Employing different loss functions for the classification of images via supervised learning

Radu Ioan Boţ [*]   André Heinrich [†]   Gert Wanka [‡]

May 8, 2013

Abstract. Supervised learning methods are powerful techniques to learn a function from a given set of labeled data, the so-called training data. In this paper the support vector machines approach is applied to an image classification task. Starting with the corresponding Tikhonov regularization problem, reformulated as a convex optimization problem, we introduce a conjugate dual problem to it and prove that, whenever strong duality holds, the function to be learned can be expressed via the dual optimal solutions. Corresponding dual problems are then derived for different loss functions. The theoretical results are applied by numerically solving a classification task using high dimensional real-world data in order to obtain optimal classifiers. The results demonstrate the excellent performance of support vector classification for this particular problem.

Keywords. machine learning, Tikhonov regularization, conjugate duality, image classification

AMS subject classification. 47A52, 90C25, 49N15

1   Introduction

Supervised learning methods such as Support Vector Machines for classification and regression belong to the class of kernel based methods that have become, especially in the last decade, a popular approach for learning functions from a given set of labeled data. They have wide fields of application such as image and text classification (cf. [6,8]), computational biology (cf. [13]) or time series forecasting and credit scoring (cf. [9,26]) and have proven to provide very good results.

This article originates from a real-world problem a supplier of the automotive industry was faced with, namely the task of establishing a computer-aided quality control of manufactured devices. These devices are photographed directly at the end of the manufacturing process and the idea was to perform quality control based on these images. Due to the promising performance of Support Vector Machines in the field of classification tasks, we applied this approach for this concrete image classification problem

---

[*]Department of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany, e-mail: radu.bot@mathematik.tu-chemnitz.de. On leave to Faculty of Mathematics, University of Vienna.

[†]Department of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany, e-mail: andre.heinrich@mathematik.tu-chemnitz.de.

[‡]Department of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany, e-mail: gert.wanka@mathematik.tu-chemnitz.de.

and, as we show here, we obtained excellent results concerning the test set classification errors.

To this end we reformulate the general Tikhonov regularization problem (cf. [25]), to which the supervised learning problem gives rise, as a convex (not necessarily differentiable) optimization problem, construct a conjugate dual to it (see, for instance, [2]), prove under suitable qualification conditions the existence of strong duality and express the optimal solutions of the primal problem via the ones of the dual. This has as consequence the formulation of the decision function to be learned by means of the optimal solutions of the dual. Hence for the specific learning task one only has to numerically solve the dual problem, which, different to the primal one, can be mainly equivalently formulated as a convex differentiable optimization problem.

The convex duality theory was used in the context of regularized supervised learning models in [15] as well. Theoretical statements on the excess classification errors for Gaussian kernels and general convex loss functions were provide in [31], while a general theory for making comparisons of excess risks of different loss functions was established in [20]. From computational point of view, one can notice in the last years an increase in the role played by the algorithms for supervised machine learning based on general iterative schemes designed for solving nonsmooth convex optimization problems. They can be seen as a theoretically well-funded alternative to some early algorithms with a consistent heuristical component (see, for instance, [8]). In this context we mention the computational approaches relying on the alternating direction method of multipliers (ADMM) (see, for instance, [5, 12]), on the use of convex smoothing techniques (see [19, 32]) and on some primal-dual proximal splitting methods (see [3]), all of them relying on quite intricate implementations. The approach introduced in this paper relies on the idea that for the most popular loss functions used in supervised classification one only needs to solve the dual conjugate problem, assuming the minimization of a convex differentiable function over the nonnegative orthant, and which can be done by standard optimization routines. The very low misclassification errors obtained when applying our approach to the considered data set certify its good performance.

The paper is organized as follows. In Section 2 the general regularization problem is introduced and it is stated as an equivalent convex optimization problem. A Fenchel-type dual problem to it is provided and, under a suitable weak qualification condition, the existence of strong duality for this primal-dual pair is proved, which gives rise to the formulation of necessary and sufficient optimality conditions. In Section 3 the general theory from the previous section is employed for several particular loss functions. An application of the theoretical results to a high dimensional image classification task is done in Section 4, allowing an analysis of the suitability of choosing one of the considered loss functions for this particular problem. A conclusive section closes the paper.

2   Theoretical considerations

Given a set of *training data* $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and corresponding labels $y_i \in \{-1, +1\}$, $i = 1, \ldots, n$, grouping the data into two different classes, a common approach for learning a classifier based on the Structural Risk Minimization Principle is to apply *Support Vector Machines (SVM)* techniques for classification. These su-

pervised learning methods were investigated in detail by Vapnik in [27]. Considering $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^d \times \{+1, -1\}$ the training set, the aim of the SVM approach is to find a function f belonging to $\mathcal{F}$, a space of real-valued functions defined on $\mathbb{R}^d$ enhanced with some *a priori* information, that correctly classifies new data into one of the two classes.

A so-called *loss function* $v : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, assumed to be proper and convex in its first variable, enables to impose a penalty for predicting $f(x_i)$, where the true value is y_i, for $i = 1, \ldots, n$. One of the common assumptions on f is *smoothness*, which guarantees that two similar inputs correspond to two similar outputs. In order to control it, one needs to consider a *smoothness functional* $\Omega : \mathcal{F} \to \mathbb{R}$ (cf. [25]), having the desired characteristic of taking high values for non-smooth functions and low values for smooth functions.

Hence, the desired function f will be the optimal solution of the *Tikhonov regularization problem*

$$\inf_{f \in \mathcal{F}} \left\{ C \sum_{i=1}^{n} v(f(x_i), y_i) + \frac{1}{2}\Omega(f) \right\}, \tag{1}$$

where $C > 0$ is the so-called *regularization parameter* controlling the tradeoff between the loss function and the smoothness functional (see [4]).

For different choices of the loss function and of the smoothness functional (1) gives rise to different problem formulations in supervised learning. When taking, for instance, the norm square as smoothing functional, one can get for different choices of the loss function the classical Support Vector Machines approaches for Classification (cf. [18, 22, 28]) and Regression (cf. [27]), while, when varying the smoothness functional, several L_1-norm and L_∞-norm approaches (cf. [14]), but also the popular lasso methods (cf. [23, 24]) arise as particular instances of (1).

