Advances in Kernel Methods

Support Vector Learning

edited by
Bernhard Schölkopf
Christopher J.C. Burges
Alexander J. Smola

The MIT Press
Cambridge, Massachusetts
London, England

Contents

19 Combining Support Vector and Mathematical Programming Methods for Classification

Kristin P. Bennett
Mathematical Sciences Department
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
bennek@rpi.edu
http://www.math.rpi.edu/~bennek

We examine the relationship between Support Vector Machines (SVM) for classification and a family of mathematical programming methods (MPM) primarily stemming from Mangasarian's Multisurface Method of Pattern Recognition. MPM and SVM share the same canonical form allowing the two approaches to be easily combined. We show how the dissimilarities of the MPM and SVM approaches have been used to generate two new methods for nonlinear discrimination: support vector decision trees and multicategory learning. Support vector decision trees are decision trees in which each decision is a support vector machine. Multicategory learning is an approach for handing classification problems with more than two classes. In computational studies, altering the original MPM to include principles of statistical learning theory almost always improved generalization. We also show how mathematical programming models and tools allowed us to develop rapidly a practical approach to solving a transduction problem using the theoretical principles of overall risk minimization. The basic of idea of transduction is to predict the class of a given set of unlabeled testing points without first estimating the classification function on the labeled training set. A semi-supervised SVM that includes both labeled training data and unlabeled test data is formulated as a mixed-integer program. Commercial optimization packages are used to solve moderately sized problems. Computational results indicate that the semi-supervised approach did improve generalization on many problems and never performed significantly worse than the baseline supervised SVM.

19.1 Introduction

In this chapter we investigate the relationship between Support Vector Machines (SVM) for classification and a family of mathematical programming methods primarily stemming from Mangasarian's *Multisurface Method of Pattern Recognition* (MSM) (Mangasiarian, 1965; Mangasarian, 1968). We focus on this family of mathematical programming methods (hereafter referred to as MPM) out of the many existing optimization-based classification methods in the literature because they are closely related to SVM. MPM and SVM were developed independently but they share the same canonical form. Thus a great potential exists for interaction between the two approaches. By combining statistical learning theory concepts and SVM ideas, such as kernels, with MPM, potentially many new SVM methods can be derived. Also, model formulation ideas from MPM can be used to develop more rapidly new algorithms based on statistical learning theory.

We begin with an overview of two optimization-based methods for classification: MSM and the *Robust Linear Programming* (RLP) method (Bennett and Mangasarian, 1992). Prior reviews cover how MPM are used for classification, clustering, and function approximation (Mangasarian, 1997; Bradley et al., 1998). Potential for integration of SVM and MPM exists in all these areas. In this chapter we will concentrate on the classification problem only. By starting with the linear classification case, we can see the common roots of MPM and SVM and where the methodologies branched in different directions. By examining dissimilarities, new opportunities for integrated approaches become apparent. Specifically in this chapter we will examine how MPM and SVM can be combined on two problems: nonlinear discrimination via decision trees and multicategory classification. Then we will illustrate how an idea from statistical learning theory on transduction, *Overall Risk Minimization* (Vapnik, 1979), can be quickly converted into a practical algorithm using ideas from MPM. These results are drawn primarily from existing work (Bennett et al., 1998; Bredensteiner and Bennett, 1998; Bennett and Demiriz, 1998). The primary goal of this chapter is to illustrate the current and potential integration of MPM and SVM by making the MPM work accessible in a common format and pointing to future possibilities.

