ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Convex Optimization for the Design of Learning Machines

K. Pelckmans, J.A.K. Suykens, B. De Moor

K.U. Leuven, ESAT-SCD-SISTA
Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium
Email: {kristiaan.pelckmans,johan.suykens}@esat.kuleuven.be

Abstract. This paper reviews the recent surge of interest in convex optimization in a context of pattern recognition and machine learning. The main thesis of this paper is that the design of task-specific learning machines is aided substantially by using a convex optimization solver as a back-end to implement the task, liberating the designer from the concern of designing and analyzing an ad hoc algorithm. The aim of this paper is twofold: (i) it phrases the contributions of this ESANN 2007 special session in a broader context, and (ii) it provides a road-map to published results in this context.

1 Introduction

Recently, techniques of Convex Optimization (CO) take a more prominent place in learning approaches, as pioneered by the work on Support Vector Machines (SVMs) and other regularization based learning schemes. Duality theory has played an important role in the development of so-called kernel machines, while the fact of uniqueness of the optimal solution has permitted theoretical as well as practical breakthroughs. A third main advantage of using CO tools in research on learning problems is that it permits fast prototyping of learning algorithms without the need for designing an appropriate training procedure explicitly. In this special session we discuss advances and new insights in this area. This paper restricts attention to the case of independently and identically distributed (i.i.d.) observations. Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ be i.i.d. samples of an unknown underlying and fixed joint distribution F_{xy}.

The special session 'Convex Optimization for the Design of Learning Machines' at ESANN 2007 is conceived much in the same spirit as [5]. This session contains 4 contributions. In [7], this overview paper is complemented with a review of recent advances in the use of SDP techniques in learning (especially results of [32, 8]). The paper [2] advances research in pattern recognition with data consisting of interval knowledge as described in [38], using an approach motivated from interval algorithmic. In [14], a fast approximation is described to perform leave-one-out crossvalidation for model selection using the kernel probit model. The authors of [56] reformulate the task of (kernel) canonical correlation analysis (kCCA) which is usually solved as an generalized eigenvalue problem as a convex SDP problem, working towards a link with the maximal margin classifier and SVMs.

The paper is organized as follows. Section 2 discusses the close relationship between estimation and optimization. Section 3 reviews new insights regarding the relation of convex optimization on the one hand, and SVMs and kernel machines on the other. Section 4 discusses the use of convex techniques for solving hard combinatorial problems. Section 5 gives insight and important pointers regarding the design of appropriate solvers.

2 Convex Optimization and Estimation

Convex optimization problems are defined in general as optimization problems in terms of a vector of unknowns $x \in \mathbb{R}^v$ which can be written as

$$\min_x f_0(x) \quad \text{s.t.} \quad \begin{cases} f_k(x) = 0 & \forall k = 1, \ldots, n_K \\ f_l(x) \leq 0 & \forall l = n_K + 1, \ldots, n_K + n_L, \end{cases} \tag{1}$$

where the cost function $f_0 : \mathbb{R}^v \to \mathbb{R}$ is a convex function The equality constraints $f_k : \mathbb{R}^v \to \mathbb{R}$ for all $k = 1, \ldots, n_K$ are linear in terms of x, and the inequality constraints $f_l : \mathbb{R}^v \to \mathbb{R}$ for all $l = n_K + 1, \ldots, n_K + n_L$ are convex ([11], Sect 4.2). Among the principal advantages of such convex optimization problems are (i) that the optimal solution takes place for a unique (or convex set of) vector(s) $\hat{x}$, and (ii) that the special structure can be exploited resulting in highly efficient solvers. Let '$Q = Q^T \succeq 0$' denote that $Q \in \mathbb{R}^{v \times v}$ is positive semi-definite symmetric matrix, i.e. $\forall x \in \mathbb{R}^v$ one has that $x^T Q x \geq 0$. Some common standard convex programs are defined, and for each one, specialized and highly efficient algorithms were established.

- If no inequalities occur and f_0 is convex and quadratical in its argument then the problem is referred to as a *least squares* problem (LS) (i.e. $n_L = 0$ and $\exists Q \succeq 0, Q \in \mathbb{R}^{v \times v}, p \in \mathbb{R}^v$ such that $f_0(x) = x^T Q x + p^T x + q$). Such a problem can be solved uniquely by solving a set of linear equations in case Q is positive definite.

