Classification with reject option

Radu Herbei and Marten H. Wegkamp *

June 21, 2005

Abstract

This paper studies two-class (or binary) classification of elements X in $\mathbb{R}^k$ that allows for a reject option. Based on n independent copies of the pair of random variables (X, Y) with $X \in \mathbb{R}^k$ and $Y \in \{0, 1\}$, we consider classifiers $f(X)$ that render three possible outputs: 0, 1 and R. The option R expresses doubt and is to be used for few observations that are hard to classify in an automatic way.

Chow (1970) derived the optimal rule minimizing the risk $\mathbb{P}\{f(X) \neq Y,\ f(X) \neq R\} + d\mathbb{P}\{f(X) = R\}$. This risk function subsumes that the cost of making a wrong decision equals 1 and that of utilizing the reject option is d. We show that the classification problem hinges on the behavior of the regression function $\eta(x) = \mathbb{E}(Y|X = x)$ near d and $1 - d$. (Here $d \in [0, 1/2]$ as the other cases turn out to be trivial.)

Classification rules can be categorized into plug-in estimators and empirical risk minimizers. Both types are considered here and we prove that the rates of convergence of the risk of any estimate depends on $\mathbb{P}\{|\eta(X) - d| \leq \delta\} + \mathbb{P}\{|\eta(X) - (1 - d)| \leq \delta\}$ and on the quality of the estimate for η or an appropriate measure of the size of the class of classifiers, in case of plug-in rules and empirical risk minimizers, respectively.

We extend the mathematical framework even further by differentiating between costs associated with the two possible errors: predicting $f(X) = 0$ whilst $Y = 1$ and predicting $f(X) = 1$ whilst $Y = 0$. Such situations are common in, for instance, medical studies where misclassifying a sick patient as healthy is worse than the opposite.

Running title: Classification with reject option

MSC2000 Subject classification: Primary 62C05 ; secondary 62G05, 62G08.

Keywords and phrases: Bayes classifiers, classification, empirical risk minimization, margin condition, plug-in rules, reject option.

*Research supported in part by the National Science Foundation under Grant DMS-0406049.

1 Introduction

Pattern recognition is about classifying an observation that takes values in some feature space $\mathcal{X}$ as coming from a fixed number of classes, say $0, 1, \ldots, M$. The simplest framework is that of binary classification ($M = 1$) with $\mathcal{X} = \mathbb{R}^k$. It is not assumed that an observation $X = x$ fully determines the label y; the same x may give rise to different labels. Based on a collection of labelled observations (x_i, y_i), the statistician's task is to form a classifier $f : \mathbb{R}^k \to \{0, 1\}$ which represents her guess of the label Y of a future observation X. This framework is known as supervised learning in the literature. The classifier

$$f(x) \quad = \quad \begin{cases} 0 & \text{if } \mathbb{P}\{Y = 0|X = x\} \geq \mathbb{P}\{Y = 1|X = x\} \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

has the smallest probability of error, see, for example, Devroye, Györfi, and Lugosi (1996, Theorem 2.1, page 10). In this paper the classifiers are allowed to report "I don't know" expressing doubt, if the observation x is too hard to classify. This happens when the conditional probability

$$\eta(x) = \mathbb{P}\{Y = 1|X = x\} \tag{2}$$

is close to $1/2$. Indeed, if

$$\mathbb{P}\{Y = 0|X = x\} = \mathbb{P}\{Y = 1|X = x\} = 1/2, \tag{3}$$

then we might just as well toss a coin to make a decision. The main purpose of supervised pattern recognition or machine learning is to classify the majority of future observations in an automatic way. However, allowing for the reject option ("I don't know") besides taking a hard decision (0 or 1) is of great importance in practice, for instance, in case of medical diagnoses. Nevertheless, this option is often ignored in the statistical literature. Chow (1970), Ripley (1995) and recently Freund, Mansour, and Schapire (2004) are notable exceptions. Some references in the engineering literature are Fumera and Roli (2003), Fumera, Roli, and Giacinto (2000), Golfarelli, Maio, and Maltoni (1997), and Hansen, Liisberg, and Salomon (1997).

Chow (1970), see Ripley (1995, Chapter 2) for a more general overview, put forth the decision theoretic framework. Let $f : \mathbb{R}^k :\to \{0, 1, R\}$ be a classifier with a reject option, where the interpretation of the output R is of being in doubt and taking no decision. The misclassification probability is

$$\mathbb{P}\{f(X) \neq Y, f(X) \neq R\}$$

and reject or doubt probability is

$$\mathbb{P}\{f(X) = R\}.$$

Assuming that the cost of making a wrong decision is 1 and that of utilizing the reject option is $d > 0$, the appropriate risk function to employ is

$$d\mathbb{P}\{f(X) = R\} + \mathbb{P}\{f(X) \neq Y, f(X) \neq R\}. \tag{4}$$

Chow (1970) shows that the optimal rule minimizing the risk (4) is

$$f^*(x) \quad = \quad \begin{cases} 0 & \text{if } 1 - \eta(x) > \eta(x) \text{ and } 1 - \eta(x) > 1 - d \\ 1 & \text{if } \eta(x) > 1 - \eta(x) \text{ and } \eta(x) > 1 - d \\ R & \text{if } \max(\eta(x), 1 - \eta(x)) \leq 1 - d \end{cases} \tag{5}$$

which we will refer to as the Bayes rule with reject option. According to this rule, we should

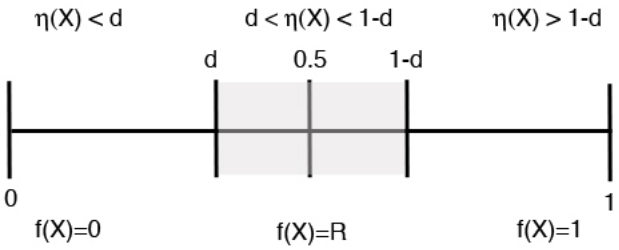

Figure 1: Bayes rule (with reject option).

never invoke the reject option if $d \geq 1/2$ and we should always reject if $d = 0$. For this reason we can restrict ourselves to the cases $0 \leq d \leq 1/2$ and we denote the relevant risk function

(4) by

$$L_d(f) = d\mathbb{P}\{f(X) = R\} + \mathbb{P}\{f(X) \neq Y, f(X) \neq R\}. \tag{6}$$

The Bayes rule (5) simplifies to

$$f^*(x) \quad = \quad \begin{cases} 0 & \text{if } \eta(x) < d \\ 1 & \text{if } \eta(x) > 1 - d \\ R & \text{otherwise,} \end{cases} \tag{7}$$

see Figure 1, and we denote its risk by

$$L_d^* = L_d(f^*) = \min_{f:\mathbb{R}^k \to \{0,1,R\}} L_d(f). \tag{8}$$

The case $d = \frac{1}{2}$ reduces to the classical situation without the reject option and the Bayes classifier (7) reduces to (1). In the remainder of the paper we demonstrate that the behavior of $\eta(x)$ near the value $1/2$ and more generally in the interval $(d, 1-d)$ is of no real importance; the classification problem hinges on what happens outside this interval, especially at the values d and $1-d$.

The paper is organized as follows. Section 2 discusses plug-in rules based on the Bayes rule (7). These rules are called this way since they replace the regression function $\eta(x)$ by an estimate $\widehat{\eta}(x)$ in formula (7). Besides introducing the reject option, we extend the existing theory for plug-in rules [Devroye, Györfi and Lugosi (1996, Theorem 2.2)] since our bound depends explicitly on both the difference $|\widehat{\eta}(X) - \eta(X)|$ and the behavior of $\eta(X)$ near the values d and $1-d$. We show that very fast rates are possible under reasonable margin conditions, extending a recent result by Audibert and Tsybakov (2005) to our more general framework. We illustrate the theory with an application to speech recognition in Section 3.

Section 4 extends the existing theory of empirical risk minimizers by allowing for the reject option. Here an estimate is found by minimizing the empirical counterpart of the risk (6) over an entire class of classifiers $\mathcal{F}$. We demonstrate that the rates of the risk (6) of the resulting minimizers to the Bayes risk L_d^* depends on the metric entropy of (a transformed class of) $\mathcal{F}$ and on the behavior of $\eta(X)$ near the values d and $1-d$. Again our results are in line with the recent developments of the theory for $d = 1/2$ [see, for example, Boucheron,

Bousquet, and Lugosi (2004b), Massart and Nédélec (2003), Tsybakov (2004), Tarigan and Van de Geer (2004) and Tsybakov and Van de Geer (2003)] and extend the theory to the general case $0 \leq d \leq 1/2$.

Section 5 pushes the theory even further. We differentiate between misclassification costs of the cases $Y = 1$ & $f(X) = 0$ and $Y = 0$ & $f(X) = 1$, a situation common in, for instance, medical studies where misclassifying a sick patient as healthy is worse than the opposite. The risk function (6) is changed to accommodate for this differentiation and the results obtained in Section 2 for the plug-in estimates are extended in a straightforward way.