Whenever possible we will use the notation in chapter 1 with a few exceptions. In the original SVM, the separating plane is defined as $\mathbf{w} \cdot \mathbf{x} + b = 0$. To make the notation consistent with MPM literature we will use $\mathbf{w} \cdot \mathbf{x} - \gamma = 0$. The forms are exactly equivalent with $\gamma = -b$. For clarity in some problem formulations, we must divide the training points into their respective classes. For $\Psi \geq 2$ classes, we will denote the classes as $A^i,\ i = 1, \ldots, \Psi$. For example, in the two-class case, the sets are defined as $A^1 := \{\mathbf{x}_i \mid i \in A^1\} := \{\mathbf{x}_i \mid (\mathbf{x}_i, y_i),\ y_i = 1,\ i = 1, \ldots, \ell\}$ and $A^2 := \{\mathbf{x}_i \mid i \in A^2\} := \{\mathbf{x}_i \mid (\mathbf{x}_i, y_i),\ y_i = -1,\ i = 1, \ldots, \ell\}$. The cardinality of the set A^i is denoted by $\ell_i = |A^i|$.

19.2 Two MPM Methods for Classification

Mangasarian's *Multisurface Method of Pattern Recognition* (Mangasiarian, 1965; Mangasarian, 1968) is very similar in derivation to the Generalized Portrait Method of Vapnik and Chervonenkis (1974). Mangasarian proposed finding a linear discriminant for linearly separable problems by solving the following optimization problem:

$$\begin{aligned} \max_{\mathbf{w},\alpha,\beta} \quad & \alpha - \beta \\ \text{subject to} \quad & \mathbf{w}\cdot\mathbf{x}_i \geq \alpha, \quad i \in A^1 \\ & \mathbf{w}\cdot\mathbf{x}_j \leq \beta, \quad j \in A^2 \\ & \|\mathbf{w}\| = 1 \end{aligned} \tag{19.1}$$

Multisurface Method

The problem constructs two parallel supporting planes, one supporting each class, and then maximizes the separation margin between the planes. The final optimal plane, $\mathbf{w}\cdot\mathbf{x} = \frac{(\alpha+\beta)}{2}$, is the same as that found by the Generalized Portrait Method. Problem (19.1) is difficult to solve as formulated since the constraint, $\|\mathbf{w}\| = 1$, is nonconvex. The SVM formulation (1.9) — (1.10) can be viewed as a transformation of (19.1) in which $\alpha - \beta = 2$ and and $\|\mathbf{w}\|_2$ is minimized. In MSM, Mangasarian proposed using the infinity-norm of $\mathbf{w}$, $\|\mathbf{w}\|_\infty = \max_{i=1,\ldots,n} |\mathbf{w}_i|$, instead of the 2-norm. Then by solving the $2N$ linear programs (LPs), the optimal solution can be found in polynomial time. In each LP one component w_d of the weight vector $\mathbf{w}$ is fixed to either 1 or -1 forcing the constraint $\|\mathbf{w}\|_\infty = 1$ to be satisfied. Thus the first $d = 1, \ldots, N$ linear programs are:

$$\begin{aligned} \max_{\mathbf{w},\alpha,\beta} \quad & \alpha - \beta \\ \text{subject to} \quad & \mathbf{w}\cdot\mathbf{x}_i \geq \alpha, \quad & i \in A^1 \\ & \mathbf{w}\cdot\mathbf{x}_j \leq \beta, \quad & j \in A^2 \\ & -1 \leq \mathbf{w}_i \leq 1, \quad & i = 1, \ldots, N \\ & \mathbf{w}_d = 1 \end{aligned} \tag{19.2}$$

The second set of N LPs consists of Problem (19.2) with $\mathbf{w}_d = -1$ replacing the constraint $\mathbf{w}_d = 1$. The solution to the LP with maximal objective value is the optimal solution of Problem (19.1) with $\|\mathbf{w}\|_\infty = 1$ (Mangasarian et al., 1990).

Unlike the Generalized Portrait Method, MSM also works for the linearly inseparable case. After training, the half-space $\mathbf{w}\cdot\mathbf{x} > \beta$ will contain only training points in A^1; the half-space $\mathbf{w}\cdot\mathbf{x} < \alpha$ will contain only training points in A^2; and the remaining margin may contain a mixture of points from both classes. By using Problem (19.2) recursively on points falling in the margin, MSM constructs a piecewise-linear discriminant function such as in figure 19.1.