- If Q is not positive semi-definite, the problem is unbounded. If $\{f_k\}_{1 \leq k \leq n_K}$ and $\{f_l\}_{n_K+1 \leq l \leq n_L}$ are linear in the argument x, the problem is a *linear programming problem* (LP).

- If both $\{f_k\}_{1 \leq k \leq n_K}$ and $\{f_l\}_{n_K < l \leq n_K + n_L}$ are linear in the argument x, and f_0 is quadratical (i.e. $f_0(x) = x^T Q x + p^T x$ with $Q \succeq 0$) the problem is a *quadratical programming problem* (QP). If Q is not positive semi-definite, the problem is in general NP hard.

- If the problem contains quadratical inequality constraints, i.e. $\exists R \in \mathbb{R}^{m \times v}, s, q \in \mathbb{R}^m$ and $t \in \mathbb{R}$ such that $f_l(x) = \|Rx - s\|_2 \leq q^T x + t$ for $n_K + 1 \leq i \leq n_K + n_L$, the problem is an instance of a *second order cone programming problem* (SOCP) [36].

- If the problem has a linear matrix inequality (LMI) in the form $X \succeq 0$ (where $X = X^T \in \mathbb{R}^{v \times v}$ denotes a matrix affinely depending on the unknowns) one has an instance of a *semidefinite programming problem* (SDP) [59].

There exist many other standard convex problems as geometric programming (see [10] and references), semi-infinite programming, stochastic and robust programming (see [1] and references), and parametric programming (see [26] and references). The task of checking whether a specific optimization problem is convex is in general a NP-hard problem. In [28], a principled approach is presented to assist in the conversion of a task to a convex problem by a method called disciplined programming, constituting of a theoretical road map and a software tool assisting in such task.

The point of view taken in machine learning and empirical (structural) risk minimization differs from the perspective of classical maximum likelihood techniques by centralizing the concept of prediction and generalization, rather than trying to recover the probabilistic mechanism underlying the data [60]. In the context of empirical risk minimization and statistical learning, convexity of the chosen loss function can be exploited to improve on the theoretical analysis, see e.g. [65].

3 Support Vector Machines and Kernel Machines

3.1 Support Vector Machine Classifiers

The advent of large margin classifiers as the Support Vector Machine boosted interest in the practice and theory of convex optimization in the context of pattern recognition and the learning methodology in general, see e.g. [60, 53]. In particular, the theory of Lagrange duality was nicely integrated with the device of Mercer kernels in order to translate linear techniques to a context of nonlinear estimation. For completeness, we review the by now classical derivation [18] as it introduces principal elements of this class of techniques. Let the class of possible outcomes of the technique (the hypothesis class) be defined as

$$\mathcal{H} = \{\text{sign}(w^T \varphi(x)), w \in \mathbb{R}^{d_\varphi}\}, \tag{2}$$

where $\varphi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\varphi}$ is an a-priori fixed (for the moment) mapping of the data to a feature space of dimension d_φ which can be potentially infinite. Remark that we omit the intercept term b for brevity of the discussion, despite its practical usefulness. The rule $\text{sign}(w^T x)$ obtaining the largest signed margin of a sample (x_i, y_i) to the hyperplane $\{x; w^T\varphi(x) + b = 0\}$ - formalized as $\frac{y_i(w^T\varphi(x_i))}{\|w\|_2}$ - is obtained by solving the following optimization problem

$$\max_w \min_{i=1,\ldots,n} \frac{y_i(w^T\varphi(x_i))}{\|w\|_2}. \tag{3}$$

This can be reformulated as a convex quadratic programming problem

$$\min_w \frac{1}{2} w^T w \quad \text{s.t.} \quad y_i(w^T\varphi(x_i)) \geq 1 \;\; \forall i = 1, \ldots, n. \tag{4}$$

Extension to the unfeasable case, i.e. the case where no such positive margin exists, is formulated as follows. Let $C \geq 0$ be a constant trade-off parameter.