In the following the function f is assumed to be an element of the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ induced by a continuous *kernel function* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (cf. [1]) which we assume to be *symmetric* and *finitely positive semidefinite*. The kernel k is said to be symmetric if $k(x, y) = k(y, x)$ for all $x, y \in \mathbb{R}^d$. A symmetric kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, which for all $m \geq 1$ and all finite sets $\{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ fulfills $\sum_{i,j=1}^{m} a_i a_j k(x_i, x_j) \geq 0$ for every arbitrary $a \in \mathbb{R}^d$ is called *finitely positive semidefinite* (cf. [18]).

Hence, the kernel function k can be decomposed as $k(x, y) = \langle \phi(x), \phi(y) \rangle_k$, where $\langle \cdot, \cdot \rangle_k$ denotes the *inner product* of $\mathcal{H}_k$ and $\phi : \mathbb{R}^d \to \mathcal{H}_k$ is a so-called *feature map*. The *representer theorem* (cf. [30]) ensures that for every minimizer f of (1) there exists a vector $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ such that

$$f(\cdot) = \sum_{i=1}^{n} c_i k(\cdot, x_i), \tag{2}$$

where the upper index ^{T} transposes a column vector on $\mathbb{R}^n$ into a row one and viceversa. For $i = 1, \ldots, n$ the vectors x_i with the property that the corresponding coefficient c_i is not equal to zero are the so-called *support vectors*. The classification is realized by the *sign-function*, i.e. for a given data point x the predicted value is equal to the sign

3

of $f(x)$ for $f(x) \neq 0$, whereas for $f(x) = 0$ we have to specify the allocation to one of the two classes.

The existence of such a representation is essential for the purpose of this paper. Finally, we define the smoothness functional Ω to be $\Omega(f) = \|f\|_k^2$ for $f \in \mathcal{H}_k$, where $\|\cdot\|_k$ denotes the norm on $\mathcal{H}_k$. The *Gram matrix* of k with respect to the set $\{x_1, \ldots, x_n\}$ is denoted by $K \in \mathbb{R}^{n \times n}$, being the matrix with entries $K_{ij} := k(x_i, x_j)$, $i, j = 1, \ldots, n$. Obviously, K is symmetric and positive semidefinite. On $\mathbb{R}^d$ we consider the Euclidean norm, while for two vectors $x, y \in \mathbb{R}^d$ we denote by $x^T y$ their scalar product. Taking $c \in \mathbb{R}^n$ to be the vector corresponding to representation (2), the smoothness functional becomes $\Omega(f) = \|f\|_k^2 = c^T K c$ and for $i = 1, \ldots, n$ it holds $f(x_i) = \sum_{j=1}^n c_j K_{ij} = (Kc)_i$. Thus we can rewrite optimization problem (1) equivalently as

$$(P_{gen}) \qquad \inf_{c \in \mathbb{R}^n} \left\{ C \sum_{i=1}^n v((Kc)_i, y_i) + \frac{1}{2} c^T K c \right\}.$$

Due to the nature of the loss function, this optimization problem is convex and in general not differentiable. In order to overcome this disadvantage, we provide a conjugate dual problem to it, prove the existence of strong duality and express the optimal solutions of (P_{gen}) via the ones of the dual. These considerations make much sense, especially when the dual problem is easier to solve than the primal one, which is actually the case for the majority of the loss functions used for supervised classification problems.

In order to make the paper self-contained, we introduce first some notions and results. For a nonempty set $D \subseteq \mathbb{R}^n$ we denote by $\mathrm{ri}(D)$ the *relative interior* of the set D, that is the interior of D relative to its affine hull. The indicator function of D is defined as

$$\delta_D : \mathbb{R}^n \to \overline{\mathbb{R}}, \quad \delta_D(x) = \begin{cases} 0, & \text{if } x \in D, \\ +\infty, & \text{otherwise.} \end{cases}$$

For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ we denote its *effective domain* by $\mathrm{dom}\, f = \{x \in \mathbb{R}^n : f(x) < +\infty\}$ and say that f is *proper* if $\mathrm{dom}\, f \neq \emptyset$ and $f > -\infty$. The *(Fenchel-Moreau) conjugate function* of f is $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$, defined by $f^*(p) = \sup_{x \in \mathbb{R}^n} \{p^T x - f(x)\}$. For all $x, p \in \mathbb{R}^n$ we have the following relation, known as the *Young-Fenchel inequality*, $f(x) + f^*(p) - p^T x \geq 0$. For $x \in \mathbb{R}^n$ with $f(x) \in \mathbb{R}$ we denote by $\partial f(x) := \{p \in \mathbb{R}^n : f(y) - f(x) \geq p^T(y - x) \ \forall y \in \mathbb{R}^n\}$ the *(convex) subdifferential of f at x*. Otherwise, we assume by convention that $\partial f(x) = \emptyset$. For $x \in \mathbb{R}^n$ with $f(x) \in \mathbb{R}$, one has that

$$p \in \partial f(x) \Leftrightarrow f(x) + f^*(p) = p^T x. \tag{3}$$

The *epigraph* of f is $\mathrm{epi}\, f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq r\}$ and f is said to be *convex*, if $\mathrm{epi}\, f$ is a convex set, while f is said to be *lower semicontinuous* if $\mathrm{epi}\, f$ is a closed set. Having a convex set D and a function $f : D \to \mathbb{R}$, we say that f is *strictly convex on D*, if

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y) \ \forall x, y \in D, x \neq y, \ \forall \lambda \in (0, 1)$$

and that f is *strongly convex on D*, if there exists $\mu > 0$ such that

$$f(\lambda x + (1 - \lambda)y) + \lambda(1 - \lambda)\mu \|x - y\|^2 \leq \lambda f(x) + (1 - \lambda)f(y) \ \forall x, y \in D \ \forall \lambda \in (0, 1).$$

When $K \in \mathbb{R}^{n \times n}$ is a given matrix, we denote by $\operatorname{Im} K := \{Kx : x \in \mathbb{R}^n\}$. Further, for $x \in \mathbb{R}$ we define $x_+ := \max\{0, x\}$.