Our interest in MSM is primarily historical, because of the similarities to SVM and because MSM set the pattern of how later MPM would address nonlinearly separable problems. While MSM was successfully used in an initial automated breast cancer diagnosis system at the University of Wisconsin-Madison (Wolberg

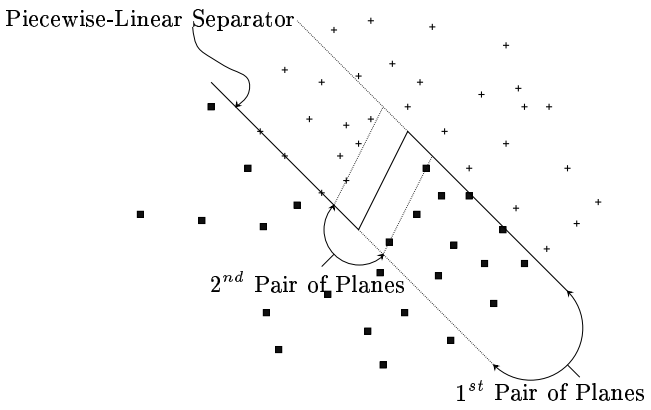

Figure 19.1 Piecewise-linear discriminant constructed by MSM

and Mangasarian, 1990; Mangasarian et al., 1990), the method performs poorly on noisy data sets since it minimizes the largest error in each class. The idea of maximizing the margin of separation was used in both the Generalized Portrait method and MSM, but different norms were used. For the nonlinear case, MSM was applied recursively to yield a piecewise-linear discriminant function. Although Mangasarian did observe that nonlinear discriminants could also be constructed by mapping the input attributes to a higher dimensional space, he made no mention of kernel-based methods.

To make MSM more tolerant of noise, Bennett and Mangasarian proposed the Robust Linear Programming method (RLP) (Bennett and Mangasarian, 1992) using the following linear program:

$$\begin{array}{ll} \min\limits_{\mathbf{w},\xi,\gamma} & \sum\limits_{i=1}^{l} \delta_i \xi_i \\ \text{subject to} & y_i(w \cdot \mathbf{x}_i - \gamma) + \xi_i \geq 1 \\ & \xi_i \geq 0 \quad i = 1, .., \ell \end{array} \tag{19.3}$$

Robust Linear Programming Method

where $\delta_i > 0$ is the fixed misclassification cost associated with point $\mathbf{x}_i$. The original RLP method used $\delta_i = \frac{1}{|A^1|}$ for points in A^1 and $\delta_i = \frac{1}{|A^2|}$ for points in A^2. These choices of δ_i ensure that the meaningless null solution $\mathbf{w} = 0$ is never the unique minimum of Problem (19.3). Smith (1968) proposed Problem (19.3) with $\delta_i = \frac{1}{\ell}$, but $w = 0$ may be the unique optimal solution of the Smith formulation.

Of course RLP is identical to the Soft Margin Hyperplane formulation (1.37) except for the absence of the capacity control term, $\|w\|$, which maximizes the margin of separation. If the 2-norm objective term, $\|w\|_2$, is added to RLP, the result is the original quadratic program for SVM (1.9) — (1.10). From Mangasarian and Meyer (1979) we know that there exists a constant $\bar{C}$ such that for any $C > \bar{C}$, the optimal solution of SVM (1.9) — (1.10) is also an optimal solution of RLP with $\delta_i = 1$. If the solution of RLP is not unique, then the SVM solution with

sufficiently large C will be the optimal solution of RLP with the least 2-norm of $\mathbf{w}$. If the 1-norm objective term, $\|\mathbf{w}\|_1$, is added, RLP can be generalized to construct a SVM variation with 1-norm capacity control:

