European Journal of Operational Research 213 (2011) 395–404

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

Optimization problems in statistical learning: Duality and optimality conditions ☆

Radu Ioan Boţ *, Nicole Lorenz

Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany

ARTICLE INFO

Article history:
Received 19 March 2009
Accepted 9 March 2011
Available online 16 March 2011

Keywords:
Machine learning
Tikhonov regularization
Convex duality theory
Optimality conditions

ABSTRACT

Regularization methods are techniques for learning functions from given data. We consider regularization problems the objective function of which consisting of a cost function and a regularization term with the aim of selecting a prediction function f with a finite representation $f(\cdot)=\sum_{i=1}^{n}c_i k(\cdot,X_i)$ which minimizes the error of prediction. Here the role of the regularizer is to avoid overfitting. In general these are convex optimization problems with not necessarily differentiable objective functions. Thus in order to provide optimality conditions for this class of problems one needs to appeal on some specific techniques from the convex analysis. In this paper we provide a general approach for deriving necessary and sufficient optimality conditions for the regularized problem via the so-called conjugate duality theory. Afterwards we employ the obtained results to the Support Vector Machines problem and Support Vector Regression problem formulated for different cost functions.

1. Some elements of statistical learning

Support Vector Machines are techniques for solving problems of learning from a given example data set based on the *Structural Risk Minimization Principle* and they were first mentioned by Vapnik in [22]. The reader is also referred to [21,23] for a deeper insight into this field.

Evgeniou, Pontil and Poggio distinguish in [8] between two types of statistical learning problems: the *Support Vector Machines Regression* problem (SVMR) and the *Regularization Networks* (RN). The problems belonging to the first class have as possible application the approximation and determination of a function by means of a data set. We deal here with a particular case of this problem, the so-called *Support Vector Machines Classification* (SVMC).

Consider a given set with n training data $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, where $X_i\in\mathbb{R}^k$ and $Y_i\in\mathbb{R}, i=1,\ldots,n$, and let $\mathfrak{F}$ be a space of functions defined on $\mathbb{R}^k$ with real values. The SVMC problem looks for a function $f\in\mathfrak{F}$ such that for a previously unknown value X the function f predicts the value Y. The penalty for predicting $f(X_i)$ having as true value Y_i for $i=1,\ldots,n$ is measured by a so-called *cost function* $v:\mathbb{R}^2\to\overline{\mathbb{R}}$.

The problem of finding an optimal function f in $\mathfrak{F}$ is *ill-posed* since there are infinitely many solutions. In order to get a *well-posed* problem, and, consequently, to be able to provide a particular solution, we need some additional *a priori* information about f. A common one is the assumption that the function f is *smooth*, in other words, two similar inputs correspond to two similar outputs. In this way one is able to control the complexity of f. To this aim one has to introduce a *regularization term* $\frac{\lambda}{2}\Omega(f)$ (cf. [2,3,20]), where the *regularization parameter* $\lambda>0$ controls the tradeoff between the cost function and the regularizer Ω (cf. [25]). In this context Ω is also called *smoothness functional* and has the desired characteristic of taking high values for non-smooth functions and low values for smooth functions. The following *Tikhonov regularization problem* arises

$$\inf_{f\in\mathfrak{F}}\left\{\sum_{i=1}^{n}v(f(X_i),Y_i)+\frac{\lambda}{2}\Omega(f)\right\}, \tag{1}$$

the objective function of which being called *regularization functional*.

