Classification with support vector machines and Kolmogorov-Smirnov bounds

Lev V. Utkin[a], Frank P.A. Coolen[b]

[a]*Department of Industrial Control and Automation, St. Petersburg State Forest Technical University, Russia*
[b]*Department of Mathematical Sciences, Durham University, UK*

Abstract

This paper presents a new statistical inference method for classification. Instead of minimizing a loss function that solely takes residuals into account, it uses the Kolmogorov-Smirnov bounds for the cumulative distribution function of the residuals, as such taking conservative bounds for the underlying probability distribution for the population of residuals into account. The loss functions considered are based on the theory of support vector machines. Parameters for the discriminant functions are computed using a minimax criterion and for a wide range of popular loss functions the computations are shown to be feasible based on new optimisation results presented in this paper. The method is illustrated in examples, both with small simulated data sets and with real-world data.

Key words: Classification, imprecise probability, Kolmogorov-Smirnov bounds, minimax, support vector machines

1. Introduction

A main goal of statistical machine learning is prediction of an unobserved output value y based on an observed input vector $\mathbf{x}$, which requires estimation of a predictor function f from training data consisting of pairs $(\mathbf{x}, y)$. Two major topics in statistics which fit into the statistical machine learning framework are regression analysis and classification. In regression analysis, one typically aims at estimation of a real-valued function based on a finite set of observations with random noise. In classification, the output variable is in one of a finite number of classes[1] and the main task is to classify the output y corresponding to each input $\mathbf{x}$ into one of the classes by means of a discriminant function. Many methods have been proposed for solving machine learning problems, but these are mostly based on rather restrictive assumptions, for example assuming the availability of a large amount of training

Email addresses: lev.utkin@mail.ru (Lev V. Utkin), frank.coolen@durham.ac.uk (Frank P.A. Coolen)

[1]Often two classes, to which attention is restricted in this paper; generalization is possible but not addressed here.

data, a known probability distribution for the random noise, or that all observations are point-valued ('precise'). Such assumptions are typically not fully satisfied in applications. In this paper, a general framework for classification is presented that allows such important aspects to be incorporated without additional assumptions, instead it uses the framework of imprecise probability [4, 29] and it can be used for a wide variety of inferences, models and real-world situations. The method presented is nonparametric and can deal with relatively few training data. It combines the use of Kolmogorov-Smirnov bounds for the distribution of residuals with the use of support vector machines. The method is only presented for precise data, but in principle it can be extended to deal with imprecise data, this is a topic of ongoing investigations and will be reported on elsewhere.

The novel approach for constructing a class of machine learning models and methods proposed in this paper uses risk functionals as in [16] and sets of probability distributions as in [32]. The starting point is a set of probability distributions related to the training data, which can just be a small amount of data, and this set can be generated by a variety of inferential methods and is assumed to be bounded by some lower and upper CDFs. Such sets of probability distributions are also called p-boxes [6]. In the classification application considered in this paper, these bounds for the set of probability distributions depend on the unknown parameter of the discriminant function, because the sets of probability distributions considered are for the random residuals and as such they depend on the model parameter. It should be noted that the considered set of distributions is not the set of parametric distributions having the same parametric form as the bounding distributions, but it is the set of all possible distributions restricted by the lower and upper bounds. This is an important feature of the proposed approach in this paper.

Traditionally, machine learning methods have used a variety of simplifying assumptions in order to maintain acceptable computational effort required for implementation. The fact that the bounds for the set of probability distributions considered in the classification problems depend on the model parameter makes it clear that any optimisation of risk functionals over the whole set of probability distributions is likely to require an enormous computational effort. In this paper it will be shown that, for a wide range of popular risk functions, computation is feasible due to new results for the optimisation.

Generally, to fit a parametric classification model the value of the parameter is computed by minimising a risk functional defined by the combination of a certain loss function and a probability distribution for the random noise [11, 28]. When using a set of probability distributions instead of a precise distribution, we can choose a single distribution from this set which maximises or minimises the risk functional; the probability distribution maximising (minimising) the risk functional corresponds to the minimax (minimin) strategy. These cases can be called the 'pessimistic' and 'optimistic' decisions, respectively. The main problem in finding these two ('extreme' or 'optimal') precise distributions is that, like the bounds of the corresponding set of distributions, they depend on the unknown classification model parameter which has to be computed. For the classification scenarios considered in this paper, we restrict attention to the minimax strategy, which is often considered to be appealing from decision-theoretic perspective. We will identify the optimal probability dis-

tributions corresponding to the minimax strategy as functions of the unknown parameter only, which enables us to substitute them into the expression for the risk functional and to compute the optimal model parameter by minimising the risk measure over the set of possible values for the parameter.

The sets of probability distributions can be constructed from training data by a variety of statistical inference methods, including imprecise ('generalized') Bayesian inference models [14, 17, 29, 30], nonparametric predictive inference [1, 3] or belief functions [5, 6, 13, 21]. The approach has recently been used in regression modelling with precise statistical data using Kolmogorov-Smirnov (KS) confidence bounds [27] and also includes imprecise Bayesian normal regression [25]. In this paper, we focus explicitly on the use of extended support vector machines (SVMs) [11, 28] to construct sets of probability distributions, as SVMs are popular tools in machine learning. It will be interesting to implement the general approach presented here with a wide range of methods for constructing the sets of probability distributions and to compare the resulting inferences, for example also with regard to the effect of parameters such as the chosen confidence level if Kolmogorov-Smirnov bounds are used; this is left as an important topic for future research.

Section 2 presents the standard classification problem considered in this paper, which is extended by considering a set of probability distributions in Section 3. Kolmogorov-Smirnov bounds are introduced in Section 4, where also their application in classification is presented. Section 5 presents the important results on optimisation that make implementation of the method possible, it is explicitly presented for the popular hinge loss function. Section 6 combines support vector machines and the use of the Kolmogorov-Smirnov bounds for classification, which is illustrated in several examples, both with small simulated data sets and real-world data, in Section 7. In Section 8 some concluding remarks are made.

2. The standard classification problem

Suppose we are given a training set

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n) \in \mathbb{R}^m \times \{-1, +1\}. \tag{1}$$

Here $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ is some nonempty set of the patterns or examples; $y_1, ..., y_n$ are labels or outputs representing the observations of classes $y = -1$ and $y = 1$. The binary-classification problem is to derive a unique separating function that maximizes the margin between the two classes.

It is supposed that the number of elements in the training set belonging to the class y is n_y and their indices form the set of indices $N(y)$, so $n_{-1} + n_1 = n$ and

$$N(y) = \{i : y_i = y\}.$$

The classification problem is usually characterized by an unknown cumulative distribution function (CDF) $F_0(\mathbf{x}, y)$ on $\mathbb{R}^m \times \{-1, +1\}$ defined by the training set or examples $\mathbf{x}_i$ and their corresponding class labels y_i.

The main problem is to find a decision function $g(\mathbf{x})$ which accurately predicts the class label y of any example $\mathbf{x}$ that may or may not belong to the training set. In other words, we seek a function g that minimizes the classification error, which is given by the probability that $g(\mathbf{x}) \neq y$. One of the possible approaches for solving the problem is the discriminant function approach which uses a real-valued function $f(\mathbf{x})$, called the discriminant function, whose sign determines the class label prediction: $g(\mathbf{x}) = \mathrm{sgn}(f(\mathbf{x}))$. The discriminant function $f(\mathbf{x})$ may be parametrized with some parameters $\mathbf{w} = (w_0, w)$, $w = (w_1, ..., w_m)$, that are determined from the training examples by means of a learning algorithm. In particular, the function $f(\mathbf{x})$ may be linear, i.e. $f(\mathbf{x}, \mathbf{w}) = \langle w, \mathbf{x} \rangle + w_0$. We also introduce the notation $x_i^{(k)}$ for the i-th element of the vector $\mathbf{x}_k$.