The dual problem to (P_{gen}) which we consider here is a Fenchel-type dual problem (cf. [4]) and it is formulated as

$$(D_{gen}) \qquad \sup_{\substack{p \in \mathbb{R}^n, \\ p = (p_1, \dots, p_n)^T}} \left\{ -C \sum_{i=1}^{n} \left(v(\cdot, y_i)\right)^* \left(-\frac{p_i}{C}\right) - \frac{1}{2} p^T K p \right\}. \qquad (4)$$

Let us denote by $\operatorname{val}(P_{gen})$ the optimal objective value of the primal problem (P_{gen}) and by $\operatorname{val}(D_{gen})$ the optimal objective value of its dual problem (D_{gen}). First of all, we show that for the minimization problem (P_{gen}) and its dual problem (D_{gen}) weak duality holds.

Theorem 1. *For (P_{gen}) and (D_{gen}) weak duality holds, i. e. $\operatorname{val}(P_{gen}) \geq \operatorname{val}(D_{gen})$.*

Proof. Let be $c \in \mathbb{R}^n$ and $p = (p_1, \dots, p_n)^T \in \mathbb{R}^n$. Then it holds, according to Young-Fenchel inequality and due to the positive semidefiniteness of K, that

$$
\begin{aligned}
0 \leq\ & C \left[\sum_{i=1}^{n} v((Kc)_i, y_i) + \sum_{i=1}^{n} \left(v(\cdot, y_i)\right)^* \left(-\frac{p_i}{C}\right) + \sum_{i=1}^{n} (Kc)_i \frac{p_i}{C} \right] \\
& + \frac{1}{2}(c - p)^T K (c - p) \\
=\ & C \sum_{i=1}^{n} v((Kc)_i, y_i) + C \sum_{i=1}^{n} \left(v(\cdot, y_i)\right)^* \left(-\frac{p_i}{C}\right) + p^T (Kc) \\
& + \frac{1}{2} c^T K c + \frac{1}{2} p^T K p - p^T (Kc) \\
=\ & C \sum_{i=1}^{n} v((Kc)_i, y_i) + \frac{1}{2} c^T K c + C \sum_{i=1}^{n} \left(v(\cdot, y_i)\right)^* \left(-\frac{p_i}{C}\right) + \frac{1}{2} p^T K p
\end{aligned}
$$

and therefore

$$C \sum_{i=1}^{n} v((Kc)_i, y_i) + \frac{1}{2} c^T K c \geq -C \sum_{i=1}^{n} \left(v(\cdot, y_i)\right)^* \left(-\frac{p_i}{C}\right) - \frac{1}{2} p^T K p,$$

i. e. $\operatorname{val}(P_{gen}) \geq \operatorname{val}(D_{gen})$. $\qquad \square$

By introducing the functions $v_i : \mathbb{R}^n \to \overline{\mathbb{R}}$, $v_i(z) = v(z_i, y_i)$, $i = 1, \dots, n$, the problem (P_{gen}) can equivalently be written as

$$(P_{gen}) \qquad \inf_{c \in \mathbb{R}^n} \left\{ C \sum_{i=1}^{n} v_i(Kc) + \frac{1}{2} c^T K c \right\}.$$

In order to ensure strong duality for the primal-dual pair $(P_{gen}) - (D_{gen})$, we impose the following *qualification condition*

$$(QC) \qquad \operatorname{Im} K \cap \prod_{i=1}^{n} \operatorname{ri}(\operatorname{dom} v(\cdot, y_i)) \neq \emptyset.$$

5

Theorem 2. *If (QC) is fulfilled, then it holds $\mathrm{val}(P_{gen}) = \mathrm{val}(D_{gen})$ and (D_{gen}) has an optimal solution.*

Proof. We notice first that

$$\mathrm{val}(P_{gen}) = \inf_{c \in \mathbb{R}^n} \left\{ \left(\sum_{i=1}^{n} Cv_i \right)(Kc) + \frac{1}{2} c^T Kc \right\}.$$

Denoting by $g : \mathbb{R}^n \to \mathbb{R}$, $g(c) = \frac{1}{2} c^T Kc$, and by taking into consideration that $\mathrm{dom}\left(\sum_{i=1}^{n} Cv_i \right) = \prod_{i=1}^{n} \mathrm{dom}\, v(\cdot, y_i)$, one has

$$K(\mathrm{ri}(\mathrm{dom}\, g)) \cap \mathrm{ri}\left(\sum_{i=1}^{n} Cv_i \right) = \mathrm{Im}\, K \cap \prod_{i=1}^{n} \mathrm{ri}(\mathrm{dom}\, v(\cdot, y_i)) \neq \emptyset.$$

This means that $\mathrm{val}(P_{gen}) < +\infty$. Moreover, there exists a $\bar{p} = (\bar{p}_1, \dots, \bar{p}_n)^T \in \mathbb{R}^n$ such that (see [2, Theorem 2.1])

$$\mathrm{val}(P_{gen}) = \sup_{p \in \mathbb{R}^n} \left\{ -\left(\sum_{i=1}^{n} Cv_i \right)^{*}(-p) - g^{*}(Kp) \right\}$$

$$= -\left(C \sum_{i=1}^{n} v_i \right)^{*}(-\bar{p}) - g^{*}(K\bar{p}).$$

Since for $q \in \mathbb{R}^n$,

$$g^{*}(q) = \begin{cases} \frac{1}{2} q^T K^{-} q, & \text{if } q \in \mathrm{Im}\, K, \\ +\infty, & \text{otherwise,} \end{cases}$$

where K^{-} is the Moore-Penrose pseudo-inverse of K, it holds

$$g^{*}(K\bar{p}) = \frac{1}{2}(K\bar{p})^T K^{-}(K\bar{p}) = \frac{1}{2} \bar{p}^T KK^{-}K\bar{p} = \frac{1}{2} \bar{p}^T K\bar{p}$$

and, so,

$$\mathrm{val}(P_{gen}) = -C\left(\sum_{i=1}^{n} v_i \right)^{*}\left(-\frac{1}{C} \bar{p} \right) - \frac{1}{2} \bar{p}^T K\bar{p}.$$