Further let $\mathfrak{H}_k$ be a *Reproducing Kernel Hilbert Space* (RKHS) introduced by a *kernel function* $k:\mathbb{R}^{k\times k}\to\mathbb{R}$ (cf. [1]). In the following we ask f to be an element of $\mathfrak{H}_k$. Moreover, we assume that k is *symmetric*, namely that $k(x,y)=k(y,x)$ for $x,y\in\mathbb{R}^k$. The kernel function k introduces a *kernel matrix* $K\in\mathbb{R}^{n\times n}$, where $k(X_i,X_j)=K_{ij}$ for $i,j=1,\ldots,n$. In this context K, which is a symmetric matrix, is said to be the *Gram matrix of k with respect to $X_1,\ldots,X_n$*. A symmetric kernel function $k:\mathbb{R}^{k\times k}\to\mathbb{R}$ which for all $n\geqslant 1$ and all finite sets $\{X_1,\ldots,X_n\}\subset\mathbb{R}^k$ fulfills

☆ Research partially supported by DFG (German Research Foundation), project WA 922/1–3.
* Corresponding author. Tel.: +49 37153134463.
E-mail addresses: radu.bot@mathematik.tu-chemnitz.de (R.I. Boţ), nicole.lorenz@mathematik.tu-chemnitz.de (N. Lorenz).

0377-2217/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2011.03.021

$\sum_{i,j=1}^{n} a_i a_j k(X_i, X_j) \geqslant 0$ for every arbitrary $a \in \mathbb{R}^n$ is called *finitely positive semidefinite kernel* (cf. [19]). One can easily see that such a kernel function gives rise to a positive semidefinite Gram matrix K. On the other hand, it is worth noticing that (see [19, Theorem 3.11]) a k which is either continuous or has a finite domain can be decomposed as $k(x,y) = \langle \Phi(x), \Phi(y) \rangle$, where $\Phi : \mathbb{R}^k \to F$ is a *feature map* and F a Hilbert space, if and only if it is finitely positive semidefinite.

It is well-known that when having a symmetric finitely positive definite kernel k and a corresponding Gram matrix one can find a RKHS $\mathfrak{H}_k$ induced by it, such that the so-called *reproducing property*, namely that $f(x) = \langle f(\cdot), k(x,\cdot) \rangle$ for all $x \in \mathbb{R}^k$, is fulfilled (cf. [1]). Shawe-Taylor and Cristianini have shown in [19] that one can construct a RKHS $\mathfrak{H}_k$ even for a symmetric finitely positive semidefinite kernel function such that the reproducing property is valid. More than that, via the so-called *representer theorem* (cf. [25]) one has that for every minimizer f of (1) there exists $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ such that

$$f = \sum_{j=1}^{n} c_j k(\cdot, X_j). \tag{2}$$

This is the setting considered in this paper and in the following we additionally assume that for $f \in \mathfrak{H}_k$ the smoothness functional is defined as $\Omega(f) = \|f\|_k^2$, where $\|\cdot\|_k$ is the norm in $\mathfrak{H}_k$. If for $f \in \mathfrak{H}_k$ the vector $c \in \mathbb{R}^n$ is the one that comes from the representation given in (2), then $\Omega(f) = \|f\|_k^2 = c^T K c$ and for all $i = 1, \ldots, n$ it holds $f(X_i) = \sum_{j=1}^{n} c_j K_{ij} = (Kc)_i$. Thus the optimization problem (1) can be equivalently written as

$$\inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} v((Kc)_i, Y_i) + \frac{\lambda}{2} c^T K c \right\}. \tag{3}$$

Unfortunately, the most popular and most efficient cost functions used in the literature on machine learning fail to be differentiable (see, for instance, [8,16,18]). This causes some difficulties when trying to furnish optimality conditions for the above problem. On the other hand, these functions turn out to be convex in the first variable and, consequently, problem (3) becomes a convex optimization problem. In the following section we provide a general approach for deriving optimality condition for problem (3) by means of the conjugate duality theory in convex optimization. The optimality conditions for (3) will be expressed as systems of nonlinear equations involving the conjugates of the cost functions or, alternatively, via convex subdifferential formulae. As a byproduct we extend in this way the approach presented in [14], where when dealing with problem (3) the authors impose invertibility for K. We show that, in spite of the fact that we avoid this assumption, one can deliver handleable optimality conditions for (3), only by exploiting the very strong results of the convex analysis.

