The Canadian Journal of Statistics
Vol. 34, No. 4, 2006, Pages 709–721
La revue canadienne de statistique

Classification with reject option

Radu HERBEI and Marten H. WEGKAMP

Key words and phrases: Bayes classifiers; classification; empirical risk minimization; margin condition; plug-in rules; reject option.

MSC 2000: Primary 62C05; secondary 62G05, 62G08.

Abstract: The authors study binary classification that allows for a reject option in which case no decision is made. This reject option is to be used for those observations for which the conditional class probabilities are close and as such are hard to classify. The authors generalize existing theory for both plug-in rules and empirical risk minimizers to this setting.

La classification à clause de renvoi

Résumé : Les auteurs s'intéressent à une méthode de classification dichotomique permettant de laisser en suspens le classement de certaines observations dont les probabilités conditionnelles d'appartenance à l'une ou l'autre classe sont si proches qu'un choix est difficile à faire. Les auteurs étudient dans ce contexte les propriétés des règles de substitution et des règles de classement minimisant le risque empirique.

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 $\mathcal{X} = \mathbf{R}^k$. 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 independent observations (x_i, y_i), the statistician's task is to form a classifier $f: \mathbf{R}^k \to \{0, 1\}$ which represents his/her guess of the label Y for a future observation X. This framework is known as supervised learning in the literature. The classifier

$$f^*(x) := \begin{cases} 0 & \text{if } \mathrm{P}\{Y = 0 \,|\, X = x\} \geq \mathrm{P}\{Y = 1 \,|\, X = x\}, \\ 1 & \text{otherwise} \end{cases}$$

has the smallest probability of error $\mathrm{P}\{f(X) \neq Y\}$, see, for example, Devroye, Györfi & Lugosi (1996, p. 10, Th. 2.1). 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

$$\eta(x) := \mathrm{P}\{Y = 1 \,|\, X = x\}$$

is close to $1/2$. Indeed, if $\mathrm{P}\{Y = 0 \,|\, X = x\} = \mathrm{P}\{Y = 1 \,|\, X = x\} = 1/2$, 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 making a hard decision (0 or 1) is of great importance in practice. Nevertheless, this option is often ignored in the statistical literature. Ripley (1996) and recently Freund, Mansour & Schapire (2004) are notable exceptions. Some references in the engineering literature are Chow (1970), Fumera & Roli (2003), Fumera, Roli & Giacinto (2000), Golfarelli, Maio & Maltoni (1997), Györfi, Györfi & Vajda (1978) and Hansen, Liisberg & Salomon (1997).

We follow the decision theoretic framework of Chow (1970); see Ripley (1996, ch. 2) for a more general overview. Let $f: \mathbf{R}^k :\to \{0, 1, ®\}$ be a classifier with a reject option, where the interpretation of the output ® is of being in doubt and making no decision. The misclassification probability is $\mathrm{P}\{f(X) \neq Y, f(X) \neq ®\}$ and reject or doubt probability is $\mathrm{P}\{f(X) = ®\}$.

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

$$L_d(f) := d\,\mathrm{P}\{f(X) = ®\} + \mathrm{P}\{f(X) \neq Y, f(X) \neq ®\}.$$

Chow (1970) shows that the optimal rule minimizing this risk $L_d(f)$ assigns 0, 1 or ® depending on which of $\eta(x)$, $1 - \eta(x)$ or d is smallest. According to this rule, we should never invoke the reject option if $d \geq 1/2$ and we should always reject if $d = 0$. For this reason we restrict ourselves to the cases $0 \leq d \leq 1/2$. The rejection cost d can be viewed as an upper bound on the conditional probability of misclassification (given X) that is considered tolerable. The Bayes rule with reject option is then

$$f_d^*(x) := \begin{cases} 0 & \text{if } \eta(x) < d, \\ 1 & \text{if } \eta(x) > 1 - d, \\ ® & \text{otherwise,} \end{cases}$$

and we denote its risk by $L_d^* := L_d(f_d^*) = \mathrm{E}\min\{\eta(X), 1 - \eta(X), d\}$. The case $d = 1/2$ reduces to the classical situation without the reject option and the Bayes classifier f_d^* reduces to f^*.

The paper is organized as follows. Section 2 discusses plug-in rules that replace the regression function $\eta(x)$ by an estimate $\hat{\eta}(x)$ in the formula for f_d^* above. Besides introducing the reject option, we extend the existing theory for plug-in rules (Devroye, Györfi & Lugosi 1996, Th. 2.2) since our bound depends explicitly on both the difference $|\hat{\eta}(X) - \eta(X)|$ and the behaviour 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 & Tsybakov (2005) to our more general framework.

Section 3 studies classifiers that minimize the empirical counterpart of the risk $L_d(f)$ over a class of classifiers $\mathcal{F}$. We demonstrate that the rates of convergence of the risk $L_d(\hat{f}_{\mathrm{ERM}})$ of the resulting minimizers $\hat{f}_{\mathrm{ERM}}$ to the Bayes risk L_d^* depend on the metric entropy of (a transformed class of) $\mathcal{F}$ and on the behaviour of $\eta(X)$ near the values d and $1 - d$. Our results are in line with the recent theoretical developments of standard binary classification (see, e.g., Boucheron, Bousquet & Lugosi 2005, Massart & Nédélec 2006, Tsybakov 2004, Tarigan & van de Geer 2004 and Tsybakov & van de Geer 2005) and extend the theory to the general case $0 \leq d \leq 1/2$.

Section 4 allows for different misclassification costs of the cases $\{Y = 1,\ f(X) = 0\}$ and $\{Y = 0,\ f(X) = 1\}$, a situation common in medical studies where misclassifying a sick patient as healthy is worse than the opposite. The risk function $L_d(f)$ is changed to accommodate this differentiation and the results obtained in Section 2 for the plug-in estimates are generalized to this situation.

