Proceedings of Machine Learning Research 116:66–80, 2020 Machine Learning for Health (ML4H) at NeurIPS 2019

Minimax Classifier with Box Constraint on the Priors

Cyprien Gilet GILET@I3S.UNICE.FR
Université Côte d'Azur, CNRS, I3S laboratory, Sophia-Antipolis, France

Susana Barbosa SUDOCARMO@GMAIL.COM
Université Côte d'Azur, CNRS, IPMC laboratory, Sophia-Antipolis, France

Lionel Fillatre LIONEL.FILLATRE@I3S.UNICE.FR
Université Côte d'Azur, CNRS, I3S laboratory, Sophia-Antipolis, France

Editors: Adrian V. Dalca, Matthew Mcdermott, Emily Alsentzer, Sam Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones

Abstract

Learning a classifier in safety-critical applications like medicine raises several issues. Firstly, the class proportions, also called priors, are in general imbalanced or uncertain. Secondly, the classifier must consider some bounds on the priors taking the form of box constraints provided by experts. Thirdly, it is also necessary to consider any arbitrary loss function given by experts to evaluate the classification decision. Finally, the dataset may contain both categorical and numerical features. To deal with both categorical and numerical features, the numerical attributes are discretized. When considering only discrete features, we propose in this paper a box-constrained minimax classifier which addresses all the mentioned issues. We derive a projected subgradient algorithm to compute this classifier. The convergence of this algorithm is established. We finally perform experiments on the Framingham heart database for illustrating the relevance of our algorithm in health care field.

1. Introduction

Context and problem statement The task of supervised classification is becoming increasingly promising in medicine fields such as medical diagnosis or health care. However, in such applications, we often have to face four difficulties. Firstly, the training set is generally imbalanced, i.e., the classes are not equally represented. In this case, minimizing the empirical risk leads the classifier to minimize the class-conditional risks of the classes with the largest number of samples. A minority class with just a small number of occurrences will tend to have a large class-conditional risk (Elkan, 2001). Furthermore, when some classes contain only a small number of samples, we can not claim that the class proportions of the training set correspond to the true state of nature. A classifier fitted on such a training set may have a poor performance on the test set (Poor, 1994). Secondly, experts in the application domain are generally able to provide us with some bounds on the class proportions. For example, in case of a medical disease, it is reasonable to bound the maximum frequency of a given disease. We can expect that the bound will improve the performance of a classifier. Thirdly, the experts can require the use of a specific loss function for evaluating the classification decisions. For example, if the classifier confuses a throat infection with a cold, the consequences are not the same as confusing a throat infection with

a lung cancer. Finally, we often have to deal with both numeric and categorical features. Many works have shown that the discretization of the numerical features can lead to results with better accuracy (Dougherty et al., 1995; Peng et al., 2009; Yang and Webb, 2009; García et al., 2016; Lustgarten et al., 2008). In this paper, we consider that the numerical features are discretized such that the classifier must only process discrete features. The goal of this paper is to build a classifier which addresses these four issues.

Related works A common approach to deal with imbalanced datasets is to balance the data by resampling the training set. But this approach may increase the misclassification risk when classifying some test samples which are imbalanced. Another common approach is the cost sensitive learning (Ávila Pires et al., 2013; Drummond and C. Holte, 2003) which aims at optimizing the cost of class misclassifications in order to counterbalance the number of occurrences of each class. However, this approach transforms the loss function provided by the experts, and these costs are generally difficult to tune. The task of learning the class-proportions which maximize the minimum empirical risk was already studied in past years. A pioneering work on the minimax criterion in the field of machine learning is (Cannon et al., 2002). This work studies the generalization error of a minimax classifier but it does not give any method to compute it. In (Kaizhu et al., 2004), the authors proposed the Minimum Error Minimax Probability Machine for the task of binary classification only. The extension to multiple classes is difficult. This method is very close to (Kaizhu et al., 2006). The Support Vector Machine (SVM) classifier can also be tuned in order to minimize the maximum class-conditional risks. The study proposed in (Davenport et al., 2010) is limited to the linear classifiers (using or not a feature mapping) and to the classification problems between only two classes. In (Farnia and Tse, 2016), the authors proposed an approach which fits a decision rule by learning the probability distribution which minimizes the worst-case of misclassification over a set of distributions centered at the empirical distribution. When the class-conditional distributions of the training set belong to a known parametric family of probability distributions, the competitive minimax approach can be an interesting solution (Feder and Merhav, 2002). Finally, in (Guerrero-Curieses et al., 2004), the authors proposed an interesting fixed-point algorithm based on generalized entropy and strict sense Bayesian loss functions. This approach alternates a resampling step of the learning set with an evaluation step of the class-conditional risk, and it leads to estimate the least-favorable priors. However, the fixed-point algorithm needs the minimax rule to be an equalizer rule. We can show that this assumption is in general not satisfied when considering discrete features.

Contributions In this paper, we propose a new method for computing the minimax classifier addressing all the previously mentioned issues. It is well known that the usual minimax classifier aims at finding the priors which maximize the minimum empirical risk over the probabilistic simplex (Poor, 1994). These class proportions are called the least favorable priors. They are generally very difficult to obtain as underlined in (Fillatre and Nikiforov, 2012) and (Fillatre, 2017). However, as discussed in (Alaiz-Rodríguez et al., 2007), it appears that sometimes a minimax classifier can be too pessimistic since its associated least favorable priors might be too far from the state of nature, and the risk of misclassifications becomes too high. In this case, our approach is suitable to consider some box constraints on the priors in order to find an acceptable trade-off between addressing the priors issues and satisfying

an acceptable risk. The resulting decision rule is the box-constrained minimax classifier. The contributions of the paper are the following. First, we calculate the optimal minimum empirical risk of the training set, also called the empirical Bayes risk. Second, we show that the empirical Bayes risk is a non-differentiable concave multivariate piecewise affine function with respect to the priors. The box-constrained minimax classifier is obtained by seeking at the maximum of the empirical Bayes risk over the box-constrained region. Third, we derive a projected subgradient algorithm which finds the least favorable proportions over the box-constrained simplex. In section 2, we present the box-constrained minimax classifier. In section 3, we study the empirical Bayes risk. Section 4 proposes an optimization algorithm to compute the box-constrained minimax classifier. Section 5 proposes some numerical experiments on the Framingham Heart dataset (University et al., From 1948). Finally, Section 6 concludes the paper.

