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

Radu Ioan Boţ * André Heinrich † Gert Wanka ‡

July 20, 2011

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 the 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 special 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. [4,6]), computational biology (cf. [9]) or time series forecasting and credit scoring (cf. [7, 13]) 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.

*Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany, e-mail: radu.bot@mathematik.tu-chemnitz.de.

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

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

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

To this end we reformulate the general Tikhonov regularization problem (cf. [12]), 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 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 opportunity 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* $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ and corresponding labels $y_i \in \{-1, +1\}$, $i = 1, \dots, n$, 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 supervised learning methods were investigated in detail by Vapnik in [14]. Considering $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^d \times \{+1, -1\}$ 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 $i = 1, ..., n$. 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. [12]), 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 regular-*

ization 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 [3]). 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 definite*. 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 $\sum_{i,j=1}^{m}a_ia_jk(x_i,x_j)>0$ for every arbitrary $a\in\mathbb{R}^d\setminus\{0\}$, is called *finitely positive definite* (cf. [11]).

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. [16]) 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

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

For $i=1,...,n$ 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 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 definite. Taking $c\in\mathbb{R}^n$ to be the vector corresponding to representation (2), the smoothness functional becomes $\Omega(f)=\|f\|_k^2=c^TKc$ and for $i=1,...,n$ it holds $f(x_i)=\sum_{j=1}^{n}c_jK_{ij}=(Kc)_i$. Thus we can rewrite optimization problem (1) equivalently as

$$(P_{gen})\qquad \inf_{c\in\mathbb{R}^n}\left\{C\sum_{i=1}^{n}v((Kc)_i,y_i)+\frac{1}{2}c^TKc\right\}. \tag{3}$$

Due to the nature of the loss function, this problem is mainly a convex and not necessarily differentiable optimization problem. 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.

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, where the upper index ^{T} transposes a column vector into a row one and viceversa. 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

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

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

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

For $i = 1, ..., n$ we denote by $\mathrm{Pr}_i : \mathbb{R}^n \to \mathbb{R}$ the *projection function* defined as $\mathrm{Pr}_i(x_1, ..., x_n) = x_i$. 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. [3]) and it is formulated as

$$(D_{gen}) \qquad \sup_{\substack{P \in \mathbb{R}^n, \\ P = (P_1, ..., P_n)^T}} \left\{ -C \sum_{i=1}^{n} \Big(v(\cdot, y_i)\Big)^* \Big(- \frac{P_i}{C}\Big) - \frac{1}{2} P^T K P \right\}. \tag{4}$$

Let us denote by $v(P_{gen})$ the optimal objective value of the primal problem (P_{gen}) and by $v(D_{gen})$ 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. $v(P_{gen}) \geq v(D_{gen})$.*

Proof. Let be $c \in \mathbb{R}^n$ and $P = (P_1, \ldots, P_n)^T \in \mathbb{R}^n$. Then it holds, according to Young-Fenchel inequality and due to the positive definiteness of K, that

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

and therefore

$$C\sum_{i=1}^{n} v\big((Kc)_i, y_i\big) + \frac{1}{2}c^T Kc \geq -C\sum_{i=1}^{n} \big(v(\cdot, y_i)\big)^*\Big(-\frac{P_i}{C}\Big) - \frac{1}{2}P^T KP,$$

i. e. $v(P_{gen}) \geq v(D_{gen})$. □

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

$$(P_{gen}) \qquad \inf_{c\in\mathbb{R}^n}\left\{C\sum_{i=1}^{n} v_i(Kc) + \frac{1}{2}c^T Kc\right\}. \tag{5}$$

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

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

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

Proof. We notice first that

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

Let be $c' \in \mathbb{R}$ such that $c' \in \cap_{i=1}^{n} \operatorname{ri}(\operatorname{dom} v(\cdot, y_i))$, which means that for all $i = 1, ..., n$ it holds $(c', \ldots, c')^T \in (\operatorname{Pr}_i)^{-1}(\operatorname{ri}(\operatorname{dom} v(\cdot, y_i))) \neq \emptyset$. Thus, according to [10, Theorem 6.7], one has that $\operatorname{ri}(\operatorname{dom} v_i) = \operatorname{ri}\left((\operatorname{Pr}_i)^{-1}(\operatorname{dom} v(\cdot, y_i))\right) = (\operatorname{Pr}_i)^{-1}\big(\operatorname{ri}(\operatorname{dom} v(\cdot, y_i))\big)$ for all $i = 1, ..., n$, hence

$$(c', \ldots, c')^T \in \bigcap_{i=1}^{n} \operatorname{ri}(\operatorname{dom} v_i) = \operatorname{ri}\left(\bigcap_{i=1}^{n} \operatorname{dom} v_i\right). \tag{6}$$

This means that $v(P_{gen}) < +\infty$.