2 Plug-in rules

In this section we consider the plug-in rule

$$\widehat{f}(x) \quad = \quad \begin{cases} 0 & \text{if } \widehat{\eta}(x) < d \\ 1 & \text{if } \widehat{\eta}(x) > 1 - d \\ R & \text{otherwise} \end{cases} \tag{9}$$

based on any estimate $\widehat{\eta}$ of the regression function

$$\eta(x) = \mathbb{P}\{Y = 1 | X = x\} = \mathbb{E}\{Y | X = x\}$$

and the form (7) of the optimal classifier. Our main result of this section, Theorem 2 below, shows that the difference between the risk $L_d(\widehat{f})$ of the estimate $\widehat{f}$ and the optimal (Bayes) risk L_d^*, the

$$\Delta_d(\widehat{f}) = L_d(\widehat{f}) - L_d^* \tag{10}$$

depends on the following two criteria:

(1) How well does $\widehat{\eta}(X)$ estimate $\eta(X)$? and

(2) What is the behavior of $\eta(X)$ near d and $1 - d$?

First we prove an auxiliary result which rewrites $\Delta_d(\widehat{f})$ into a convenient form.

Lemma 1. *For any $0 \leq d \leq 1/2$, we have*

$$\begin{aligned}\Delta_d(\widehat{f}) &= \mathbb{E}\,|d-\eta(X)|\left(\mathbb{1}_{\{f^*(X)=0,\,\widehat{f}(X)\neq f^*(X)\}}+\mathbb{1}_{\{\widehat{f}(X)=0,\,\widehat{f}(X)\neq f^*(X)\}}\right)+\\ &\quad \mathbb{E}\,|1-d-\eta(X)|\left(\mathbb{1}_{\{f^*(X)=1,\,\widehat{f}(X)\neq f^*(X)\}}+\mathbb{1}_{\{\widehat{f}(X)=1,\,\widehat{f}(X)\neq f^*(X)\}}\right). \qquad (11)\end{aligned}$$

Proof. We first write the difference

$$\begin{aligned}\Delta_d(\widehat{f}) &= d\left(\mathbb{P}\left\{\widehat{f}(X)=R\right\}-\mathbb{P}\left\{f^*(X)=R\right\}\right)+\\ &\quad \left(\mathbb{P}\left\{\widehat{f}(X)=1,Y=0\right\}-\mathbb{P}\left\{f^*(X)=1,Y=0\right\}\right)+\\ &\quad \left(\mathbb{P}\left\{\widehat{f}(X)=0,Y=1\right\}-\mathbb{P}\left\{f^*(X)=0,Y=1\right\}\right).\end{aligned}$$

Observe that

$$\begin{aligned}&\mathbb{P}\left\{\widehat{f}(X)=1,Y=0\right\}+\mathbb{P}\left\{\widehat{f}(X)=0,Y=1\right\}\\ &=\mathbb{E}(1-\eta(X))\mathbb{1}_{\left\{\widehat{f}(X)=1\right\}}+\mathbb{E}\eta(X)\mathbb{1}_{\left\{\widehat{f}(X)=0\right\}}\\ &=\mathbb{E}(1-\eta(X))\mathbb{1}_{\left\{\widehat{f}(X)=1\right\}}-\mathbb{E}\eta(X)\mathbb{1}_{\left\{\widehat{f}(X)=1\right\}}\\ &\quad -\mathbb{E}\eta(X)\mathbb{1}_{\left\{\widehat{f}(X)=R\right\}}+\mathbb{E}\eta(X)\\ &=\mathbb{E}(1-2\eta(X))\mathbb{1}_{\left\{\widehat{f}(X)=1\right\}}-\mathbb{E}\eta(X)\mathbb{1}_{\left\{\widehat{f}(X)=R\right\}}+\mathbb{E}\eta(X).\end{aligned}$$

Combining the two preceding displays, we obtain

$$\begin{aligned}\Delta_d(\widehat{f}) &= \mathbb{E}(1-2\eta(X))\left(\mathbb{1}_{\{\widehat{f}(X)=1\}}-1_{\{f^*(X)=1\}}\right)+\\ &\quad \mathbb{E}\,(d-\eta(X))\left(\mathbb{1}_{\{\widehat{f}(X)=R\}}-\mathbb{1}_{\{f^*(X)=R\}}\right) \qquad (12)\\ &= \mathbb{E}(d-\eta(X)+1-d-\eta(X))\left(\mathbb{1}_{\{\widehat{f}(X)=1\}}-1_{\{f^*(X)=1\}}\right)+\\ &\quad \mathbb{E}\,(d-\eta(X))\left(\mathbb{1}_{\{\widehat{f}(X)=R\}}-\mathbb{1}_{\{f^*(X)=R\}}\right)\end{aligned}$$

Recall the definition of f^* given by (7), and split the indicator functions using disjoint events,

and the above becomes

$$
\begin{aligned}
&= \mathbb{E}\left(|d-\eta(X)|+|1-d-\eta(X)|\right)\times \\
&\quad \times\left(\mathbb{1}_{\{\widehat{f}(X)=1,\, f^*(X)=0\}}+\mathbb{1}_{\{\widehat{f}(X)=0,\, f^*(X)=1\}}\right)+ \\
&\quad \mathbb{E}\,|d-\eta(X)|\left(\mathbb{1}_{\{\widehat{f}(X)=0,\, f^*(X)=R\}}+\mathbb{1}_{\{\widehat{f}(X)=R,\, f^*(X)=0\}}\right)+ \\
&\quad \mathbb{E}\,|1-d-\eta(X)|\left(\mathbb{1}_{\{\widehat{f}(X)=1,\, f^*(X)=R\}}+\mathbb{1}_{\{\widehat{f}(X)=R,\, f^*(X)=1\}}\right) \\
&= \mathbb{E}\,|d-\eta(X)|\left(\mathbb{1}_{\{f^*(X)=0,\, \widehat{f}(X)\neq f^*(X)\}}+\mathbb{1}_{\{\widehat{f}(X)=0,\, \widehat{f}(X)\neq f^*(X)\}}\right)+ \\
&\quad \mathbb{E}\,|1-d-\eta(X)|\left(\mathbb{1}_{\{f^*(X)=1,\, \widehat{f}(X)\neq f^*(X)\}}+\mathbb{1}_{\{\widehat{f}(X)=1,\, \widehat{f}(X)\neq f^*(X)\}}\right).
\end{aligned}
$$

The proof is complete.

□

This lemma clearly confirms that the Bayes rule with reject option f^* defined in (7) minimizes the risk $L_d(f)$ as already shown in Chow (1970). Indeed, the right-hand side in (11) is non-negative and equals zero if and only if $\widehat{f}=f^*$.

Theorem 2. *Let* $0\le d\le 1/2$ *and* $\delta\ge 0$, *and*

$$
P_d(\delta)=\mathbb{P}\left\{|d-\eta(X)|\le\delta\right\}+\mathbb{P}\left\{|1-d-\eta(X)|\le\delta\right\}. \tag{13}
$$

We have

$$
\Delta_d(\widehat{f})\le\inf_{\delta\ge 0}\left\{2(1-d)\mathbb{P}\left\{|\eta(X)-\widehat{\eta}(X)|>\delta\right\}+\delta P_d(\delta)\right\}. \tag{14}
$$

Proof. We first recall from (12) in the proof of Lemma 1 that

$$
\begin{aligned}
\Delta_d(\widehat{f}) &= \mathbb{E}\,|d-\eta(X)|\left(\mathbb{1}_{\{f^*(X)=0,\, \widehat{f}(X)\neq f^*(X)\}}+\mathbb{1}_{\{\widehat{f}(X)=0,\, \widehat{f}(X)\neq f^*(X)\}}\right)+ \\
&\quad \mathbb{E}\,|1-d-\eta(X)|\left(\mathbb{1}_{\{f^*(X)=1,\, \widehat{f}(X)\neq f^*(X)\}}+\mathbb{1}_{\{\widehat{f}(X)=1,\, \widehat{f}(X)\neq f^*(X)\}}\right).
\end{aligned}
$$

Let F_1, F_2 denote the following events

$$\begin{aligned} F_1 &= \{\eta(X) < d < \widehat{\eta}(X)\} \cup \{\widehat{\eta}(X) < d < \eta(X)\} \\ F_2 &= \{\eta(X) < 1-d < \widehat{\eta}(X)\} \cup \{\widehat{\eta}(X) < 1-d < \eta(X)\} \end{aligned}$$

The first term in the above expression can be bounded as follows:

$$\begin{aligned} &\mathbb{E}\,|d-\eta(X)|\left(\mathbb{1}_{\{f^*(X)=0,\,\widehat{f}(X)\neq f^*(X)\}} + \mathbb{1}_{\{\widehat{f}(X)=0,\,\widehat{f}(X)\neq f^*(X)\}}\right) \\ &= \mathbb{E}\,|d-\eta(X)|\,\mathbb{1}_{F_1}\left(\mathbb{1}_{\{|\eta-\widehat{\eta}|>\delta\}} + \mathbb{1}_{\{|\eta-\widehat{\eta}|\leq\delta\}}\right) \\ &\leq \mathbb{E}\,|d-\eta(X)|\,\mathbb{1}_{\{|\eta-\widehat{\eta}|>\delta\}} + \mathbb{E}\,|d-\eta(X)|\,\mathbb{1}_{\{|d-\eta(X)|\leq\delta\}} \\ &\leq (1-d)\mathbb{P}\left\{|\eta-\widehat{\eta}|>\delta\right\} + \delta\mathbb{P}\left\{|d-\eta(X)|\leq\delta\right\} \end{aligned}$$

Similarly, one can bound

$$\begin{aligned} &\mathbb{E}\,|1-d-\eta(X)|\left(\mathbb{1}_{\{f^*(X)=1,\,\widehat{f}(X)\neq f^*(X)\}} + \mathbb{1}_{\{\widehat{f}(X)=1,\,\widehat{f}(X)\neq f^*(X)\}}\right) \\ &= \mathbb{E}\,|1-d-\eta(X)|\,\mathbb{1}_{F_2}\left(\mathbb{1}_{\{|\eta-\widehat{\eta}|>\delta\}} + \mathbb{1}_{\{|\eta-\widehat{\eta}|\leq\delta\}}\right) \\ &\leq (1-d)\mathbb{P}\left\{|\eta-\widehat{\eta}|>\delta\right\} + \delta\mathbb{P}\left\{|1-d-\eta(X)|\leq\delta\right\} \end{aligned}$$

We conclude the proof by combining the two preceding displays.