All proofs are relegated to the Appendix.

2. PLUG-IN RULES

In this section we consider the plug-in rule

$$\hat{f}_{\mathrm{PI}}(x) := \begin{cases} 0 & \text{if } \hat{\eta}(x) < d, \\ 1 & \text{if } \hat{\eta}(x) > 1 - d, \\ ® & \text{otherwise} \end{cases}$$

based on some estimate $\hat{\eta}(x)$ of the regression function $\eta(x)$ and the form of the optimal classifier f_d^*. Our main result of this section, Theorem 2 below, shows that the regret

$$\Delta_d(\hat{f}_{\mathrm{PI}}) := L_d(\hat{f}_{\mathrm{PI}}) - L_d^*$$

depends on how well $\hat{\eta}(X)$ estimates $\eta(X)$ and the behaviour of $\eta(X)$ near d and $1-d$. First we prove an auxiliary result which rewrites $\Delta_d(\hat{f})$ into a convenient form and which is an improvement of its counterpart in the standard binary classification setting (see, for example, Devroye, Györfi & Lugosi 1996, p. 16, Th. 2.2,). In the sequel, $\hat{\eta}$ is assumed to be independent of (X, Y).

LEMMA 1. *For any* $0 \le d \le 1/2$, *we have*

$$\begin{aligned}\Delta_d(\hat{f}_{\mathrm{PI}}) &= \mathrm{E}|d-\eta(X)|\big(\mathbb{I}_{\{f^*(X)=0,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}+\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=0,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}\big)\\ &\quad+\mathrm{E}|1-d-\eta(X)|\big(\mathbb{I}_{\{f^*(X)=1,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}+\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=1,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}\big).\end{aligned}$$

This lemma clearly confirms that the Bayes rule with reject option f_d^* minimizes the risk $L_d(f)$ as already shown in Chow (1970). Indeed, the right-hand side in the preceding display is nonnegative and equals zero if and only if $\hat{f}_{\mathrm{PI}} = f^*$.

THEOREM 2. *Let* $0 \le d \le 1/2$ *and define*

$$P_d(\delta) := \mathrm{P}\{|d-\eta(X)| \le \delta\} + \mathrm{P}\{|1-d-\eta(X)| \le \delta\},$$

for all $\delta \ge 0$. *Then*

$$\Delta_d(\hat{f}_{\mathrm{PI}}) \le \inf_{\delta\ge 0}\{2(1-d)\mathrm{P}\{|\eta(X)-\hat{\eta}(X)|>\delta\}+\delta P_d(\delta)\}.$$

This theorem extends the recent result by Audibert & Tsybakov (2005). Their inequality (3.2) without the peeling device can be simplified to, in absence of the reject option ($d = \frac{1}{2}$),

$$\Delta_{\frac{1}{2}}(\hat{f}_{\mathrm{PI}}) \le 2\delta\mathrm{P}\left\{0 < \left|\eta(X)-\frac{1}{2}\right| \le \delta\right\} + \mathrm{P}\{|\hat{\eta}(X)-\eta(X)| \ge \delta\}.$$

Theorem 2 indicates that fast rates (faster than $n^{-1/2}$) can be achieved using plug-in estimates. We briefly discuss two situations:

ASSUMPTION (A1). *There exists a* $\delta_0 > 0$ *such that* $P_d(\delta_0) = 0$.

ASSUMPTION (A2). *There exist* $A < \infty$, $\alpha \ge 0$ *such that* $P_d(\delta) \le A\delta^\alpha$ *for all* $\delta > 0$.

Condition (A1) means that $\eta(x)$ stays away from the values d and $1-d$. In this case very fast rates for $\Delta_d(\hat{f}_{\mathrm{PI}})$ are possible, depending on the smoothness of η only. This condition is used by Massart & Nédélec (2006) in the context of standard binary classification without the reject option ($d = 1/2$).

Condition (A2), analogous to Tsybakov's margin condition (see 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 $1-d$ for large values of α. We illustrate this by assuming that the probability $r_n(\delta) = \mathrm{P}\{|\hat{\eta}(X)-\eta(X)| \ge \delta\}$ is of the form $C_0\exp(-C_1 n^\gamma \delta^2)$ for some positive constants C_0, C_1 and γ. Typically, γ will depend on the degree of smoothness of η and the dimension k of the feature space. Condition (A2) ensures that $2\delta P_d(\delta) \le 2A\delta^{1+\alpha}$. Theorem 2 then guarantees that $\Delta_d(\hat{f})$ is bounded above by $r_n(\delta) + 2A\delta^{1+\alpha}$. Choosing $\delta = \log(n)/(C_1 n^{\gamma/2})$, we obtain that for some positive constant C_2,

$$\Delta_d(\hat{f}) \le C_2\{\log(n)/n^{\gamma/2}\}^{1+\alpha}.$$

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 (A1). This is the optimal situation where the fastest rates can be guaranteed. See Audibert & Tsybakov (2005) for the corresponding situation without the reject option.

3. EMPIRICAL RISK MINIMIZATION

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of the pair $(X, Y) \in \mathbf{R}^k \times \{0, 1\}$. This section discusses minimization of the empirical counterpart

$$\hat{L}_d(f) := \frac{d}{n} \sum_{i=1}^{n} \mathbb{I}_{\{f(X_i)=®\}} + \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{f(X_i)\neq Y_i, f(X_i)\neq ®\}}$$

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

$$\Delta_d(\hat{f}_{\mathrm{ERM}}) \le C_0 \inf_{f \in \mathcal{F}} \Delta_d(f) + R_n$$

for some constant $C_0 \ge 1$. The remainder R_n depends on the sample size n, the margin condition on $\eta(x)$ and the size of the class $\mathcal{F}$. For a general theory of empirical risk minimizers, we refer to, e.g., Anthony & Bartlett (1999), Bartlett & Mendelson (2003), Bartlett, Bousquet & Mendelson (2005), Boucheron, Bousquet & Lugosi (2004 and 2005), Devroye, Györfi & Lugosi (1996), Massart (2006), and van de Geer (2000) .

