framework of divergence estimation.

In Chapter 4 the semi-supervised class-prior estimation problem was considered. This problem was solved by matching a model based on the the training class-conditional densities to the test input density under some divergence measure. The f-divergences and the L_2 distance were investigated as divergence measures. Using the Pearson divergence and L_2 distance resulted in computationally

simple estimators of the class prior. Experimentally, these direct divergence estimators led to more accurate estimates of the class prior than existing indirect methods. Furthermore, these more accurate class-prior estimates, in turn, led to a higher classification accuracy when the classifier was adapted for a change in class priors.

In Chapter 5 the problem of clustering a dataset was considered. In this setup, it was assumed that two datasets with different class priors are available. It was shown that by directly estimating the sign of the density difference, a labeling can be obtained. Theoretical analysis revealed that the asymptotic convergence rate of our method is of the same order as the fully supervised case. However, our method does not require any labeled samples.

In Chapter 6 the problem of estimating the class-prior in positive-only labeled data was considered. To solve this problem, the idea of partially matching a scalar multiplied by a probability distribution to a probability distribution was introduced. Using the Pearson divergence, a simple and direct estimator for the class prior was obtained. Analysing the existing method with the framework introduced here showed that the existing method can also be interpreted as matching the Pearson divergence. However, in the existing method, this Pearson divergence was estimated in an indirect manner. Experimental results showed that the proposed method gives a more accurate estimate of the class-prior than the existing indirect method.

Although the proposed solutions for estimating the class prior in a semi-supervised case and from positive-only labeled data worked well, the proposed framework is perhaps more important. By posing existing machine learning problems as divergence minimization problems direct divergence estimators can be used. By directly estimating these divergences, algorithms that outperform existing indirect methods can be obtained.

7.2 Future problems

In this final section, we discuss several important problems for the future.

7.2.1 Investigation of density-difference divergences

A drawback of f-divergences is that they are defined in terms of the density ratio. This density ratio may not be bounded, making estimation of these divergences difficult in many instances. To solve this problem, a new class of divergences was introduced in Section 2.3 that is defined in terms of the density difference. These divergences are a generalization of the L_2 distance between probability distributions (Sugiyama et al., 2012c, 2013c).

There are several aspects of these divergences that still have to be investigated. The first is the identification of new functions $\psi(t)$ that define the divergence and the properties of these new functions. Secondly, the estimation aspect is important. For many functions, the last term in Eq. (3.6) can not be calculated analytically, making the estimation difficult.

The question of reducing variance is also important. For the L_2 distance between probability densities this aspect has already been investigated, leading to the *constrained least-squares density difference* formulation (Nguyen et al., 2012).

7.2.2 Reduction of bias in class-prior estimation from positive and unlabeled data

In Chapter 6 the problem of estimating the class-prior from positive and unlabeled data was discussed. The framework for estimating the class prior by partially matching a function to a density was introduced. We showed that the existing method can also be interpreted as matching distributions under a divergence. In that section, we showed that both methods are biased if the class-conditional densities significantly overlap.

A possible approach to reduce this bias is to consider other non-f-divergences as measures of similarity. The motivation for this is that we used an f-divergence to match a non-density to a density: $\theta p(\boldsymbol{x}|y=1)$ was matched to $p(\boldsymbol{x})$ under an f-divergence, but $\theta p(\boldsymbol{x}|y=1)$ is not a density (unless $\theta = 1$).

Following this line of reasoning, we consider the following penalized version

of the L_1 distance[1]

$$f(t) = \begin{cases} -(t-1) & 0 \leq t \leq 1, \\ \infty & t > 1, \\ \infty & t < 0. \end{cases}$$

The result for the ideal case (when the densities are known) is illustrated in Figure 7.1. As can be seen from this graph, the minimum for the penalized divergence is closer to the true value than the unpenalized case.

The penalized distance can be estimated from samples according to

$$\mathrm{D}(\theta) = \sup_{\boldsymbol{\alpha}} \quad \theta \frac{1}{n'} \sum_{j=1}^{n'} \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_j) - \frac{1}{n} \sum_{i=1}^{n} \max \left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i), -1 \right).$$

The max operation is due to the fact that the conjugate of $f(t)$ is $f^*(z) = \max(z, -1)$. An ℓ_2 regularizer can be added to the above with the dual purpose of regularization and ensuring that the problem is bounded. By calculating the dual, it can be shown that the above can be solved analytically.