□

This theorem generalizes and improves Theorem 2.2 in Devroye, Györfi and Lugosi (1996). Their result states that in absence of the reject option ($d = \frac{1}{2}$),

$$\Delta_{\frac{1}{2}}(\widehat{f}) = \mathbb{E}|1-2\eta(X)|\mathbb{1}_{\{\widehat{f}(X)\neq f^*(X)\}} \leq 2\mathbb{E}|\widehat{\eta}(X)-\eta(X)|.$$

Note that the inequality on the right does not reveal the crucial behavior of $\eta(X)$ near $1/2$. Theorem 2 above does emphasize the importance of $\eta(X)$ near d and $1-d$ through the probability $P_d(\delta)$.

Remark. Theorem 2 indicates that fast rates (faster than $n^{-1/2}$) can be achieved using plug-in estimates. We briefly discuss two situations:
(a) There exists a $\delta_0 > 0$ such that $P_d(\delta_0) = 0$.

(b) There exist $C < \infty$, $\alpha \geq 0$ such that $P_d(\delta) \leq C\delta^\alpha$ for all $\delta > 0$.

The first condition means that for some δ_0 and all $x \in \mathcal{X}' \subseteq \mathbb{R}^k$ with $\mathbb{P}\{X \in \mathcal{X}'\} = 1$,

$$|d - \eta(x)| \geq \delta_0 \qquad \text{and} \qquad |1 - d - \eta(x)| \geq \delta_0,$$

that is, $\eta(x)$ stays away from the values d and $1 - d$. In this case very fast rates for $\Delta_d(\widehat{f})$ are possible, depending on the smoothness of η only. This condition is used by Nédélec and Massart (2003)in the context of binary classification without the reject option ($d = 1/2$).

The second condition, analogous to *Tsybakov's* margin condition [cf. Tsybakov (2004)], again in the context of binary classification without the reject option ($d = 1/2$), means that $\eta(X)$ puts little probability mass around d and $1 - d$. We illustrate this by assuming that the probability

$$r_n(\delta) = \mathbb{P}\left\{|\widehat{\eta}(X) - \eta(X)| \geq \delta\right\}$$

is of the form $C_0 \exp(-C_1 n^\gamma \delta^2)$ for some C_0, C_1 and $\gamma > 0$. Typically, γ will depend on the degree of smoothness of η and the dimension k of the feature space. Condition (b) ensures that $P_d(\delta) \leq C\delta^{1+\alpha}$. Theorem 2 guarantees that $\Delta_d(\widehat{f})$ is bounded above by $r_n(\delta) + C\delta^{1+\alpha}$. Choosing $\delta = \log(n)/(C_1 n^{\gamma/2})$, we obtain that for some $C' > 0$,

$$\Delta_d(\widehat{f}) \leq C' n^{-(1+\alpha)\gamma/2} \cdot \log^{1+\alpha}(n).$$

The two extreme cases are $\alpha = 0$ and $\alpha = +\infty$. The case $\alpha = 0$ does not impose any restrictions on η and it guarantees only the slowest possible rates. Since no structure is imposed on η, it takes into account the worst possible scenario [i.e., the worst distribution of the pair (X, Y)]. The case $\alpha = +\infty$ on the other hand imposes a lot of structure and it corresponds to situation (a). This is the optimal situation where the fastest rates can be guaranteed. See Audibert and Tsybakov (2005) for the corresponding situation without the reject option.

3 Illustrations

As an illustration to the plug-in rules with reject option, we consider two popular classification methods: *kernel rules* and *logistic regression* and apply the above to a (functional) dataset

that comprises 100 digitized voice recordings of the words *boat* (55 times) and *goat* (45 times). We refer the reader to Biau, Bunea, and Wegkamp (2005) for a complete description of the data. We represent the data as $\mathcal{D} = \{(X_i, Y_i) \in \mathbb{R}^k \times \{0,1\}, i = 1 \dots n\}$, where $n = 100$, X_i is a curve (digitized recording) and Y_i is the word membership of X_i; $Y_i = 0(1)$ corresponds to *boat* (*goat*). We compute for each X_i its Fourier coefficients $(X_{i1}, X_{i2}, \dots)$ and select the first k of them (effective dimension) to represent the curve. Classification is further performed on the coefficients $(X_{i1}, X_{i2}, \dots, X_{ik}) \in \mathbb{R}^k$. Let $\mathcal{D}_{-i}$ be obtained by removing the i-th case (X_i, Y_i) and let $\widehat{\eta}_{-i}$ be a generic estimator of η using $\mathcal{D}_{-i}$. For each $i = 1 \dots n$, we construct the plug-in classifiers

$$\widehat{f}_{-i}(x) \quad = \quad \begin{cases} 0 & \text{if } \widehat{\eta}_{-i}(x) \leq 0.5 \\ 1 & \text{if } \widehat{\eta}_{-i}(x) > 0.5 \end{cases}$$

and estimate the error rate by

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \neq \widehat{f}_{-i}(X_i)\}}. \tag{15}$$

In what follows we briefly recall the two standard rules mentioned above, estimate the various (smoothing) parameters required, then introduce the reject option, and study the performance of the rejection mechanism based on the error-reject trade-off.

Logistic regression. In the classical setup of the *logistic regression* rule, the conditional probability is estimated as

$$\widehat{\eta}_{\widehat{\beta}}(x) = \left(1 + e^{-\widehat{\beta}^T x}\right)^{-1} \qquad \text{for } x \in \mathbb{R}^k$$

We select the effective dimension k to minimize the error rate (15). The choice $\widehat{k} = 19$ yields the best error rate of 0.20.

Kernel rules. For a given kernel $K : \mathbb{R}^k \to \mathbb{R}$ and a bandwidth $h > 0$, an estimate of the a posteriori probability η is

$$\widehat{\eta}_{\mathrm{K,h}}(x) = \frac{\displaystyle\sum_{i=1}^{n} \mathbb{1}_{\{Y_i=1\}} K\left(\frac{x - X_i}{h}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)} \tag{16}$$

The estimated effective dimension and bandwidth, minimizing the error rate (15), are $\widehat{k} = 43$ $\widehat{h} = 0.003$, yielding an error rate of 0.13.

In both settings, we introduce the reject option, and the corresponding classifiers

$$\widehat{f}(x) \quad = \quad \begin{cases} 0 & \text{if } \widehat{\eta}(x) < d \\ 1 & \text{if } \widehat{\eta}(x) > 1 - d \\ R & \text{if } d \leq \widehat{\eta}(x) \leq 1 - d \end{cases}$$

$\widehat{\eta}$ being replaced by the logistic regression - and kernel estimators. Estimates of the error rate $\mathcal{E} = \mathbb{P}\{Y \neq \widehat{f}(X), \widehat{f}(X) \neq R\}$, reject rate $\mathcal{R} = \mathbb{P}\{\widehat{f}(X) = R\}$ and the risk function (4) as functions of the rejection threshold d are computed using the leave-one-out procedure described above.

For the first situation, *logistic regression rule*, the left panel of Figure 2 presents the error rate (dashed) and reject rate (solid) as functions of the rejection threshold d. The right panel of Figure 2 contains the misclassification rate (dashed) and the risk (4) (solid) as functions of d. As Chow (1970) points out, the performance of this procedure can be judged by the error-reject trade-off. The left panel in figure 3 shows that employing a reject option in this situation does not seem too effective: The error rate $\mathcal{E} = \mathcal{E}(\mathcal{R})$, as a function of the reject rate $\mathcal{R}$, decreases at a nearly constant rate of roughly 0.2. Perhaps a more interesting plot, that of the estimated error-reject ratio, is presented in the right panel of Figure 3. The error-reject ratio is the proportion of rejected items that would have been misclassified, that is,

$$\mathcal{E} - \mathcal{R} \text{ ratio} = \frac{\mathcal{E}(0) - \mathcal{E}(\mathcal{R})}{\mathcal{R}} \qquad \text{for } \mathcal{R} > 0.$$

We refer to Hansen, Liisberg, and Salomon (1997) for a related discussion in the context of handwritten digit recognition. As one can see, invoking the reject option is worst initially ($\mathcal{E} - \mathcal{R}$ ratio of 0 at small reject rates, d close to $\frac{1}{2}$), and it improves afterwards, with a maximum of 30% rejected curves that would normally be incorrectly classified.