$$\begin{array}{ll} \min\limits_{\mathbf{w},\gamma,s,\xi} & \lambda \|w\|_1 + (1-\lambda)\sum\limits_{i=1}^{\ell}\delta_i\xi_i \\ \text{subject to} & y_i[\mathbf{w}\cdot\mathbf{x}_i - \gamma] \geq 1-\xi_i \\ & \xi_i \geq 0, \quad i = 1,\ldots,\ell \end{array} \tag{19.4}$$

where $\delta_i > 0$ is a fixed misclassification cost associated with point $\mathbf{x}_i$ and $\lambda \in (0,1)$ is the relative weight on the margin maximization term. If λ is close to 1 more emphasis is placed on obtaining a large margin. If λ is close to 0 then the emphasis is on reducing the misclassification error. This problem is equivalent to following parametric linear program:

Primal RLP with capacity control

$$\begin{array}{ll} \min\limits_{\mathbf{w},\gamma,s,\xi} & \lambda\sum\limits_{j=1}^{N}s_j + (1-\lambda)\sum\limits_{i=1}^{\ell}\delta_i\xi_i \\ \text{subject to} & y_i[\mathbf{w}\cdot\mathbf{x}_i - \gamma] \geq 1-\xi_i \\ & \xi_i \geq 0, \quad i = 1,\ldots,\ell \\ & -s_j <= \mathbf{w}_j <= s_j, \quad j = 1,\ldots,N \end{array} \tag{19.5}$$

A commercial linear programming package such as CPLEX (CPL, 1994), based on simplex or interior point algorithms, can be used to solve very efficiently the dual RLP problem (Murthy, 1983):

Dual RLP with capacity control

$$\begin{array}{ll} \min\limits_{\alpha} & \sum\limits_{i=1}^{\ell}\alpha_i \\ \text{subject to} & -(1-\lambda)\mathbf{e} \leq \sum\limits_{i=1}^{\ell} y_i\alpha_i\mathbf{x}_i \leq (1-\lambda)\mathbf{e} \\ & \sum\limits_{i=1}^{\ell} y_i\alpha_i = 0 \\ & 0 \leq \alpha_i \leq \delta_i\lambda, \quad i = 1,\ldots,\ell \end{array} \tag{19.6}$$

where $\mathbf{e}$ is an N-dimensional vector of ones. The optimal $\mathbf{w}$ and γ are the Lagrangian multipliers of the constraints of Problem (19.6). Most linear programming packages provide the both the optimal primal and dual solutions.

RLP with 1-norm capacity control has been investigated in several papers (Bennett and Bredensteiner, 1998; Bredensteiner, 1997; Bradley and Mangasarian, 1998a; Bennett et al., 1998). Adding capacity control to RLP has been found empricially to improve generalization. Also there is no empirical evidence that either the 1-norm or 2-norm formulation produces superior generalization. It is an open question what the theoretical generalization differences are. In this chapter, we will refer to the 1-norm form as RLP and the 2-norm form as SVM. Other names,

such as Linear Programming Support Vector Machine, have used (Bradley and Mangasarian, 1998a). One benefit of SVM over RLP is that the kernels can easily be introduced into the dual problem in order to make nonlinear discriminants as discussed in chapter 1. But new approaches for incorporating kernels into linear programming based methods are being developed such as those in chapter 18. One major benefit of RLP over SVM is dimensionality reduction. Both RLP and SVM minimize the magnitude of the weights $\mathbf{w}$. But RLP forces more of the weights to be 0. This sparsity characteristic of the 1-norm compared to the 2-norm is also used in Basis Pursuit (S. S. Chen, 1996). A second benefit of RLP over SVM is that it can be solved using linear programming instead of quadratic programming. State-of-the-art general-purpose linear program solvers are more efficient, more robust, and capable of solving larger problems than are quadratic program solvers. If the original training data is sparse, the resulting LP formulation will be sparse and typical linear program solvers are constructed to exploit any sparsity. Even for sparse training data, the Hessian of the SVM quadratic program can become very dense. Dense quadratic programs are more difficult. The greater effectiveness of linear versus quadratic programming algorithms is definitely true for general-purpose solvers but optimization methods adapted to SVM problem structure such as the ones discussed in this book and in (Bradley and Mangasarian, 1998b) may help alleviate this difference. Other linear programming formulations such as those of Glover (1990) are also popularly used.