Given the training data, the linear discriminant training problem is to minimize the following risk measure [28]

$$R(\mathbf{w}) = \int_{\mathbb{R}^m \times \{-1,1\}} L(f(\mathbf{x}, \mathbf{w}), y) \mathrm{d}F_0(\mathbf{x}, y).$$

Here the loss function $L(\mathbf{x}, y)$ usually takes a positive value when the sign of the discriminant function (the class label prediction) does not coincide with the class label y. The minimization of the risk measure is carried out over the parametric class of functions $f(\mathbf{x}, \mathbf{w})$. In other words, the optimisation process is aimed at finding the function $f(\mathbf{x}, \mathbf{w})$ which provides the minimum of $R(\mathbf{w})$ such that $R(\mathbf{w}_{\mathrm{opt}}) = \min_{\mathbf{w}} R(\mathbf{w})$. We assume that the vector of parameters $\mathbf{w}$ takes values from a set Λ.

3. The classification problem under a set of probability distributions and the minimax strategy

The loss function depends on the separating function $f(\mathbf{x}, \mathbf{w})$ which is a function of the random variable $\mathbf{w}$ and as such can be considered to be a random variable itself. In order to be short, we will write f instead of $f(\mathbf{x}, \mathbf{w})$ in equations below. It is difficult to consider the CDFs of many variables. Therefore, we do not construct the CDFs $F_0(\mathbf{x}, y)$, $y = 1, 2$, on the basis of the training set. Instead, we consider the CDFs $F(f(\mathbf{x}, \mathbf{w}) \mid y)$, $y = 1, 2$, which will be written as $F(f, y)$ for ease of notation. The notation $F(f(\mathbf{x}, \mathbf{w}) \mid y)$ does not mean that the vector of parameters $\mathbf{w}$ is a random variable. It means that we consider the CDF of the discriminant function under fixed parameters $\mathbf{w}$.

If we replace the CDF $F_0(\mathbf{x}, y)$ by the CDF $F(f, y)$ and the loss function $L(\mathbf{x}, y)$ by the loss function $L(f, y)$, we can rewrite the risk functional as

$$R(\mathbf{w}) = \int_{\mathbb{R} \times \{-1,1\}} L(f, y) \mathrm{d}F(f, y).$$

Moreover, we represent the joint probability as $F(f, y) = F(f|y) \cdot P(y)$. If we assume for simplicity that for every example $\mathbf{x}$ there exists a value f, then $P(y)$ is the prior probability that an arbitrary point $\mathbf{x}$ from $\mathbb{R}^m$ belongs to the class y.

Let us rewrite the risk functional taking into account the two possible values of y,

$$R(\mathbf{w}) = P(-1)R_{-1}(\mathbf{w}) + P(1)R_{+1}(\mathbf{w}).$$

Here

$$R_{-1}(\mathbf{w}) = \int_{\mathbb{R}} L(f, y)\mathrm{d}F(f| - 1),$$

$$R_{+1}(\mathbf{w}) = \int_{\mathbb{R}} L(f, y)\mathrm{d}F(f|1).$$

Suppose that the distributions F are unknown. However, we assume that some lower and upper bounds for a set $\mathcal{F}(y)$ of the CDFs $F(f|y)$ are known with some accuracy (e.g. a pre-specified level of confidence), and they are $\underline{F}(f|y)$ and $\overline{F}(f|y)$, respectively. Let

$$\mathcal{F}(y) = \{F(f|y) \mid \forall f \in \mathbb{R},\ \underline{F}(f|y) \leq F(f|y) \leq \overline{F}(f|y)\}.$$

In other words, there is an unknown precise "true" CDF $F(f|y) \in \mathcal{F}(y)$ for every $y \in \{-1, +1\}$, but we do not know it and only know that, with the given level of accuracy (or 'confidence'), it belongs to the set $\mathcal{F}(y)$. As mentioned before, the set $\mathcal{F}(y)$ is not a set of parametric distributions having the same parametric form as the bounding distributions, but it is the set of all possible distributions restricted by the lower and upper bounds.

One of the possible ways to fit the classification model, that is to determine values for the parameters of the discriminant function, is by using the minimax (pessimistic) strategy. According to the minimax strategy, we select a probability distribution from the set $\mathcal{F}(-1)$ and a probability distribution from the set $\mathcal{F}(+1)$ such that the risk measures $R_{-1}(\mathbf{w})$ and $R_{+1}(\mathbf{w})$ achieve their maximum for every fixed w. It should be noted that the "optimal" probability distributions may be different for different values of parameters w. This implies that the corresponding "optimal" probability distributions depend on w. The minimax strategy can be explained in a simple way. We do not know a precise probability distribution F and every distribution from $\mathcal{F}$ can be selected. Therefore, we should take the "worst" distribution providing the largest value of the risk functional[2]. The minimax criterion can be interpreted as an insurance against the worst case because it aims at minimizing the expected loss in the least favorable case [19].

Since the sets $\mathcal{F}(-1)$ and $\mathcal{F}(1)$ are obtained independently for $y = -1$ and $y = 1$, respectively, the maximum value of the risk functional $R(\mathbf{w})$ is

$$\overline{R}(\mathbf{w}) = P(-1) \max_{F(f|-1)\in\mathcal{F}(-1)} R_{-1}(\mathbf{w}) + P(1) \max_{F(f|1)\in\mathcal{F}(1)} R_{+1}(\mathbf{w}).$$

The minimax risk functional with respect to the minimax strategy is now of the form:

$$\overline{R}(\mathbf{w}_{\mathrm{opt}}) = \min_{w\in\Lambda} \overline{R}(\mathbf{w}).$$

[2]This criterion of decision making can be regarded as the well-known Γ-minimax [2, 10, 24]. However, it is often given in terms of utilities and is usually called Γ-maximin.

Let us consider in detail the first problem, that is $\max_{F(f \mid -1)\in\mathcal{F}(-1)} R_{-1}(\mathbf{w})$. Most loss functions $L(f,-1)$ applied in classification are increasing in f. This implies that the upper bound for $R_{-1}(\mathbf{w})$, i.e. the maximum of $R_{-1}(\mathbf{w})$ over all distributions from $\mathcal{F}(-1)$ is achieved at the distribution $\underline{F}(f|1)$ (see, for instance, [31]). Hence,

$$\overline{R}_{-1}(\mathbf{w}) = \int_{\mathbb{R}} L(f,-1)\mathrm{d}\underline{F}(f|-1).$$

In the same way we can consider the second problem, $\max_{F(f|1)\in\mathcal{F}(1)} R_{+1}(\mathbf{w})$. Most loss functions $L(f,1)$ are decreasing in f. Therefore, the upper bound for $R_{+1}(\mathbf{w})$ is achieved at the distribution $\overline{F}(f|1)$. This implies that

$$\overline{R}_{+1}(\mathbf{w}) = \int_{\mathbb{R}} L(f,1)\mathrm{d}\overline{F}(f|1).$$

So, restricting attention henceforth to loss functions $L(f,-1)$ which are increasing in f and $L(f,1)$ which are decreasing in f, the upper bound for the risk functional $R(\mathbf{w})$ is of the form

$$\overline{R}(\mathbf{w}) = P(-1)\int_{\mathbb{R}} L(f,-1)\mathrm{d}\underline{F}(f|-1) + P(1)\int_{\mathbb{R}} L(f,1)\mathrm{d}\overline{F}(f|1). \tag{2}$$

It is important to define suitable CDFs $\overline{F}(f|y)$ and $\underline{F}(f|y)$ to define $\mathcal{F}(y)$ based on the available information. We propose the use of the Kolmogorov-Smirnov bounds, as explained in the following section. In addition to their appeal as generally valid nonparametric bounds for the CDF, it will be shown in this paper that they can be implemented in our approach as the optimisation problems involved can be solved without imposing substantial computational difficulties.