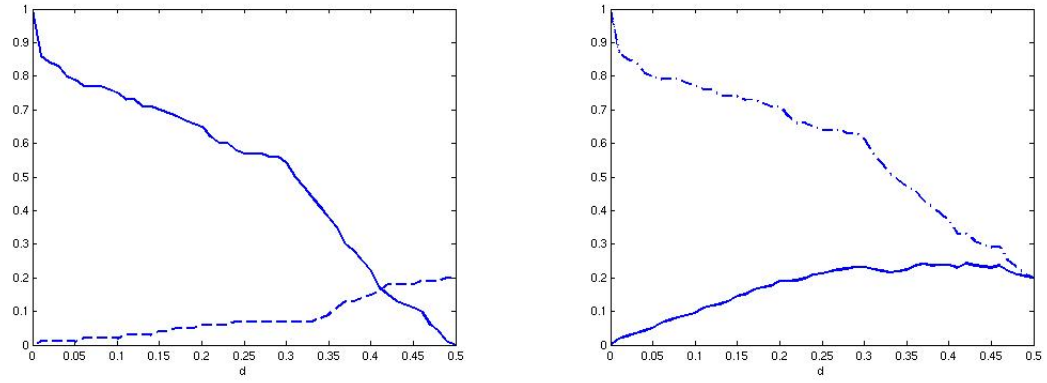

Figure 2: Logistic regression rule. *Left:* Error rate $\mathbb{P}\{Y \neq f(X), f(X) \neq R\}$ (dashed curve) and rejection rate $\mathbb{P}\{f(X) = R\}$ (solid curve) versus rejection cost d. *Right:* Risk function $\mathbb{P}\{Y \neq f(X), f(X) \neq R\} + d\mathbb{P}\{f(X) = R\}$ (solid curve) and misclassification rate $\mathbb{P}\{Y \neq f(X)\}$ (dashed curve) versus rejection cost d.

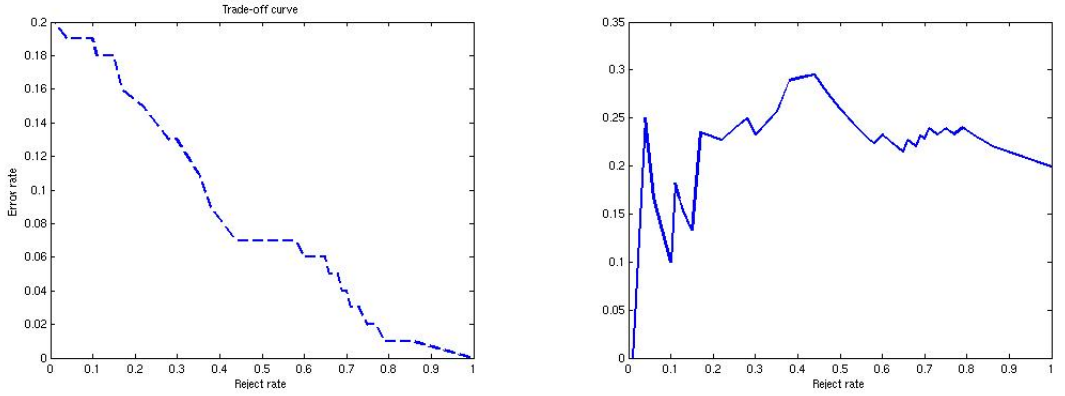

Figure 3: Logistic regression rule. *Left:* trade-off curve. *Right:* error-reject ratio curve.

The situation changes dramatically by using a *kernel rule*. The error rate is reduced almost by a factor of 2, with a dramatic reduction in the reject rate at small values of d (see

Figure 4). The right panel of Figure 5 shows a perfect rejection mechanism (an estimated $\mathcal{E} - \mathcal{R}$ ratio equal to 1) at small rejection rates (d close to $\frac{1}{2}$) with a decline afterwards. The trade-off curve (left panel of Figure 5), shows a significant improvement as well, at relatively small rejection rates ($10\% - 20\%$, the already low error rate can be further reduced significantly.

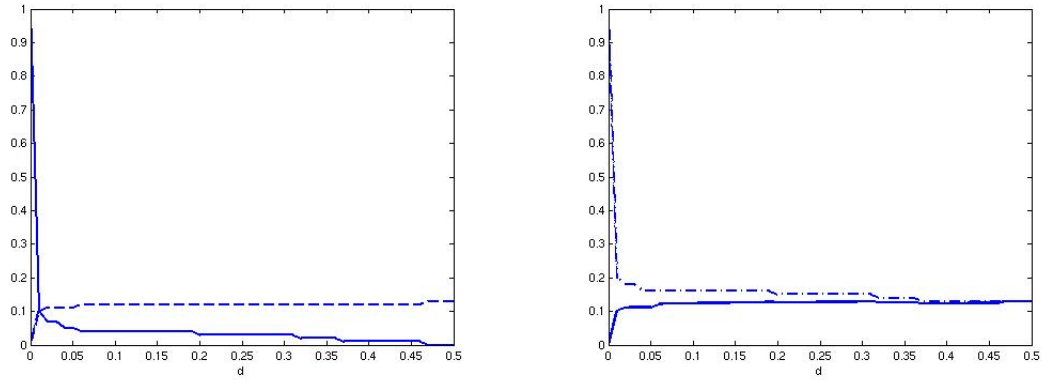

Figure 4: Kernel rule. *Left:* Error rate $\mathbb{P}\{Y \neq f(X), f(X) \neq R\}$ (dashed curve) and rejection rate $\mathbb{P}\{f(X) = R\}$ (solid curve) against rejection cost d. *Right:* Risk function (solid curve) $\mathbb{P}\{Y \neq f(X), f(X) \neq R\} + d\mathbb{P}\{f(X) = R\}$ and misclassification rate (dashed curve) $\mathbb{P}\{Y \neq f(X)\}$ against rejection cost d.

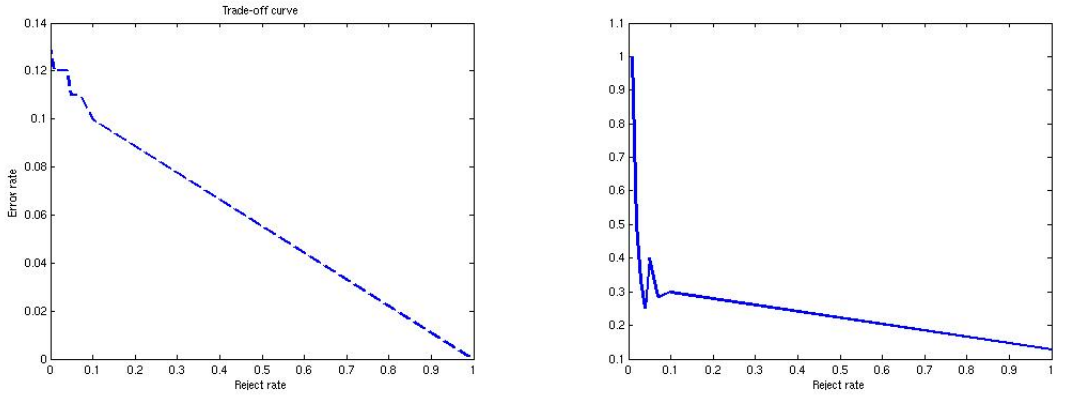

Figure 5: Kernel rule. *Left:* trade-off curve. *Right:* Error-reject ratio.

4 Empirical risk minimization

This section discusses minimization of the empirical counterpart

$$\widehat{L}_d(f) = \frac{d}{n}\sum_{i=1}^{n} \mathbb{1}\{f(X_i) = R\} + \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i, f(X_i) \neq R\} \tag{17}$$

of the risk $L_d(f)$ defined in (6) over a set $\mathcal{F}$ of classifiers $f : \mathbb{R}^k \to \{0, 1, R\}$ with rejection option R. Let $\widehat{f} \in \mathcal{F}$ be the minimizer of $\widehat{L}_d(f)$. The aim is to establish an oracle inequality for the regret $\Delta_d(\widehat{f}) = L_d(\widehat{f}) - L_d^*$ of the form

$$\Delta_d(\widehat{f}) \leq C \inf_{f \in \mathcal{F}} \Delta_d(f) + R_n$$

for some constant $C \geq 1$. The remainder R_n depends on the sample size n, the margin condition on $\eta(x)$ and the size of the class $\mathcal{F}$. We refer to e.g. Anthony and Bartlett (1999), Bartlett and Mendelson (2003), Bartlett, Bousquet, and Mendelson (2004), Boucheron, Bousquet, and Lugosi (2004a and 2004b), Devroye, Györfi, and Lugosi (1996), Massart (2003), and Van de Geer (2000) for a general theory of empirical risk minimizers.

We introduce a bit more notation. Since we are interested in $\Delta_d(f)$ rather than $L_d(f)$, we introduce the loss function

$$\begin{aligned} g_{f,d}(x, y) \;=\; & d\left(\mathbb{1}_{\{f(x)=R\}} - \mathbb{1}_{\{f^*(x)=R\}}\right) + \\ & \left(\mathbb{1}_{\{f(x)\neq y, f(x)\neq R\}} - \mathbb{1}_{\{f^*(x)\neq y, f^*(x)\neq R\}}\right) \end{aligned} \tag{18}$$

so that

$$\mathbb{E} g_{f,d}(X, Y) = \Delta_d(f). \tag{19}$$

We define

$$\widehat{\Delta}_d(f) = \frac{1}{n}\sum_{i=1}^{n} g_{f,d}(X_i, Y_i). \tag{20}$$

We demonstrated in Section 2 that the degree of difficulty of the classification problem depends heavily on the behavior of $\eta(X)$ near d and $1 - d$ for the plug-in estimate. The same conclusion holds for empirical risk minimizers. It turns out [see, for example, Boucheron,

Bousquet, and Lugosi (2004a and b), Bartlett and Mendelson (2003), and Massart and Nédélec (2003), all for the case $d = 1/2$ only] that the way $\mathbb{E}g_{f,d}^2(X)$ relates to $\mathbb{E}g_{f,d}(X,Y)$ is an important ingredient for the rate of the remainder term R_n. Bartlett and Mendelson (2003) call a class $\mathcal{F}$ that satisfies

$$\mathbb{E}g_{f,d}^2(X,Y) \leq B\left\{g_{f,d}(X,Y)\right\}^{\beta} \quad \text{for all } f \in \mathcal{F}, \tag{21}$$

a Bernstein(β, B)-class. We first obtain this link under the two scenarios [situation (a) and (b)] considered previously in Section 2.