The described regularization framework includes many well-known learning methods. Depending on the application one can use different cost functions (see for instance [8,14] for several examples). In section 3 we consider some particular instances of the *Support Vector Machines Classification* problem, namely when the output Y takes values in $\{+1, -1\}$. In this case we speak about a *(binary) classification problem*. In particular we deal with the *hinge loss* (or *soft margin*) (cf. [7,22]) $v^{hl} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $v^{hl}(a, Y) = (1 - (a + b)Y)_+$, for $b \in \mathbb{R}$, but also with the *generalized hinge loss* (cf. [5]) $v^{ghl} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $v^{ghl}(a, Y) = (1 - (a + b)Y)_+^u$, where $u > 1$ is given.

In section 4 we turn our attention to the *Support Vector Regression* problem, which is characterized by the fact that the output Y may take arbitrary real values. In this context we deal with the following *extended loss* function $v^{el} : \mathbb{R} \times \mathbb{R} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, $v^{el}(a, Y) = \delta_{[-\varepsilon, \varepsilon]}(Y - a)$, where $\varepsilon > 0$, as well as with a *generalization of Vapnik's ε-insensitive loss* introduced by Smola, Schölkopf and Müller in [18], which we describe in detail in subsection 4.2. Especially by means of the extended loss we succeed in underlining the role of the regularity conditions when providing optimality conditions even in the context of machine learning. Obviously, via the general approach from section 2 one can consider also other cost functions suitable for the classification and regression problem.

It is worth to notice that in the investigations made in the sections 3 and 4 we take advantage of the convexity properties of cost functions involved. This fact allows us to employ the convex duality theory and to make use of the well-developed convex subdifferential calculus. On the other hand, this approach suggests the possibility to use nonsmooth and nonconvex cost functions in statistical learning. In order to provide optimality conditions for the optimization problems arising in this way, one could apply the calculus formulae which exist in the literature for different subdifferentials. In a first step one could consider locally Lipschitz cost functions in connection with the Clarke subdifferential (cf. [6]), but also some more general classes of functions in connection with some appropriate subdifferential notions, as one can find in [10].

The paper is closed by a conclusive section.

2. Notation and preliminary results

For two vectors $x, y \in \mathbb{R}^n$ we denote by $x^T y$ their *scalar product*, where the upper index ^{T} transposes a column vector into a row one and viceversa. By e_i, $i = 1, \ldots, n$, we denote the ith *unit-vector* in $\mathbb{R}^n$. For a nonempty set $D \subseteq \mathbb{R}^n$ we denote by $\delta_D : \mathbb{R}^n \to \overline{\mathbb{R}}$ the *indicator function* of D, which is defined by $\delta_D(x) = 0$ if $x \in D$, being equal to $+\infty$, otherwise. Further, by $\mathrm{ri}\,(D)$ we denote the *relative interior* of the set D, that is the interior of D relative to its affine hull. 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 $f^*(p) = \sup_{x \in \mathbb{R}^n} \{p^T x - f(x)\}$. We have the following relation, known as the *Young-Fenchel inequality*, $f(x) + f^*(p) - p^T x \geqslant 0$ and this is true for all $x, p \in \mathbb{R}^n$. For $x \in \mathbb{R}^n$ with $f(x) \in \mathbb{R}$ we denote by $\partial f(x) := \{p \in \mathbb{R}^n : f(y) - f(x) \geqslant p^T(y - x)\ \forall y \in \mathbb{R}^n\}$ 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

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

For a linear mapping $K : \mathbb{R}^n \to \mathbb{R}^m$ we denote by $\mathrm{Im}(K) := \{Kx : x \in \mathbb{R}^n\}$ the *image* of K. Further, for $x \in \mathbb{R}$ we define $x_+ := \max(0, x)$.

In order to develop a duality theory and to formulate necessary and sufficient optimality conditions for problem (3), we treat first, by means of some techniques from the convex analysis, the following optimization problem

$$(P) \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^{l} v_i(Kc) + g(c) \right\},$$