As from (QC) one has $\cap_{i=1}^{n} \mathrm{ri}(\mathrm{dom}\, v_i) = \prod_{i=1}^{n} \mathrm{ri}(\mathrm{dom}\, v(\cdot, y_i)) \neq \emptyset$, it follows (cf. [16]) that there exist $\bar{p}^i \in \mathbb{R}^n$, $i = 1, \dots, n$, with $\sum_{i=1}^{n} \bar{p}^i = \bar{p}$, such that

$$\left(\sum_{i=1}^{n} v_i \right)^{*}\left(-\frac{1}{C} \bar{p} \right) = \sum_{i=1}^{n} v_i^{*}\left(-\frac{1}{C} \bar{p}^i \right)$$

and, therefore,

$$\mathrm{val}(P_{gen}) = -C \sum_{i=1}^{n} v_i^{*}\left(-\frac{1}{C} \bar{p}^i \right) - \frac{1}{2} \left(\sum_{i=1}^{n} \bar{p}^i \right)^T K \left(\sum_{i=1}^{n} \bar{p}^i \right).$$

Further, for all $i = 1, \dots, n$, it holds

$$v_i^{*}\left(-\frac{1}{C} \bar{p}^i \right) = \sup_{z \in \mathbb{R}^n} \left\{ -\frac{1}{C} (\bar{p}^i)^T z - v(z_i, y_i) \right\} = \begin{cases} \left(v(\cdot, y_i) \right)^{*}\left(-\frac{\bar{p}_i^i}{C} \right), & \text{if } \bar{p}_j^i = 0, \forall j \neq i, \\ +\infty, & \text{otherwise.} \end{cases}$$

Since the optimal objective value of (P_{gen}) is finite, it holds $\bar{p}_i = \bar{p}_i^i$ for $i = 1, ..., n$ and

$$\text{val}(P_{gen}) = -C \sum_{i=1}^{n} \left(v(\cdot, y_i) \right)^* \left(-\frac{\bar{p}_i}{C} \right) - \frac{1}{2} \bar{p}^T K \bar{p}.$$

This, along with the weak duality theorem, provides the desired result, $\bar{p}$ being an optimal solution to (D_{gen}). $\qquad\square$

The next theorem furnishes the necessary and sufficient optimality conditions for the primal-dual pair $(P_{gen}) - (D_{gen})$.

Theorem 3. *Let (QC) be fulfilled. If $\bar{c} \in \mathbb{R}^n$ is an optimal solution to (P_{gen}), then there exists an optimal solution $\bar{p} \in \mathbb{R}^n$ to (D_{gen}) such that*

(i) $-\frac{\bar{p}_i}{C} \in \partial v(\cdot, y_i)((K\bar{c})_i),\ i = 1, \ldots, n;$

(ii) $K(\bar{c} - \bar{p}) = 0.$

Vice-versa, if for $\bar{c} \in \mathbb{R}^n$ and $\bar{p} \in \mathbb{R}^n$ the statements (i) and (ii) are fulfilled, then $\bar{c} \in \mathbb{R}^n$ is an optimal solution to (P_{gen}), $\bar{p} \in \mathbb{R}^n$ is an optimal solution to (D_{gen}) and $\text{val}(P_{gen}) = \text{val}(D_{gen})$.

Proof. Assuming that $\bar{c} \in \mathbb{R}^n$ is an optimal solution to (P_{gen}), from Theorem 2 we get the existence of an optimal solution $\bar{p} \in \mathbb{R}^n$ to (D_{gen}) such that

$$C \left[\sum_{i=1}^{n} v((K\bar{c})_i, y_i) + \sum_{i=1}^{n} \left(v(\cdot, y_i) \right)^* \left(-\frac{\bar{p}_i}{C} \right) + \sum_{i=1}^{n} (K\bar{c})_i \frac{\bar{p}_i}{C} \right]$$
$$+ \frac{1}{2} \bar{c}^T K \bar{c} + \frac{1}{2} \bar{p}^T K \bar{p} - \bar{p}^T K \bar{c} = 0. \qquad (5)$$

By the Young-Fenchel inequality one has

$$v((K\bar{c})_i, y_i) + (v(\cdot, y_i))^* \left(\frac{\bar{p}_i}{C} \right) \geq (K\bar{c})_i \frac{\bar{p}_i}{C} \ \forall i = 1, \ldots, n,$$

and

$$\frac{1}{2} \bar{c}^T K \bar{c} + \frac{1}{2} \bar{p}^T K \bar{p} \geq \bar{p}^T K \bar{c}.$$

Hence, these inequalities must be fulfilled as equations, which are further equivalent to

$$\begin{cases} v((K\bar{c})_i, y_i) + (v(\cdot, y_i))^* \left(\frac{\bar{p}_i}{C} \right) = (K\bar{c})_i \frac{\bar{p}_i}{C} \ \forall i = 1, \ldots, n, \\ \frac{1}{2}(\bar{c} - \bar{p})^T K (\bar{c} - \bar{p}) = 0. \end{cases}$$

Thus $\bar{c} - \bar{p}$ is a global minimum of the convex function $p \mapsto 1/2 p^T K p$, which means that the second statement in the relations above is nothing else but $K(\bar{c} - \bar{p}) = 0$. Statement (i) follows from the characterization of the subdifferential given in (3).

Vice-versa, if for $\bar{c} \in \mathbb{R}^n$ and $\bar{p} \in \mathbb{R}^n$ the statements (i) and (ii) are fulfilled, then (5) holds, which implies in light of Theorem 1 that $\bar{c} \in \mathbb{R}^n$ is an optimal solution to (P_{gen}), $\bar{p} \in \mathbb{R}^n$ is an optimal solution to (D_{gen}) and $\text{val}(P_{gen}) = \text{val}(D_{gen})$. $\qquad\square$

Remark 1. If K is positive definite, then, due to the fact that $v(\cdot, y_i)$ is proper and convex for all $i = 1, ..., n$, the qualification condition (QC) is automatically fulfilled. Thus, according to Theorem 3, $\bar{c} \in \mathbb{R}^n$ is an optimal solution for (P_{gen}) if and only if there exists an optimal solution $\bar{p} \in \mathbb{R}^n$ to (D_{gen}) such that

(i) $-\frac{\bar{p}_i}{C} \in \partial v(\cdot, y_i)((K\bar{c})_i)$, $i = 1, \ldots, n$;

(ii) $\bar{c} = \bar{p}$.