Lemma 3. *Let $f : \mathbb{R}^k \to \{0, 1, R\}$ be a classifier with a reject option, and let $0 \leq d \leq \frac{1}{2}$. Assume that for some $\delta_0 > 0$,*

$$\mathbb{P}\left\{\, |d - \eta(X)| \geq \delta_0,\ |1 - d - \eta(X)| \geq \delta_0 \,\right\} = 1.$$

Then $\mathbb{E}g_{f,d}(X,Y) \geq s\mathbb{E}g_{f,d}^2(X,Y)$, where $s = \delta_0/(1+d)^2$.

Proof. Since f and d are fixed, we write g in place of $g_{f,d}$. Observe that

$$\begin{aligned}
s^{-1}\mathbb{E}g(X,Y) &= s^{-1}\Delta_d(f) \\
&= s^{-1}\mathbb{E}|d - \eta(X)|\left(\mathbb{1}_{\{f(X)=0, f(X)\neq f^*(X)\}} + \mathbb{1}_{\{f^*(X)=0, f(X)\neq f^*(X)\}}\right) + \\
&\quad s^{-1}\mathbb{E}|1 - d - \eta(X)|\left(\mathbb{1}_{\{f(X)=1, f(X)\neq f^*(X)\}} + 1_{\{f^*(X)=1, f(X)\neq f^*(X)\}}\right) \\
&\geq \mathbb{E}\,(1+d)^2\left(\mathbb{1}_{\{f(X)=0, f(X)\neq f^*(X)} + \mathbb{1}_{f^*(X)=0, f(X)\neq f^*(X)\}}\right) + \\
&\quad (1+d)^2\,\mathbb{E}\left(1_{\{f(X)=1, f(X)\neq f^*(X)\}} + 1_{\{f^*(X)=1, f(X)\neq f^*(X)\}}\right) \\
&= (1+d)^2\,\mathbb{E}\left(\mathbb{1}_{\{f(X)\neq f^*(X)\}} + \mathbb{1}_{\{f(X)=1, f^*(X)=0} + \mathbb{1}_{f(X)=0, f^*(X)=1\}}\right) \\
&\geq (1+d)^2\,\mathbb{P}\{f(X) \neq f^*(X)\} \\
&\geq \mathbb{E}g^2(X,Y),
\end{aligned}$$

where the last inequality follows from the bound

$$\begin{aligned}
|g(x,y)| &\leq d\left(\mathbb{1}_{\{f(x)=R,\, f(x)\neq f^*(x)\}} + \mathbb{1}_{\{f^*(x)=R,\ f(x)\neq f^*(x)\}}\right) + \\
&\quad \left(\mathbb{1}_{\{f(x)=1,\, f^*(x)=0\}} + \mathbb{1}_{\{f(x)=0,\, f^*(x)=1\}} + \mathbb{1}_{\{f(x)=R,\, f(x)\neq f^*(x)\}} + \mathbb{1}_{\{f^*(x)=R,\, f(x)\neq f^*(x)\}}\right) \\
&= (1+d)\left(\mathbb{1}_{\{f(x)=R,\, f(x)\neq f^*(x)\}} + \mathbb{1}_{\{f^*(x)=R,\, f(x)\neq f^*(x)\}}\right) + \\
&\quad (1+d)\left(\mathbb{1}_{\{f(x)=1,\, f^*(x)=0\}} + \mathbb{1}_{\{f(x)=0,\, f^*(x)=1\}}\right) \\
&\leq (1+d)\,\mathbb{1}_{\{f(x)\neq f^*(x)\}}.
\end{aligned}$$

This proves the lemma.

□

Lemma 4. *Let $f : \mathbb{R}^k \to \{0, 1, R\}$ be a classifier with a reject option, and let $0 \leq d \leq \frac{1}{2}$. Assume that there exist $A > 0$ and $\alpha \geq 0$ such that for all $t \geq 0$,*

$$\mathbb{P}\{|1 - d - \eta(X)| \leq t\} \leq At^\alpha \quad \text{and} \quad \mathbb{P}\{|d - \eta(X)| \leq t\} \leq At^\alpha. \tag{22}$$

Then, for some finite constant C depending on A, α and d,

$$\mathbb{E}g_{f,d}^2(X,Y) \leq C\left\{\mathbb{E}g_{f,d}(X,Y)\right\}^{\alpha/(\alpha+1)}. \tag{23}$$

The case $\alpha = 0$ imposes no restriction on η and leads to slow rates. On the other extreme, $\alpha = +\infty$ corresponds to the situation in Lemma 3 and leads to very fast rates. Tsybakov (2004) made the pertinent observation that faster rates than $n^{-1/2}$ can be obtained for empirical risk minimizers $\widehat{f}$ even in the nontrivial case of $L^* \neq 0$ - under assumption (22) in case $d = 1/2$.

Proof of Lemma 4. Define the events E_1, E_2, E_3 and E_4 by

$$\begin{aligned}
E_1 &= \{f^*(X) = 1, f(X) \neq f^*(X)\}, & E_2 &= \{f(X) = 1, f(X) \neq f^*(X)\}, \\
E_3 &= \{f^*(X) = 0, f(X) \neq f^*(X)\}, & E_4 &= \{f(X) = 0, f(X) \neq f^*(X)\}.
\end{aligned}$$

Lemma 1 implies that

$$\Delta_d(f) \;=\; \mathbb{E}\,|1-d-\eta(X)|\,(\mathbb{1}_{E_1}+\mathbb{1}_{E_2}) + \mathbb{E}\,|d-\eta(X)|\,(\mathbb{1}_{E_3}+\mathbb{1}_{E_4}). \tag{24}$$

The first term on the right can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\,|1-d-\eta(X)|\,\mathbb{1}_{E_1} &\geq t\mathbb{P}\left(E_1 \cap \{|1-d-\eta(X)|>t\}\right)\\
&= t\mathbb{P}\{|1-d-\eta(X)|>t\} - t\mathbb{P}\left(E_1^c \cap \{|1-d-\eta(X)|>t\}\right)\\
&\geq t\left\{(1-At^\alpha) - \mathbb{P}(E_1^c)\right\}\\
&= t\left\{\mathbb{P}(E_1) - At^\alpha\right\}.
\end{aligned}$$

The other three terms in (24) can be handled in a similar way and we obtain

$$\begin{aligned}
\Delta_d(f) &\geq t\left(\{\mathbb{P}(E_1\cup E_2\cup E_3\cup E_4) - 4At^\alpha\right\}\\
&\geq t\left(\mathbb{P}\{f(X)\neq f^*(X)\} - 4At^\alpha\right).
\end{aligned}$$

Choosing

$$t = \left(\frac{\mathbb{P}\{f(X)\neq f^*(X)\}}{8A}\right)^{1/\alpha}$$

we find that

$$\Delta_d(f) \geq \frac{\mathbb{P}^{\frac{1+\alpha}{\alpha}}\{f(X)\neq f^*(X)\}}{2(8A)^{1/\alpha}},$$

or

$$\mathbb{P}\{f(X)\neq f^*(X)\} \leq \left(2(8A)^{1/\alpha}\Delta_d(f)\right)^{\frac{\alpha}{1+\alpha}},$$

and we obtain that

$$\begin{aligned}
\mathbb{E}g^2(X,Y) &\leq (1+d)^2\,\mathbb{P}\{f(X)\neq f^*(X)\}\\
&\leq (1+d)^2\left(2(8A)^{1/\alpha}\Delta_d(f)\right)^{\frac{\alpha}{1+\alpha}}\\
&= C\left\{\mathbb{E}g(X,Y)\right\}^{\frac{\alpha}{1+\alpha}},
\end{aligned}$$

for

$$C = 2^{\frac{\alpha}{1+\alpha}} (1+d)^2 (8A)^{1/(1+\alpha)}.$$

This concludes the proof. □

These two preliminary results allow us to formulate the first theorem. It assumes minimization over a finite set of classifiers. This is for instance the case when we select tuning parameters such as a bandwidth or dimension over a finite grid as in Biau, Bunea, and Wegkamp (2005).