4. Kolmogorov-Smirnov bounds and optimal lower and upper CDFs

The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible [33]. It is particularly attractive to allow the class of admissible distributions for a statistical problem to be the class of all distributions, without assuming a specific functional form. Given an independent and identically distributed (i.i.d.) sample $Z_1, ..., Z_n$ with CDF $F(z) = P(Z \leq z)$ on the real line, we can estimate F in the framework of nonparametric methods with the empirical distribution function F_n as the CDF that puts mass $1/n$ at each data point Z_i,

$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \leq z),$$

where

$$I(Z_i \leq z) = \begin{cases} 1, & Z_i \leq z, \\ 0, & Z_i > z. \end{cases}$$

According to well-known properties of the empirical CDF, such as the Glivenko-Cantelli theorem and the Dvoretzky-Kiefer-Wolfowitz inequality, the empirical CDF converges to the true CDF as $n \to \infty$. Roughly speaking, this means that we can learn distributions to any required level of accuracy just by collecting enough data. However, in practice the amount of statistical data is often limited. One of the ways for taking into account the scarcity of statistical data and for constructing bounds for the set of probability distributions is by using the Kolmogorov-Smirnov (KS) confidence limits for the CDF, which can be regarded as distribution-free bounds around the empirical CDF. If we assume that $F(z)$ is some unknown true probability distribution of Z, then we can choose a critical value of the test statistic $d_{n,1-\gamma}$ such that the band $[F_n(z) - d_{n,1-\gamma}, F_n(z) + d_{n,1-\gamma}]$ will contain $F(z)$ entirely with probability $1-\gamma$, which is to be interpreted as a confidence statement in the frequentist statistical framework. Let $k_{1-\gamma}$ be the $(1-\gamma)$-quantile of the Kolmogorov distribution. Then we can write for large values of n

$$d_{n,1-\gamma} \approx k_{1-\gamma}/\sqrt{n}.$$

The values of $k_{1-\gamma}$ for different γ can e.g. be found in [12]. If the values of n are rather small (say $n \leq 10$), then there is another expression for $d_{n,1-\gamma}$ [12]

$$d_{n,1-\gamma} \approx k_{1-\gamma}\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right)^{-1}.$$

Denoting $\upsilon = d_{n,1-\gamma}$ for short, we finally write the following bounds for some unknown CDF $F(z)$:

$$\max(F_n(z) - \upsilon, 0) \leq F(z) \leq \min(F_n(z) + \upsilon, 1). \tag{3}$$

It can be seen from the above inequality that the left tail of the upper bound for the CDF is υ for $z = \zeta \to -\infty$. Similarly, the right tail of the lower bound for the CDF is $1 - \upsilon$ for $z = \xi \to \infty$. These tails reflect the lack of information beyond the minimal and maximal observations in the data set, together with the fact that no further assumptions are made for the CDF in the nonparametric framework. We introduce notations ζ and ξ indicating some boundary points for the lower and upper distribution functions, respectively, in order to consider the tails later in the paper. Allowing probability υ to be assigned to ζ, the upper bound $\min(F_n(z)+\upsilon, 1)$ can be considered to be a CDF itself, and similarly allowing probability υ to be assigned to ξ enables the lower bound $\max(F_n(z)-\upsilon, 0)$ to be considered to be a CDF; this will be done henceforth in this paper. It is important to emphasize that these introduced ζ and ξ, and indeed the probabilities assigned to them in this construction, do not influence the inferences in this paper, as will be explained in Section 5.

We now consider the use of the KS bounds in classification, where we construct the KS bounds for every class y. Denote the critical value of the KS statistic for class y by υ_y, then

$$\underline{F}(f|y) = \max(F_n(f) - \upsilon_y, 0),$$

$$\overline{F}(f|y) = \min(F_n(f) + \upsilon_y, 1).$$

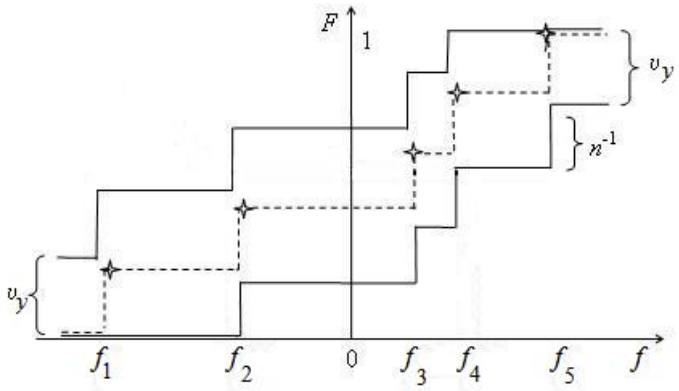

Figure 1: The empirical CDF and its Kolmogorov-Smirnov bounds

The corresponding lower and upper probability density functions (PDFs), $\underline{p}(f|y)$ and $\overline{p}(f|y)$, respectively, are weighted sums of Dirac functions where the number of terms and their weights are different and depend on υ_y. The following cases of υ can be considered.

Case 1: If $\upsilon_y < 1/n_y$, $\upsilon_y < 1/2$, then

$$\underline{p}(f|y) = \left(\frac{1}{n_y} - \upsilon_y\right)\delta(f - f_1) + \frac{1}{n_y}\sum_{i\in N(y)\setminus\{1\}}\delta(f - f_i) + \upsilon_y\delta(f - \xi),$$

$$\overline{p}(f|y) = \upsilon_y\delta(f - \zeta) + \frac{1}{n_y}\sum_{i\in N(y)\setminus\{n_y\}}\delta(f - f_i) + \left(\frac{1}{n_y} - \upsilon_y\right)\delta(f - f_{n_y}).$$

In this lower PDF the left-most point, denoted by f_1, has a smaller weight than the other points. In this upper PDF the right-most point, denoted by f_{n_y}, has a smaller weight than the other points.

Case 2: If $1/n_y \le \upsilon_y < 2/n_y$, $\upsilon_y < 1/2$, then

$$\underline{p}(f \mid y) = \left(\frac{2}{n_y} - \upsilon_y\right)\delta(f - f_2) + \frac{1}{n_y}\sum_{i\in N(y)\setminus\{1,2\}}\delta(f - f_i) + \upsilon_y\delta(f - \xi),$$

$$\overline{p}(f \mid y) = \upsilon_y\delta(f - \zeta) + \frac{1}{n_y}\sum_{i\in N(y)\setminus\{n_y,n_y-1\}}\delta(f - f_i) + \left(\frac{2}{n_y} - \upsilon_y\right)\delta(f - f_{n_y-1}).$$

This case is illustrated in Figure 1. Continuing in the same way, we get the further cases which can be presented as follows.

Case b_y: If $(b_y - 1)/n_y \le \upsilon_y < b_y/n_y$, $\upsilon_y < 1/2$, then

$$\underline{p}(f \mid y) = \left(\frac{b_y}{n_y} - \upsilon_y\right)\delta(f - f_{b_y}) + \frac{1}{n_y}\sum_{i\in N(y)\setminus\{1,2,\ldots,b_y\}}\delta(f - f_i) + \upsilon_y\delta(f - \xi),$$

$$\overline{p}(f \mid y) = \upsilon_y \delta(f-\zeta) + \frac{1}{n_y} \sum_{i \in N(y) \setminus \{n_y, n_y-1, \ldots, n_y-b_y\}} \delta(f-f_i) + \left(\frac{b_y}{n_y} - \upsilon_y\right) \delta(f - f_{n_y-b_y+1}).$$

Here the KS band is as follows. The lower CDF has $b_y - 1$ jumps equal to 0, one jump (the left one) equal to $b_y/n_y - \upsilon_y$ and the other jumps all equal to $1/n_y$. The upper CDF also has $b_y - 1$ jumps equal to 0, one jump (the right one) equal to $b_y/n_y - \upsilon_y$ and all other jumps equal to $1/n_y$. This is a very important peculiarity of the lower and upper CDFs, which will be used later. Every considered case is defined by the relationship of the values n_y and υ_y. We can find the value b_y as a function of n_y and υ_y, namely, b_y is an integer satisfying the two-sided inequality

$$\upsilon_y n_y \leq b_y \leq \upsilon_y n_y + 1.$$

If we substitute the above PDFs into expression (2) for the risk measure, then we can get the upper risk measure by the minimax strategy. However, the main difficulty in using the above "optimal" PDFs for computing the upper risk functional is that we do not know the order of points $f_1, ..., f_n$ because every point depends on the unknown classification parameters w. We know only the weights of all points, or the sizes of all jumps so to say. Nevertheless, this is very useful information which will be used later.