$$\min_{w,e} \frac{1}{2} w^T w + C \sum_{i=1}^{n} e_i \quad \text{s.t.} \quad y_i(w^T\varphi(x_i)) \geq 1 - e_i, e_i \geq 0, \;\; \forall i = 1, \ldots, n, \tag{5}$$

where $\{e_i\}_{i=1}^n$ is a set of so-called slack variables. This formulation is known as the *primal formulation* of the support vector machine, its Lagrange dual problem provides additional insight into the problem formulation. Let the Lagrangian be $\mathcal{L}(\alpha, \beta, w, e) = \frac{1}{2} w^T w + C \sum_{i=1}^n e_i - \sum_{i=1}^n \alpha_i \left(y_i(w^T \varphi(x_i)) - 1 + e_i \right)$ $- \sum_{i=1}^n \beta_i e_i$ with positive multipliers $\{\alpha_i \geq 0\}_{i=1}^n$ and $\{\beta_i \geq 0\}_{i=1}^n$ Now, it is a classical result that '$\min_{w,e} \max_{\alpha_i \beta_i \geq 0} \mathcal{L}(\alpha, \beta, w, e)$', corresponds with the solution to the primal problem (5). By Slater's condition (see e.g. [11], Chapter 5 and references) this problem is equivalent to '$\max_{\alpha_i \beta_i \geq 0} \min_{w,e} \mathcal{L}(\alpha, \beta, w, e)$'. The first order condition for optimality $\partial \mathcal{L}(\alpha, \beta, w, e)/\partial w = 0$ gives the equality $w = \sum_{i=1}^n \alpha_i y_i \varphi(x_i)$; and from $\partial \mathcal{L}(\alpha, \beta, w, e)/\partial e_i = 0$ one has $C = \alpha_i + \beta_i$ for all $i = 1, \ldots, n$. Let $\mathbf{y} = (y_1, \ldots, y_n)^T \in \{-1, 1\}^n$ be a vector. This results in the dual problem after eliminating the primal variables w, e_i and the multipliers β_i:

$$\min_{\alpha} \frac{1}{2} \alpha^T (\Omega \circ \mathbf{y}\mathbf{y}^T) \alpha - \mathbf{y}^T \alpha \quad \text{s.t.} \quad 0 \leq \alpha_i \leq C \quad \forall i = 1, \ldots, n, \tag{6}$$

where $(\Omega \circ \mathbf{y}\mathbf{y}^T) \in \mathbb{R}^{n \times n}$ is defined as $(\Omega \circ \mathbf{y}\mathbf{y}^T)_{ij} = \varphi(x_i)^T \varphi(x_j) y_i y_j$. Now from the first order condition for optimality w.r.t. w, it turns out that one can evaluate the optimal estimate $\text{sign}(\hat{w}^T \varphi(x))$ as

$$\text{sign}(\hat{w}^T \varphi(x)) = \text{sign}\left(\sum_{i=1}^n \alpha_i y_i \varphi(x_i)^T \varphi(x) \right). \tag{7}$$

This reasoning results in additional insights in the structure and geometric meaning of the training algorithm, as e.g. characterization of sparseness of the result [60, 53], a geometrical interpretation in terms of the reduced convex hull as in [39], a sensitivity interpretation of the Lagrange multipliers [41], see also [26, 11] for a generic discussion. Observing that the optimum to (5) can be found and can be evaluated (7) entirely in terms of the dual variables α and the inner products $\varphi(x)^T \varphi(x')$, allows the application of the kernel trick. Therefor, one defines a suitable positive definite Mercer kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ which is guaranteed to represent an innerproduct of an appropriate feature map (or $\exists \varphi : \mathbb{R}^d \to \mathbb{R}^{d_\varphi}$ such that $K(x, x') = \varphi(x)^T \varphi(x')$). The key motivation is that this trick liberates one from the task to design explicitly an appropriate feature map. In general, linear methods which permit the kernel trick as above, result in an instance of a kernel method.

An important new direction in this research [29] is the explicit calculation of the set of solutions (usually the regularization trade-off parameter C) which are obtained by varying a design-parameter as in parametric programming (see e.g. [26]). This research on the so-called regularization path (instantiated in [22]) is a further step towards the integration of the design of learning algorithms and algorithms for convex optimization.