Remark 2. If K is positive definite, then the function g is strongly convex (on $\mathbb{R}^n$). Consequently, if $v(\cdot, y_i)$, $i = 1, \ldots, n$, is, additionally, lower semicontinuous, the optimization problem (P_{gen}) has a *unique optimal solution* (see, for instance, [7, Satz 6.33]). Further, due to the fact that $p \mapsto \frac{1}{2}p^T K p$ is strictly convex (on $\mathbb{R}^n$), one can see that the dual problem (D_{gen}) has at most one optimal solution. Consequently, due to Remark 1, whenever K is positive definite and $v(\cdot, y_i)$ is lower semicontinuous, for $i = 1, ..., n$, then in order to solve (P_{gen}) one can equivalently solve (D_{gen}) which in this case has an unique optimal solution $\bar{p}$, this being also the unique optimal solution of (P_{gen}).

3   Some classical loss functions as particular cases

In this section we deal with particular instances of the general model described in the previous one and construct, for three particular loss functions, the corresponding dual problems. We employed the three dual problems in concretely solving a classification problem on a data set of images, as we will show in Section 4.

3.1   Hinge loss

The first loss function we consider here is the *hinge loss* $v_{hl} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, defined as

$$v_{hl}(a, y) = (1 - ay)_+ \, ,$$

which is a proper, convex and lower semicontinuous function in its first component, while (QC) is obviously fulfilled. The primal optimization problem (P_{gen}) becomes in this case

$$(P_{hl}) \qquad \inf_{c \in \mathbb{R}^n} \left\{ C \sum_{i=1}^{n} \left(1 - (Kc)_i y_i \right)_+ + \frac{1}{2} c^T K c \right\}.$$

To obtain the dual problem (D_{hl}) of (P_{hl}) (cf. (4)) for this special loss function, we use the Lagrange technique in order to calculate the conjugate function of $v_{hl}(\cdot, y_i)$, for

$i = 1, ..., n$. For $z \in \mathbb{R}$ and $i = 1, ..., n$ we have

$$-\left(v_{hl}(\cdot, y_i)\right)^*(z) = -\sup_{a \in \mathbb{R}} \{za - (1 - ay_i)_+\} = \inf_{\substack{a, t \in \mathbb{R}, \\ t \geq 0, \, t \geq 1 - ay_i}} \{-za + t\}$$

$$= \sup_{k \geq 0, \, r \geq 0} \left\{ \inf_{a, t \in \mathbb{R}} \{-za + t + k(1 - ay_i - t) - rt\} \right\}$$

$$= \sup_{k \geq 0, \, r \geq 0} \left\{ \inf_{a \in \mathbb{R}} \{-za - kay_i\} + \inf_{t \in \mathbb{R}} \{t - kt - rt\} + k \right\}$$

$$= \sup_{\substack{k \geq 0, \, r \geq 0, \\ k + r = 1, \\ z + ky_i = 0}} k = \sup_{\substack{k \in [0,1], \\ k = -zy_i}} k = \begin{cases} -zy_i, & \text{if } zy_i \in [-1, 0], \\ -\infty, & \text{otherwise.} \end{cases}$$

Note that in the calculations above we used the fact that the labels y_i, $i = 1, \ldots, n$, can only take the values +1 or -1 for the binary classification task we consider in this paper (cf. Section 4). With the above formula we obtain the following dual problem

$$(D_{hl}) \qquad \sup_{\substack{p \in \mathbb{R}^n, \\ p_i y_i \in [0, C], \, i = 1, \ldots, n}} \left\{ \sum_{i=1}^{n} p_i y_i - \frac{1}{2} p^T K p \right\}$$

or, equivalently,

$$(D_{hl}) \qquad \inf_{\substack{p \in \mathbb{R}^n, \\ p_i y_i \in [0, C], \, i = 1, \ldots, n}} \left\{ \frac{1}{2} p^T K p - \sum_{i=1}^{n} p_i y_i \right\}.$$

By defining the vector $\alpha = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$, $\alpha_i := p_i y_i$, $i = 1, \ldots, n$, the dual problem can equivalently be written as

$$(D_{hl}) \qquad \inf_{\alpha_i \in [0, C], \, i = 1, \ldots, n} \left\{ \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K_{ij} - \sum_{i=1}^{n} \alpha_i \right\},$$

a representation which is recognized to be the commonly used form of the dual problem to (P_{hl}) in literature.

3.2 Generalized hinge loss

Beside the hinge loss, the binary image classification task has been performed for two other loss functions, as we point out in Section 4. They both represent particular instances of the *generalized hinge loss* $v_{ghl}^u : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$v_{ghl}^u(a, y) = (1 - ay)_+^u,$$

where $u > 1$. The generalized hinge loss function is proper, convex and lower semi-continuous in its first component, too, while the qualification condition (QC) is again obviously fulfilled. The primal problem this loss function gives rise to reads

$$(P_{ghl}^u) \qquad \inf_{c \in \mathbb{R}^n} \left\{ C \sum_{i=1}^{n} (1 - (Kc)_i y_i)_+^u + \frac{1}{2} c^T K c \right\}.$$

To obtain its dual problem we need the conjugate function of $v_{ghl}^u(\cdot, y_i)$ for $i = 1, ..., n$. For all $z \in \mathbb{R}$ and all $i = 1, ..., n$ we have

$$-(v_{ghl}^u(\cdot, y_i))^*(z) = -\sup_{a \in \mathbb{R}} \{za - (1 - ay_i)_+^u\} = \inf_{\substack{a, t \in \mathbb{R}, \\ t \geq 1 - ay_i}} \left\{-za + t^u + \delta_{[0, +\infty)}(t)\right\}.$$

By taking into account that the function $t \mapsto t^u + \delta_{[0, +\infty)}(t)$ is convex, we can make again use of Lagrange duality, which provides the following formula for the conjugate of $v_{ghl}^u(\cdot, y_i)$ for $i = 1, ..., n$ and $z \in \mathbb{R}$

$$
\begin{aligned}
-(v_{ghl}^u(\cdot, y_i))^*(z) &= \sup_{k \geq 0} \left\{\inf_{a \in \mathbb{R}, t \geq 0} \left\{-za + t^u + k(1 - ay_i - t)\right\}\right\} \\
&= \sup_{k \geq 0} \left\{\inf_{a \in \mathbb{R}} \{-za - kay_i\} + \inf_{t \geq 0} \{t^u - kt\} + k\right\} \\
&= \sup_{\substack{k \geq 0, \\ k = -zy_i}} \left\{(1 - u) \left(\frac{k}{u}\right)^{\frac{u}{u-1}} + k\right\} \\
&= \begin{cases} (1 - u) \left(\frac{-zy_i}{u}\right)^{\frac{u}{u-1}} - zy_i, & \text{if } zy_i \leq 0, \\ -\infty, & \text{otherwise.} \end{cases}
\end{aligned}
$$