Since we are interested in $\Delta_d(f)$ rather than $L_d(f)$, we define the loss function

$$g_{f,d}(x, y) := d\big(\mathbb{I}_{\{f(x)=®\}} - \mathbb{I}_{\{f^*(x)=®\}}\big) + \big(\mathbb{I}_{\{f(x)\neq y, f(x)\neq ®\}} - \mathbb{I}_{\{f^*(x)\neq y, f^*(x)\neq ®\}}\big)$$

so that $\mathrm{E} g_{f,d}(X, Y) = \Delta_d(f)$. We demonstrated in Section 2 that the degree of difficulty of the classification problem depends heavily on the behaviour 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, e.g., Boucheron, Bousquet & Lugosi 2004, 2005; Bartlett & Mendelson 2003; Massart & Nédélec 2006, all for the case $d = 1/2$ only) that the way $\mathrm{E} g_{f,d}^2(X)$ relates to $\mathrm{E} g_{f,d}(X, Y)$ is an important ingredient for the rate of the remainder term R_n. Bartlett & Mendelson (2003) call a class $\mathcal{F}$ that satisfies

$$\mathrm{E} g_{f,d}^2(X, Y) \le B\big\{\mathrm{E} g_{f,d}(X, Y)\big\}^{\beta} \quad \text{for all } f \in \mathcal{F}$$

a Bernstein(β, B)-class. We first obtain this link under the two scenarios, Conditions (A1) and (A2), considered previously in Section 2.

LEMMA 3. *Let $f: \mathbf{R}^k \to \{0, 1, ®\}$ be a classifier with a reject option, and let $0 \le d \le \frac{1}{2}$. Assume that Condition* (A1) *holds, i.e., $P_d(\delta_0) = 0$ for some $\delta_0 > 0$. Then $\mathrm{E} g_{f,d}(X, Y) \ge s \mathrm{E} g_{f,d}^2(X, Y)$, where $s = \delta_0/(1 + d)^2$.*

LEMMA 4. *Let $f: \mathbf{R}^k \to \{0, 1, ®\}$ be a classifier with a reject option, and let $0 \le d \le \frac{1}{2}$. Assume that Condition* (A2) *holds, i.e. $\exists A > 0$, $\alpha \ge 0$ such that for all $t \ge 0$, $P_d(t) \le A t^{\alpha}$. Then*

$$\mathrm{E} g_{f,d}^2(X, Y) \le 2^{\alpha/(1+\alpha)} (1 + d)^2 (8A)^{1/(1+\alpha)} \big\{\mathrm{E} g_{f,d}(X, Y)\big\}^{\alpha/(\alpha+1)}.$$

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 $\hat{f}$ even in the nontrivial case of $L^* \neq 0$—under assumption (A2) in case $d = 1/2$.

These two preliminary results allow us to formulate the first theorem concerning empirical risk minimizers. It assumes minimization over a finite set of classifiers as is the case when we select tuning parameters such as a bandwidth or a dimension over a finite grid (possibly depending on the sample size n).

THEOREM 5. *Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a finite collection of classifiers $f: \mathbf{R}^k \to \{0, 1, ®\}$ with reject option ®. Assume that Condition* (A2) *holds. Then there exists a constant $C < \infty$ (depending on α, A and d) such that for all $n \geq 1$ and $\rho > 0$,*

$$\mathrm{E}\Delta_d(\hat{f}_{\mathrm{ERM}}) \leq (1+\rho) \min_{f \in \mathcal{F}} \Delta_d(f) + C\rho^{-\frac{\alpha}{2+\alpha}} \left(\frac{(1+\rho)^2 \log M}{n} \right)^{\frac{1+\alpha}{2+\alpha}},$$

where α is defined in Condition (A2).

If the stronger assumption (A1) holds, it follows easily from the proof of Theorem 5 that

$$\mathrm{E}\Delta_d(\hat{f}_{\mathrm{ERM}}) \leq (1+\rho) \min_{f \in \mathcal{F}} \Delta_d(f) + C' \frac{(1+\rho)^2 \log M}{\rho n}$$

for some finite constant C'. The next result extends Theorem 5 in that it allows for infinite classes $\mathcal{F}$.

THEOREM 6. *Assume that Condition* (A2) *holds and set $\beta = \alpha/(1+\alpha)$. Let $\mathcal{F}$ be a collection of classifiers $f: \mathbf{R}^k \to \{0, 1, ®\}$ with reject option ®. Assume that the classes of sets $\{x \in \mathbf{R}^k : f(x) = 1\}$, $\{x \in \mathbf{R}^k : f(x) = 0\}$ and $\{x \in \mathbf{R}^k : f(x) = ®\}$, indexed by $f \in \mathcal{F}$, are VC classes with finite VC-dimension V. Then for all $\rho > 0$ and n large enough, there exists a finite constant C'' depending on A, α, ρ and V, such that*

$$\mathrm{E}\Delta_d(\hat{f}_{\mathrm{ERM}}) \leq (1+\rho) \inf_{f \in \mathcal{F}} \Delta_d(f) + C'' n^{-\frac{1+\alpha}{2+\alpha}}.$$

This result extends the result by Massart & Nédélec (2006), 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 loss functions that allow for faster computation.