Theorem 5. *Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a finite collection of classifiers $f : \mathbb{R}^k \to \{0, 1, R\}$ with reject option R. For $0 \le d \le 1/2$ and $\alpha \ge 0$ fixed, let η satisfy condition (22) in Lemma 4. Then the minimizer $\widehat{f} = \widehat{f}(d, \mathcal{F})$ of the empirical risk $\widehat{L}(f)$ defined in (17) satisfies the following oracle inequality: There exists a constant $C < \infty$ such that for all $n \ge 1$ and $\rho > 0$,*

$$\mathbb{E}\Delta_d(\widehat{f}) \le (1+\rho) \min_{f \in \mathcal{F}} \Delta_d(f) + C \left(\frac{\log M}{\rho n} \right)^{\frac{1+\alpha}{2+\alpha}} \tag{25}$$

Proof. Set $\beta = \alpha/(1+\alpha)$. Since $\widehat{f}$ minimizes $\widehat{\Delta}_d(f)$, we find that for any $f \in \mathcal{F}$ and $\rho > 0$,

$$\begin{aligned} \Delta_d(\widehat{f}) &= (1+\rho)\widehat{\Delta}_d(\widehat{f}) + \left\{ \Delta_d(\widehat{f}) - (1+\rho)\widehat{\Delta}_d(\widehat{f}) \right\} \\ &\le (1+\rho)\widehat{\Delta}_d(f) + \left\{ \Delta_d(\widehat{f}) - (1+\rho)\widehat{\Delta}_d(\widehat{f}) \right\} \end{aligned}$$

and consequently

$$\mathbb{E}\Delta_d(\widehat{f}) \le (1+\rho) \min_{f \in \mathcal{F}} \Delta_d(f) + \mathbb{E}\left\{ \Delta_d(\widehat{f}) - (1+\rho)\widehat{\Delta}_d(\widehat{f}) \right\}.$$

Since we have

$$\begin{aligned} \Delta_d(\widehat{f}) - (1+\rho)\widehat{\Delta}_d(\widehat{f}) &\le \max_{1 \le j \le M} \Delta_d(f_j) - (1+\rho)\widehat{\Delta}_d(f_j) \\ &= \max_{1 \le j \le M} \mathbb{E} g_{f_j,d}(X,Y) - (1+\rho)\frac{1}{n}\sum_{i=1}^{n} g_{f_j,d}(X,Y), \end{aligned}$$

we find by Bernstein's inequality that there exists a $c_0 < \infty$ such that for all $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left\{\Delta_d(\widehat{f}) - (1+\rho)\widehat{\Delta}_d(\widehat{f}) \geq \delta\right\} &\leq \sum_{j=1}^{M} \exp\left(-c_0 \frac{n\rho^{-1}\left\{\delta + \Delta_d(f_j)\right\}^2}{\delta + \Delta_d(f_j) + \Delta_d(f_j)^\beta}\right) \\
&\leq \sum_{j=1}^{M} \exp\left(-c_0 \frac{n\rho^{-1}\left\{\delta + \Delta_d(f_j)\right\}^2}{\left\{\delta + \Delta_d(f_j)\right\} \vee \left\{\delta + \Delta_d(f_j)\right\}^\beta}\right) \\
&\leq M \exp\left(-c_0 n \rho^{-1} \delta^{2-\beta}\right).
\end{aligned}$$

The proof of the theorem follows from a simple integration argument.

□

The second result extends Theorem 5 in that it allows for infinite classes $\mathcal{F}$.

Theorem 6. *Let $\alpha \geq 0$, $0 \leq d \leq 1/2$ be fixed, and set $\beta = \alpha/(1+\alpha)$. Let $\mathcal{F}$ be a collection of classifiers $f : \mathbb{R}^d \to \{0, 1, R\}$ with reject option R. Let $\mathcal{G}$ be the class of loss functions $g_{f,d}$ induced by $f \in \mathcal{F}$ and*

$$\mathcal{G}(R) = \left\{g \in \mathcal{G} : \mathbb{E}g^2(X, Y) \leq R\right\}.$$

For

$$\Psi(\delta) \geq \int_0^\delta \sqrt{H_2(x, \mathcal{G}(\delta))}\, dx \vee \delta, \tag{26}$$

we assume that $\delta^{-2}\Psi(\delta^{2\beta})$ is a non-increasing function in δ. Set $\bar{\Delta} = \inf_f \Delta_d(f)$, and let δ_n be

$$\sqrt{n}(\bar{\Delta} + \delta_n)^2 \geq c_0 \Psi((\bar{\Delta} + \delta_n)^\beta). \tag{27}$$

Let η satisfy condition (22) in Lemma 4. Then the minimizer $\widehat{f} = \widehat{f}(d, \mathcal{F})$ of the empirical risk $\widehat{L}(f)$ defined in (17) satisfies the following oracle inequality: There exist positive constants $c, C < \infty$ such that for all $n \geq 1$ and $\delta > \delta_n$,

$$\mathbb{P}\left\{\Delta_d(\widehat{f}) \geq 2 \min_{f \in \mathcal{F}} \Delta_d(f) + \delta\right\} \leq C \exp\left(-cn\delta^{2-\beta}\right). \tag{28}$$

Proof. Assume without loss of generality that the minimum of $\Delta_d(f)$ is attained for some

$\bar{f} \in \mathcal{F}$, and write $\bar{\Delta} = \Delta_d(\bar{f})$. Since $\widehat{\Delta}_d(\widehat{f}) \leq \widehat{\Delta}_d(\bar{f})$, we find that

$$\begin{aligned}
\mathbb{P}\left\{\Delta_d(\widehat{f}) \geq 2\bar{\Delta} + \delta\right\} &= \mathbb{P}\left\{\sup_{f\in\mathcal{F}:\ \Delta_d(f)\geq 2\bar{\Delta}+\delta} \widehat{\Delta}_d(\bar{f}) - \widehat{\Delta}_d(f) \geq 0\right\} \\
&\leq \sum_{j=1}^{\infty} \mathbb{P}\left\{\sup_{f\in\mathcal{F}_j} \widehat{\Delta}_d(\bar{f}) - \widehat{\Delta}_d(f) \geq 0\right\} \\
&= \sum_{j=1}^{\infty} \mathbb{P}\left\{\sup_{f\in\mathcal{F}_j} (\widehat{\Delta}_d - \Delta_d)(\bar{f}) - (\widehat{\Delta}_d - \Delta_d)(f) - (\Delta_d(f) - \bar{\Delta}) \geq 0\right\}
\end{aligned}$$

where

$$\mathcal{F}_j = \left\{f \in \mathcal{F} : 2^j\delta \leq \Delta_d(f) - 2\bar{\Delta} \leq 2^{j+1}\delta\right\}.$$

Consequently, we need to bound probabilities

$$P_j = \mathbb{P}\left\{\sup_{f\in\mathcal{F}_j} (\widehat{\Delta}_d - \Delta_d)(\bar{f}) - (\widehat{\Delta}_d - \Delta_d)(f) \geq \bar{\Delta} + 2^j\delta\right\}.$$

Observe further that for each $f \in \mathcal{F}_j$,

$$\mathbb{E}(g_{\bar{f}} - g_f)^2(X, Y) \lesssim (\bar{\Delta} + 2^j\delta)^{\beta},$$

by the condition on η. Invoke Lemma 8.5 of Van de Geer (2000, page 130) and assumption (27) and the conclusion follows. See the reasoning at Mohammadi (2004, pp 57 – 59) for details. □

In the typical case where the classes of sets

$$\{x \in \mathbb{R}^d :\ f(x) = 1\}, \qquad \{x \in \mathbb{R}^d :\ f(x) = 0\} \qquad \text{and} \qquad \{x \in \mathbb{R}^d :\ f(x) = R\},$$

indexed by $f \in \mathcal{F}$, are VC classes with finite VC-dimension V, the preceding result yields the oracle inequality

$$\mathbb{E}\Delta_d(\widehat{f}) \leq 2 \inf_{f\in\mathcal{F}} \Delta_d(f) + C\left(\frac{V}{n}\right)^{\frac{1+\alpha}{2+\alpha}}, \qquad \text{for some } C < \infty.$$

This extends the result by Massart and Nédélec (2003), who obtain a similar result for the case $d = 1/2$ (no reject option) only and who assume in addition that the Bayes classifier f^*

belongs to the set $\mathcal{F}$. Future work is needed to see whether it is possible to employ convex risk functions that allow for faster computation.

To conclude this section we present yet another illustration, as a small application of Theorem 5. We use the *kernel rule* (16) and the dataset *boat-goat* described in Section 3.

We split the data $\mathcal{D}$ into a training set $\mathcal{D}_1$ and $\mathcal{D}_2$, each of size 50 and consider, separately, two choices for d: $d = .25$ and $d = .50$. For each pair (k, h), we use the training data $\mathcal{D}_1$ to estimate η and the testing data $\mathcal{D}_2$ to estimate the risk

$$\widehat{L}_{h,k,b} = \frac{d}{50} \sum_{(X_i,Y_i)\in\mathcal{D}_2} \mathbb{1}_{\{\widehat{f}_{k,h}(X)=R\}} + \frac{1}{50} \sum_{(X_i,Y_i)\in\mathcal{D}_2} \mathbb{1}_{\{Y\neq\widehat{f}_{k,h}(X), f_{k,h}(X)\neq R\}}.$$

We estimate the smoothing parameters (h, k) by the minimizers $(\widehat{k}_b, \widehat{h}_b)$ of the empirical risk above over the grid $\{1, 2, \ldots, 100\} \times (0, 1]$. This data-splitting procedure introduces additional variability due to the randomness of the split. To reduce that, we suggest to repeat the above procedure $B = 41$ times by taking various random splits of $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$, and to choose $\widehat{k} = \text{median}(\widehat{k}_b)$ and $\widehat{h} = \text{median}(\widehat{h}_b)$. Finally, we consider another $B' = 200$ partitions, and estimate the final risk using a simple sample average

$$\widehat{L} = \frac{1}{B_1} \sum_{b=1}^{B_1} \widehat{L}_{\widehat{h},\widehat{k},b}.$$

We present two situations ($d = 0.5$ - no reject option, and $d = 0.25$). At the fist stage of this procedure, $B = 41$ splits are used to select the tuning parameters. We obtained $\widehat{h} = 0.003$ and $\widehat{k} = 51$ in case $d = 0.25$, and $\widehat{h} = 0.003$ and $k = 41$ in case $d = 0.5$. Figure 6, left and middle panels, depicts the corresponding histograms. The second stage uses $B' = 200$ partitions of the data $\mathcal{D}$ to estimate the risk. The right panels in figure 6 show histograms for the estimated risks. The average risks for the two situations considered were 0.1665 corresponding to $d = 0.5$ (no reject option) and 0.1520 for $d = 0.25$. The results are consistent with those presented in Section 3.