Hence, the corresponding dual problem to (P_{ghl}^u) looks like

$$(D_{ghl}^u) \qquad \sup_{\substack{p_i \in \mathbb{R}, \\ p_i y_i \geq 0, i=1...,n}} \left\{\frac{1 - u}{(Cu^u)^{\frac{1}{u-1}}} \sum_{i=1}^n (p_i y_i)^{\frac{u}{u-1}} + \sum_{i=1}^n p_i y_i - \frac{1}{2} p^T K p\right\}.$$

Formulated as an infimum problem, (D_{ghl}^u) becomes

$$(D_{ghl}^u) \qquad \inf_{\substack{p_i \in \mathbb{R}, \\ p_i y_i \geq 0, i=1...,n}} \left\{\frac{1}{2} p^T K p + \frac{u - 1}{(Cu^u)^{\frac{1}{u-1}}} \sum_{i=1}^n (p_i y_i)^{\frac{u}{u-1}} - \sum_{i=1}^n p_i y_i\right\},$$

while, by taking $\alpha = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$, $\alpha_i := p_i y_i$, $i = 1, \ldots, n$, one obtains for it the following equivalent formulation

$$(D_{ghl}^u) \qquad \inf_{\alpha_i \geq 0, i=1,...,n} \left\{\frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K_{ij} + \frac{u - 1}{(Cu^u)^{\frac{1}{u-1}}} \sum_{i=1}^n \alpha_i^{\frac{u}{u-1}} - \sum_{i=1}^n \alpha_i\right\}.$$

This problem give rise for $u = 2$ to

$$(D_{ghl}^2) \qquad \inf_{\alpha_i \geq 0, i=1,...,n} \left\{\frac{1}{2} \left(\sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K_{ij} + \frac{1}{2C} \sum_{i=1}^n \alpha_i^2\right) - \sum_{i=1}^n \alpha_i\right\}$$

and for $u = 3$ to

$$(D_{ghl}^3) \qquad \inf_{\alpha_i \geq 0, i=1,...,n} \left\{\frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K_{ij} + \frac{2}{\sqrt{27C}} \sum_{i=1}^n \alpha_i^{\frac{3}{2}} - \sum_{i=1}^n \alpha_i\right\},$$

which are the situations that we employ, along the one corresponding to the hinge loss, in Section 4 for solving the classification task.

Remark 3. The problems (D_{hl}) and (D^2_{ghl}) are convex quadratic optimization problems with affine inequality constraints and they can be solved by making use of one of the standard solvers which exist for this class of optimization problems. This is not anymore the case for (D^3_{ghl}), which is however a convex optimization problem. Thus one can use for solving it instead one of the standard solvers for convex differentiable optimization problems with affine inequality constraints. In order to solve both the quadratic and the non-quadratic optimization problems, we applied appropriate optimization routines from the MATLAB$^{\circledR}$ [11] optimization package.

4 Application to image classification

In this section we describe the data for which the classification task, based on the approach described above, has been performed. Furthermore, we illustrate how the data has been preprocessed and give numerical results for the problems (D_{hl}), (D^2_{ghl}) and (D^3_{ghl}) arising when considering the different loss functions investigated in Section 3. These investigations can be extended by considering other loss functions and by calculating the corresponding dual problems. The only assumption we need for the former is convexity and lower semicontinuity in the first component, which the majority of the popular loss functions (with the exception of the $0 - 1$-*loss*) fulfill.

4.1 Training data

The available data were photographs of components used in the automotive industry, taken by a camera that is an internal part of the machine that produces these items. The overall task is to decide whether a produced component is fine or has to be considered as defective. In particular, a component is considered to be fine if a wire has been brazed correctly onto an attachment and it is defective otherwise. Consequently, a binary classification problem arises, where the label $+1$ denotes the class of components that are fine and the label -1 denotes the class of components that are defective. In other words, the goal of the classification task is to distinguish good joints from bad joints.

There was a total number of 4740 photographs of the components available, represented as gray scale images of size 200×50 pixels. Consisting of 2416 images of class $+1$ and 2324 images of class -1 the data set was nearly balanced. Since each pixel of the 8-bit gray-scale image represents a specific shade of gray, we assigned to it a value between 0 to 255, where the value equals 0 if the pixel is purely black and 255 if the pixel is purely white, respectively. Figure 4.1 shows four example images, two of each class.

4.2 Preprocessing

In order to be able to use the images for the classification task, we first transformed them into vectors. First, each of the images has been represented as a matrix $M_t \in \mathbb{R}^{200 \times 50}$, $M_t = (m^t_{i,j})^{200,50}_{i,j=1}$, $t = 1, \ldots, 4740$, with entries $m^t_{ij} \in \{0, 1, \ldots, 255\}$, $i = 1, \ldots, 200$, $j = 1, \ldots, 50$. By simply concatenating the rows of the matrix M_t, we obtained a vector
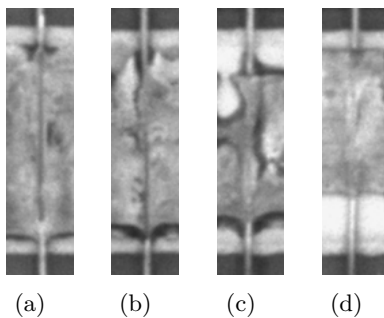
(a)   (b)   (c)   (d)

Figure 4.1: Example images of good ((a), (b)) and bad ((c), (d)) joints.

m_t representing image t, i.e.

$$m_t = (m_{11}^t, \ldots, m_{1\,200}^t, \ldots, m_{50\,1}^t, \ldots, m_{50\,200}^t)^T = (m_{t\,1}, \ldots, m_{t\,10\,000})^T \in \mathbb{R}^{10\,000}.$$