2. Principle of box-constrained minimax classifier

Given $K \geq 2$ the number of classes, let $\mathcal{Y} = \{1, \ldots, K\}$ be the set of class labels and $\hat{\mathcal{Y}} = \mathcal{Y}$ the predicted labels. Let $\mathcal{X}$ be the space of all feature values. Let $L : \mathcal{Y} \times \hat{\mathcal{Y}} \to [0, +\infty)$ be the loss function such that, for all $(k, l) \in \mathcal{Y} \times \hat{\mathcal{Y}}$, $L(k, l) := L_{kl}$ corresponds to the loss, or the cost, of predicting the class l whereas the real class is k. For example, the $L_{0\text{-}1}$ loss function is defined by $L_{kk} = 0$ and $L_{kl} = 1$ when $k \neq l$. Given a multiset $\{(Y_i, X_i), i \in \mathcal{I}\}$ containing a number m of labeled learning samples, the task of supervised classification is to learn a decision rule $\delta : \mathcal{X} \to \hat{\mathcal{Y}}$ which assigns each sample $i \in \mathcal{I}$ to a class $\hat{Y}_i \in \hat{\mathcal{Y}}$ from its feature vector $X_i := [X_{i1}, \ldots, X_{id}] \in \mathcal{X}$ composed of d observed features, and such that δ minimizes the empirical risk $\hat{r}(\delta) = \frac{1}{m} \sum_{i \in \mathcal{I}} L(Y_i, \delta(X_i))$ (Vapnik, 1999; Hastie et al., 2009; Duda et al., 2000). As explained in (Poor, 1994), this risk can be written as

$$\hat{r}\left(\delta_{\hat{\pi}}\right) \;=\; \sum_{k \in \mathcal{Y}} \hat{\pi}_k \, \hat{R}_k\left(\delta_{\hat{\pi}}\right), \tag{1}$$

where $\hat{\pi} = [\hat{\pi}_1, \ldots, \hat{\pi}_K]$ corresponds to the class proportions of the training set satisfying, for all $k \in \mathcal{Y}$, $\hat{\pi}_k = \frac{1}{m} \sum_{i \in \mathcal{I}} \mathbb{1}_{\{Y_i = k\}}$, [1] and where $\hat{R}_k\left(\delta_{\hat{\pi}}\right)$ corresponds to the empirical class-conditional risk associated to class k defined as

$$\hat{R}_k\left(\delta_{\hat{\pi}}\right) \;=\; \sum_{l \in \hat{\mathcal{Y}}} L_{kl} \, \hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k). \tag{2}$$

Here, $\hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k)$ denotes the empirical probability for the classifier δ to assign the class l given that the true class is k. Note that in (1) and (2), the notation $\delta_{\hat{\pi}}$ means that the decision rule δ was fitted under the priors $\hat{\pi}$. More generally, we will use the notation δ_π to denote that the decision rule δ was fitted under the priors π, for any π in the K-dimensional probabilistic simplex $\mathbb{S}$ defined by $\mathbb{S} := \{\pi \in [0, 1]^K : \sum_{k=1}^K \pi_k = 1\}$. In the following, $\Delta := \{\delta : \mathcal{X} \to \hat{\mathcal{Y}}\}$ denotes the set of all possible classifiers.

1. The indicator function of event E is denoted $\mathbb{1}_{\{E\}}$.

2.1. Minimax classifier principle

Let $\{(Y'_i, X'_i), i \in \mathcal{I}'\}$ be the multiset containing a number m' of test samples satisfying the unknown class proportions $\pi' = [\pi'_1, \ldots, \pi'_K]$. The classifier $\delta_{\hat{\pi}}$ fitted with the samples $\{(Y_i, X_i), i \in \mathcal{I}\}$ is then used to predict the classes Y'_i of the test samples $i \in \mathcal{I}'$ from their associated features $X'_i \in \mathcal{X}$. As described in (Poor, 1994), the risk of misclassification with respect to the classifier $\delta_{\hat{\pi}}$ and as a function of π' is defined as $\hat{r}(\pi', \delta_{\hat{\pi}}) = \sum_{k \in \mathcal{Y}} \pi'_k \hat{R}_k(\delta_{\hat{\pi}})$. Figure 1, left, illustrates the risk $\hat{r}(\pi', \delta_{\hat{\pi}})$ for $K = 2$. In this case, it can be rewritten as

$$\hat{r}\left(\pi', \delta_{\hat{\pi}}\right) \;=\; \pi'_1 \hat{R}_1(\delta_{\hat{\pi}}) + \pi'_2 \hat{R}_2(\delta_{\hat{\pi}}) \;=\; \pi'_1 \left(\hat{R}_1(\delta_{\hat{\pi}}) - \hat{R}_2(\delta_{\hat{\pi}})\right) + \hat{R}_2(\delta_{\hat{\pi}}). \tag{3}$$

It is then clear that $\hat{r}(\pi', \delta_{\hat{\pi}})$ is a linear function of π'_1. It is easy to verify that the maximum value of $\hat{r}(\pi', \delta_{\hat{\pi}})$ is $M(\delta_{\hat{\pi}}) := \max\{\hat{R}_1(\delta_{\hat{\pi}}), \hat{R}_2(\delta_{\hat{\pi}})\}$. Since $M(\delta_{\hat{\pi}})$ is larger than $\hat{r}(\pi', \delta_{\hat{\pi}})$, it involves that the risk of the classifier can change significantly when π' differs from $\hat{\pi}$. More generally, for K classes, the maximum risk which can be attained by a classifier when π' is unknown is $M(\delta_{\hat{\pi}}) := \max\{\hat{R}_1(\delta_{\hat{\pi}}), \ldots, \hat{R}_K(\delta_{\hat{\pi}})\}$. Hence, a solution to make a decision rule $\delta_{\hat{\pi}}$ robust with respect to the class proportions π' is to fit $\delta_{\hat{\pi}}$ by minimizing $M(\delta_{\hat{\pi}})$. As explained in (Poor, 1994), this minimax problem is equivalent to consider the following optimization problem:

$$\delta^B_{\bar{\pi}} \;=\; \operatorname*{argmin}_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\pi, \delta_\pi) \;=\; \operatorname*{argmin}_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\delta_\pi). \tag{4}$$

As shown in (Ferguson, 1967), the famous Minimax Theorem establishes that

$$\min_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\delta_\pi) \;=\; \max_{\pi \in \mathbb{S}} \min_{\delta \in \Delta} \hat{r}(\delta_\pi). \tag{5}$$

This theorem holds because our classification problem involves only discrete features. In the following, given $\pi \in \mathbb{S}$, we define $\delta^B_\pi := \operatorname{argmin}_{\delta \in \Delta} \hat{r}(\delta_\pi)$ as the optimal Bayes classifier for a given prior π. Hence, according to (5), provided that we can calculate δ^B_π for any $\pi \in \mathbb{S}$, the optimization problem (4) is equivalent to calculate the least favorable priors $\bar{\pi} = \operatorname{argmax}_{\pi \in \mathbb{S}} \hat{r}(\delta^B_\pi)$. The minimax classifier $\delta^B_{\bar{\pi}}$ is the Bayes classifier calculated with the prior $\bar{\pi}$.