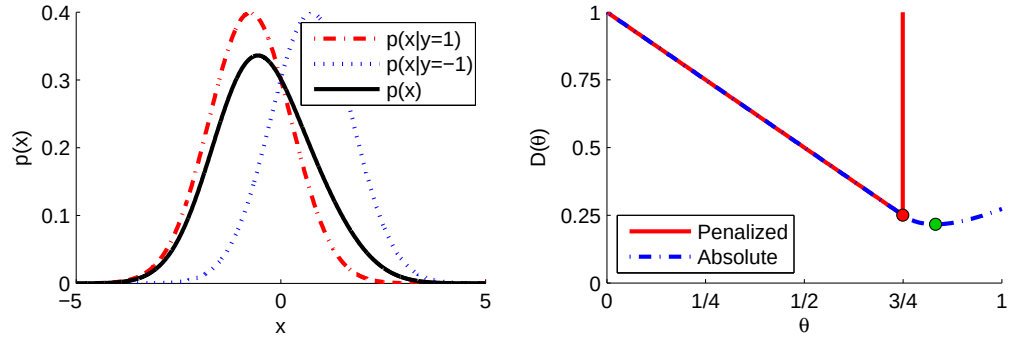
7.2.3 Statistical guarantees for class-prior estimation

An important aspect is investigating several statistical aspects of class-prior estimation. Currently, the divergence estimation for estimating the class-priors achieves the optimal convergence rate, however, no guarantees are given on the accuracy of the estimate of the class prior. The formulation of estimating the class-prior using the L_2-distance in Section 4.5 would perhaps be more amendable to such a statistical analysis due to the simplicity of the estimator.

7.2.4 Class-prior change model for time-series data

In this thesis a fixed change in the class-prior between the training and test phase was considered. However, many time series problems may be characterized by

[1]The L_1 distance was considered here for the sake of simplicity since it leads to a drastically simpler optimization problem.



(a) (b)

Figure 7.1: Example of partial matching for the L_1 distance and a penalized L_1 distance in a highly overlapped case. The true underlying densities are given in 7.1(a), where the unknown class prior is $p(y = 1) = \frac{3}{4}$. The L_1 distance and penalized distance for $\theta p(\boldsymbol{x}|y = 1)$ to $p(\boldsymbol{x})$ is given in 7.1(b). As can be seen from this, the bias of the penalized formulation is much smaller.

a continuous change in class priors. For example, the number of outliers in an industrial process may vary by time. Such a scenario can be modeled by a class prior that changes at each time step.

Assume that the sample $\boldsymbol{x}_t$ drawn at timestep t is drawn according to

$$\boldsymbol{x}_t \overset{\text{i.i.d.}}{\sim} p_t(\boldsymbol{x}),$$

where

$$p_t(\boldsymbol{x}) = p\left(\boldsymbol{x}|y = 1\right) p_t(y = 1) + p\left(\boldsymbol{x}|y = -1\right) \left[1 - p_t(y = 1)\right].$$

Comparing samples at different timesteps give,

$$p_t(\boldsymbol{x}) - p_{t-n}(\boldsymbol{x}) = \underbrace{\left[p_t(y = 1) - p_{t-n}(y = 1)\right]}_{\text{Depends on } t} \underbrace{\left[p\left(\boldsymbol{x}|y = 1\right) - p\left(\boldsymbol{x}|y = -1\right)\right]}_{\text{Depends on } \boldsymbol{x}}.$$

Therefore, the difference between densities factorizes such that one factor varies with t and the factor that varies with $\boldsymbol{x}$. By using an appropriate model of the class-prior and density difference, the time-varying class priors may be estimated.

Different variations on this theme of modeling problems based on a change in class priors can also be investigated.

7.2.5 Semi-supervised classification via small-support assumption

In many problems, for instance image-category identification, we wish to only recognize one specific class from among several different classes. For instance, we may have a large set of images, but only want to select images of the category 'cat' from that dataset. In the supervised setup, we can view this as a classification problem where the one class $y = 1$ consists of 'cat' training images and the other class $y = -1$ consists of images of all other classes (e.g. 'dog', 'car', etc...).

Unlabeled samples can be assigned labels by

$$\widehat{y} = \text{sign}\left[p(\boldsymbol{x}|y=1)p(y=1) - p(\boldsymbol{x}|y=-1)p(y=-1)\right]. \tag{7.1}$$

The above expression must be estimated from the dataset.

In practical problems, however, often only a small number of labeled samples may be available. The small number of samples may not be descriptive of the negative class. This is because the negative class consists of many different categories that are combined. An example of such a situation is illustrated in Figure 7.2.

Often we have a large unlabeled dataset in addition to a labeled dataset. To use this additional data, most semi-supervised learning methods make an assumption on the data such as the cluster assumption, smoothness assumption or manifold assumption (Chapelle et al., 2006). These assumptions may or may not hold depending on the dataset.