5. DISTINCT MISCLASSIFICATION ERRORS

Assume that making the error predicting $f(X) = 0$ while $Y = 1$ is more costly than making the error predicting $f(X) = 1$ while $Y = 0$. This situation arises often in practice for instance in risk management or medical diagnostics. In order to accommodate for this, we consider the more general risk function

$$L_{d,\theta}(f) := d\,\mathrm{P}\{f(X) = ®\} + \mathrm{P}\{Y = 1, f(X) = 0) + \theta \mathrm{P}(Y = 0, f(X) = 1\},$$

with $\theta < 1$. The optimal rule minimizing this new risk assigns 0, 1, or ® depending on which of the following is smallest: $\eta(x)$, $\theta\{1 - \eta(x)\}$, or d and it equals

$$f^*_{d,\theta}(x) := \begin{cases} 0 & \text{if } \eta(x) < \min\{d, \theta/(1+\theta)\}, \\ 1 & \text{if } \eta(x) > \max\{1 - d/\theta, \theta/(1+\theta)\}, \\ ® & \text{if } d \leq \eta(x) \leq 1 - d/\theta. \end{cases}$$

Because $\theta < 1$, we consider $d \leq \theta/(1+\theta)$ only since we would never invoke the reject option for larger values of d. Thus we have

$$f^*_{d,\theta}(x) = \begin{cases} 0 & \text{if } \eta(x) < d, \\ 1 & \text{if } \eta(x) > 1 - d/\theta, \\ \text{®} & \text{if } d \leq \eta(x) \leq 1 - d/\theta. \end{cases}$$

We denote the Bayes error by $L^*_{d,\theta} := L^*_{d,\theta}(f^*_{d,\theta}) = \mathrm{E}\min\{\eta(X), \theta(1-\eta(X)), d\}$ and define, for any classifier f, the regret $\Delta_{d,\theta}(f) := L_{d,\theta}(f) - L^*_{d,\theta}$.

LEMMA 7. *Let $\hat{f}$ be the plug-in rule that replaces η by an estimate $\hat{\eta}$ in the Bayes rule $f^*_{d,\theta}$ above. For any $\theta < 1$ and $d \leq \theta/(1+\theta)$, we have*

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

The equivalent of Theorem 2 is the following result.

THEOREM 8. *Let $\hat{f}$ be the plug-in rule that replaces η by an estimate $\hat{\eta}$ in the Bayes rule $f^*_{d,\theta}$ above. Let $\theta < 1$, $d \leq \theta/(1+\theta)$ and define*

$$P_{d,\theta}(\delta) := \mathrm{P}\{|d - \eta(X)| \leq \delta\} + \mathrm{P}\left\{\left|1 - \frac{d}{\theta} - \eta(X)\right| \leq \delta\right\}$$

for all $\delta \geq 0$. We have

$$\Delta_d(\hat{f}) \leq \inf_{\delta \geq 0}\left\{\left(2 - d - \frac{d}{\theta}\right)\mathrm{P}\{|\eta(X) - \hat{\eta}(X)| > \delta\} + \delta P_{d,\theta}(\delta)\right\}.$$

6. ILLUSTRATIONS

As an illustration to the plug-in rule with reject option, we consider two classification examples, corresponding to the case of equal misclassification costs ($\theta = 1$) and that of distinct misclassification costs ($\theta \neq 1$). Before we proceed, a few important observations should be made. In both situations the cost θ and the rejection cost d play an important role. The purpose of the following illustrations is to present and discuss the roles of d and θ under our setting, in particular their effect on the misclassification rates. Throughout the present work d and θ are fixed quantities and not parameters to be estimated by the user. They are specified beforehand, independently of the data and the classification procedure used. This is crucial for the setting of the paper and the risk functions L_d and $L_{d,\theta}$ are meaningful only in this setting. We understand that for each given problem, the user has to determine these costs, a task which is not necessarily trivial. In a medical setting when classifying whether a disease is present or absent, the reject option presumably leads to quantifiable costs for additional tests and perhaps delays in applying the treatment. It may improve the chance of making the right decision, while making a mistake may result in other costs associated with the patient being treated for a nonexistent disease or not treated at all. Also, under the second setup ($\theta \neq 1$), when a patient is erroneously classified as being healthy (a situation that is now considered worse than the opposite), late treatment could potentially lead to a different set of quantifiable costs, leading to a numerical value of θ. Similarly, when a loan application is classified as approved or not, a mistake may result in quantifiable losses (interest,

bank charges, costs associated with recuperating the money), while costs related to employing a reject option may be more affordable (a few extra labor hours) yet it may result in making the right decision. All these situations lead to different ways of selecting (determining) d or θ.

Our first illustration focuses on the case when errors have equal costs. We considered the well known Pima Indian diabetes data set (Ripley 1996, p. 114) and used a local linear regression rule to perform the classification. In a preprocessing stage we filtered the data set by eliminating cases with missing observations. We kept a total of $n = 395$ records. We fitted a logistic regression model to the remaining data and selected *gulose, bmi, diabetes* and *age* as significant covariates. For each case, we further fitted a leave-one-out local linear regression model using the *locfit package* (with the default options) available in *R*. The left panel of Figure 1 displays estimates of the error rate $\mathrm{P}(Y \neq f(X), f(X) \neq \text{®})$ (dashed) and the reject rate $\mathrm{P}(f(X) = \text{®})$ (solid). Under no rejection ($d = 1/2$), the largest error rate is about 22%. As Chow (1970) points out, the advantages of classifying with a reject option as well as the performance of the classification procedure itself can be judged by the error-reject trade-off that is displayed on the right panel of Figure 1. In this example we observe that a small reject rate ($10\% - 15\%$) reduces the error rate to $14\% - 16\%$.