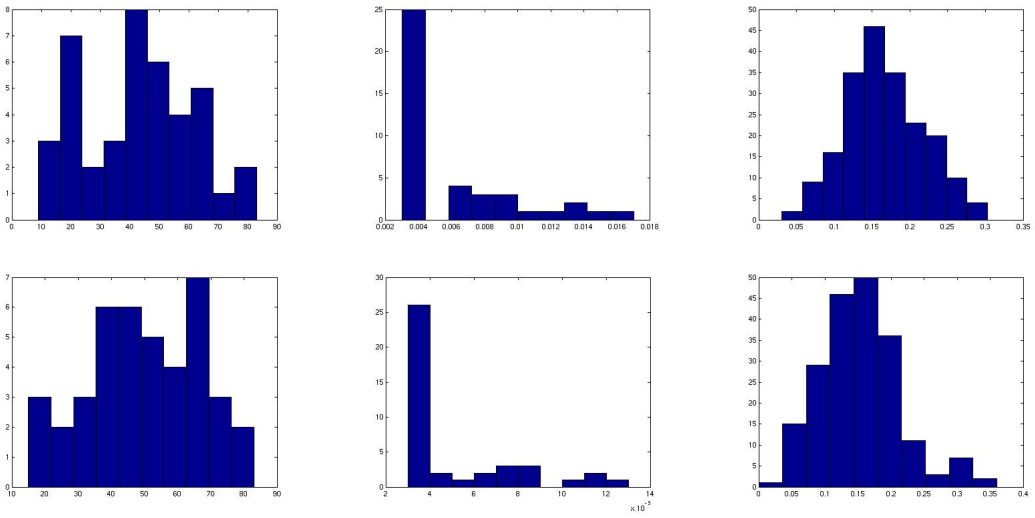

Figure 6: Histograms of the tuning parameters $\widehat{h}$ (*left*), $\widehat{k}$ (*middle*) and risk $\widehat{L}$ (*right*) with $d = 0.5$ (*top*) and $d = 0.25$ (*bottom*).

5 Differentiation between misclassification errors

In what follows, for a given case $(X, Y) \in \mathbb{R}^k \times \{0, 1\}$ and a classifier f, we say that we make a *type 1 error* if $Y = 0$ and $f(X) = 1$ and a *type 2 error* if $Y = 1$ and $f(X) = 0$. Often, in practice, type 1 errors are more costly than type 2 errors or vice-versa. It is the case in the medical field for example where misclassifying a sick patient as healthy is, in general, far worse than the reverse. In order to accommodate for that, we consider a more general risk function

$$L_{d,\theta}(f) = d\mathbb{P}(f(X) = R) + \mathbb{P}(Y = 1, f(X) = 0) + \theta\mathbb{P}(Y = 0, f(X) = 1) \tag{29}$$

where θ represents the ratio of the costs of the two types of error. The optimal rule minimizing the risk (29) is:

$$f^*_{d,\theta}(x) \quad = \quad \begin{cases} 0 & \text{if } 1 - \eta(x) > \eta(x) \text{ and } 1 - \eta(x) > 1 - d \\ 1 & \text{if } \eta(x) > 1 - \eta(x) \text{ and } \eta(x) > 1 - \frac{d}{\theta} \\ R & \text{if } d \leq \eta(x) \leq 1 - \frac{d}{\theta} \end{cases}$$

Following the same reasoning as before we consider $d \leq \frac{1}{2}$ and without loss of generality

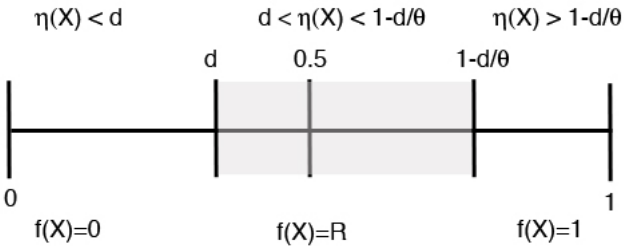

Figure 7: Bayes rule, weighted errors.

we can assume that $\theta > 1$. Let $L^*_{d,\theta}$ denote the Bayes error,

$$L^*_{d,\theta} = L^*_{d,\theta}(f^*_{d,\theta}) = \min_{f:\mathbb{R}^d \to \{0,1,R\}} L_{d,\theta}(f).$$

and for any classifier f, the regret is:

$$\Delta_{d,\theta}(f) = L_{d,\theta}(f) - L^*_{d,\theta}$$

Lemma 7. *For any $0 \leq d \leq 1/2$, and $\theta > 1$, we have*

$$\begin{aligned}\Delta_{d,\theta}(\widehat{f}) &= \mathbb{E}\,|d - \eta(X)| \left(\mathbb{1}_{\{f^*(X)=0,\, \widehat{f}(X)\neq f^*(X)\}} + \mathbb{1}_{\{\widehat{f}(X)=0,\, \widehat{f}(X)\neq f^*(X)\}}\right) + \\ &\quad \mathbb{E}\theta \left|1 - \frac{d}{\theta} - \eta(X)\right| \left(\mathbb{1}_{\{f^*(X)=1,\, \widehat{f}(X)\neq f^*(X)\}} + \mathbb{1}_{\{\widehat{f}(X)=1,\, \widehat{f}(X)\neq f^*(X)\}}\right). \quad (30)\end{aligned}$$

Proof. Following the proof of Lemma 1, we find that

$$\begin{aligned}\Delta_{d,\theta}(\widehat{f}) &= d\left(\mathbb{P}\left\{\widehat{f}(X) = R\right\} - \mathbb{P}\left\{f^*(X) = R\right\}\right) + \\ &\quad \theta\left(\mathbb{P}\left\{\widehat{f}(X) = 1, Y = 0\right\} - \mathbb{P}\left\{f^*(X) = 1, Y = 0\right\}\right) + \\ &\quad \left(\mathbb{P}\left\{\widehat{f}(X) = 0, Y = 1\right\} - \mathbb{P}\left\{f^*(X) = 0, Y = 1\right\}\right).\end{aligned}$$

Recall that

$$\begin{aligned}\mathbb{P}\left\{\widehat{f}(X) = 1, Y = 0\right\} &= \mathbb{E}(1 - \eta(X))\mathbb{1}_{\left\{\widehat{f}(X)=1\right\}} \\ \mathbb{P}\left\{\widehat{f}(X) = 0, Y = 1\right\} &= \mathbb{E}(\eta(X))(1 - \mathbb{1}_{\left\{\widehat{f}(X)=1\right\}} - \mathbb{1}_{\left\{\widehat{f}(X)=R\right\}})\end{aligned}$$

Thus,

$$\begin{aligned}\Delta_{d,\theta}(\widehat{f}) &= \mathbb{E}(\theta - \theta\eta(X) - \eta(X))\left(\mathbb{1}_{\{\widehat{f}(X)=1\}} - 1_{\{f^*(X)=1\}}\right) + \\ &\quad \mathbb{E}\left(d - \eta(X)\right)\left(\mathbb{1}_{\{\widehat{f}(X)=R\}} - \mathbb{1}_{\{f^*(X)=R\}}\right) \\ &= \mathbb{E}\left[d - \eta(X) + \theta\left(1 - \frac{d}{\theta} - \eta(X)\right)\right]\left(\mathbb{1}_{\{\widehat{f}(X)=1\}} - 1_{\{f^*(X)=1\}}\right) + \\ &\quad \mathbb{E}\left(d - \eta(X)\right)\left(\mathbb{1}_{\{\widehat{f}(X)=R\}} - \mathbb{1}_{\{f^*(X)=R\}}\right)\end{aligned} \tag{31}$$

Just like before we split each indicator using disjoint events, and combining the above display with (30), this yields:

$$\begin{aligned}&= \mathbb{E}\left(|d - \eta(X)| + \theta\left|1 - \frac{d}{\theta} - \eta(X)\right|\right) \times \\ &\quad \times \left(\mathbb{1}_{\{\widehat{f}(X)=1,\, f^*(X)=0\}} + \mathbb{1}_{\{\widehat{f}(X)=0,\, f^*(X)=1\}}\right) + \\ &\quad \mathbb{E}\,|d - \eta(X)|\left(\mathbb{1}_{\{\widehat{f}(X)=0,\, f^*(X)=R\}} + \mathbb{1}_{\{\widehat{f}(X)=R,\, f^*(X)=0\}}\right) + \\ &\quad \mathbb{E}\theta\left|1 - \frac{d}{\theta} - \eta(X)\right|\left(\mathbb{1}_{\{\widehat{f}(X)=1,\, f^*(X)=R\}} + \mathbb{1}_{\{\widehat{f}(X)=R,\, f^*(X)=1\}}\right)\end{aligned}$$

$$\begin{aligned}&= \mathbb{E}\,|d - \eta(X)|\left(\mathbb{1}_{\{f^*(X)=0,\, \widehat{f}(X)\neq f^*(X)\}} + \mathbb{1}_{\{\widehat{f}(X)=0,\, \widehat{f}(X)\neq f^*(X)\}}\right) + \\ &\quad \mathbb{E}\theta\left|1 - \frac{d}{\theta} - \eta(X)\right|\left(\mathbb{1}_{\{f^*(X)=1,\, \widehat{f}(X)\neq f^*(X)\}} + \mathbb{1}_{\{\widehat{f}(X)=1,\, \widehat{f}(X)\neq f^*(X)\}}\right)\end{aligned}$$

which concludes the proof. □

The equivalent of Theorem 2 is the following result.