There are many extensions of the basic MSM and RLP methods. The papers (Mangasarian, 1997; Bradley et al., 1998; Bredensteiner, 1997) all contain interesting reviews. For example, two related problems are *feature selection*: constructing the best linear discriminant using the minimum number of attributes; and *misclassification minimization*: explicitly minimizing the number of points misclassified (Bradley and Mangasarian, 1998a; Bradley et al., 1995; Bredensteiner and Bennett, 1997; Bennett and Bredensteiner, 1997). Both problems require minimization of a metric that counts the number of nonzero components of a vector. Both are NP-Hard problems (Amaldi and Kann, 1998, 1995), but approximate answers may be found using nonconvex optimization techniques. Work on maximum feasible subsystems of linear relations can also be applied to these problems (Amaldi and Kann, 1998). These techniques are potentially applicable to SVM-related problems as well.

MPM and SVM have significantly differed in their approach to nonlinear discrimination and multicategory discrimination. For nonlinear discrimination, SVM perform linear discrimination in a higher-dimensional space using kernels to make the problem tractable. Starting with MSM, the primary MPM approach has been to use many linear discriminants to construct piecewise-linear discriminants via a decision tree. In section 19.3 we will investigate how SVM can be combined with the MPM-based decision tree algorithms. Note that there are some MPM that do perform nonlinear mappings into higher dimensional space, most notably the polynomial neural network approaches of Roy et al. (1993, 1995); Roy and Mukhopadhyay (1997). In section 19.4, we examine the SVM and MPM approaches to multi-

category discrimination. Multicategory discrimination is the problem of classifying points with more than two classes. The two approaches are combined to yield new methods.

19.3 Nonlinear Separation via Decision Trees

The primary MPM approach for nonlinear separation has been to construct piecewise-linear discriminant functions. These functions are decision trees. This approach can also be used with SVM. The original MSM can be viewed as producing a decision tree with specialized structure. RLP (19.3) has also been successfully used in decision tree algorithms (Bennett, 1992). Here we consider decision trees in which each decision is a support vector machine. Recent results on applying learning theory to decision trees show that there is a tradeoff between the structural complexity of a tree, i.e. the depth and number of nodes, and the complexity of the decisions that are used (Golea et al., 1998; Shawe-Taylor and Cristianini, 1998). We also know that for a given tree structure and empirical risk, decisions with larger margins should produce better generalization (Shawe-Taylor and Cristianini, 1998). So statistical learning theory suggests that using SVM in decision trees is a good idea for generalization. Another benefit is that the decision tree structure provides valuable information about a problem beyond class membership. The decision tree produces potentially interpretable rules, the attributes selected for the decisions indicate which attributes are important, and the leaf nodes cluster the data in potentially meaningful ways. For large data sets, trees with simple decisions based on one attribute can be enormous. Using more powerful decisions, we can construct trees with a much simpler structure. SVM can be regarded as decision trees with one decision but that single decision is largely a black box. By using a linear SVM with 1-norm capacity control (RLP) to construct each decision, a linear rule based on only the necessary attributes will be produced. The 2-norm SVM usually is a function of all the attributes. This attribute reduction is essential in many practical applications. What we want is something in between a very large univariate decision tree and a single nonlinear support vector machine. The ideal decision tree should generalize well, select the only relevant attributes, and provide information about the properties of the underlying data relevant to the application.

Support Vector Decision Trees

In this section we will examine the Support Vector Decision Tree algorithm (SVDT) and its successful application to a database marketing problem. Full details of this work can be found in (Bennett et al., 1998). SVDT uses Dual RLP (19.6) to construct simple decision trees with excellent dimensionality reduction.