3.2 Variations on the Theme

In [55], a least squares reformulation of the SVM is elaborated (Least Squares Support Vector Machine classifier or LS-SVM classifier). The hypothesis class

$\mathcal{H} = \{\operatorname{sign}(w^T\varphi(x) + b), w \in \mathbb{R}^{d_\varphi}, b \in \mathbb{R}\}$ including an intercept term is considered. The optimality behind the LS-SVM is as follows

$$\min_{w,b,e} \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} e_i^2 \quad \text{s.t.} \quad y_i(w^T\varphi(x_i) + b) = 1 - e_i, \quad \forall i = 1, \ldots, n. \tag{8}$$

As previously, the Lagrange dual problem can be derived, and the criterion (8) can be optimized consequaently by solving the following linear system for the unknowns $\alpha = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

$$\left[\begin{array}{c|c} 0 & \mathbf{y}^T \\ \hline \mathbf{y} & (\Omega \circ \mathbf{y}\mathbf{y}^T) + I_N/\gamma \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 1_n \end{array}\right]. \tag{9}$$

Analogous to (7), the optimum to (8) can be evaluated using the solution to (9) can be evaluated in a point $x \in \mathbb{R}^d$ as $\operatorname{sign}\left(\sum_{i=1}^n \alpha_i y_i \varphi(x_i)^T \varphi(x) + \hat{b}\right)$ where $\hat{\alpha}$ and $\hat{b}$ solves (9). In analogy to the least squares approach underlying a variety of estimation problems, the above formulation was found to underly a broad range of kernel tasks as kernel PCA, kernel CCA, and lends itselves especially as a simple and flexible formulation for modeling situations with increasing complexity [55]. A range of formulations can be derived by choosing an appropriate convex function $L : \mathbb{R} \rightarrow \mathbb{R}$ as

$$\min_{w,b,e} \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} L(e_i) \quad \text{s.t.} \quad y_i(w^T\varphi(x_i) + b) = 1 - e_i, \quad \forall i = 1, \ldots, n. \tag{10}$$

The standard SVM can be recovered by choice of the Hinge function $L(e) = (1 - e)_+$ (with $(z)_+$ returning the positive part of its argument). The L_2-SVM as formulated in [60] takes $L(e) = (1 - e)_+^2$. This generic formulation in terms of L attracted especially interest in a context of reproducing kernel Hilbert spaces, see e.g. [24]. Discussion of a variety of convex choices for L was given in [65].

In the ν-SVM as introduced in [50, 52] and further discussed in [12], the following objective is proposed

$$\min_{w,b,\rho,e} \frac{1}{2} w^T w - \nu\rho + C \sum_{i=1}^{n} e_i \quad \text{s.t.} \quad y_i(w^T\varphi(x_i) + b) \geq \rho - e_i, e_i \geq 0, \rho \geq 0 \quad \forall i, \tag{11}$$

where from the conditions for optimality it follows that ν is an upper-bound to the fraction of training (margin) errors. Moreover, ν is a lower-bound on the number of support vectors (i.e. the number $|\{\alpha_i > 0\}_i|$). This mechanisms allow for a further integration of results in learning theory employing a compression argument with the practice of SVMs. In addition to the aforementioned formulations, there exists a range of different methods as kernel fisher discriminant analysis, and a class of methods based on (penalized) maximum likelihood. Kernel logistic regression [66] is amongst the most popular of those, and can be cast as a convex geometric programming problem [10], but as well the kernel probit model [14] is advantageous in many cases.

Different variations on the themes were proposed for handling multiclass classification tasks, where the output could take a different label $y \in \{1, \ldots, D\}$. A classical approach uses multi-class encoding schemes, but more direct formulations were also framed as a convex optimization problem, including [19]. Recently [63] rephrases the multiclass SVM as an SDP, allowing for a straightforward extension to the semi- or unsupervised case. Ordinal regression handles the case where the outputs $y \in \{1, \ldots, D\}$ have an ordering, but have no associated metric. This task were cast in the framework of SVMs as in [31, 17, 25]. Here the key issue is that any two sample points have an intrinsic order which is to be predicted accurately with the learning machine. In particular [31] reframes the problem as a regular SVM where a samples x is replaced by a couple of samples (x, x') with corresponding output y indicating whether x is preferential over x' or vice versa. Optimizing the related area under the ROC (Receiver Operator Characteristics) curve is described e.g. in [15].