5. Risk functional and extreme points

Suppose that we do not know the optimal CDFs for writing the upper risk measure. However, we exactly know that this CDF is a step-function with jumps (possibly of size 0) at n points. Let $T(k, y)$ be a subset of the index set $N(y)$ consisting of k elements. If every jump of the optimal CDF (so weight of the corresponding optimal PDF) for a given y is of the size h_i, then we can write the following constraints

$$\max(k/n_y - \upsilon_y, 0) \leq \sum_{i \in T(k,y)} h_i \leq \min(k/n_y + \upsilon_y, 1),$$
$$\forall T(k, y) \subseteq N(y),\ y \in \{-1, +1\}.$$

It follows from $\sum_{i \in N(y)} h_i = 1$ that $\sum_{i \in T(k,y)} h_i \leq 1$ for all $T(k, y) \subseteq N(y)$. This implies that the above constraints can be rewritten as the following system of linear inequalities for every y

$$k/n_y - \upsilon_y \leq \sum_{i \in T(k,y)} h_i \leq k/n_y + \upsilon_y, \tag{4}$$

$$\sum_{i \in T(k,y)} h_i \geq 0,\ \sum_{i \in N(y)} h_i = 1,\ k = 1, ..., n_y. \tag{5}$$

Denote $h(y) = (h_i, i \in N(y))$ and suppose that the above constraints produce a set of probability distributions $\mathcal{H}(y)$, that is every distribution $h(y)$ from $\mathcal{H}(y)$ satisfies all the

constraints (4)-(5). Then the upper bound for the risk functional can be found as a solution to the following linear programming problem:

$$\overline{R}(\mathbf{w}) = \sum_{y=-1,+1} P(y) \max_{h(y)\in\mathcal{H}(y)} \sum_{i\in N(y)} h_i \cdot L(f_i, y),$$

subject to $h(y) \in \mathcal{H}(y)$.

Hence, the optimal values of the parameter vector w are computed by minimizing the upper bound $\overline{R}(\mathbf{w})$ over w, i.e.

$$\overline{R}(\mathbf{w}_{\text{opt}}) = \min_{w\in\Lambda} \left(\sum_{y=-1,+1} P(y) \max_{h(y)\in\mathcal{H}(y)} \sum_{i\in N(y)} h_i \cdot L(f_i, y) \right). \tag{6}$$

It is very important here that constraints (4)-(5) do not depend on w. This peculiarity allows us to reduce the set of optimisation problems to a single optimisation problem by using the extreme points.

We introduce two optimisation variables

$$G(y) = \max_{h(y)\in\mathcal{H}(y)} \sum_{i\in N(y)} h_i \cdot L(f_i, y),\ y = -1, +1.$$

We rewrite problem (6) as

$$\overline{R}(\mathbf{w}_{\text{opt}}) = \min_{w\in\Lambda} \{P(-1)G(-1) + P(1)G(1)\}, \tag{7}$$

subject to

$$G(y) \geq \sum_{i\in N(y)} h_i \cdot L(f_i, y),\ \forall h(y) \in \mathcal{H}(y),\ y = -1, +1.$$

The above optimisation problem contains infinitely many constraints, namely one constraint for every probability distribution $h(y) \in \mathcal{H}(y)$. In order to overcome this difficulty, note that the set of distributions $\mathcal{H}$ can be viewed as a simplex in a finite dimensional space. According to well-known results from linear programming theory, the objective function for a fixed value w attains its maximum at an extreme point of the unit simplex of the dimension n_y. Since the set $\mathcal{E}(\mathcal{H}(y))$ of extreme points is finite, this implies that the infinite set of constraints is reduced to some finite set. Finally, we can rewrite the constraints as

$$G(y) \geq \sum_{i\in N(y)} h_i \cdot L(f_i, y),\ \forall h(y) \in \mathcal{E}(\mathcal{H}(y)),\ y = -1, +1. \tag{8}$$

The next task is to find the set $\mathcal{E}(\mathcal{H}(y))$ of extreme points. This is also a hard problem. However, we know from the form of the optimal CDFs or PDFs (see Case b_y) that the probability distribution $h(y)$ can be represented as a sum of sizes of jumps of CDFs (see

the previous section), i.e. it has $b_y - 1$ zero-valued points, one point of the value $b_y/n_y - \upsilon_y$ and $n_y - b_y$ points of value $1/n_y$. The sum of all these values is $1 - \upsilon_y$. It is not 1 because there is a point ξ for the lower CDF or ζ for the upper CDF with the jump of size υ_y. However, this point does not depend on the parameters w and can be removed from the consideration. Indeed, let us write the constraints, for instance, for $y = -1$ as follows

$$G(-1) \geq \upsilon_{-1} L(\xi, -1) + \sum_{i \in N(y)} h_i L(f_i, -1).$$

We now introduce a new variable $G^*(-1) = G(-1) - \upsilon_{-1} L(\xi, -1)$. By substituting it into the objective function, we can see that the location of the optimal value $\overline{R}(\mathbf{w}_{\text{opt}})$ is not affected by the term $-P(-1)\upsilon_{-1} L(\xi, -1)$, because this term does not depend on w. Therefore, we can remove it from the optimisation problem.

The obtained distributions $h(y)$ are just extreme points belonging to $\mathcal{E}^*(\mathcal{H}) \subset \mathcal{E}(\mathcal{H})$. Other extreme points belonging to $\mathcal{E}(\mathcal{H}) \backslash \mathcal{E}^*(\mathcal{H})$ are not interesting for us because they definitely cannot provide larger values of $\sum_{i \in N(y)} h_i \cdot L(f_i, y)$.

So, we get an optimisation problem whose solution depends on the loss function $L(f_i, y)$. For specific loss functions this argument can be developed further. We illustrate this for the so-called hinge loss function, which is a popular and flexible loss function for SVMs [28] and which is of the form

$$L(f,y) = \max(0, 1 - yf). \tag{9}$$

The hinge loss function will be used in the examples in Section 7. After substituting the hinge loss function into constraints (8) and by using the expression for the upper risk functional (7), we get the optimisation problem

$$\overline{R}(\mathbf{w}_{\text{opt}}) = \min_{w \in \Lambda} \left\{ P(-1)G(-1) + P(1)G(1) \right\}, \tag{10}$$

subject to

$$G(y) \geq \sum_{i \in N(y)} h_i \max(0, 1 - yf_i), \ \forall h(y) \in \mathcal{E}^*(\mathcal{H}(y)), \ y = -1, +1. \tag{11}$$

Let us introduce new non-negative variables $\xi_i \geq 0$ such that $\xi_i = 1 - y_i f(\mathbf{x}_i, \mathbf{w})$ if $1 - y_i f(\mathbf{x}_i, \mathbf{w}) \geq 0$, and $\xi_i = 0$ if $1 - y_i f(\mathbf{x}_i, \mathbf{w}) < 0$. Then we can rewrite constraints (11) as

$$G(y) \geq \sum_{i \in N(y)} h_i \xi_i, \ \forall h(y) \in \mathcal{E}^*(\mathcal{H}(y)), \ y = -1, +1, \tag{12}$$

$$\xi_i \geq 1 - y_i f(\mathbf{x}_i, \mathbf{w}), \ \xi_i \geq 0, \ i = 1, ..., n, \ y = -1, +1. \tag{13}$$

So we arrive at only a single optimisation problem with a finite number of constraints, which shows that the method presented in this paper is computationally feasible for such a loss function. Moreover, if the function f is linear, then we have a linear programming problem.