2.2. Benefits of Box-constrained minimax classifier

Sometimes, the minimax classifier may appear too pessimistic since the least favorable priors $\bar{\pi}$ may be too far from the priors $\hat{\pi}$ of the training set, and experts may consider that the class proportions $\bar{\pi}$ is unrealistic. For example in Figure 1, right, let us suppose that the proportions of class 1 are bounded between $a_1 = 0.1$ and $b_1 = 0.4$. If we look at the point b_1, it is clear that the classifier $\delta^B_{\bar{\pi}}$ fitted for the class proportions $\hat{\pi}_1$ of the training set is very far from the minimum empirical Bayes risk $\hat{r}\left(\pi', \delta^B_{\pi'}\right)$. The minimax classifier $\delta^B_{\bar{\pi}}$ is more robust and the box-constrained minimax classifier $\delta^B_{\pi^\star}$ has no loss. If we look now at the point a_1, the minimax classifier is disappointing but the loss of the box-constrained minimax classifier is still acceptable. In other words, the box-constrained minimax classifier seems to provide us with a reasonable trade-off between the loss of performance and the robustness to the prior change. To our knowledge, the concept of box-constrained minimax classifier

has not been studied yet. More generally, in the case where we bound independently each class proportion, we therefore consider the box-constrained simplex

$$\mathbb{U} := \mathbb{S} \cap \mathbb{B}, \tag{6}$$

where $\mathbb{B} := \{\pi \in \mathbb{R}^K : \forall k = 1, \ldots, K,\ 0 \leq a_k \leq \pi_k \leq b_k \leq 1\}$ is the box constraint which delimits independently each class proportion. Hence, to compute the box-constrained minimax classifier with respect to $\mathbb{B}$, we consider the minimax problem $\delta^B_{\pi^\star} = \operatorname{argmin}_{\delta \in \Delta} \max_{\pi \in \mathbb{U}} \hat{r}(\delta_\pi)$, and according to (5), provided that we can calculate δ^B_π for any $\pi \in \mathbb{U}$, this problem leads to the optimization problem

$$\pi^\star \ = \ \arg\max_{\pi \in \mathbb{U}} \hat{r}(\delta^B_\pi). \tag{7}$$

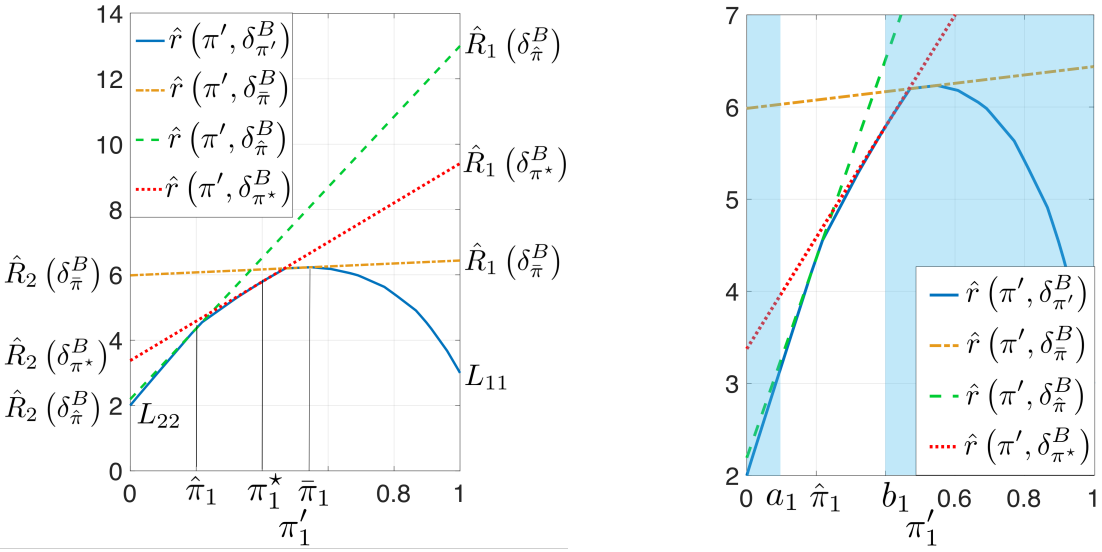

Figure 1: Comparison between the empirical Bayes classifier $\delta^B_{\hat{\pi}}$, the minimax classifier $\delta^B_{\bar{\pi}}$ and the box-constrained minimax classifier $\delta^B_{\pi^\star}$.

3. Discrete empirical Bayes risk

This section defines the empirical Bayes risk and studies its behavior as a function of the priors.

3.1. Empirical Bayes risk for the training set prior

For all $k \in \mathcal{Y}$, let $\mathcal{I}_k = \{i \in \mathcal{I} : Y_i = k\}$ be the set of learning samples from the class k, and $m_k = |\mathcal{I}_k|$ the number of samples in $\mathcal{I}_k$. Thus with these notations and in link with (2), we can write

$$\hat{\mathbb{P}}(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k) \ = \ \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{\delta_{\hat{\pi}}(X_i) = l\}}. \tag{8}$$

Since each feature X_{ij} is discrete, it takes on a finite number of values t_j. It follows that the feature vector $X_i := [X_{i1}, \ldots, X_{id}]$ takes on a finite number of values in the finite set $\mathcal{X} = \{x_1, \ldots, x_T\}$ where $T = \prod_{j=1}^{d} t_j$. Each vector x_t can be interpreted as a "profile vector" which characterizes the samples. Let us note $\mathcal{T} = \{1, \ldots, T\}$ the set of indices. Let us define for all $k \in \mathcal{Y}$ and for all $t \in \mathcal{T}$,

$$\hat{p}_{kt} = \frac{1}{m_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{X_i = x_t\}} \tag{9}$$

the probability estimate of observing the features profile $x_t \in \mathcal{X}$ with the class label k. In the context of statistical hypothesis testing theory, (Schlesinger and Hlavác, 2002) calculates the risk of a statistical test with discrete inputs. We can extend this calculation to the empirical risk of a classifier $\delta_{\hat{\pi}} \in \Delta$ with discrete features in the context of machine learning, and in the next Theorem, we show that we can compute the non-naïve empirical Bayes classifier $\delta^B_{\hat{\pi}}$ which minimizes (1) over the training set.