Denote by $\mathcal{D} = \{(m_t, y_t), t = 1, \ldots, 4740\} \subset \mathbb{R}^{10\,000} \times \{-1, +1\}$ the set of all data available. Following [10], due to numerical reasons, but also in order to guarantee an equal participation of the training image in the learning process, the data has been normalized by dividing each data point by the quantity $(\frac{1}{4740} \sum_{t=1}^{4740} \|m_t\|^2)^{\frac{1}{2}}$. Despite the fact that nowadays computations can in fact be performed for $10\,000-$dimensional vectors, we found it desirable to reduce their dimension to a value for which computations can be performed comparatively fast, especially concerning the calculation of the kernel matrix and the value of the decision function. For that reason, a so-called *feature ranking* (see [21]) was performed, by assigning a score to each pixel indicating its relevance for distinguishing between the two classes. Therefore, for the set of input data $D = \{m_1, \ldots, m_{4740}\}$ we defined the sets

$$D^+ := \{m_t \in D : y_t = +1 \} \text{ and } D^- := \{m_t \in D : y_t = -1\}$$

For both of these data sets, we calculated the *mean* μ_i,

$$\mu_i(D^+) = \frac{1}{|D^+|} \sum_{m_j \in D^+} m_{ji}, \quad \mu_i(D^-) = \frac{1}{|D^-|} \sum_{m_j \in D^-} m_{ji}, \quad i = 1, \ldots, 10\,000,$$

and the *variance* σ_i^2,

$$\sigma_i^2(D^+) = \frac{1}{|D^+|} \sum_{m_j \in D^+} (m_{ji} - \mu_i(D^+))^2, \quad \sigma_i^2(D^-) = \frac{1}{|D^-|} \sum_{m_j \in D^-} (m_{ji} - \mu_i(D^-))^2,$$

$i = 1, \ldots, 10\,000$, for each separate pixel of the images in the sets D^+ and D^-. Here, the values $|D^+|$ and $|D^-|$ denote the number of elements contained in the corresponding sets D^+ and D^-, respectively. The score S_i for the $i-$th pixel has been then calculated by

$$S_i(D) = \frac{(\mu_i(D^+) - \mu_i(D^-))^2}{\sigma_i^2(D^+) + \sigma_i^2(D^-)} \text{ for } i = 1, \ldots, 10\,000.$$

By applying this approach to the data set of images (cf. Figure (4.1)), we determined a score for each pixel, indicating its *relevance* for the classification task. Figure 4.2 plots the scores that have been assigned to the separate pixels. Finally, we have chosen only the pixels with a score greater or equal 0.1 in order to reduce the dimension of the input data. This approach provided a number of 4398 pixel relevant for the classification task.
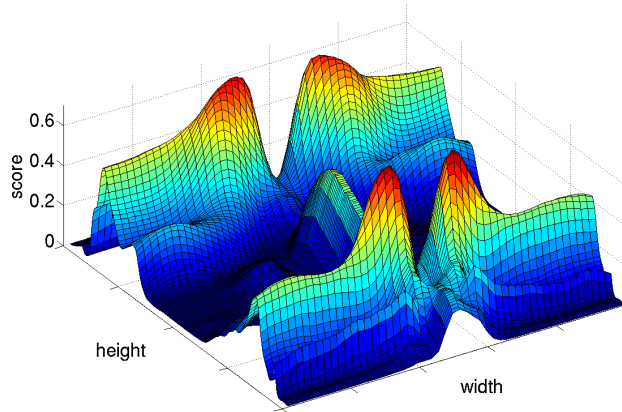

Figure 4.2: Visualization of the scores of the pixels.

4.3   Numerical results

To obtain a classifier numerical tests were performed for the three choices of the loss function discussed in the previous section, namely the *hinge loss* $v_{hl}(a, y) = (1 - ay)_+$ and the *generalized hinge loss* $v_{ghl}^u(a, y) = (1 - ay)_+^u$ for $u = 2$ and $u = 3$ and the corresponding three dual problems (D_{hl}), (D_{ghl}^2) and, respectively, (D_{ghl}^3) were used. As kernel function the *Gaussian RBF kernel*

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

with parameter $\sigma > 0$ was chosen. This gave rise to a positive definite Gram matrix K and, therefore, according to Remark 2, an optimal solution $\bar{p} := (\bar{p}_1, ..., \bar{p}_n)^T$ of the dual was an optimal solution of the primal, too. In this way the components of this vector provided the decision function we looked for when considering the classification task. Since the regularization parameter C and the kernel parameter σ were unknown and had to be determined by the user, first, a *10-fold cross validation* was performed for each of the three loss functions and for each combination (C, σ) from a given set of values for each parameter. The whole data set was split into *ten* disjoint and equally sized subsets resulting in ten folds, each of them consisting of 474 input data points. The average *classification error* over all ten test folds for each parameter combination and for each loss function was computed, giving information about the corresponding best combination of parameters C and σ. Table 4.1 shows the average classification errors over ten folds for a selection of tested parameter combinations.

13

loss function	C	σ			
		0.1	0.5	1	10
hinge loss	1	0.2321	0.3376	0.4220	49.030
	10	0.1899	0.2321	0.3165	0.6752
	100	0.1899	0.1688	0.2532	0.4220
	1000	0.1899	0.2110	0.3587	0.2954
quadratic hinge loss	1	0.2110	0.2743	0.3376	2.1100
	10	0.2110	0.2110	0.2532	0.4642
	100	0.1899	0.1688	0.2954	0.3587
	1000	0.1899	0.2110	0.3165	0.3376
cubic hinge loss	1	0.2110	0.2532	0.2954	1.0972
	10	0.1899	0.2321	0.3376	0.4431
	100	0.1899	0.1899	0.3165	0.3376
	1000	0.1899	0.2110	0.3165	0.3587

Table 4.1: Average classification errors over ten folds in percentage of the number of images contained in the test sets.

As one can see, the classification errors are remarkably small for all loss functions and for nearly all combinations of the *kernel parameter σ* and the *regularization parameter C*. There is an average misclassification rate of only up to 1% of the images contained in the test sets. The smallest errors occur for the combination $C = 100$ and $\sigma = 0.5$ for all loss functions. Taking this parameter combination as the optimal one, one obtains a number of 151 support vectors for the *hinge loss* function, i. e. only 3.2% of the images of the whole training data set are needed for the decision function. Concerning the *quadratic hinge loss*, we obtained 178 support vectors which is just a little more than for the usual hinge loss function. For the *cubic hinge loss* a total of 2207 support vectors was obtained, which is nearly the half of the full training set.