6. SVM for Kolmogorov-Smirnov classification

According to the method of classification modelling developed for the minimax strategy using the Kolmogorov-Smirnov bounds, the optimal parameters $w_{\rm opt}$ of the discrimant function can be computed by solving the optimisation problem with objective function (10) and constraints (12)-(13). The problem is easy to solve when the function $f(\mathbf{x}, \mathbf{w})$ is linear. In order to consider possible non-linear cases of the function f, we study how to apply the well-known SVM method for getting the optimal parameters $w_{\rm opt}$.

First, we assume that the function f is linear and of the form $f(\mathbf{x}, \mathbf{w}) = \langle w, \mathbf{x} \rangle + w_0$. Denote the k-th probability distribution $h(y)$ from $\mathcal{E}^*(\mathcal{H}(y))$ by $h^{(k)}(y)$ and the total number of the distributions in $\mathcal{E}^*(\mathcal{H}(y))$ by N_y. It should be noted that N_y depends only on two parameters n_y and υ_y. Let us add the standard Tikhonov regularization term $\frac{1}{2} \langle w, w \rangle$ (this is the most popular penalty or smoothness term) [23] to the objective function (10) and the constant "cost" parameter C. The smoothness (Tikhonov) term can be regarded as a constraint which enforces uniqueness by penalizing functions with wild oscillation and effectively restricting the space of admissible solutions (we refer to [7] for a detailed analysis of regularization methods). This leads to the quadratic programming problem

$$\overline{R}(\mathbf{w}_{\rm opt}) = \min \left(\frac{1}{2} \langle w, w \rangle + C \cdot P(-1)G(-1) + C \cdot P(1)G(1) \right), \tag{14}$$

subject to

$$G(y) \geq \sum_{i \in N(y)} h_i \xi_i, \ \forall h(y) \in \mathcal{E}^*(\mathcal{H}(y)), \ y = -1, 1, \tag{15}$$

$$\xi_i \geq 1 - y_i f(\mathbf{x}_i, \mathbf{w}), \ i = 1, ..., n, \ y = -1, 1, \tag{16}$$

$$\xi_i \geq 0, \ i = 1, ..., n, \ y = -1, 1. \tag{17}$$

Instead of minimizing the primary objective function (14), a dual objective function, the so-called Lagrangian, can be formed of which the saddle point is the optimum. The Lagrangian is

$$\begin{aligned} L = & \frac{1}{2} \langle w, w \rangle + C \cdot P(-1)G(-1) + C \cdot P(1)G(1) - \sum_{i=1}^{n} \eta_i \xi_i \\ & - \sum_{i=1}^{n} \varphi_i \left(\xi_i - 1 + y_i \langle w, \mathbf{x}_i \rangle + y_i w_0 \right) \\ & - \sum_{y=-1,+1} \sum_{k=1}^{N_y} t_k(y) \left(G(y) - \sum_{i \in N(y)} h_i^{(k)} \xi_i \right). \end{aligned}$$

Here $\eta_i, \varphi_i, t_k(y)$, $i = 1, ..., n$, $k = 1, ..., N_y$, $y = -1, 1$, are Lagrange multipliers. Hence, the dual variables have to satisfy positivity constraints $\eta_i \geq 0, \varphi_i \geq 0, t_k(y) \geq 0$ for all i, k, and

y. The saddle point can be found by setting the derivatives equal to zero

$$\partial L/\partial w_0 = \sum_{i=1}^{n} \varphi_i y_i = 0, \tag{18}$$

$$\partial L/\partial w_j = w_j - \sum_{i=1}^{n} \varphi_i y_i x_i^{(j)} = 0,\ j = 1, ..., m, \tag{19}$$

$$\partial L/\partial \xi_i = -\eta_i - \varphi_i + \sum_{k=1}^{N_{-1}} t_k(-1) h_i^{(k)} = 0,\ i \in N(-1), \tag{20}$$

$$\partial L/\partial \xi_i = -\eta_i - \varphi_i + \sum_{k=1}^{N_{+1}} t_k(1) h_i^{(k)} = 0,\ i \in N(1), \tag{21}$$

$$\partial L/\partial G(y) = C \cdot P(y) - \sum_{k=1}^{N_y} t_k(y) = 0,\ y \in \{-1, 1\}. \tag{22}$$

It follows from (22) that

$$C \cdot P(y) G(y) - G(y) \sum_{k=1}^{N_y} t_k(y) = G(y) \left(C \cdot P(y) - \sum_{k=1}^{N_y} t_k(y) \right) = 0.$$

Using (18) we can now simplify the objective function as

$$L = \frac{1}{2} \langle w, w \rangle - \sum_{i=1}^{n} \eta_i \xi_i - \sum_{i=1}^{n} \varphi_i \left(\xi_i - 1 + y_i \langle w, \mathbf{x}_i \rangle \right) + \sum_{y=-1,+1} \sum_{k=1}^{N_y} t_k(y) \left(\sum_{i \in N(y)} h_i^{(k)} \xi_i \right).$$

It follows from (20) and (21) that

$$\begin{aligned} L &= \frac{1}{2} \langle w, w \rangle - \sum_{i=1}^{n} \xi_i \left[\eta_i + \varphi_i - \sum_{y=-1,+1} \sum_{k=1}^{N_y} t_k(y) h_i^{(k)} \right] - \sum_{i=1}^{n} \varphi_i \left(y_i \langle w, \mathbf{x}_i \rangle - 1 \right) \\ &= \frac{1}{2} \langle w, w \rangle + \sum_{i=1}^{n} \varphi_i \left(y_i \langle w, \mathbf{x}_i \rangle - 1 \right). \end{aligned}$$

After substituting (19) into the Lagrangian, we finally get the dual optimisation problem:

$$\text{maximize} \quad L = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \varphi_i \varphi_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{n} \varphi_i,$$

subject to

$$\sum_{i=1}^{n} \varphi_i y_i = 0,\ \sum_{k=1}^{N_y} t_k(y) = C \cdot P(y),\ y \in \{-1, 1\},$$

$$0 \le \varphi_i \le \sum_{k=1}^{N_{-1}} t_k(-1) h_i^{(k)}, \quad i \in N(-1),$$

$$0 \le \varphi_i \le \sum_{k=1}^{N_{+1}} t_k(1) h_i^{(k)}, \quad i \in N(1).$$

It is very interesting to note that the objective function does not differ from the objective function obtained in the standard SVM with the empirical probability distribution for the examples, i.e. by exploiting the empirical risk functional. However, the main difference is in the constraints. Instead of the constant "cost" parameter C in the right sides of the inequalities for the φ_i, we use the available information about the "extreme" probability distributions $h^{(k)}$ from $\mathcal{E}(\mathcal{H}(y))$.

The function $f(\mathbf{x}_i, \mathbf{w})$ can be rewritten in terms of Lagrange multipliers as

$$f = \sum_{i=1}^{n} y_i \varphi_i \langle \mathbf{x}_i, \mathbf{x} \rangle + w_0.$$

This is the so-called support vector expansion, i.e. w can be completely described as a linear combination of modified training data $\mathbf{x}_i$, $i = 1, ..., n$ [22].