Theorem 1 *The empirical Bayes classifier $\delta^B_{\hat{\pi}}$ fitted on the training set with the class proportions $\hat{\pi} \in \mathbb{S}$ is*

$$\delta^B_{\hat{\pi}} : X_i \mapsto \arg\min_{l \in \hat{\mathcal{Y}}} \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{X_i = x_t\}}. \tag{10}$$

Its associated empirical Bayes risk is $\hat{r}\left(\delta^B_{\hat{\pi}}\right) = \sum_{k \in \mathcal{Y}} \hat{\pi}_k \hat{R}_k\left(\delta^B_{\hat{\pi}}\right)$, where the empirical class-conditional risk is

$$\hat{R}_k\left(\delta^B_{\hat{\pi}}\right) = \sum_{t \in \mathcal{T}} \sum_{l \in \hat{\mathcal{Y}}} L_{kl}\, \hat{p}_{kt}\, \mathbb{1}_{\{\lambda_{lt} = \min_{q \in \hat{\mathcal{Y}}} \lambda_{qt}\}},\ \forall k \in \mathcal{Y}, \tag{11}$$

and $\lambda_{lt} = \sum_{k \in \mathcal{Y}} L_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}$ for all $l \in \hat{\mathcal{Y}}$ and all $t \in \mathcal{T}$.

Proof *The proof is omitted for the lack of space.* ■

According to Theorem 1, the non-naïve Bayes classifier $\delta^B_{\hat{\pi}}$ is easily calculable in the case of discrete features since we only need to compute the probabilities $\hat{p}_{kt}$ and the priors $\hat{\pi}_k$. This classifier outperforms, on the training set, any more advanced classifiers like deep learning based classifiers.

3.2. Empirical Bayes risk extended to any prior over the simplex

Since we can only consider the samples from the training set, the probabilities $\hat{p}_{kt}$ defined in (9) are assumed to be estimated once for all. Indeed, the statistical estimation theory (Rao, 1973) has established that the estimates $\hat{p}_{kt}$ correspond to the maximum likelihood estimates of the true probabilities p_{kt} for all couples $(k, t) \in \mathcal{Y} \times \mathcal{T}$. By estimating these probabilities with the full training set, we get the best unbiased estimate with the smallest variance. This paper assumes that these class-conditional probabilities are representative of the test set. However, as explained in Section 2, we can not be confident in the class proportions estimate $\hat{\pi}_k$. For this reason, the empirical Bayes risk must be viewed as a function of the class proportions.

Let us denote δ_π^B the empirical Bayes classifier fitted on a training set with the class proportions $\pi \in \mathbb{S}$, keeping unchanged the class-conditional probabilities $\hat{p}_{kt}$:

$$\delta_\pi^B : X_i \mapsto \arg\min_{l \in \hat{\mathcal{Y}}} \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} L_{kl}\, \pi_k\, \hat{p}_{kt}\, \mathbb{1}_{\{X_i = x_t\}}. \tag{12}$$

From Theorem 1, it follows that the minimum empirical Bayes risk extended to any prior π is given by the function $V : \mathbb{S} \to [0, 1]$ defined by

$$V(\pi) \;=\; \hat{r}\left(\delta_\pi^B\right) \;=\; \sum_{k \in \mathcal{Y}} \pi_k\, \hat{R}_k\left(\delta_\pi^B\right), \tag{13}$$

$$\text{where} \quad \hat{R}_k\left(\delta_\pi^B\right) = \sum_{t \in \mathcal{T}} \sum_{l \in \hat{\mathcal{Y}}} L_{kl}\, \hat{p}_{kt}\, \mathbb{1}_{\left\{\sum_{k \in \mathcal{Y}} L_{kl}\, \pi_k\, \hat{p}_{kt} = \min_{q \in \hat{\mathcal{Y}}} \sum_{k \in \mathcal{Y}} L_{kq}\, \pi_k\, \hat{p}_{kt}\right\}}, \quad \forall k \in \mathcal{Y}. \tag{14}$$

The function $V : \pi \mapsto V(\pi)$ gives the minimum value of the empirical Bayes risk when the class proportions are π and the class-conditional probabilities $\hat{p}_{kt}$ remain unchanged. In other words, a classifier can be said robust to the priors if its risk remains very close to $V(\pi)$ whatever the value of π.

It is well known in the literature (Poor, 1994; Duda et al., 2000) that the Bayes risk, as a function of the priors, is concave over $\mathbb{S}$. The following proposition shows that this result holds when considering the empirical Bayes risk (13), and studies the differentiability of V over $\mathbb{S}$. Let us note that these results hold over the box-constrained probabilistic simplex $\mathbb{U}$ since $\mathbb{U} \subset \mathbb{S}$.

Proposition 2 *The empirical Bayes risk $V : \pi \mapsto V(\pi)$ is a concave multivariate piecewise affine function over $\mathbb{S}$ with a finite number of pieces. Finally, if there exist $\pi, \pi' \in \mathbb{S}$ and $k \in \mathcal{Y}$ such that $\hat{R}_k\left(\delta_\pi^B\right) \neq \hat{R}_k\left(\delta_{\pi'}^B\right)$, then V is non-differentiable over the simplex $\mathbb{S}$.*

Proof *The proof is omitted for the lack of space.* ■

According to (13), the optimization problem (7) is equivalent to the optimization problem

$$\pi^\star \;=\; \arg\max_{\pi \in \mathbb{U}} V(\pi). \tag{15}$$

We have established in proposition 2 that $V : \pi \mapsto V(\pi)$ is concave and non-differentiable over $\mathbb{U}$ provided that there exist $\pi, \pi' \in \mathbb{U}$, $k \in \mathcal{Y}$ such that $\hat{R}_k\left(\delta_\pi^B\right) \neq \hat{R}_k\left(\delta_{\pi'}^B\right)$. It is therefore necessary to develop an optimization algorithm adapted to both the non-differentiability of V and the domain $\mathbb{U}$.