Recent advances were made towards the formulation and analysis of learning machines for dealing with structured input- as well as structured output data. Structured input data (i.e. data which do not necessary have a natural embedding in an Euclidean space) are classically dealt with by the adoption of a proper kernel function (see e.g. [53] and references). Learning and prediction in the context of structured outputs is tackled in a variety of ways, see e.g. [48] A specific case is found in learning based on observational data with specific structure includes the cases where (a) data in the form of intervals is available [38, 2], (b) data is known to be perturbed [33] using robust programming [1], and (c) input observations contain missing values [42].

3.3 Regression

This subsection reviews a number of popular approaches to function approximation and regression based on the methodology of SVMs. The following formulation extends the method of SVMs to the regression case (SVR)[60]

$$\min_{w,b,e} \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{n} e_i \quad \text{s.t.} \quad -\epsilon - e_i \le w^T \varphi(x_i) + b - y_i \le \epsilon + e_i, e_i \ge 0 \ \forall i = 1, \ldots, n. \tag{12}$$

Arguably more natural is the use of a least squares loss, its extension towards ridge regression and smoothing splines, or a dual kernel ridge regression [49], and embodied in a context of LS-SVMs where L is the convex squared loss:

$$\min_{w,b,e} \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{n} L(e_i) \quad \text{s.t.} \quad w^T \varphi(x_i) + b - y_i = e_i, \quad \forall i = 1, \ldots, n \tag{13}$$

A robust version for handling outliers using Huber's loss function is described in [37]. A variation on the same theme is found in the estimation of linear and kernel based quantile regression as in [34], which targets the conditional quantile functions of a joint distribution, generalizing the conditional mean (as in regression).

3.4 Unsupervised and Semi-supervised Learning

While previous subsections discuss in general the case where one wants to learn a relation between two random variables (x, y) given a sample $\{(x_i, y_i)\}_{i=1}^n$ (the so-called supervised case), convex optimization techniques become also pervasive in the unsupervised case where the overall goal is to find 'structure' in the data $\{x_i\}_{i=1}^n$ without referring to a specific output variable. In [45], a method based on L_1 estimation and the regularization path was presented to cast methods as k-means and hierarchical clustering as a convex problem which can be solved using a QP.

One of the earliest extensions of the methodology of SVMs towards unsupervised learning was support vector clustering [4] and [62]. In the formulation of support vector data description (SVDD) [57] finding the minimum radius hypersphere enclosing the most part of the data, and in the one-class SVM [51], one tries to recover clusters in the data by separating the data in a feature space from the origin, see also [53]. In maximal margin clustering [63], one tackles the combinatorial problem of recovering a labeling of the data realizing a maximal margin using an SDP relaxation.

In [20], an SDP approach was given for detecting the structure in data much alike principal component, but additionally realizing sparseness in the result resulting in interpretability of the result as well as computational advantages ('sparse PCA'). In [56], a reformulation of CCA is given in terms of an SDP, and the relation with a maximal margin classifier as the SVM was given.

The task of semi-supervised learning [60] amounts to learning a classification rule where one aims at optimizing the generalization performance using unlabeled datasamples besides the given labeled training set. In general, the problem boils down to a combinatorial problem, but convex relaxations were devised using SDP programs as in [8], much in the same spirit as the seminal paper [27] devising an SDP relaxation for the NP hard MAXCUT problem with statistical guarantees on the approximation. In [63], an SDP approach to an unsupervised SVM formulation is discussed. In co-training and SVM-2k one extends this approach by exploiting the extra information available in the data as multiple views of the same unlabeled data reflect the desired classification. Transductive inference [60] amounts to finding an optimal classifier for the case one knows beforehand on which points the classifier is to be evaluated. It was conjectured [60] that this task is formally '*less complex*' than the inductive case were a classifier is to be learned for general future use. The relation of problems of transductive inference on graphs was related to the classical MAXFLOW-MINCUT algorithm (which is a special case of a linear program) [9]. A similar task is studied in [43] where a linear programming relaxation was used for transductive inference over a weighted graph.