Since the unlabeled data drawn from $p(\boldsymbol{x})$ is available, a class label can be assigned with

$$\begin{aligned}\widehat{y} &= \text{sign}\left(2p(\boldsymbol{x}|y=1)p(y=1) - p(\boldsymbol{x})\right), \\ &= \text{sign}\left(2p(\boldsymbol{x}|y=1)p(y=1) - \left[p(\boldsymbol{x}|y=1)p(y=1) + p(\boldsymbol{x}|y=-1)p(y=-1)\right]\right).\end{aligned} \tag{7.2}$$

In situations where $p(\boldsymbol{x}|y=-1)$ is a combination of different underlying classes, estimating the above may be more accurate. To clarify the assumption: The as-

sumption is that the positive class-conditional density has a small support and the negative class-conditional density has a wide support.

This idea may be incorporated by combining the loss functions for estimating Eq. (7.2) and Eq. (7.2).

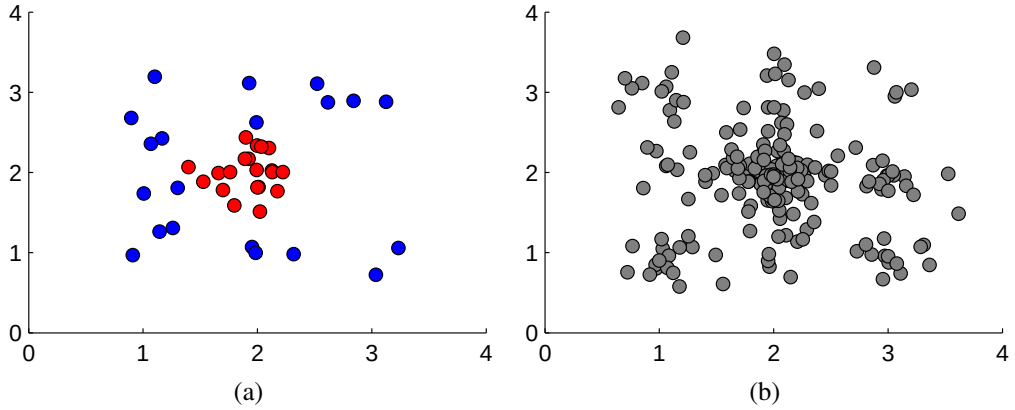


Figure 7.2: Example of the problem described in Sec. 7.2.5. Samples from $y = -1$ is from many different underlying classes and has a wide support.

Bibliography

S. M. Ali and S. D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28: 131–142, 1966.

N.H. Anderson, P. Hall, and D.M. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1): 41–54, 1994.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, 2006.

G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 9999:2973–3009, 2010.

K.M. Borgwardt, A. Gretton, M.J. Rasch, H.P Kriegel, B. Schölkopf, and A.J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787.

Y. S. Chan and H. T. Ng. Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics*, pages 89–96, 2006.

C.C. Chang and C.J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

S. Clémençon, N. Vayatis, and M. Depecker. AUC optimization and the two-sample problem. In *Advances in Neural Information Processing Systems 22*, pages 360–368. 2009.

R. Collobert, F.H. Sinz, J. Weston, and L. Bottou. Trading convexity for scalability. In *ICML*, pages 201–208, 2006.

C. Cortes and M. Mohri. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems 16*, pages pp. 313–320. MIT Press, Cambridge, MA, 2004.

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 442–450. 2010.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2: 229–318, 1967.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, 39(1):1–38, 1977.

M. F. Duarte and Y. H. Hu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(7):826–838, 2004.

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Wiley, New York, NY, USA, 2nd edition, 2001.

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

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220. ACM, 2008.

A. Gretton, K.M. Borgwardt, M. Rasch, B. Schölkopf, and A.J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

A. Gretton, K Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, Cambridge, MA, 2008.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012a.

A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu. Optimal kernel choice for large-scale two-sample tests. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1214–1222. 2012b.

P. Hall. On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 147–156, 1981.

P. Hall and M.P. Wand. On nonparametric discrimination using density differences. *Biometrika*, 75(3):541–547, 1988.

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and semiparametric models*. Springer, 2004.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, USA, 2001.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47 (1):153–161, 1979.

S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Inlier-based outlier detection via direct density ratio estimation. In F. Giannotti, D. Gunopulos,

F. Turini, C. Zaniolo, N. Ramakrishnan, and X. Wu, editors, *Proceedings of IEEE International Conference on Data Mining (ICDM2008)*, pages 223–232, Pisa, Italy, Dec. 15–19 2008.

J. Huang, A.J. Smola, A. Gretton, K.M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 601–608. MIT Press, Cambridge, MA, 2007.

J.K. Hunter and B. Nachtergaele. *Applied Analysis*. World Scientific Inc. Co., River Edge, NY, USA, 2001.