In order to compare the overall performance of the resulting classifier for different choices of the loss function we performed a *nested cross validation* (cf. [17, 29]), too. In this way one obtains an unbiased estimate of the true classification error for each model. More precisely, we implemented a so-called *two nested 10-fold cross validation*, i. e. for an outer loop the whole set of images was split into ten disjoint sets used as test sets to obtain the classification error. For each test set the remaining data again was split into ten disjoint sets used in the inner loop. Then, via the 10-fold cross validation approach described above the best values for the optimal kernel parameter and the regularization parameter observed in the experiment were chosen. Once these

loss function	hinge loss	quadratic hinge loss	cubic hinge loss
average test error	0.050	0.041	0.046

Table 4.2: The average misclassification rate obtained via the two nested 10-fold cross validation for each loss function.

parameters are determined, they were used for training a classifier on the whole set and for testing it on the remaining test set. As result we got the average test error over ten test sets for each choice of the loss function, which provided a characteristic for comparing the performance of these classifiers. The results are shown in Table 4.2 and indicate that the quadratic hinge loss function works best for this image classification task in the sense that it yields the lowest misclassification rates.

5 Conclusions

This paper aimed at solving an image classification task involving high dimensional real-world data. For this purpose, starting with the general Tikhonov regularization problem, reformulated as a convex optimization problem, we introduced a Fenchel-type dual problem for it and proved the existence of strong duality. This gave us the possibility to express the decision function for the classification problem via the dual optimal solutions. For three particular loss functions the corresponding dual problems have been calculated and for each of them numerical tests have been performed. The obtained results confirm the applicability of the support vector technique for classification tasks based on real-world data.

Acknowledgements

References

[1] Aronszajn N., Theory of reproducing kernels, T. Am. Math. Soc., 1950, 686, 337–404

[2] Boţ R.I., Conjugate Duality in Convex Optimization, Lecture Notes in Econom. and Math. Systems, 637, Springer, Berlin Heidelberg, 2010

[3] Boţ R.I., Csetnek E.R., Heinrich A., On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems, 2012, arXiv:1303.2875 [math.OC]

[4] Boţ R.I., Lorenz N., Optimization problems in statistical learning: Duality and optimality conditions, Eur. J. Oper. Res., 2011, 213(2), 395–404

[5] Boyd S., Parikh N., Chu E., Peleato B., Eckstein J., Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn., 2010, 3(1), 1–122

[6] Chapelle O., Haffner P., Vapnik V.N., Support vector machines for histogramm-based image classification, IEEE T. Neural Networ., 1999, 10, 1055–1064

[7] Geiger C., Kanzow, C., Theorie und Numerik restringierter Optimierungsaufgaben, Springer, Berlin Heidelberg New York, 2002

[8] Joachims T., Learning to Classify Text using Support Vector Machines: Methods, Theory and Algorithms, Kluwer Academic Publishers, Boston Dordrecht London, 2002

[9] Kim K., Financial time series forecasting using support vector machines, Neurocomputing, 2003, 55(1-2), 307–319

[10] Lal T.N., Chapelle O., Schölkopf B., Combining a filter method with SVMs. In: I. Guyon, S. Gunn, M. Nikravesh and L. A. Zadeh (Eds.), Feature Extraction: Foundations and Applications, 2006, Springer, Berlin Heidelberg, 439–445

[11] MATLAB 8.0, The MathWorks Inc., 2012, Natick, Natick, Massachusetts

[12] Mota J.F.C., Xavier J.M.F., Aguiar P.M.Q., Püschel M., D-ADMM: A communication-efficient distributed algorithm for separable optimization, 2012, arXiv:1202.2805 [math.OC]

[13] Noble W.S., Support vector machine application in computational biology, In: B. Schölkopf, K. Tsuda, J.-P. Vert (Eds.), Kernel Methods in Computational Biology, 2004, The MIT Press, 71–92

[14] Pedroso J.P., Murata N., Support vector machines with different norms: motivation, formulations and results, Patter Recogn. Lett., 2001, 22(12), 1263–1272

[15] Rifkin R.M., Lippert R.A., Value regularization and Fenchel duality, J. Mach. Learn. Res., 2007, 8, 441–479

[16] Rockafellar R.T., Convex Analysis, 1970, Princeton University Press

[17] Ruschhaupt M., Huber W., Poustka A., Mansmann U., A compendium to ensure computational reproducibility in high-dimensional classification tasks, Stat. Appl. Gen. Mo. B., 2004, 3(7), Article 37

[18] Shawe-Taylor J., Cristianini N., Kernel Methods for Pattern Analysis, 2004, Cambridge University Press

[19] Sra S., Nowozin S., Wright S.J., Optimization for Machine Learning, 2011, The MIT Press

[20] Steinwart I., How to compare different loss functions and their risks, Constr. Approx., 2007, 26(2), 225–287

[21] Stoppiglia H., Dreyfus G., Dubois R., Oussar Y., Ranking a random feature for variable and feature selection, J. Mach. Learn. Res., 2003, 3, 1399–1414

[22] Suykens J.A.K., Vanderwalle J., Least squares support vector machines classifiers, Neural Process. Lett., 1999, 9(3), 293–300

[23] Tibshirani R., Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B, 1996, 58(1), 267–288

[24] Tibshirani R., Saunders M., Rosset S., Zhu J., Kneight K., Sparsity and smoothness via the fused lasso, J. R. Stat. Soc. Ser. B, 2005, 67(1), 91–108

[25] Tikhonov A.N., Arsenin V.Y., Solutions of Ill-posed Problems, 1977, W.H. Winston, Washington D.C.

[26] Van Gestel T., Baesens B., Garcia J., Van Dijcke P., A support vector machine approach to credit scoring, Bank -en Finan., 2003, 2, 73–82

[27] Vapnik V.N., The Nature of Statistical Learning Theory, 1995, Springer, New York

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

[29] Varma S., Simon R., Bias in error estimation when using cross-validation for model selection, BMC Bioinformatics, 2006, 7, Article 91

[30] Wahba G., Spline Models for Observational Data, 1990, Series in Applied Mathematics, 59, SIAM, Philadelphia

[31] Xiang D.-H., Zhou D.-X., Classification with Gaussians and convex loss, J. Mach. Learn. Res., 2009, 10, 1447–1468

[32] Ying Y., Huang K., Campbell C., Sparse metric learning via smooth optimization, In: Bengio Y., Schuurmans D., Lafferty J., Williams C., Culotta A., Advances in Neural Information Processing Systems 22, 2009, The NIPS Foundation, 2205–2213