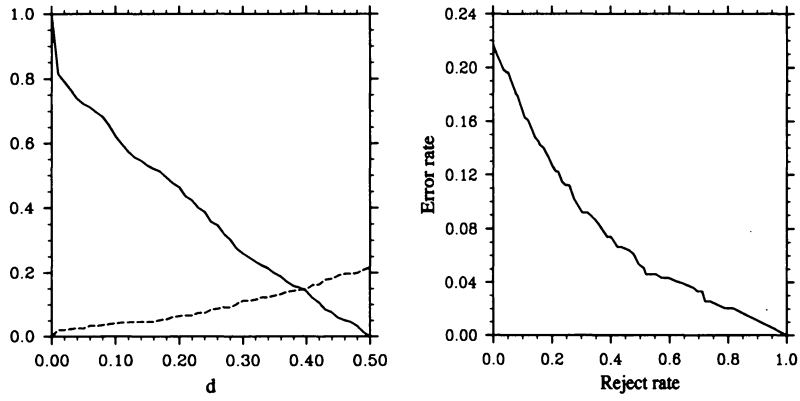

FIGURE 1: Pima data set; left panel: error rate (dashed) and reject rate (solid) as functions of the reject cost d; right panel: trade-off curve.

For the second illustration, we considered 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 ($Y = 0$) and presence ($Y = 1$) of heart disease. The description of the data set informed us that the cost of misclassifying a sick person as healthy is five times higher than the reverse, and thus we set $\theta = 1/5$. The data were processed in a similar way as described above for the Pima data set and we selected *sex, chest pain type, resting blood pressure, maximum heart rate achieved, number of major vessels colored by fluorosopy* and *thal* as significant covariates. We performed the same classification procedure as before; the results are summarized in Figure 2. The left panel displays the estimated error rates (misclassifying a sick person as healthy – dashed and the reverse – dotted) and the estimated reject rate as functions of the reject cost d. The two trade-off curves are presented in the right panel of Figure 2. We notice that if the reject option is not used the most costly mistake has a very low rate of about 4%, while classifying a healthy person as sick occurred in 24% of the cases. Introducing the reject option allows for a further reduction in the error rates, more significantly for the less costly error. A significant error rate reduction when classifying a healthy person as sick requires in this case a high reject rate.

In conclusion, under a standard binary classification setup, when mistakes are too costly, error rates are reduced by using a reject option.

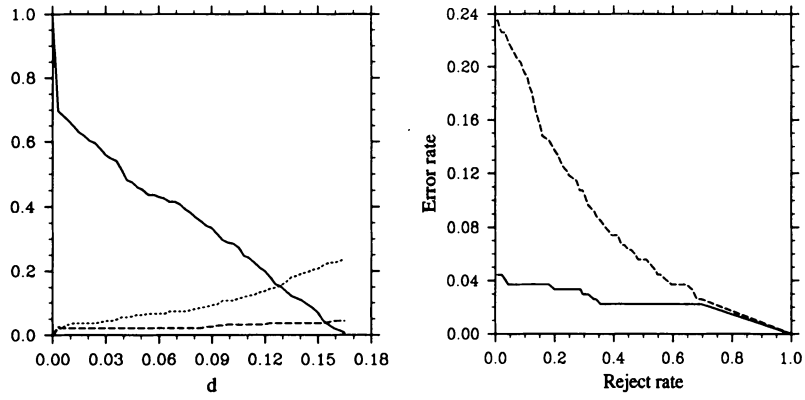

FIGURE 2: Heart data set; left panel: error rates ($\mathrm{P}(f(X) = 0, Y = 1)$ – dashed and $\mathrm{P}(f(X) = 1, Y = 0)$ – dotted) and reject rate (solid) as functions of the reject cost d; right panel: trade-off curves ($\mathrm{P}(f(X) = 0, Y = 1)$ versus Reject rate – solid and $\mathrm{P}(f(X) = 1, Y = 0)$ versus Reject rate – dashed).

APPENDIX

Proof of Lemma 1. Observe that

$$\begin{aligned}\Delta_d(\hat{f}_{\mathrm{PI}}) &= d\big(\mathrm{P}\{\hat{f}_{\mathrm{PI}}(X) = ®\} - \mathrm{P}\{f^*(X) = ®\}\big) \\ &\quad + \big(\mathrm{P}\{\hat{f}_{\mathrm{PI}}(X) = 1, Y = 0\} + \mathrm{P}\{\hat{f}_{\mathrm{PI}}(X) = 0, Y = 1\}\big) \\ &\quad - \big(\mathrm{P}\{f^*(X) = 1, Y = 0\} + \mathrm{P}\{f^*(X) = 0, Y = 1t\}\big)\end{aligned}$$

and, for any classifier with reject option f,

$$\begin{aligned}&\mathrm{P}\{f(X) = 1, Y = 0\} + \mathrm{P}\{f(X) = 0, Y = 1\} \\ &\qquad = \mathrm{E}(1 - \eta(X))\mathbb{I}_{\{f(X)=1\}} + \mathrm{E}\eta(X)\mathbb{I}_{\{f(X)=0\}} \\ &\qquad = \mathrm{E}(1 - 2\eta(X))\mathbb{I}_{\{f(X)=1\}} - \mathrm{E}\eta(X)\mathbb{I}_{\{f(X)=®\}} + \mathrm{E}\eta(X).\end{aligned}$$

Using the above for $\hat{f}_{\mathrm{PI}}$ and f^*, the regret becomes

$$\begin{aligned}\Delta_d(\hat{f}_{\mathrm{PI}}) &= \mathrm{E}(1 - 2\eta(X))\big(\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=1\}} - \mathbb{I}_{\{f^*(X)=1\}}\big) \\ &\quad + \mathrm{E}(d - \eta(X))\big(\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=®\}} - \mathbb{I}_{\{f^*(X)=®\}}\big) \\ &= \mathrm{E}(d - \eta(X) + 1 - d - \eta(X))\big(\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=1\}} - \mathbb{I}_{\{f^*(X)=1\}}\big) \\ &\quad + \mathrm{E}(d - \eta(X))\big(\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=®\}} - \mathbb{I}_{\{f^*(X)=®\}}\big).\end{aligned}$$