Let us consider how the above problem can be modified in the "precise" case when we have one precise nonparametric probability distribution for every y and $v = 0$. In this case, we write $N_y = 1$, $t_1(y) = C \cdot P(y)$, $h_i^{(1)} = 1/n_y$, if $i \in N(y)$. Then the constraints for φ_i and φ_i^* become

$$0 \le \varphi_i \le C \cdot P(y)/n_y, \quad i \in N(y).$$

If $P(y) = n_y/n$, then the constraints are rewritten as

$$0 \le \varphi_i \le C/n, \quad i = 1, ..., n.$$

This indeed gives the standard SVM. So, we get the SVM approach under the minimax strategy taking into account the Kolmogorov-Smirnov bounds.

Before we illustrate our approach in several examples, we briefly consider the possibility to deal with non-linear discriminant functions. It should be noted that the vectors $\mathbf{x}_i$ only appear in the dual optimisation problem via the dot product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. This implies that we can replace $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ by the so-called kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ which has to satisfy some special properties. This is a direct way for incorporating the non-linearity into the classification problem. In this case, the discriminant function is of the form

$$f = \sum_{i=1}^{n} y_i \varphi_i K(\mathbf{x}_i, \mathbf{x}_j) + w_0.$$

We do not study the kernels and their use in classification problems in detail in this paper, because their use does not directly depend on the Kolmogorov-Smirnov bounds

and there are a lot of books and papers devoted to this important topic, see for example [15, 20, 34].

A main difficulty of the dual optimization problem is that it requires large memory which strongly depends on the number of training examples, but does not significantly depend on the number of features. The total number of "optimal" probability distributions in $\mathcal{E}^*(\mathcal{H}(y))$ is

$$N_y = n \cdot \binom{n}{b_y}.$$

We have $N(-1) + N(1) + 3$ constraints, but the number of variables in every constraint is $n + N_{-1} + N_{+1}$. In order to overcome the above difficulty and to simplify the dual quadratic programming problem we prove next that it can be decomposed into $N_{-1}N_{+1}$ simple optimization problems. Let us write some of the Karush-Kuhn-Tucker complementarity conditions

$$t_k(y) \left(G(y) - \sum_{i \in N(y)} h_i^{(k)} \xi_i \right) = 0.$$

It follows from the definition of $G(y)$ that it takes the largest value of $\sum_{i \in N(y)} h_i^{(k)} \xi_i$. If we assume that all points $\mathbf{x}_i$ are different, then $G(y) = \sum_{i \in N(y)} h_i^{(k)} \xi_i$ for some $k = s_y$. Hence, all values $t_k(y)$ are 0 except for the case $k = s_y$. According to the condition $\sum_{k=1}^{N_y} t_k(y) = C \cdot P(y)$, we conclude that $t_{s_y} = C \cdot P(y)$. Returning to the constraints to the dual optimization problem and using the fact that $t_k(y) = 0$ for all $k \neq s_y$ and $t_k = C \cdot P(y)$ for $k = s_y$, we rewrite the problem as

$$\text{maximize} \quad L(s_{-1}, s_{+1}) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \varphi_i \varphi_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{n} \varphi_i,$$

subject to

$$\sum_{i=1}^{n} \varphi_i y_i = 0,$$

$$0 \leq \varphi_i \leq C \cdot P(-1) h_i^{(s_{-1})}, \quad i \in N(-1),$$

$$0 \leq \varphi_i \leq C \cdot P(+1) h_i^{(s_{+1})}, \quad i \in N(1).$$

So, we get a simple optimization problem with the same objective function and simple constraints. It has the form of the standard weighted SVM with weights $h_i^{(s_{-1})}$ and $h_i^{(s_{+1})}$. One can see that the number of variables is n, and the number of constraints is $N(-1) + N(1) + 1$. However, we do not know the values s_{-1} and s_{+1}. Therefore, we solve the problem for every pair (s_{-1}, s_{+1}) such that $s_y = 1, ..., N_y$. The largest value of $L(s_{-1}, s_{+1})$ corresponds to optimal values of s_{-1} and s_{+1}. By using the derived optimization problem, we reduce the memory requirements, but increase the execution time for solving many optimization problems.

7. Examples

We illustrate the method proposed in this paper via several examples, all computations have been performed using the statistical software R [18]. In all these examples the hinge loss function (9) has been used, together with a linear separation function and KS bounds corresponding to $\gamma = 0.1$. We investigate the performance of the proposed method and compare it with the standard SVM approach by considering the accuracy (ACC), which is the proportion of correctly classified cases on a sample of data and is often used to quantify the predictive performance of classification methods. So, ACC is an estimate of a classifier's probability of a correct response, and it is an important statistical measure of the performance of a binary classification test. ACC can formally be written as

$$ACC = \frac{N_T}{N},$$

where N_T is the number of test data for which the predicted class for an example coincides with its true class, and N is the total number of test data. We will denote the accuracy measure for the proposed KS-bounds minimax strategy as ACC_{minimax} and for the standard SVM as ACC_{standard}.

Example 1
We first consider the performance of our method for a small example with simulated data. We generated two subsets of examples corresponding to different classes such that the number of examples of both classes are $n_{-1} = n_1 = 8$. Every example is defined by two features ($m = 2$). Values of features are generated in accordance with the normal probability distribution with for class $y = -1$ mean values $m_1(-1) = m_2(-1) = 5$ and for class $y = 1$ mean values $m_1(1) = m_2(1) = 8$, with the subscript indicating the specific feature. The standard deviation is $\sigma_i(y) = 2$ for both classes and both features.

Using the minimax strategy presented in this paper, and assuming that the separating function is linear, we get for $\gamma = 0.1$, $k_{1-\gamma} = 1.22$ and $v_y = 0.408$, the separation function shown in Fig. 2, where the triangle markers correspond to examples from the class $y = -1$ and the crosses correspond to examples from the class $y = 1$. If instead we use the standard SVM method for these data, we obtain the separation function shown in Fig. 3.

It should be remarked that the KS-bounds method with the minimax criterion, as presented in this paper, explicitly takes the example values with largest residuals (hence the misclassified examples which are furthest from the separating line) into account while it may neglect some points close to the line, whereas the standard SVM method takes all values of residuals of misclassified examples into account. So, the minimax method aims at minimal total distance to the line of misclassified points furthest away from it, which in this example leads to quite a different separating line than derived at by the standard SVM method.

To investigate the predictive performance of these two separating functions, we generated 100 test examples from the same normal distributions with parameters $m_i(-1) = 5$,

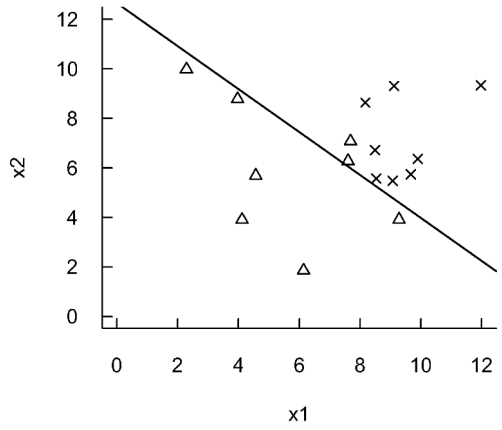

Figure 2: The separation line by the minimax strategy (Example 1)

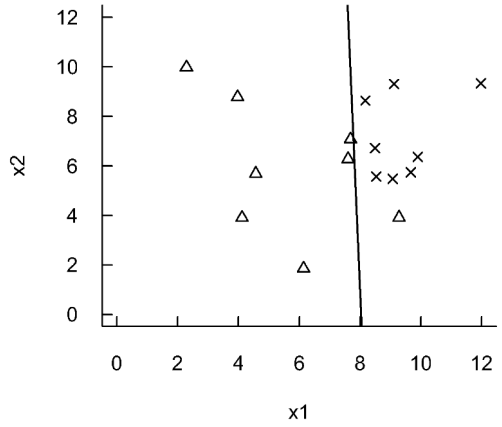

Figure 3: The separation line by the standard SVM (Example 1)