4. Maximization over the box-constrained probabilistic simplex

We are interested in solving the optimization problem (15). In order to compute the least favorable priors $\pi^\star$ which maximize V over the box-constrained simplex $\mathbb{U}$ in the general case where V is non-differentiable, we propose to use a projected subgradient algorithm based on (Alber et al., 1998) and following the scheme

$$\pi^{(n+1)} \;=\; \mathrm{P}_{\mathbb{U}}\left(\pi^{(n)} + \frac{\gamma_n}{\eta_n}\, g^{(n)}\right). \tag{16}$$

In (16), at each iteration n, $g^{(n)}$ denotes a subgradient of V at $\pi^{(n)}$, γ_n denotes the subgradient step, $\eta_n = \max\{1, \|g^{(n)}\|_2\}$, and $\mathrm{P}_{\mathbb{U}}$ denotes the projection onto the box-constrained simplex $\mathbb{U}$. Let us note that this algorithm also holds in the case where the condition "for all $(\pi, \pi', k) \in \mathbb{U} \times \mathbb{U} \times \mathcal{Y}, \hat{R}_k\left(\delta_\pi^B\right) = \hat{R}_k\left(\delta_{\pi'}^B\right)$" is satisfied, i.e. the function V is affine over $\mathbb{U}$. Theorem 3 establishes the convergence of the iterates (16) to a solution $\pi^\star$ of (15).

Theorem 3 *Given $\pi \in \mathbb{U}$, the vector $\hat{R}\left(\delta_\pi^B\right) := \left[\hat{R}_1\left(\delta_\pi^B\right), \ldots, \hat{R}_K\left(\delta_\pi^B\right)\right] \in \mathbb{R}^K$ composed of the class-conditional risks is a subgradient of the empirical Bayes risk $V : \pi \mapsto V(\pi)$ at the point π. Moreover, when $g^{(n)} = \hat{R}\left(\delta_{\pi^{(n)}}^B\right)$ and the sequence of steps $(\gamma_n)_{n\geq 1}$ satisfies*

$$\inf_{n\geq 1} \gamma_n > 0, \quad \sum_{n=1}^{+\infty} \gamma_n^2 < +\infty, \quad \sum_{n=1}^{+\infty} \gamma_n = +\infty, \tag{17}$$

the sequence of iterates following the scheme (16) converges to a solution $\pi^\star$ of (15), whatever the initialization $\pi^{(1)} \in \mathbb{S}$.

Proof *The proof is a consequence of Theorem 1 in (Alber et al., 1998).* ■

Remark 4 *When the empirical Bayes risk V is not zero everywhere, the subgradient $\hat{R}\left(\delta_{\pi^\star}^B\right)$ at the box-constrained minimax optimum does not vanish, otherwise the associated risk $V(\pi^\star)$ would be null too. This would contradict the fact that $\pi^\star$ is a solution of (15). In this case, the sequence of iterates (16) with $g^{(n)} = \hat{R}\left(\delta_{\pi^{(n)}}^B\right)$ at each step is infinite.*

According to Remark 4, we need to consider a stopping criterion. We propose to follow (Boyd et al., 2003) which shows that the difference between the box-constrained minimax risk $V(\pi^\star) = \max_{\pi\in\mathbb{U}} V(\pi)$ and the worst empirical Bayes risk computed until the iteration N is bounded by:

$$\left| \max_{n\leq N} \left\{ V\left(\pi^{(n)}\right) \right\} - V(\pi^\star) \right| \leq \frac{\rho^2 + \sum_{n=1}^N \gamma_n^2}{2\sum_{n=1}^N \gamma_n}, \tag{18}$$

where ρ is a constant satisfying $\|\pi^{(1)} - \pi^\star\|_2 \leq \rho$. Since (18) converges to 0 as $N \to \infty$, we can choose a small tolerance $\varepsilon > 0$ as a stopping criterion. Moreover in (16), to perform the exact projection onto the box-constrained probabilistic simplex $\mathbb{U}$ at each iteration n, we propose to consider the algorithm provided by (Rutkowski, 2017). The procedure for computing our box-constrained minimax classifier is summarized in the step by step Algorithm 1 in Appendix A.

5. Numerical experiments

Dataset description For illustrating the interest of our box-constrained minimax classifier in health care field, we applied our algorithm to the Framingham Heart database (University et al., From 1948). This database contains the clinical observations of 3,658 individuals (after removing individuals with missing values) who have been followed for 10 years. The objective of the Framingham study was to predict the development of a Coronary Heart

Disease (CHD) within 10 years based on $d = 15$ observed features measured at inclusion. We therefore have $K = 2$ classes, with class 2 corresponding to individuals who have developed a CHD, and class 1 corresponding to the others. Among the 15 features, 7 are categorical (*sex, education, smoking status, previous history of stroke, diabetes, hypertension, antihypertensive treatment*) and 8 are numeric (*age, number of cigarettes per day, cholesterol levels, systolic blood pressure, diastolic blood pressure, heart rate, body mass index (BMI), glycemia*). The dataset is imbalanced: $\hat{\pi} = [0.85, 0.15]$, which means that 15% of the individuals have developed a CHD within 10 years. For this experiment, we considered the $L_{0\text{-}1}$ loss function.

Features discretization In order to apply our algorithm, we need to discretize the numerical features. To this aim, many methods can be applied as explained in (Dougherty et al., 1995; Peng et al., 2009). We can use supervised discretization methods such as (Kerber, 1992; Liu and Setiono, 1995; Kurgan and Cios, 2004), or unsupervised methods such as the Kmeans algorithm (MacQueen, 1967). Here we decided to quantize the features using the Kmeans algorithm with a number $T \geq K$ of centroids. In other words, each real feature vector $X_i \in \mathbb{R}^d$ composed of all the features is quantized with the index of the centroid closest to it, i.e., $Q(X_i) = j$ where $Q : \mathbb{R}^d \mapsto \{1, \ldots, T\}$ denotes the Kmeans quantizer and j is the index of the centroid of the cluster in which X_i belongs to. The choice of T is important since it has an impact on the generalization error of the classifier. When a classifier is fitted with respect to a given training dataset (the whole training dataset or just a group of training subsets when cross-validation is employed), the best choice of T is estimated by using a 10-fold cross-validation procedure. In other words, at each iteration of this cross-validation, we perform the Kmeans quantizer with different values of T for discretizing the features and, for each T, we compute the training risk and the validation risk. We then compare the average training risk with the average validation risk and we choose T such that the validation risk does not exceed the training risk by more than 1%. An example of this procedure is given in Figure 3, left.