3.5 Model Structures and Prior Knowledge

A fundamental advantage of the methology of converting learning tasks into standard convex optimization problems is that the incorporation of more refined model structures or structural prior knowledge can be introduced into an specific formulation straightforwardly. Additive models consisting of $d > 0$ components take the following form

$$f(x) = \sum_{l=1}^{d} w_l^T \varphi(x^l), \tag{14}$$

where x^l denote the lth feature of x. This model class is naturally approached from a RKHS perspective as in [61], or from an optimization context as in [41] where in both cases the relationship with the superposition of kernels is made explicit. From a statistical perspective, (generalized) additive models are discussed in [30]. Extensions were also described for learning semiparametric models taking the form $f(x) = w^T\varphi(x) + \sum_{l=1}^{L} b_l x^l$ with parameters $\{b_l\}_{l=1}^{L}$, see e.g. [54, 41]. The case were the additive noise in the regression case takes a prespecified coloring scheme was delt with in [23], and the relation of the coloring scheme with the design of the kernel was further made explicit in [41]. Other cases where the prior structural knowledge is imposed on the learning task is considered in [37] (for structural inequalities), and in [41] (for monotonicity constraints).

3.6 Input Selection and Model Selection

Problems of model selection constitute one of the major research challenges in the field of machine learning and pattern recognition. Except for the formulation of a proper model selection criterion, a main concern - at least in practice - is to find a good procedure for the selection of design-parameters (or hyper-parameters) optimally with respect to the model selection criterion. We start the discussion with the largely unsolved problem of input selection. Here one tries to find a subset of the features $\{1, \ldots, d\}$ which are significant for the task at hand. In the context of empirical risk minimization, one often resorts to the weaker problem of trying to find a (small) subset of inputs which make up a classifier which predicts (almost) as good as a black-box model. An interesting convex approach was found in the use of the convex L_1 norm resulting in sparseness amongst the optimal coefficients. This sparseness is then interpreted as an indication of non-relevant features. The LASSO [58, 22] was amonst the first to advocate this approach, but also the literature on basis pursuit [16] and compressed sensing [21] employs a similar strategy. In [6], a SOCP was emploid for the sake of feature selection in a linear model.

In the techniques for learning the kernel instantiated in [32] one translates the positive-definite restriction of an appropriate Mercer kernel in a positive definite constraint of the kernel matrix, resulting in an SDP. The problem of choosing a (subset of) a set of kernels appropriate for a learning task can be tackled in a variety of ways as e.g. described in [40] (giving a theoretical account), [3] (using SOCPs), and [41] (using a QP).

The task of evaluating fastly the model selection criterion of leave-one-out crossvalidation received considerable attention. It was shown that computation could be performed efficiently using the Sherman-Morrison-Woodbury formula, as in [13] (for LS-SVM regression). In [44], the authors propose an alternative scheme for parametrizing the regularization trade-off, resulting in a convex problem of setting this trade-off with respect to a model selection criterion as crossvalidation. In the follow-up paper [46], a convex approach towards setting the regularization parameter in schemes as ridge regression, smoothing splines and LS-SVMs is discussed.

4 Generic and Task Specific Solvers

To make the discussion complete, we provide some pointers to relevant literature giving insight in the problem of solving a convex optimization problem. Research on efficient solvers for generic convex optimization constitute a broad research area. Solver procedures for convex problems with inequality constraints can roughly be classified as primal-dual methods, interior-point methods, active set methods and many others, see e.g. [11] for a review. This research culminated in a set of highly efficient (commercial) general purpose software tools including CPLEX, MOSEK, LOQO or SeDuMi.

The success of kernel machines as SVMs incurred a surge of literature discussing routines for optimizing a convex optimization having a task-specific structure. In particular, the case of QPs with a quadratical objective function, a set of box-constraints and a single equality constraint solving an SVM is investigated thoroughly, resulting in fast and efficient routines. A major class constitute of the so-called decomposition techniques as instantiated in [47]. Those method exploit the sparseness in the QP much similar in spirit as the active set solvers, for a discussion see e.g. [35] who implemented the technique in one of the most popular software tools `LIBSVM`. Other software tools include `SVMlight` or `Torch`. Of importance is the numerical stability of the algorithms (in terms of generalization of the result) speed of convergence and memory requirements. A critical point in implementations concerns the choice of an appropriate stopping criterion, see e.g. [35]. Approaches for solving SVMs for huge sized datasets were described in [64], while advances for efficiently solving L_1 based formulations as the LASSO were described in [29]. In the case of LS-SVMs where the solution is given by a linear system, an accurate and numerical robust procedure was found in a conjugate gradient approach as described in [55]. This approach was implemented in the toolbox `LS-SVMlab`.