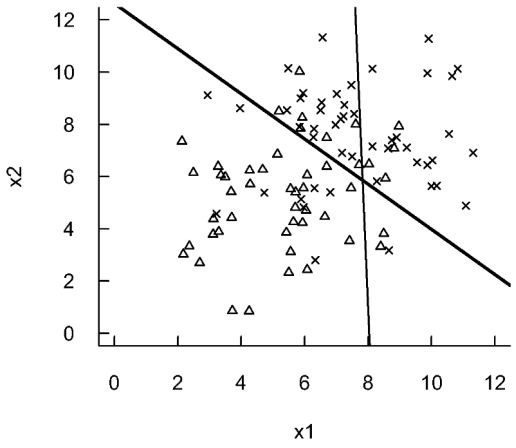

Figure 4: Predictive performance of both methods (Example 1)

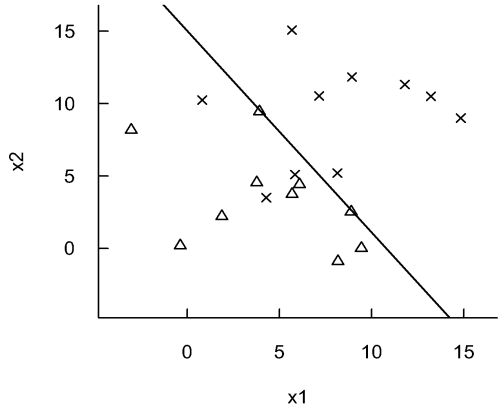

Figure 5: The separation line by the minimax strategy, $\sigma_i = 4$ (Example 1)

$m_i(1) = 8$, $\sigma_i(y) = 2$, and we applied both separating functions to them. This resulted in $ACC_{\text{minimax}} = 0.79$ and $ACC_{\text{standard}} = 0.66$, so the newly proposed method performs better in this case than the standard SVM, the data and separating lines are shown in Fig. 4.

We repeated this example with all inputs into the simulated values the same except for increased standard deviations $\sigma_i(y) = 4$, leading to more overlap of the data from the two groups. Figures 5 and 6 show again the initially simulated observations together with the minimax and SVM based separating lines. The predictive results for this example were $ACC_{\text{minimax}} = 0.80$ and $ACC_{\text{standard}} = 0.71$, as illustrated by Figure 7. It is not necessarily the fact that the minimax method again performs better that should be noted, but particularly the fact that the standard method leads to a very different separating line in both these cases while the minimax method leads to a separating line that seems pretty robust and that more naturally separates the two groups.

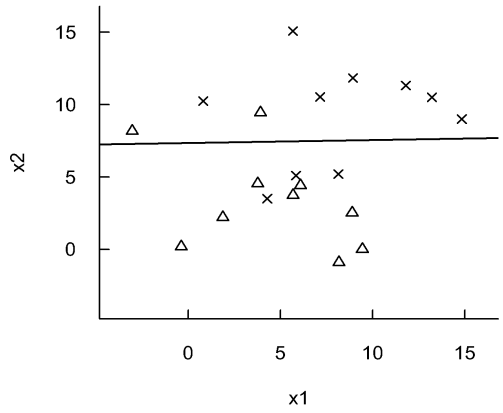

Figure 6: The separation line by the standard SVM, $\sigma_i = 4$ (Example 1)

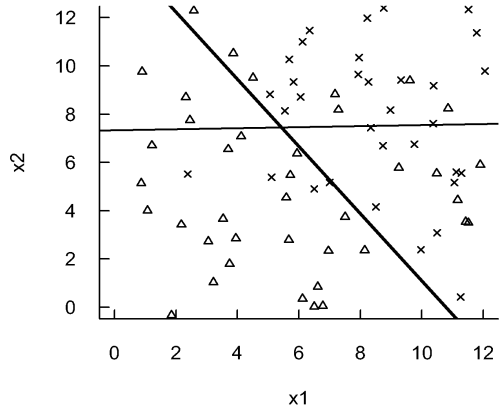

Figure 7: Predictive performance of both methods, $\sigma_i = 4$ (Example 1)

Example 2
We use "Haberman's Survival Data Set" from the UCI Machine Learning Repository [8]. The data set contains cases from a study on the survival of patients who had undergone surgery for breast cancer. The number of features (attributes) is three: age of patient at time of operation (x_1), patient's year of operation (x_2), and the number of positive axillary nodes detected (x_3). The classes are defined by the survival status of every patient ($y = -1$: the patient survived 5 years or longer; $y = 1$: the patient died within 5 year). The total number of examples is 306, of which the number of examples with $y = -1$ is 225.

First, we have randomly selected 16 examples (8 examples from each class). Using this reduced data set, the two separating functions according to our minimax method and the standard SVM approach are obtained and applied to the total data set to check their predictive performance for classification. This led to the following measures of accuracy: $ACC_{\text{minimax}} = 0.31$, $ACC_{\text{standard}} = 0.29$, hence their performance for this single case is pretty similar with only a small advantage for the minimax method. We may want to take into account the fact that numbers of examples in the two classes in this data set are different. To do so, we have constructed two separating functions based on $n_{-1} = 12$ and $n_1 = 4$ randomly selected patients, which reflects the proportions of examples in the classes. Applying these two functions to the total data set led to $ACC_{\text{minimax}} = 0.75$, $ACC_{\text{standard}} = 0.28$, which suggests a far better result for the minimax method, but of course that can be due to it being just a single application with small numbers of examples used. To get a better insight into the comparative performances of these two methods, we repeated this analysis, with the same numbers of examples but different random selections, 23 times. The results were mostly pretty close, with the averages of the ACC_{minimax} values equal to 0.54 and the average of the ACC_{standard} values equal to 0.50. Of the 23 repeats, the performance of the minimax approach was best 17 times, while the standard SVM approach was better 4 times and on two occasions both had the same numbers of correct classifications.

Example 3
As a further example we applied both our minimax method and the standard SVM method to the "Pima Indian Diabetes" data set, also from the UCI Machine Learning Repository [8]. The Pima data set has eight features ($m = 8$), with 768 training instances (examples) of which 500 are labeled as positive. Every example is characterized by numerical-valued features. We aimed at keeping close to the proportion of examples in the data set and took a total of 16 examples, with $n_{-1} = 10$ and $n_1 = 6$. This led to two separating functions, which applied to the total data set led to $ACC_{\text{minimax}} = 0.70$, $ACC_{\text{standard}} = 0.64$. We also repeated this to get a better insight into the performance of our method. We repeated this example 36 times, taking different random selections from the total data with each time $n_{-1} = 10$ and $n_1 = 6$. The average of the ACC_{minimax} values was 0.65 and the average of the ACC_{standard} values was 0.62, with the minimax method having the better performance in 25 cases and SVM performing better in 11 cases. Finally, we also applied these two methods with the very small numbers of examples $n_{-1} = 5$ and $n_1 = 3$, and we considered

10 cases, with the examples used again selected randomly from the data. The average of the ACC_{minimax} values was 0.68 and the average of the ACC_{standard} values was 0.60, with the minimax method having the better performance in 7 cases, SVM performing better in 1 case and exactly the same number of correct classifications in 2 cases. This shows in particular that the presented method can provide a reasonable performance even in cases of very small amounts of data, which was one of the main reasons to consider the nonparametric KS bounds method.

8. Concluding remarks

In this paper, a new method for constructing a classification model has been proposed, which is based on KS bounds and capable to deal with a small amount of statistical data, as frequently occurs in practice. This new method has several important features. First, it has a clear explanation and justification in the framework of decision theory as it uses a formal framework with the minimax criterion. Secondly, it allows a wide variety of inferential methods for constructing the p-boxes based on other bands, for instance, one can explore the use of Anderson-Darling bands [9]. Thirdly, the resulting statistical inferences are similar to some well-known robust statistics methods by weighting observations differently, hence the current approach provides novel formal justifications and interpretations for such statistical methods in a decision theoretic framework. A further important strength of the proposed method is the link with the SVM approach which has become very popular in the machine learning community.