Theorem 8. *Let* $0 \leq d \leq 1/2$, $\delta \geq 0$, $\theta > 1$ *and*

$$P_d(\delta) = \mathbb{P}\left\{|d - \eta(X)| \leq \delta\right\} + \mathbb{P}\left\{\left|1 - \frac{d}{\theta} - \eta(X)\right| \leq \delta\right\}. \tag{32}$$

We have

$$\Delta_d(\widehat{f}) \leq \inf_{\delta \geq 0}\left\{(2 - d - \frac{d}{\theta})\mathbb{P}\left\{|\eta(X) - \widehat{\eta}(X)| > \delta\right\} + \delta P_d(\delta)\right\}. \tag{33}$$

The proof is very similar to the proof of Theorem 2 and is for this reason omitted.

For illustration purposes we use the *heart* data set from the STATlog project. The data set along with a complete description is available online at `http://www.liacc.up.pt/ML/statlog/datasets/heart/heart.doc.html`. There are $n = 270$ cases and $k = 13$ attributes. The problem is to distinguish between absence (0) and presence (1) of heart disease. The cost of misclassifying a sick person as healthy is five times more than the reverse according to the STATlog documentation, thus we set $\theta = 5$. An analysis similar to the one described in Section 3 is performed, using a logistic regression rule. Figure 8, left panel, presents the error rates and rejection rate against the rejection threshold d. It should be noted that even at $d = \frac{1}{2}$ due to a high type 1 error cost the algorithm will reject 30% of the cases. A closer analysis of the trade-off curves, summarized in Figure 9, reveals an important improvement, though. The initial slopes of the two curves are rather steep, implying that for a slight increase in the reject rate, a significant improvement in the two error rates can be achieved. The rate of decay increases to 0 as the rejection rate gets higher.

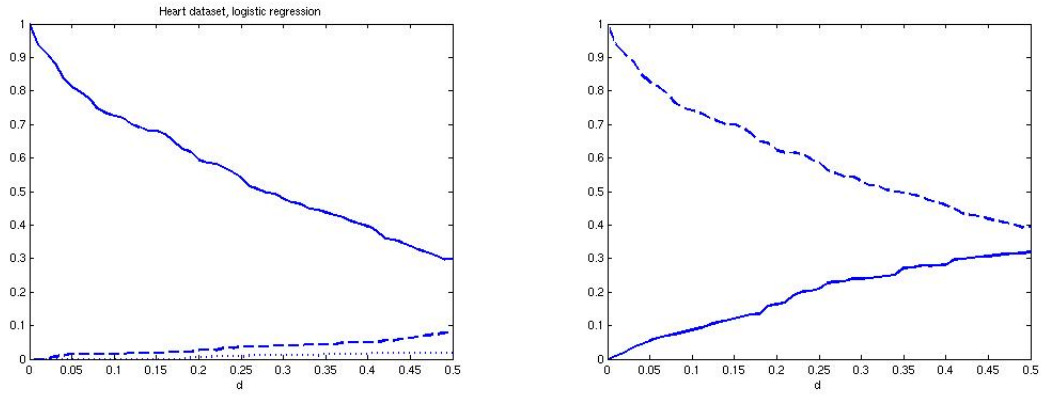

Figure 8: Logistic regression rule. *Left panel:* Error rates $\mathbb{P}\{Y = 1, f(X) = 0\}$, $\mathbb{P}\{Y = 0, f(X) = 1\}$ and rejection rate $\mathbb{P}\{f(X) = R\}$ against rejection cost d. *Right panel:* Risk function $\mathbb{P}\{Y = 1, f(X) = 0\} + \theta\mathbb{P}\{Y = 0, f(X) = 1\} + d\mathbb{P}\{f(X) = R\}$ and misclassification rate $\mathbb{P}\{Y \neq f(X)\}$ against rejection cost d.

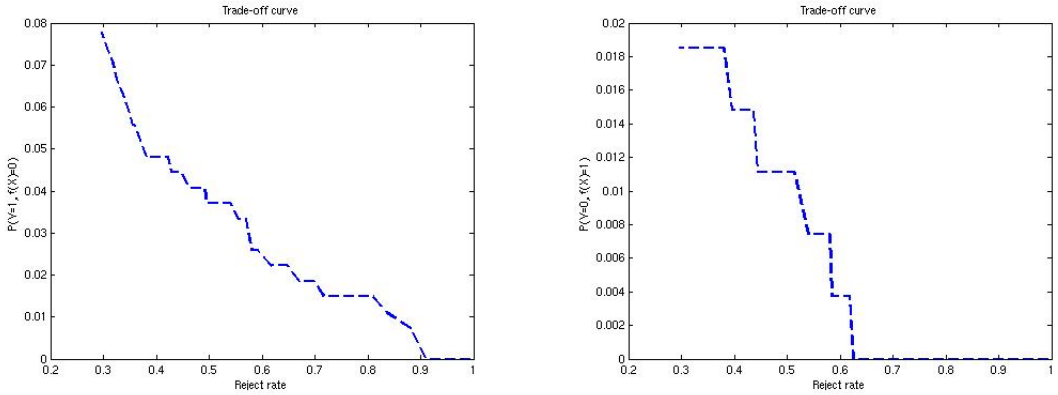

Figure 9: Logistic regression rule. Trade-off curves: $\mathbb{P}\{Y = 1, f(X) = 0\}$ against the rejection rate $\mathbb{P}\{f(X) = R\}$ (*left*) and $\mathbb{P}\{Y = 0, f(X) = 1\}$ against the rejection rate $\mathbb{P}\{f(X) = R\}$ (*right*).

References

[1] Anthony, M and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

[2] Audibert,J.-Y, Tsybakov, A. (2005). Fast convergence rates for plug-in classifiers under margin conditions. (personal communcation).

[3] Bartlett, P. L. and Mendelson, S. (2003). Empirical risk minimization. *Manuscript*, `http://users.rsise.anu.edu.au/~shahar/paper5.ps`

[4] Bartlett, P. L., Bousquet, O., and Mendelson, S. (2004) Local Rademacher complexities. *Annals of Statistics* (in press). `http://www.stat.berkeley.edu/~bartlett/papers/bbm-lrc-02b.pdf`.

[5] Biau, G., Bunea, F., and Wegkamp, M.H. (2005). Function Classification in Hilbert Spaces. *I.E.E.E. Transactions on Information Theory* (in press).

[6] Boucheron, S., Bousquet, O., and Lugosi, G. (2004b). *Theory of Classification: a Survey of Recent Advances. Manuscript.* `http://www.econ.upf.es/~lugosi/esaimsurvey.pdf`.

[7] Boucheron, S., Bousquet, O., and Lugosi, G. (2004a). Introduction to statistical learning theory. In *Advanced Lectures in Machine Learning* (O. Bousquet, U. von Luxburg, and G. Rätsch, Editors), 169–207, Springer.

[8] Clémençon, S., Lugosi, G., and Vayatis, N. (2005). Ranking and scoring using empirical risk minimization. *Manuscript.* `http://www.econ.upf.es/~lugosi/coltscoring.ps`.

[9] Chow, C.K. (1970). On optimum error and reject trade-off. *IEEE Transactions on Information Theory* **16**, 41–46.

[10] L. Devroye, L. Györfi, and G. Lugosi. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York.

[11] Freund, Y., Mansour, Y. and Schapire, R. E. (2004). Generalization bounds for averaged classifiers *Annals of Statistics*, **32**(4), 1698–1722.

[12] Fumera, G, and Roli, F. (2004). Analysis of error-reject trade-off in linearly combined multiple classifiers. *Pattern Recognition*, **37**, 1245–1265.

[13] Fumera, G., Roli, F., and Giacinto, G. (2000). Reject option with multiple thresholds. *Pattern Recognition* **33**, 2099–2101.

[14] Golfarelli, M., Maio, D., and Maltoni, D. (1997). On the error-reject trade-off in biometric verification systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(7), 786 –796.

[15] Hansen, L. K., Lissberg, C., and Salamon, P. (1997). The error-reject tradeoff. *Open Systems & Information Dynamics* **4**, 159 – 184.

[16] Massart, P. and Nédélec, E. (2003). Risk bounds for statistical learning. *Preprint, Université Paris Sud.* `http://www.math.u-psud.fr/~massart/margin.pdf`.

[17] Massart, P. (2003). St Flour Lecture Notes. *Manuscript.* `http://www.math.u-psud.fr/~massart/flour.pdf`.

[18] Mohammadi, L. (2005). *Estimation of thresholds in classification.* Ph-D thesis, University of Leiden.

[19] Ripley, B. D. (1996). *Pattern recognition and neural networks.* Cambridge University Press, Cambridge.

[20] Tarigan, B. and Van de Geer, S. A. (2004). Adaptivity of support vector machines with l_1 penalty. *Technical Report MI 2004-14, University of Leiden.*

[21] Tsybakov, A.B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, **32**(1), 135–166.

[22] Tsybakov, A. B. and Van de Geer, S. A. (2003). Square root penalty: adaptation to the margin in classification and in edge estimation. *Prépublication PMA-820, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII.*

[23] Van de Geer, S. A. (2000). *Empirical processes in M-estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.