5 Conclusion

This paper[1] reviewed recent advances on the interplay between convex optimization and the design of learning machines, and focussed in particular on SVMs and kernel machines.

[1] **Acknowledgments**. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. Research supported by BOF PDM/05/161, FWO grant V 4.090.05N, IPSI Fraunhofer FgS, Darmstadt, Germany. (Research Council KUL): GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants; (Flemish Government): (FWO): PhD/postdoc grants, projects, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, G.0302.07. research communities (ICCoS, ANMMM, MLDM); (IWT): PhD Grants,GBOU (McKnow), Eureka-Flite2 - Belgian Federal Science Policy Office: IUAP P5/22,PODO-II,- EU: FP5-Quprodis; ERNSI; - Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard. JS is a professor and BDM is a full professor at K.U.Leuven Belgium.

References

[1] F. Alizadeh and D. Goldfarb. Second-order cone programming. *Math. Program.*, 95(1):3–51, 2003.

[2] C. Angulo, D. Anguita, and L. Gonzalez. Interval disciminant analysis using support vector machines. In *Proceedings of the fifteenth European Symposium on Artificial Neural Networks*, 2007.

[3] F.R. Bach, G.R.G. Lanckriet, and M.I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

[4] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137, 2001.

[5] K.P. Bennett and E. Parrado-Hernández. The interplay of optimization and machine learning research. *Journal of Machine Learning Research*, 7:1265–1281, 2006. (Special Topic on Machine Learning and Optimization).

[6] C. Bhattacharya. Second order cone programming formulations for feature selection. *Journal of Machine Learning Research*, 5:1417–1433, november 2004.

[7] T. De Bie. Deploying sdp for machine learning. In *Proceedings of the fifteenth European Symposium on Artificial Neural Networks*, 2007.

[8] T. De Bie and N. Cristianini. Fast SDP relaxations of graph cut clustering, transduction, and other combinatorial problems. *Journal of Machine Learning Research*, 7:1409–1436, 2006.

[9] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 19–26. Morgan Kaufmann Publishers, 2001.

[10] S. Boyd, S.J. Kim, and L. Vandenberghe. A tutorial on geometric programming. *Optimization and Engineering*, to appear, 2006.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[12] P.-H. Chen, C.-J. Lin, and B. Schölkopf. A tutorial on nu-support vector machines. *Applied Stochastic Models in Business and Industry*, 21:111–136, 2005.

[13] G. C. Cawley and N. L. C. Talbot. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17(10):1467–1475, 2004.

[14] G.C. Cawley. Model selection for kernel probit regression. In *Proceedings of the fifteenth European Symposium on Artificial Neural Networks*, 2007.

[15] S. Clemencon, G. Lugosi, N. Vayatis, P. Aurer, and R. Meir. Ranking and scoring usingf empirical risk minimization, *Proceedings of COLT 2005, in LNCS Computational Learning Theory*, vol. 3559, pp.1–15, Springer, 2005.

[16] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[17] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2006.

[18] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20(3):273–297, 1995.

[19] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[20] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 2007 (Accepted for publication).

[21] D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289 – 1306, 2006.

[22] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[23] M. Espinoza, J.A.K. Suykens, and B. De Moor. Ls-svm regression with autocorrelated errors. *in Proc. of the 14th IFAC Symposium on System Identification (SYSID)*, pages 582–587, 2006. Newcastle, Australia.

[24] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

[25] Glenn Fung, Romer Rosales, and Balaji Krishnapuram. Learning rankings via convex hull separation. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 395–402. MIT Press, Cambridge, MA, 2006.

[26] T. Gal and H.J. Greenberg, editors. *Advances in Sensitivity Analysis and Parametric Programming*. Management Science. Kluwer, 1998.

[27] M.X. Goemans and D.P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995.

[28] M.C. Grant. *Disciplined Convex Programming*. PhD thesis, Stanford, Electrical Engineering, Dec. 2004.

[29] T. Hastie, S. Rosset, and R. Tibshirani. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, October 2004.

[30] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.