The small examples presented in this paper lead to careful optimism about the performance of our novel method compared to the standard SVM approach, which suggests that further investigation into this method, including the choice of the value of γ, and its application in a variety of areas might be of interest. Indeed, such further investigation sets out an important programme of research, where the proposed method is to be compared to many competing methods for classification, from the literatures of classical statistics, imprecise probability-based statistics and machine learning. Detailed investigation will mostly need to be based on extensive simulations, considering a wide variety of scenarios with sample sizes ranging from small to very large, and considering different means of comparing the classification results. Particular care will have to be taken with regard to scenarios with differing levels of agreement with assumptions underlying different methods, in particular the extent to which the usual normality assumption for residuals fits with the data generation mechanisms. Of course, if this fit is close then the method presented in this paper will be at a disadvantage, so main focus will be on how close its performance matches that of methods based on assumptions which happen to be accurate. In scenarios where such further assumptions underlying competing methods are not realistic, our method is expected to perform better yet detailed studies will be needed to reveal the actual difference in performance related to sample sizes and specific aspects of underlying distributions. Of course, in many practical situations the actual underlying scenario, in particular any suitable probability distributions, will typically be unknown, hence we would always suggest

simultaneous use of a variety of classification methods in order to compare the results; if these agree strongly then one can have confidence in the resulting inferences, while in the event that these vary widely great care is required with regard to assumptions underlying the methods used, where the method presented in this paper has the advantage of being based on relatively weak assumptions. As mentioned, the fact that the method presented in this paper has strong foundationgs and appears to perform well in the small examples presented, justifies such a further research programme which we hope to undertake in the near future.

It has been shown that the dual optimisation form for our method has the same objective function as the standard SVM form, with the difference being the addition of further constraints which are implicitly defined by the confidence level $1 - \gamma$. A key feature of SVMs is the use of kernels which are functions that transform the input data to a high-dimensional space where the learning problem is solved. Such kernel functions can be linear or nonlinear, which will allow us to significantly extend the class of discriminant functions that can be used.

The main difficulty of implementing the method is a possibly large number of extreme points $\mathcal{E}^*(\mathcal{H}(y))$ and, as a result, a large number of constraints in the primal or dual optimisation problems. However, the method is explicitly aimed at cases where one only has a small amount of statistical data, in which case the number of additional constraints is limited and the optimisation problem does not lead to major computational difficulties.

It is not difficult to study another decision criterion, for example the minimin strategy, which can be regarded as an optimistic criterion. For this criterion, the optimisation problem is easier than for maximin because we do not need to introduce new variables $G(-1)$ and $G(1)$. It is interesting to consider also so-called cautious decision making as a linear combination of the minimax and minimin strategies. A method for cautious decision making was proposed by Utkin and Augustin [26] and it can also be applied to classification problems.

It is important to note that the training set is divided into two subsets which are considered separately in the proposed classification model. We could divide the training set into several subsets in the same way. This technique allows us to simply extend the binary model to the multi-class classification.

As mentioned in the introduction to this paper, a further important problem in practical classification problems is that data are often imperfect, with some missing entries or forms of censoring. The generalization of the approach presented in this paper to deal with such problems is a major research challenge to which imprecise methods may provide exciting solutions.

Acknowledgement

We would like to express our appreciation to the anonymous referees whose very valuable comments have improved the paper.

References

[1] T. Augustin, F.P.A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124:251–272, 2004.

[2] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[3] F.P.A. Coolen. Nonparametric predictive inference. In: *International Encyclopedia of Statistical Science*, Miodrag Lovric (Ed.). Springer, 968–970, 2011.

[4] F.P.A. Coolen, M.C. Troffaes, T. Augustin. Imprecise probability. In: *International Encyclopedia of Statistical Science*, Miodrag Lovric (Ed.). Springer, 645–648. 2011.

[5] A.P. Dempster. Upper and lower probabilities induced by a multi-valued mapping. *Annales of Mathematical Statistics*, 38:325–339, 1967.

[6] S. Destercke, D. Dubois, E. Chojnacki. Unifying practical uncertainty representations - i: Generalized p-boxes. *International Journal of Approximate Reasoning*, 49:649–663, 2008.

[7] T. Evgeniou, T. Poggio, M. Pontil, A. Verri. Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38:421–432, 2002.

[8] A. Frank, A. Asuncion. UCI machine learning repository (archive.ics.uci.edu/ml/).

[9] J. Frey. Confidence bands for the cdf when sampling from a finite population. *Computational Statistics & Data Analysis*, 53:4126–4132, 2009.

[10] I. Gilboa, D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18:141–153, 1989.

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

[12] N.L. Johnson, F. Leone. *Statistics and Experimental Design in Engineering and the Physical Sciences*, volume 1. Wiley, New York, 1964.

[13] E. Kriegler, H. Held. Utilizing belief functions for the estimation of future climate change. *International Journal of Approximate Reasoning*, 39:185–209, 2005.

[14] V.J. Montgomery, F.P.A. Coolen, A.D.M. Hart. Bayesian probability boxes in risk assessment. *Journal of Statistical Theory and Practice*, 3:69–83, 2009.

[15] F.M. Mulier V. Cherkassky. *Learning from Data: Concepts, Theory, and Methods*. Wiley-IEEE Press, UK, 2007.

[16] S. Petit-Renaud, T. Denoeux. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning*, 35:1–28, 2004.

[17] E. Quaeghebeur, G. de Cooman. Imprecise probability models for inference in exponential families. In: J.-M. Bernard, T. Seidenfeld, M. Zaffalon (Eds), *Proceedings of the 4rd International Symposium on Imprecise Probabilities and Their Applications, ISIPTA'05*, pp. 287–296, Pittsburgh, Pennsylvania, July 2005. Carnegie Mellon University.

[18] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005.

[19] C.P. Robert. *The Bayesian Choice*. Springer, New York, 1994.

[20] B. Scholkopf, A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, Massachusetts, 2002.

[21] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[22] A.J. Smola, B. Scholkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

[23] A.N. Tikhonov, V.Y. Arsenin. *Solutions of Ill-Posed Problems*. W.H. Winston, Washington DC, 1977.

[24] M.C.M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45:17–29, 2007.

[25] L.V. Utkin. Regression analysis using the imprecise Bayesian normal model. *International Journal of Data Analysis Techniques and Strategies*, 2:356–372, 2010.

[26] L.V. Utkin, Th. Augustin. Efficient algorithms for decision making under partial prior information and general ambiguity attitudes. In: T. Seidenfeld F.G. Cozman, R. Nau (Eds), *Proceedings of the 4th International Symposium on Imprecise Probabilities and Their Applications, ISIPTA'05*, pp. 349–358, Pittsburgh, USA, July 2005. Carnegie Mellon University, SIPTA.

[27] L.V. Utkin, F.P.A. Coolen. On reliability growth models using Kolmogorov-Smirnov bounds. *International Journal of Performability Engineering*, 7:5–19, 2011.

[28] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[29] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[30] P. Walley. Inferences from multinomial data: Learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996.

[31] P. Walley. Measures of uncertainty in expert systems. *Artificial Intelligence*, 83:1–58, 1996.

[32] G. Walter, Th. Augustin, A. Peters. Linear regression analysis under sets of conjugate priors. In: G. de Cooman, J. Vejnarova, M. Zaffalon (Eds), *Proceedings of the Fifth International Symposium on Imprecise Probabilities and Their Applications*, pages 445–455, Prague, Czech Republic, 2007.

[33] L. Wasserman. *All of Nonparametric Statistics*. Springer, New York, 2006.

[34] A.R. Webb. *Statistical Pattern Recognition, 2nd Edition*. Wiley, 2002.