SVDT performs top down induction of decision trees (TDIDT) like many other decision tree algorithms including CHAID, CART, MSMT, C4.5, and OC1 (Breiman et al., 1984; Bennett, 1992; Quinlan, 1993; Murthy et al., 1994). The primary distinguishing factors between SVDT and other TDIDT algorithms are the type of decisions used (linear SVM) and the method of constructing the decisions (RLP (19.6)). The basic TDIDT algorithm works as follows:

Algorithm 19.3.1 BASIC TDIDT Alogrithm
Start with the root node.
While a node remains to split

- *Construct decision based on some splitting criterion.*
- *Partition the node into two or more child nodes based on decision.*

Prune the tree if necessary.

In our case the splitting criterion and method is RLP (19.6) solved using the commercial linear programming package CPLEX 5.0 (CPL, 1994).

SVDT has been applied to problems in database marketing. In database marketing problems the primary goal is not testing set accuracy. The primary goal is to produce a good rank ordering of customers. Decision tree algorithms using one attribute per decision are frequently used in database marketing. The problem is that univariate decision tree algorithms can produce very large trees with hundreds of decisions on these large marketing data sets. In SVDT, we use more powerful decisions to produce very compact trees typically with three decisions. By using RLP (19.6) with 1-norm capacity control to construct the decisions, SVDT also performs extensive dimensionality reduction. This is a property of the 1-norm formulation. The primal objective term $\|\mathbf{w}\|_1$ tends to force the weights to zero. For each decision, all of the original attributes are included in the RLP (19.6) model, and then at optimality some of the constraints will be inactive or equivalently not at bound. The Lagrangian multipliers for the constraints are the primal $\mathbf{w}$ variables. Those corresponding to the inactive constraints must be zero due to complementarity in the Karush-Kuhn-Tucker optimality conditions. So the linear program solver will automatically determine which weights/attributes can be eliminated from the problem at that decision.

The SVDT is constructed as part of the customer scoring process. The tradeoffs between training accuracy, dimensionality reduction, and capacity control are controlled by the parameter λ. We tune λ and prune the tree using a validation set. The tuning takes into account the desired goals of the models in database marketing. The customers are ranked based on the response rate of the decision tree nodes and the minimum distance of the points from the decision. The results are reported in gainscharts. The gainscharts are used to compare models for validation, to identify potential customers, and to determine expected utility or profit. In (Bennett et al., 1998) results are given on three business problems. SVDT produced simple, accurate trees that would perform excellent scoring using a small number of attributes. The trees also provide information about the structure of the problem.

On the data set Business I, SVDT produced the tree in figure 19.2. The training data consisted of 2,358 points with 612 attributes and the testing set consisted of 1,006 points. CPLEX produced the root decision in 55 CPU seconds on a Sun Ultra 1/140 with 688 Megabytes of memory. The testing set accuracy of the tree is 77.8% using three linear decisions. For comparison, C4.5 (Quinlan, 1993) produced a tree consisting of 251 univariate decisions with 66% testing set accuracy. C4.5 took 110