[31] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000. MIT Press, Cambridge, MA.

[32] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, and M.I. Jordan L. El Ghaoui. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[33] G.R.G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M.I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.

[34] Y. Li, Y. Liu, and J. Zhu. Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*, 102(477):255–268, 2007.

[35] C.-J. Lin. A formal analysis of stopping criteria of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 13:1045–1052, 2002.

[36] M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second order programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

[37] O.L. Mangasarian and D.R. Musicant. Robust linear and support vector regression. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 22(9):950–955, 2000.

[38] O.L. Mangasarian, J.W. Shavlik, and E.W. Wild. Knowledge-based kernel approximation. *Journal of Machine Learning Research*, 5:1127–1141, September 2004.

[39] M.E. Mavroforakis and S. Theodoridis. A geometric approach to support vector machine (svm) classification. *IEEE Transactions on Neural Networks*, 17(3):671– 682, 2006.

[40] Charles A. Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

[41] K. Pelckmans. *Primal-Dual kernel Machines*. PhD thesis, Faculty of Engineering, K.U.Leuven, Leuven, may. 2005. 280 p., TR 05-95.

[42] K. Pelckmans, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18:684–692, 2005.

[43] K. Pelckmans, J. Shawe-Taylor, J.A.K. Suykens, and B. De Moor. Margin based transductive graph cuts using linear programming. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, to appear*, San Juan, Puerto Rico, 2007.

[44] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Additive regularization trade-off: Fusion of training and validation levels in kernel methods. *Machine Learning*, 62(3):217–252, 2006.

[45] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Convex clustering shrinkage. *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*, 2005. Windsor, UK.

[46] K. Pelckmans, J.A.K. Suykens, and B. De Moor. A convex approach to validation-based learning of the regularization constant. *Accepted for Publication in IEEE Transactions on Neural Networks*, 2006.

[47] J. Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel methods - Support Vector Learning*, pages 185–208, 2000. MIT Press, Cambridge, MA.

[48] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, 7:1601–1626, 2006. (Special Topic on Machine Learning and Optimization).

[49] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th Int. Conf. on Machine learning(ICML'98)*, pages 515–521. Morgan Kaufmann, 1998.

[50] B. Schölkopf, P.L. Bartlett, A.J. Smola, and R. Williamson. Shrinking the tube: A new support vector regression algorithm. In Micael I. Jordan, Michael J. Kearns, Sara Solla, and David A. Cohn, editors, *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge, MA, 1998.

[51] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, and A. Smola. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

[52] B. Schölkopf, A.J. Smola, R.C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

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

[54] A.J. Smola, T.T. Friess, and B. Schölkopf. Semiparametric support vector and linear programming machines. In Micael I. Jordan, Michael J. Kearns, Sara Solla, and David A. Cohn, editors, *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge, MA, 1998.

[55] J.A.K. Suykens, T. van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[56] S. Szedmak, T. De Bie, and D.R. Hardoon. A metamorphosis of canonical correlation analysis into multivariate maximum margin learning. In *Proceedings of the fifteenth European Symposium on Artificial Neural Networks*, 2007.

[57] D.M.J. Tax and R.P.W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20(11-13):1191–1199, 1999.

[58] R.J. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[59] L. Vandenberghe and S. Boyd. Semidefinite programming'. *SIAM Review*, 38:49–95, 1996.

[60] V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.

[61] G. Wahba. *Spline models for observational data*. SIAM, 1990.

[62] J. Weston, A. Gammerman, M.O. Stitson, V. Vapnik, V. Vovk, and C. Watkins. Support vector density estimation. *Advances in Kernel methods - Support Vector Learning*, pages 293 – 305, 2000. MIT Press, Cambridge, MA.

[63] L. Xu and D. Schuurmans. Unsupervised and semi-supervised multi-class support vector machines, *In National Conference on Artificial Intelligence (AAAI-05)* 2005.

[64] L. Zanni, T. Serafini, and G. Zanghirati. Parallel software for training large scale support vector machines on multiprocessor systems. *Journal of Machine Learning Research*, 7:1467–1492, 2006. (Special Topic on Machine Learning and Optimization).

[65] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annual Statistics*, 32:56–134, 2004.

[66] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *Journal of Computational & Graphical Statistics*, 14(1):185–205, 2005.