Box-constraint generation In order to illustrate the benefits of the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ compared to the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ and the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$, we consider a box-constraint $\mathbb{B}_\beta$ centered in $\hat{\pi}$, and such that, given $\beta \in [0, 1]$,

$$\mathbb{B}_\beta = \left\{\pi \in \mathbb{R}^K : \forall k \in \mathcal{Y}, \quad \hat{\pi}_k - \rho_\beta \leq \pi_k \leq \hat{\pi}_k + \rho_\beta\right\}, \quad \rho_\beta := \beta \left\|\hat{\pi} - \bar{\pi}\right\|_\infty. \tag{19}$$

Our box-constrained probabilistic simplex is therefore $\mathbb{U}_\beta = \mathbb{S} \cap \mathbb{B}_\beta$. Thus, when $\beta = 0$, $\mathbb{B}_0 = \{\hat{\pi}\}$, $\mathbb{U}_0 = \{\hat{\pi}\}$ and $\pi^\star = \hat{\pi}$. When $\beta = 1$, $\bar{\pi} \in \mathbb{B}_1$, hence $\bar{\pi} \in \mathbb{U}_1$ and $\pi^\star = \bar{\pi}$. For the next experiment, after having estimated the proportions $\hat{\pi}$ and $\bar{\pi}$ over the main dataset, we chose $\beta = 0.5$ which results that $\mathbb{B}_{0.5} = \{\pi \in \mathbb{R}^2 : 0.68 \leq \pi_1 \leq 1,\ 0 \leq \pi_2 \leq 0.32\}$. In other words, we consider that the proportion of sick patients should not exceed 0.32%. Let us note that here and in the following, the least favorable priors $\bar{\pi}$ were estimated using our box-constrained minimax algorithm when considering $\mathbb{B} = [0, 1] \times [0, 1]$, so that $\mathbb{U} = \mathbb{S}$. The minimax classifier is a particular case of the box-constraint minimax classifier.

Results We performed a 10-fold cross-validation and we applied our box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ when considering the box $\mathbb{B}_{0.5}$ described above. We compared $\delta^{\mathrm{B}}_{\pi^\star}$ to the Logistic Regression $\delta^{\mathrm{LR}}_{\hat{\pi}}$, the Random Forest $\delta^{\mathrm{RF}}_{\hat{\pi}}$, the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$ (10), and the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$. We applied $\delta^{\mathrm{LR}}_{\hat{\pi}}$ and $\delta^{\mathrm{RF}}_{\hat{\pi}}$ to both the original dataset and the

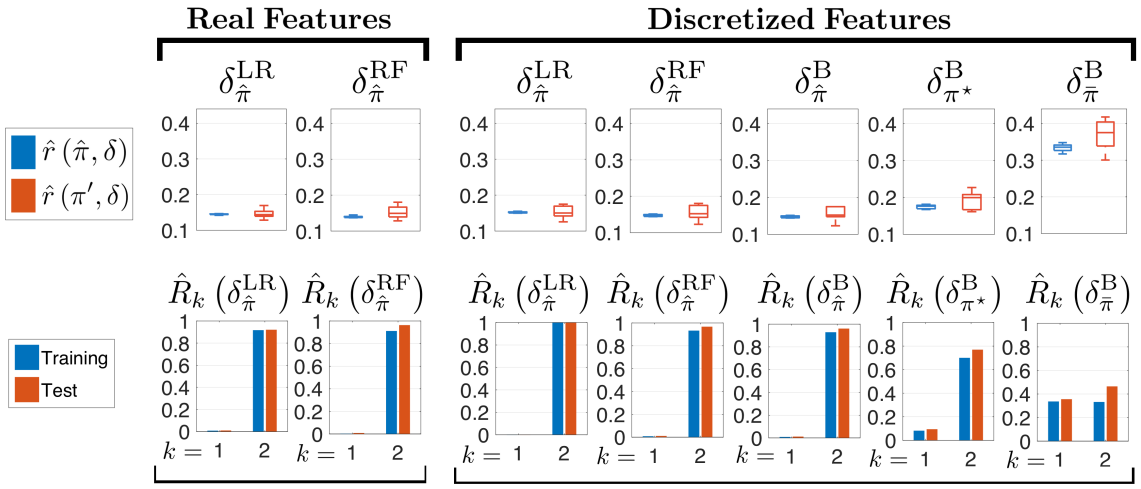

Figure 2: The boxplots (training versus test) illustrate the dispersion of the global risks of misclassification. The barplots correspond to the average class-conditional risk associated to each classifier.

discretized dataset, in order to evaluate the impact of the discretization. We can observe in Figure 2 that the performances associated to $\delta^{\mathrm{LR}}_{\hat{\pi}}$ and $\delta^{\mathrm{RF}}_{\hat{\pi}}$ are similar when considering real or discretized features. And these performances are moreover similar to the discrete Bayes classifier $\delta^{\mathrm{B}}_{\hat{\pi}}$. However, when regarding the class conditional-risks, the classifiers $\delta^{\mathrm{LR}}_{\hat{\pi}}$, $\delta^{\mathrm{RF}}_{\hat{\pi}}$ and $\delta^{\mathrm{B}}_{\hat{\pi}}$ are not satisfying when predicting accurately the patients who tend to develop a CHD. To do so, it is rather preferable to consider our minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$, even if it appears globally too pessimistic. In the case where the global risk of $\delta^{\mathrm{B}}_{\bar{\pi}}$ is not acceptable, it is therefore preferable to reduce the box-constraint area and consider the box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$, which is a trade-off between $\delta^{\mathrm{B}}_{\hat{\pi}}$ and $\delta^{\mathrm{B}}_{\bar{\pi}}$. The box-constraint area has an impact on the results and this aspect is discussed in the next paragraph. Let us note that, for the training steps of this procedure, our algorithm computed $\bar{\pi} = [0.52 \pm 0.01, 0.58 \pm 0.01]$ and $\pi^\star = [0.68 \pm 0.001, 0.32 \pm 0.001]$ such as $V(\bar{\pi}) = 0.33 \pm 0.01$ and $V(\pi^\star) = 0.28 \pm 0.01$. Finally, the results associated to the test steps presented in Figure 2 were computed when considering each whole fold test set satisfying the class proportions $\pi' = \hat{\pi}$.