Since the conclusion is obvious if $v(P_{gen}) = -\infty$ due to Theorem 1, we assume in the following that $v(P_{gen}) \in \mathbb{R}$. Denoting by $g : \mathbb{R}^n \to \mathbb{R}$, $g(c) = \frac{1}{2}c^T Kc$, and by taking into consideration that K is non-singular, we have

$$K(\mathrm{ri}(\mathrm{dom}\, g)) \cap \mathrm{ri}\left(\bigcap_{i=1}^{n} \mathrm{dom}\, v_i\right) = K(\mathbb{R}^n) \cap \mathrm{ri}\left(\bigcap_{i=1}^{n} \mathrm{dom}\, v_i\right) = \mathrm{ri}\left(\bigcap_{i=1}^{n} \mathrm{dom}\, v_i\right) \neq \emptyset.$$

We then have (see [2, Theorem 2.1]) that there exists a $\bar{P} \in \mathbb{R}^n$ such that

$$\begin{aligned} v(P_{gen}) &= \sup_{P \in \mathbb{R}^n} \Big\{ -\Big(\sum_{i=1}^{n} Cv_i\Big)^*(-P) - g^*(KP) \Big\} \\ &= -\Big(C\sum_{i=1}^{n} v_i\Big)^*(-\bar{P}) - \frac{1}{2}(-K\bar{P})^T K^{-1}(K\bar{P}) \\ &= -C\left(\sum_{i=1}^{n} v_i\right)^*\left(-\frac{1}{C}\bar{P}\right) - \frac{1}{2}\bar{P}^T K\bar{P}. \end{aligned}$$

As from (6) $\cap_{i=1}^{n} \mathrm{ri}(\mathrm{dom}\, v_i) \neq \emptyset$, it follows (cf. [10]) that there exist $\bar{P}^i \in \mathbb{R}^n$, $i = 1, ..., n$, 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,

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

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

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

Since the optimal objective value of (P_{gen}) is finite, by defining $\bar{P}_i := \bar{P}_i^i$ for $i = 1, ..., n$, one has $\sum_{i=1}^{n} \bar{P}^i = (\bar{P}_1, \ldots, \bar{P}_n)^T \in \mathbb{R}^n$ and

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

where $\bar{P} := (\bar{P}_1, \ldots, \bar{P}_n)^T$. This, along with the weak duality theorem, provides the desired result, $\bar{P}$ being an optimal solution to (D_{gen}). □

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. Then $\bar{c} \in \mathbb{R}^n$ 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*) $\bar{c} = \bar{P}$.

Proof. From Theorem 2 we get the existence of an optimal solution $\bar{P} \in \mathbb{R}^n$ to (D_{gen}) such that

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

This is equivalent to

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

and further to, since K is positive definite,

$$\begin{cases} -\frac{\bar{P}_i}{C} \in \partial v(\cdot, y_i)((K\bar{c})_i), \quad i = 1, \ldots, n, \\ \bar{c} = \bar{P}. \end{cases}$$

The opposite direction follows analogously. □

Remark 1. Since K is positive definite, the function g is strongly convex (on $\mathbb{R}^n$). Consequently, if $\cap_{i=1}^n \operatorname{dom} v(\cdot, y_i) \neq \emptyset$ and $v(\cdot, y_i)$, $i = 1, \ldots, n$, are, additionally, lower semicontinuous, the optimization problem (P_{gen}) has a *unique optimal solution* (see, for instance, [5, Satz 6.33]). Further, due to the fact that $P \mapsto \frac{1}{2}P^T KP$ is strictly convex (on $\mathbb{R}^n$), one can see that the dual problem (D_{gen}) has at most one optimal solution.

Remark 2. Due to Remark 1 and Theorem 3, if (QC) is fulfilled and $v(\cdot, y_i)$, $i = 1, \ldots, n$, are lower semicontinuous, it follows that, 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

$$v_{hl}(a, y) = (1 - ay)_+, \tag{7}$$

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

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

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

$$\begin{aligned}
-\big(v_{hl}(\cdot,y_i)\big)^*(z) &= -\sup_{a\in\mathbb{R}}\{za-(1-ay_i)_+\} = \inf_{\substack{a,t\in\mathbb{R},\\ t\geq 0,\, t\geq 1-ay_i}}\{-za+t\}\\
&= \sup_{k\geq 0,\, r\geq 0}\left\{\inf_{a,t\in\mathbb{R}}\{-za+t+k(1-ay_i-t)-rt\}\right\}\\
&= \sup_{k\geq 0,\, r\geq 0}\left\{\inf_{a\in\mathbb{R}}\{-za-kay_i\}+\inf_{t\in\mathbb{R}}\{t-kt-rt\}+k\right\}\\
&= \sup_{\substack{k\geq 0,\, r\geq 0,\\ k+r=1,\\ z+ky_i=0}} k = \sup_{\substack{k\in[0,1],\\ k=-zy_i}} k = \begin{cases} -zy_i, & \text{if } zy_i\in[-1,0],\\ -\infty, & \text{otherwise.}\end{cases}
\end{aligned}$$