T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

T. Kanamori, T. Suzuki, and M. Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012a.

T. Kanamori, T. Suzuki, and M. Sugiyama. Computational complexity of kernel-based density-ratio estimation: A condition number analysis. *Machine Learning*, 2012b. , to appear.

M. Karasuyama and M. Sugiyama. Canonical dependency analysis based on squared-loss mutual information. *Neural Networks*, 34:46–55, 2012.

Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2012.

A. Keziou. Dual representation of ϕ-divergences and applications. *Comptes Rendus Mathématique*, 336(10):857–862, 2003.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

P. Latinne, M. Saerens, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: Evidence from a multi-class problem in remote sensing. In *Proceedings of the 18th International Conference on Machine Learning*, pages 298–305, 2001.

W. Li, Q. Guo, and C. Elkan. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2):717–725, 2011.

Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in non-standard situations. *Machine Learning*, 46(1):191–202, 2002.

S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, CA, USA, 1967. University of California Press.

G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley and Sons, New York, NY, USA, 1997.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):pp. 429–443, 1997. ISSN 00018678.

T.D. Nguyen, M.C. du Plessis, and M. Sugiyama. Constrained least-squares density-difference estimation. *Under review*, 2012.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010a.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010b.

G. Niu. *Discriminative Methods with Imperfect Supervision in Machine Learning*. PhD thesis, Tokyo Institute of Technology, Tokyo, Japan, 2012.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50: 157–175, 1900.

Q. Que and M. Belkin. Inverse density as an inverse problem: the fredholm equation approach. In *Advances in Neural Information Processing Systems 26*, pages 1484–1492. 2013.

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, USA, 2009.

R. Rifkin, G. Yeo, and T. Poggio. Regularized least-squares classification. *Advances in Learning Theory: Methods, Model and Applications. NATO Science Series III: Computer and Systems Sciences*, 190:131–153, 2003.

M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14: 21–41, 2001.

M. Schmidt. minFunc—Unconstrained differentiable multivariate optimization in MATLAB, 2005.

B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

B. Schölkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor, and J.C. Platt. Support vector method for novelty detection. In *NIPS*, volume 12, pages 582–588, 1999.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000. ISSN 0162-8828.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90 (2):227 – 244, 2000.

B. W. Silverman. *Density Estimation: For Statistics and Data Analysis*. Chapman and Hall, London, UK, 1986.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research*, 2:67–93, 2002.

M. Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, May 2007.

M. Sugiyama. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, E93-D(10):2690–2701, 2010a.

M. Sugiyama. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, E93-D:2690–2701, 2010b.

M. Sugiyama. Machine learning with squared-loss mutual information. *Entropy*, 15(1):80–112, 2013.

M. Sugiyama and M. Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, Cambridge, MA, USA, 2012.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, May 2007.

M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1433–1440, Cambridge, Massachusetts, USA, 2008a. MIT Press.

M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008b.

M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011a.

M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya. On information-maximization clustering: Tuning parameter selection and analytic solution. In L. Getoor and T. Scheffer, editors, *Proceedings of 28th International Conference on Machine Learning (ICML2011)*, pages 65–72, Bellevue, Washington, USA, Jun. 28–Jul. 2 2011b.

M. Sugiyama, T. Suzuki, and T. Kanamori. Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, 2012a. , to appear.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge, UK, 2012b.

M. Sugiyama, T. Suzuki, T. Kanamori, M. C. du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 692–700, 2012c.

M. Sugiyama, N. Gang, M. Yamada, M. Kimura, and H. Hachiya. Information-maximization clustering based on squared-loss mutual information. *Neural Computation*, 2013a. to appear.

M. Sugiyama, S. Liu, M. C. du Plessis, M. Yamanaka, M. Yamada, T. Suzuki, and T. Kanamori. Direct divergence approximation between probability distributions and its applications in machine learning. *Journal of Computing Science and Engineering*, 7(2):99–111, 2013b.

M. Sugiyama, T. Suzuki, T. Kanamori, M. C. du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. *Neural Computation*, 2013c. , to appear.

T. Suzuki and M. Sugiyama. Least-squares independent component analysis. *Neural Computation*, 23(1):284–301, 2011.

T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*, 3(25):725–758, 2013.

D.M. Titterington. Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 37–46, 1983.

H.L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. Detection, Estimation, and Modulation Theory. John Wiley and Sons, New York, NY, USA, 1968.

V. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science Series. Springer, 2000. ISBN 9780387987804.

V. Vapnik, I. Braga, and R. Izmailov. Constructive setting of the density ratio estimation problem and its rigorous solution. *arXiv preprint arXiv:1306.0407*, 2013.

V.N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

A. L. Yuille and Anand Rangarajan. The concave-convex procedure (CCCP). In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 114–, New York, NY, USA, 2004. ACM. ISBN 1-58113-838-5.