Changes in the priors of the test set In order to study the robustness of each classifier when the class proportions π' of the test set are uncertain, we uniformly generated 1,000 random priors $\pi^{(s)}$, $s \in \{1, \ldots, 1000\}$, over the box-constrained simplex $\mathbb{U}_{0.5}$ using the procedure (Reed, 1974). For each repetition of the cross-validation, we generated 1000 test subsets, each one containing around 50 samples and satisfying one of the random priors $\pi^{(s)}$. Each fitted classifier was then tested when considering all the 1000 random priors uniformly dispersed over $\mathbb{U}_{0.5}$. In Figure 3, right, we observe that when the class proportions of the test set changed uniformly over $\mathbb{U}_{0.5}$, the minimax classifier $\delta^{\mathrm{B}}_{\bar{\pi}}$ was the most robust since the most stable, but it was also the most pessimistic contrary to the other classifiers. The box-constrained minimax classifier $\delta^{\mathrm{B}}_{\pi^\star}$ appears here again as a trade-off between $\delta^{\mathrm{B}}_{\bar{\pi}}$ and $\delta^{\mathrm{B}}_{\hat{\pi}}$.

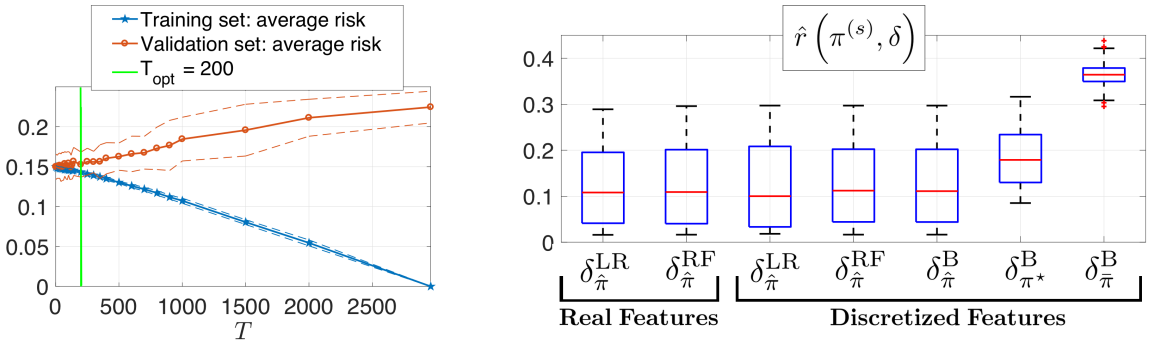

Figure 3: **Left.** Risks $\hat{r}\left(\hat{\pi}, \delta_{\hat{\pi}}^{B}\right)$ as a function of the number of centroids T. The dashed curves show the standard-deviation around the mean. **Right.** Evaluation of the robustness of each classifier when $\pi' = \pi^{(s)}$ changes over $\mathbb{U}_{0.5}$. Here, $\hat{r}(\pi^{(s)}, \delta)$ corresponds to the 10-fold cross-validation average risk associated to the test set satisfying the priors $\pi^{(s)} \in \mathbb{U}_{0.5}$, $s \in \{1, \ldots, 1000\}$.

Impact of the Box-constraint area In order to measure the impact of the box-constraint area on $\delta_{\pi^\star}^{B}$, we resized the radius ρ_β of $\mathbb{B}_\beta$ in (19) by changing the value of β from 0 to 1. Let consider the function $\psi : \Delta \to \mathbb{R}^+$ such that

$$\psi(\delta) \;=\; \max_{k \in \mathcal{Y}} \hat{R}_k(\delta) - \min_{k \in \mathcal{Y}} \hat{R}_k(\delta), \tag{20}$$

which aims at measuring how equalizer a given classifier $\delta \in \Delta$ is. In Figure 4, left, we observe that when β increases from 0 to 1, $V(\pi^\star)$ increases from $V(\hat{\pi})$ to $V(\bar{\pi})$. At the same time, in Figure 4, right, when β increases from 0 to 1, $\psi\left(\delta_{\pi^\star}^{B}\right)$ decreases from $\psi\left(\delta_{\hat{\pi}}^{B}\right)$ to $\psi\left(\delta_{\bar{\pi}}^{B}\right)$. Hence, the larger the box-constraint area is, the more equalizer the classifier $\delta_{\pi^\star}^{B}$ is, but the more pessimistic $\delta_{\pi^\star}^{B}$ becomes, since $V(\pi^\star)$ becomes much bigger than $V(\hat{\pi})$. In the case where $\delta_{\pi^\star}^{B}$ appears globally too pessimistic, it would be rather interesting to reduce the box-constraint area in order to find a trade-off between decreasing the empirical risk $V(\pi^\star)$ close enough to $V(\hat{\pi})$, and keeping an acceptable conditional risk of missing the individuals who tend to develop a CHD.

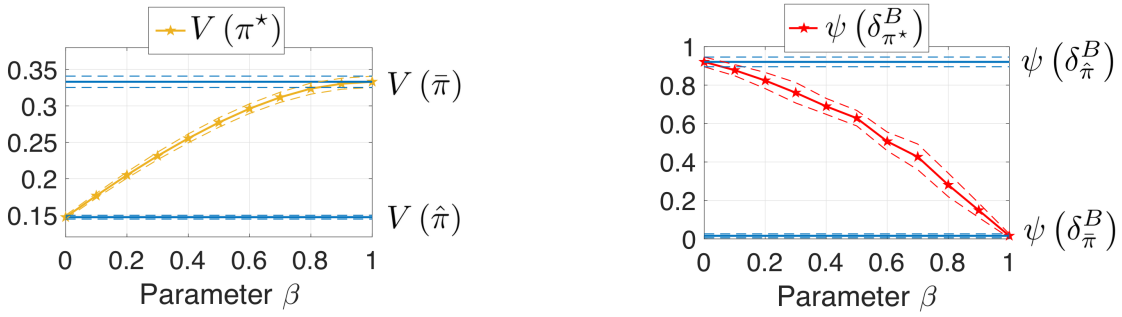

Figure 4: Impact of the box-constraint area on $\delta_{\pi^\star}^{\mathrm{B}}$ when β increases from 0 to 1 in (19), after a 10-fold cross-validation procedure. Results are presented as mean $\pm$ std.

6. Conclusion

This paper proposes a box-constrained minimax classifier which i) is robust to the imbalanced or uncertain class proportions, ii) includes some bounds on the class proportions, iii) can take into account any given loss function, and iv) is suitable for working on discrete/discretized features. In future work, we propose to investigate the robustness of the classifier with respect to the class-conditional probabilities $\hat{p}_{kt}$.

Acknowledgments

The authors thank Nicolas Glaichenhaus for his contributions and his help in this project, and the Provence-Alpes-Côte d'Azur region for its financial support.