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

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

or, equivalently,

$$(D_{hl}) \qquad \inf_{\substack{P\in\mathbb{R}^n,\\ P_i y_i\in[0,C],\, i=1,\ldots,n}}\left\{\frac{1}{2}P^T KP - \sum_{i=1}^{n}P_i y_i\right\}. \tag{8}$$

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* $v^u_{ghl} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$v^u_{ghl}(a, y) = (1 - ay)^u_+, \tag{9}$$

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

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

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

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

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 $v^u_{ghl}(\cdot, y_i)$ for $i = 1, ..., n$ and $z \in \mathbb{R}$

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

Hence, the corresponding dual problem to (P^u_{ghl}) looks like

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

Formulated as an infimum problem, (D^u_{ghl}) becomes

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

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

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

This problem give rise for $u = 2$ to

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

and for $u = 3$ to

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

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.

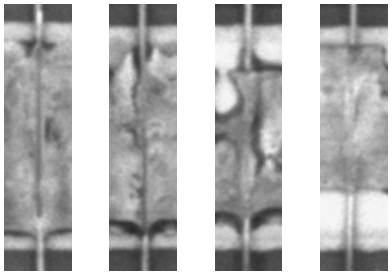
4 Application to image classification

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.

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.



(a) (b) (c) (d)

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

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 m_t representing image t, i. e.

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

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 [8], the data has been normalized by dividing each data point by the quantity $(\frac{1}{4740}\sum_{t=1}^{4740}\|m_t\|^2)^{\frac{1}{2}}$, due to numerical reasons. 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 dimension 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* 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^-. 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.

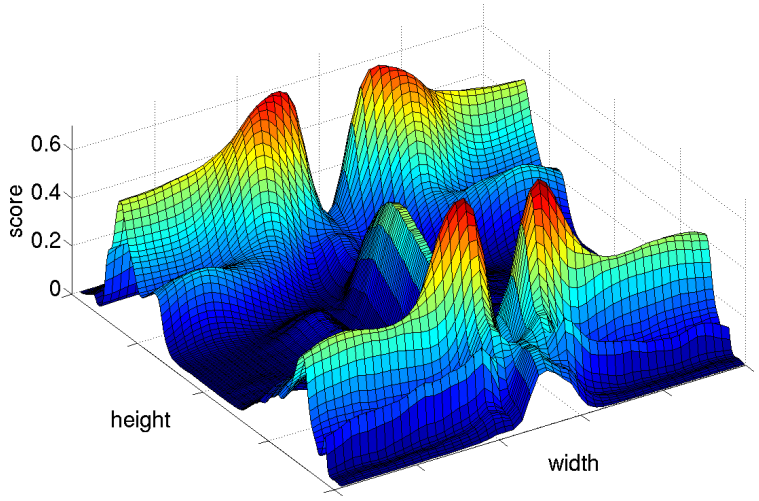


Figure 4.2: Visualization of the scores of the pixels.

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* $v^u_{ghl}(a, y) = (1 - ay)^u_+$ for $u = 2$ and $u = 3$ and the corresponding three dual problems (D_{hl}), (D^2_{ghl}) and, respectively, (D^3_{ghl}) were used. As kernel function the *Gaussian kernel function*

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

was chosen. Since the regularization parameter C and the kernel parameter σ were unknown and had to be determined by the user, 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.

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*

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.

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.

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 test have been performed. The obtained results reveal the applicability of the support vector technique for classification tasks based on real-world data.

Acknowledgements

We would like to thank Continental Automotive GmbH in Limbach-Oberfrohna for providing the data for this challenging classification problem.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society* 686:337–404, 1950.

[2] R.I. Boţ. *Conjugate Duality in Convex Optimization*. Lecture Notes in Economics and Mathematical Systems, Vol. 637, Springer-Verlag, Berlin Heidelberg, 2010.

[3] R.I. Boţ and N. Lorenz. Optimization problems in statistical learning: Duality and optimality conditions. *European Journal of Operational Research* 213(2):395–404, 2011.

[4] O. Chapelle, P. Haffner and V.N. Vapnik. Support vector machines for histogramm-based image classification. *IEEE Transactions on Neural Networks* 10:1055–1064, 1999.

[5] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer-Verlag, Berlin Heidelberg New York, 2002.

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

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

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

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

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

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

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

[13] T. Van Gestel, B. Baesens, J. Garcia and P. Van Dijcke. A support vector machine approach to credit scoring. *Bank en Financiewezen* 2:73–82, 2003.

[14] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

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

[16] G. Wahba. *Spline Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, 1990.