From the definition of f_d^*, and after splitting the indicator functions using disjoint events, we find

$$\begin{aligned}\Delta_d(\hat{f}_{\mathrm{PI}}) &= \mathrm{E}\big(|d - \eta(X)| + |1 - d - \eta(X)|\big)\big(\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=1,\, f^*(X)=0\}} + \mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=0,\, f^*(X)=1\}}\big) \\ &\quad + \mathrm{E}|d - \eta(X)|\big(\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=0,\, f^*(X)=®\}} + \mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=®,\, f^*(X)=0\}}\big) \\ &\quad + \mathrm{E}|1 - d - \eta(X)|\big(\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=1,\, f^*(X)=®\}} + \mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=®,\, f^*(X)=1\}}\big) \\ &= \mathrm{E}|d - \eta(X)|\big(\mathbb{I}_{\{f^*(X)=0,\, \hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}} + \mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=0,\, \hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}\big) \\ &\quad + \mathrm{E}|1 - d - \eta(X)|\big(\mathbb{I}_{\{f^*(X)=1,\, \hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}} + \mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=1,\, \hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}\big).\end{aligned}$$

The proof of the lemma is complete. □

Proof of Theorem 2. Recall from the proof of Lemma 1 that

$$\begin{aligned}\Delta_d(\hat{f}) &= \mathrm{E}|d-\eta(X)|\big(\mathbb{I}_{\{f^*(X)=0,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}+\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=0,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}\big)\\ &\quad+\mathrm{E}|1-d-\eta(X)|\big(\mathbb{I}_{\{f^*(X)=1,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}+\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=1,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}\big).\end{aligned}$$

Let F_1, F_2 denote the following events

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

Next we argue that

$$\begin{aligned}&\mathrm{E}|d-\eta(X)|\big(\mathbb{I}_{\{f^*(X)=0,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}+\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=0,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}\big)\\ &\quad=\mathrm{E}|d-\eta(X)|\mathbb{I}_{F_1}\big(\mathbb{I}_{\{|\eta(X)-\hat{\eta}(X)|>\delta\}}+\mathbb{I}_{\{|\eta(X)-\hat{\eta}(X)|\le\delta\}}\big)\\ &\quad\le\mathrm{E}|d-\eta(X)|\mathbb{I}_{\{|\eta(X)-\hat{\eta}(X)|>\delta\}}+\mathrm{E}|d-\eta(X)|\mathbb{I}_{\{|d-\eta(X)|\le\delta\}}\\ &\quad\le(1-d)\mathrm{P}\big\{|\eta(X)-\hat{\eta}(X)|>\delta\big\}+\delta\mathrm{P}\big\{|d-\eta(X)|\le\delta\big\}\end{aligned}$$

and

$$\begin{aligned}&\mathrm{E}|1-d-\eta(X)|\big(\mathbb{I}_{\{f^*(X)=1,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}+\mathbb{I}_{\{\hat{f}_{\mathrm{PI}}(X)=1,\,\hat{f}_{\mathrm{PI}}(X)\neq f^*(X)\}}\big)\\ &\quad=\mathrm{E}\,|1-d-\eta(X)|\,\mathbb{I}_{F_2}\big(\mathbb{I}_{\{|\eta-\hat{\eta}|>\delta\}}+\mathbb{I}_{\{|\eta(X)-\hat{\eta}(X)|\le\delta\}}\big)\\ &\quad\le(1-d)\mathrm{P}\big\{|\eta(X)-\hat{\eta}(X)|>\delta\big\}+\delta\mathrm{P}\big\{|1-d-\eta(X)|\le\delta\big\}.\end{aligned}$$

Combining the two preceding displays concludes the proof. □

Proof of Lemma 3. Since f and d are fixed, we write g in place of $g_{f,d}$. Observe that

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

where the last inequality follows from the bound

$$\begin{aligned}|g(x,y)| &\le d\big(\mathbb{I}_{\{f(x)=\text{®},\,f(x)\neq f^*(x)\}}+\mathbb{I}_{\{f^*(x)=\text{®},\,f(x)\neq f^*(x)\}}\big)\\ &\quad+\big(\mathbb{I}_{\{f(x)=1,\,f^*(x)=0\}}+\mathbb{I}_{\{f(x)=0,\,f^*(x)=1\}}\\ &\quad+\mathbb{I}_{\{f(x)=\text{®},\,f(x)\neq f^*(x)\}}+\mathbb{I}_{\{f^*(x)=\text{®},\,f(x)\neq f^*(x)\}}\big)\\ &=(1+d)\big(\mathbb{I}_{\{f(x)=\text{®},\,f(x)\neq f^*(x)\}}+\mathbb{I}_{\{f^*(x)=\text{®},\,f(x)\neq f^*(x)\}}\big)\\ &\quad+(1+d)\big(\mathbb{I}_{\{f(x)=1,\,f^*(x)=0\}}+\mathbb{I}_{\{f(x)=0,\,f^*(x)=1\}}\big)\\ &\le(1+d)\mathbb{I}_{\{f(x)\neq f^*(x)\}}.\end{aligned}$$

This proves the lemma. □

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

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

Lemma 1 implies that

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

The first term on the right is bounded by

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

The three other terms in the penultimate display can be handled in a similar way, and we obtain

$$\Delta_d(f) \geq t\big(\{\mathrm{P}(E_1 \cup E_2 \cup E_3 \cup E_4) - 4At^\alpha\} \geq t(\mathrm{P}\{f(X) \neq f^*(X)\} - 4At^\alpha\big).$$

Choosing

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

we find

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

or

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

and we obtain

$$\mathrm{E}g^2(X, Y) \leq (1 + d)^2\mathrm{P}\{f(X) \neq f^*(X)\} \leq (1 + d)^2\big(2(8A)^{1/\alpha}\Delta_d(f)\big)^{\frac{\alpha}{1+\alpha}}.$$