References

Rocío Alaiz-Rodríguez, Alicia Guerrero-Curieses, and Jesús Cid-Sueiro. Minimax regret classifier for imprecise class distributions. *Journal of Machine Learning Research*, 8: 103–130, Jan 2007.

Ya. I. Alber, A. N. Iusem, and M. V. Solodov. On the projected subgradient method for nonsmooth convex optimization in a hilbert space. *Mathematical Programming*, 81(1): 23–35, Mar 1998.

Bernardo Ávila Pires, Csaba Szepesvari, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1391–1399, Atlanta, Georgia, USA, 17–19 Jun 2013.

Stephen Boyd, Lin Xiao, and Almir Mutapcic. Lecture notes: Subgradient methods, stanford university, 2003. URL: http://web.mit.edu/6.976/www/notes/subgrad_method.pdf.

Adam Cannon, James Howse, Don Hush, and Clint Scovel. Learning with the Neyman-Pearson and min-max criteria. *Los Alamos National Laboratory, Tech. Rep. LA-UR*, pages 02–2951, 2002.

Mark A Davenport, Richard G Baraniuk, and Clayton D Scott. Tuning support vector machines for minimax and Neyman-Pearson classification. *IEEE transactions on pattern analysis and machine intelligence*, 32(10):1888–1898, 2010.

James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. *International Conference on Machine Learning*, 1995.

Chris Drummond and Robert C. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*, 2003.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, pages 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

F. Farnia and D. Tse. A minimax approach to supervised learning. In *Advances in NIPS 29*, pages 4240–4248. 2016.

Meir Feder and Neri Merhav. Universal composite hypothesis testing: A competitive minimax approach. *IEEE Transactions on information theory*, 48(6):1504–1517, 2002.

T.S. Ferguson. *Mathematical Statistics : A Decision Theoretic Approach*. Academic Press, 1967.

Lionel Fillatre. Constructive minimax classification of discrete observations with arbitrary loss function. *Signal Processing*, 141:322–330, 2017.

Lionel Fillatre and Igor Nikiforov. Asymptotically uniformly minimax detection and isolation in network monitoring. *IEEE Transactions on Signal Processing*, 60(7):3357–3371, 2012.

Salvador García, Julián Luengo, and Francisco Herrera. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, 98:1–29, 2016.

A. Guerrero-Curieses, R. Alaiz-Rodriguez, and J. Cid-Sueiro. A fixed-point algorithm to minimax learning with neural networks. *IEEE Transactions on Systems, Man and Cybernetics, Part C, Applications and Reviews*, 34(4):383–392, Nov 2004.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2nd edition, 2009.

Huang Kaizhu, Yang Haiqin, King Irwin, R. Lyu Michael, and Laiwan Chan. The minimum error minimax probability machine. *Journal of Machine Learning Research*, page 1253–1286, 2004.

Huang Kaizhu, Yang Haiqin, King Irwin, and R. Lyu Michael. Imbalanced learning with a biased minimax probability machine. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(4):913–923, Aug 2006.

Randy Kerber. Chimerge: Discretization of numeric attributes. *AAAI-92 Proceedings*, pages 123–127, 1992.

A. Lukasz Kurgan and Krysztof J. Cios. Caim discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16:145–153, 2004.

Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. *IEEE, International Conference on tools with Artificial Intelligence*, 1995.

Jonathan L. Lustgarten, Vanathi Gopalakrishnan, Himanshu Grover, and Shyam Visweswaran. Improving classification performance with discretization on biomedical datsets. *AMIA 2008 Symposium Proceedings*, pages 445–449, 2008.

James MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

Liu Peng, Wang Qing, and Gu Yujia. Study on comparison of discretization methods. *IEEE, International Conference on Artificial Intelligence and Computational Intelligence*, pages 380–384, 2009.

H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag New York, 2nd edition, 1994.

C. Radhakrishna Rao. *Linear Statistical Inference and its Applications*. Wiley, 1973.

W. J. Reed. Random points in a simplex. *Pacific J. Math.*, 54(2):183–198, 1974.

K. E. Rutkowski. Closed-form expressions for projectors onto polyhedral sets in hilbert spaces. *SIAM Journal on Optimization*, 27:1758–1771, 2017.

M.I. Schlesinger and Václav Hlavác. *Ten Lectures on Statistical and Structural Pattern Recognition*. Springer Netherlands, 1st edition, 2002.

Boston University, the National Heart Lung, and Blood Institute. The framingham heart study, From 1948. Downloaded data: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset.

Vladimir Vapnik. An overview of statistical learning theory. *IEEE transactions on Neural Networks*, 10 5:988–99, 1999.

Ying Yang and Geoffrey I. Webb. Discretization for naive-bayes learning: managing discretization bias and variance. *Machine Learning*, 74(1):39–74, Jan 2009.

Appendix A.

The procedure for computing our box-constrained minimax classifier $\delta^B_{\pi^\star}$ is summarized step by step in Algorithm 1. In practice, we choose the sequence of steps $(\gamma_n)_{n\geq 1} = 1/n$ which satisfies (17).

Algorithm 1 Box-constrained minimax classifier

Input: $(Y_i, X_i)_{i\in\mathcal{I}}$, K, N.
Compute $\pi^{(1)} = \hat{\pi}$
Compute the $\hat{p}_{kt}$'s as described in (9).
$r^\star \leftarrow 0$
$\pi^\star \leftarrow \pi^{(1)}$
for $n = 1$ **to** N **do**
  for $k = 1$ **to** K **do**
    $g_k^{(n)} \leftarrow \hat{R}_k\left(\delta^B_{\pi^{(n)}}\right)$ see (14)
  end for
  $r^{(n)} = \sum_{k=1}^K \pi_k^{(n)} g_k^{(n)}$ see (1)
  if $r^{(n)} > r^\star$ **then**
    $r^\star \leftarrow r^{(n)}$
    $\pi^\star \leftarrow \pi^{(n)}$
  end if
  $\gamma_n \leftarrow 1/n$
  $\eta_n \leftarrow \max\{1, \|g^{(n)}\|_2\}$
  $z^{(n)} \leftarrow \pi^{(n)} + \gamma_n\, g^{(n)}/\eta_n$
  $\pi^{(n+1)} \leftarrow \mathrm{P}_{\mathbb{U}}\left(z^{(n)}\right)$
end for
Output: $r^\star$, $\pi^\star$ and $\delta^B_{\pi^\star}$ provided by (12) with $\pi = \pi^\star$.