This concludes the proof. □

Proof of Theorem 5. Define the empirical counterpart of the regret $\Delta_d(f)$ by

$$\hat{\Delta}_d(f) := \frac{1}{n}\sum_{i=1}^n g_{f,d}(X_i, Y_i)$$

and set $\beta = \alpha/(1 + \alpha)$ with α as in (A2). Since $\hat{f}_{\mathrm{ERM}}$ minimizes $\hat{\Delta}_d(f)$, we find that for any $f \in \mathcal{F}$ and $\rho > 0$,

$$\begin{aligned}
\Delta_d(\hat{f}_{\mathrm{ERM}}) &= (1 + \rho)\hat{\Delta}_d(\hat{f}_{\mathrm{ERM}}) + \big\{\Delta_d(\hat{f}_{\mathrm{ERM}}) - (1 + \rho)\hat{\Delta}_d(\hat{f}_{\mathrm{ERM}})\big\} \\
&\leq (1 + \rho)\hat{\Delta}_d(f) + \big\{\Delta_d(\hat{f}_{\mathrm{ERM}}) - (1 + \rho)\hat{\Delta}_d(\hat{f}_{\mathrm{ERM}})\big\},
\end{aligned}$$

and consequently,

$$\mathrm{E}\Delta_d(\hat{f}_{\mathrm{ERM}}) \leq (1 + \rho)\min_{f\in\mathcal{F}}\Delta_d(f) + \mathrm{E}\big\{\Delta_d(\hat{f}_{\mathrm{ERM}}) - (1 + \rho)\hat{\Delta}_d(\hat{f}_{\mathrm{ERM}})\big\}.$$

Since

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

we find by the union bound, Bernstein's inequality and Lemma 4 that there exists a $\kappa < \infty$ (depending on α, A and d) such that for all $\delta > 0$,

$$\begin{aligned}
&\mathrm{P}\{\Delta_d(\hat{f}_{\mathrm{ERM}}) - (1+\rho)\hat{\Delta}_d(\hat{f}_{\mathrm{ERM}}) \geq \delta\} \\
&\quad \leq \sum_{j=1}^{M} \exp\left(-\frac{n}{2}\frac{\left[\{\delta + \rho\Delta_d(f_j)\}/(1+\rho)\right]^2}{\kappa\{\Delta_d(f_j)\}^\beta + \{\delta + \rho\Delta_d(f_j)\}/\{3(1+\rho)\}}\right) \\
&\quad \leq \sum_{j=1}^{M} \exp\left(-\frac{n}{4}\frac{\left[\{\delta + \rho\Delta_d(f_j)\}/(1+\rho)\right]^2}{\kappa\{\Delta_d(f_j)\}^\beta\}}\right) \\
&\qquad + \sum_{j=1}^{M} \exp\left(-\frac{n}{4}\frac{\left[\{\delta + \rho\Delta_d(f_j)\}/(1+\rho)\right]^2}{\{\delta + \rho\Delta_d(f_j)\}/\{3(1+\rho)\}}\right) \\
&\quad \leq M\exp\left(-\frac{\rho^\beta n\delta^{2-\beta}}{4\kappa(1+\rho)^2}\right) + M\exp\left(-\frac{3n\delta}{4(1+\rho)}\right).
\end{aligned}$$

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

Proof of Theorem 6. Assume without loss of generality that the minimum of $\Delta_d(f)$ is attained for some $\bar{f} \in \mathcal{F}$, and write $\bar{\Delta}_d := \Delta_d(\bar{f})$. Since $\hat{\Delta}_d(\hat{f}_{\mathrm{ERM}}) \leq \hat{\Delta}_d(\bar{f})$, we find that

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

where

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

Consequently, we need to bound probabilities

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

Observe further that for each $f \in \mathcal{F}_j$, there exists $\kappa' < \infty$ (depending on α, A, d and ρ) with

$$\mathrm{E}\{g_{\bar{f},d}(X,Y) - g_{f,d}(X,Y)\}^2 \leq \kappa'(\bar{\Delta}_d + 2^j\delta)^\beta,$$

by Lemma 4 and $\beta = \alpha/(1+\alpha)$ with α defined in (A2). Invoke the VC-property of the classes of sets $\{x \in \mathbf{R}^k : f(x) = 1\}$, $\{x \in \mathbf{R}^k : f(x) = 0\}$ and $\{x \in \mathbf{R}^k : f(x) = ®\}$, $f \in \mathcal{F}$, and van der Vaart and Wellner (1996, p. 136, Th. 2.6.4; p. 248, Th. 2.14.16) to get $\sum_{j\geq 1} P_j \leq C_0\exp(-c_0 n\delta^{2-\beta})$ for some finite, positive constants c_0 and C (depending on α, A, d and ρ) and n large enough. The conclusion follows after a simple integration argument.

Proof of Lemma 7. Following the proof of Lemma 1, we find that

$$\begin{aligned}
\Delta_{d,\theta}(\hat{f}) &= \theta\big(\mathrm{P}\{\hat{f}(X) = 1, Y = 0\} - \mathrm{P}\{f^*(X) = 1, Y = 0\}\big) \\
&\quad + d\big(\mathrm{P}\{\hat{f}(X) = ®\} - \mathrm{P}\{f^*(X) = ®\}\big) \\
&\quad + \big(\mathrm{P}\{\hat{f}(X) = 0, Y = 1\} - \mathrm{P}\{f^*(X) = 0, Y = 1\}\big).
\end{aligned}$$

Recall that

$$\mathrm{P}\{\hat{f}(X) = 1, Y = 0\} = \mathrm{E}(1 - \eta(X))\mathbb{I}_{\{\hat{f}(X)=1\}}$$

and

$$\mathrm{P}\{\hat{f}(X) = 0, Y = 1\} = \mathrm{E}\eta(X)\big(1 - \mathbb{I}_{\{\hat{f}(X)=1\}} - \mathbb{I}_{\{\hat{f}(X)=\textcircled{R}\}}\big)$$

Thus,

$$\begin{aligned}
\Delta_{d,\theta}(\hat{f}) &= \mathrm{E}(\theta - \theta\eta(X) - \eta(X))\big(\mathbb{I}_{\{\hat{f}(X)=1\}} - \mathbb{I}_{\{f^*(X)=1\}}\big) \\
&\quad + \mathrm{E}(d - \eta(X))\big(\mathbb{I}_{\{\hat{f}(X)=\textcircled{R}\}} - \mathbb{I}_{\{f^*(X)=\textcircled{R}\}}\big) \\
&= \mathrm{E}\left[d - \eta(X) + \theta\left(1 - \frac{d}{\theta} - \eta(X)\right)\right]\big(\mathbb{I}_{\{\hat{f}(X)=1\}} - \mathbb{I}_{\{f^*(X)=1\}}\big) \\
&\quad + \mathrm{E}(d - \eta(X))\big(\mathbb{I}_{\{\hat{f}(X)=\textcircled{R}\}} - \mathbb{I}_{\{f^*(X)=\textcircled{R}\}}\big) \\
&= \mathrm{E}\left(|d - \eta(X)| + \theta\left|1 - \frac{d}{\theta} - \eta(X)\right|\right)\big(\mathbb{I}_{\{\hat{f}(X)=1,\, f^*(X)=0\}} + \mathbb{I}_{\{\hat{f}(X)=0,\, f^*(X)=1\}}\big) \\
&\quad + \mathrm{E}|d - \eta(X)|\big(\mathbb{I}_{\{\hat{f}(X)=0,\, f^*(X)=\textcircled{R}\}} + \mathbb{I}_{\{\hat{f}(X)=\textcircled{R},\, f^*(X)=0\}}\big) \\
&\quad + \mathrm{E}\theta\left|1 - \frac{d}{\theta} - \eta(X)\right|\big(\mathbb{I}_{\{\hat{f}(X)=1,\, f^*(X)=\textcircled{R}\}} + \mathbb{I}_{\{\hat{f}(X)=\textcircled{R},\, f^*(X)=1\}}\big) \\
&= \mathrm{E}|d - \eta(X)|\big(\mathbb{I}_{\{f^*(X)=0,\, \hat{f}(X)\neq f^*(X)\}} + \mathbb{I}_{\{\hat{f}(X)=0,\, \hat{f}(X)\neq f^*(X)\}}\big) \\
&\quad + \mathrm{E}\theta\left|1 - \frac{d}{\theta} - \eta(X)\right|\big(\mathbb{I}_{\{f^*(X)=1,\, \hat{f}(X)\neq f^*(X)\}} + \mathbb{I}_{\{\hat{f}(X)=1,\, \hat{f}(X)\neq f^*(X)\}}\big)
\end{aligned}$$

which concludes the proof. □

Proof of Theorem 8. The proof is very similar to the proof of Theorem 2 and is for this reason omitted.

ACKNOWLEDGEMENTS

The authors thank the Editor, Associate Editor and the two referees for helpful comments. Wegkamp's research was supported in part by a grant from the National Science Foundation.

REFERENCES

M. Anthony & P. L. Bartlett (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

J.-Y. Audibert & A. B. Tsybakov (2005). Fast convergence rates for plug-in classifiers under margin conditions. Prépublication, Laboratoire de Probabilités et Modèles Aléatoires, Universitś Paris VI et VII–CNRS.

P. L. Bartlett & S. Mendelson (2003). Empirical risk minimization. Unpublished manuscript.

P. L. Bartlett, O. Bousquet & S. Mendelson (2005). Local Rademacher complexities, *The Annals of Statistics*, 33, 1497–1537.

S. Boucheron, O. Bousquet & G. Lugosi (2004). Introduction to statistical learning theory. In *Advanced Lectures in Machine Learning* (O. Bousquet, U. von Luxburg & G. Rätsch, eds.), Springer, New York, pp. 169–207.

S. Boucheron, O. Bousquet & G. Lugosi (2005). Theory of classification: a survey of recent advances, *ESAIM: Probability and Statistics*, 9, 323–375.

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

L. Devroye, L. Györfi & G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Y. Freund, Y. Mansour & R. E. Schapire (2004). Generalization bounds for averaged classifiers. *The Annals of Statistics*, 32, 1698–1722.

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

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

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

L. Györfi, Z. Györfi & I. Vajda (1978). Bayesian decision with rejection. *Problems of Control and Information Theory*, 8, 445–452.

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

P. Massart (2006). *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII–2003*. Springer, in press. Preprint available online in pdf format at: http://www.math.u-psud.fr/~massart/stf2003_massart.pdf

P. Massart & E. Nédélec (2006). Risk bounds for statistical learning. *The Annals of Statistics*, in press.

B. D. Ripley (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

B. Tarigan & S. A. van de Geer (2004). Adaptivity of support vector machines with l_1 penalty. Technical Report MI 2004-14, University of Leiden.

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

A. B. Tsybakov & S. A. van de Geer (2005). Square root penalty: adaptation to the margin in classification and in edge estimation, *The Annals of Statistics*, 33, 1203–1224.

S. A. van de Geer (2000). *Empirical Processes in M-estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

A. W. van der Vaart & J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

Received 23 June 2005
Accepted 8 June 2006

Radu HERBEI: herbei@stat.ohio-state.edu
Department of Statistics, Ohio State University
Columbus OH 43210-1247, USA

Marten H. WEGKAMP: wegkamp@stat.fsu.edu
Department of Statistics, The Florida State University
Tallahassee, FL 32306-4330, USA