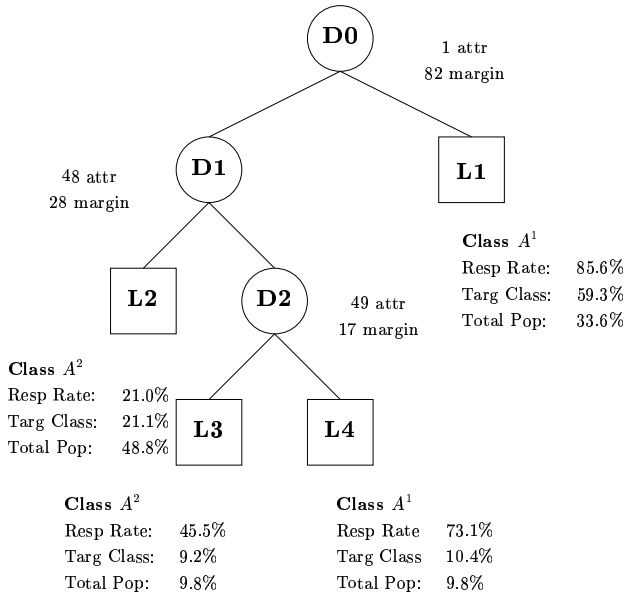

Figure 19.2 Decision tree for Business I test data, target class A^1

minutes to construct the decision tree on the same platform. More accurate SVDT trees exist but their corresponding gainscharts are not significantly different. Recall that the goal is to identify "good" customers, e.g. customers in the target class A^1. The structure of this tree is very interesting. The tree has three internal nodes. A value of $\lambda = .50$ was used at the root node D0 and a value of $\lambda = .25$ for the two subsequent decision nodes D1 and D2. In figure 19.2 each leaf node is labeled with the response rate of the target Class A^1: (|Class A^1 at node|/|all points at node|), the percentage of the target class reached at each node: (|Class A^1 at node|/|Class A^1 in population|), and the percentage of the total population reaching the node: (|all points at node|/|all points in population|). The optimal number of attributes and margin of separation ($\frac{2}{\|\mathbf{w}\|_2}$) is also given for each decision. The optimal number of attributes is the number of attributes at that decision with nonzero optimal weights. Notice that the first decision has only **one** nonzero weight and thus requires only one attribute. This simple decision, using one attribute, produces leaf L1 which reaches 59.3% of the target class with a response rate of 85.8%. Decision D1 uses 48 attributes to produce Leaf L2 that predominately contains points in A^2. The response rate of the target class is only 21.0%, and 48.9% of the total population reaches that node. So in some sense decision D0 identifies the easy points in A^1 and decision D1 separates the points easily identified as being in A^2. The points reaching decision D2 are difficult to classify. The margin of separation in decision D2 is much smaller and the accuracy of the decision is low. But decision D2 still produces useful information for scoring. We can rank customers based on the leaf of the tree that they reach. Our customer preference is L1, L4, L3, and then L2.

Decile	Class A^1 Population	Cumulative Class A^1 Population	Class A^1 Response Rate	Cumulative Response Rate	Lift
1	20.04	20.04	**97.03**	97.03	199.62
2	18.00	38.04	**88.00**	92.54	190.37
3	15.54	53.58	**76.00**	87.04	179.07
4	15.13	68.71	**73.27**	83.58	171.95
5	8.79	77.51	42.57	75.35	155.01
6	7.36	84.87	35.64	68.71	141.35
7	4.50	89.37	22.00	62.07	127.70
8	4.29	93.66	20.79	56.89	117.05
9	3.48	97.14	17.00	52.49	107.98
10	2.86	100.00	13.86	48.61	100.00

Table 19.1 Gainschart for Business I test data.

After the points are ranked, they are sorted and the results displayed in a gainschart. Each line of the gainschart contains one decile (10%) of the population. The deciles appear in order of response rate. For each decile we report the percentage of the total target class population included in that decile, the cumulative percentage of the total target class population, response rate of the target class in that decile, and the cumulative target class response rate. The last column represents lift, a measure of how much better we are doing over choosing customers at random. The lift is defined as 100*(response rate)*|Class A^1 in population|/|Class A^1 in decile|. Deciles with over 50% of Class A^1 are shown in bold. The gainschart for the Business I test data is given in table 19.1.

Once we have the final model, we construct the gainschart using the test data. The testing set response rate by decile is used to estimate the expected business response rate. So for example if we market the top 40% of the customers we could expect a response rate of 83.6% and we would reach approximately 68.7% of all possible target customers in our population. The rule of thumb in database marketing is: if in the fifth decile more than 70% of the target class is reached, then the model is successful. In our gainschart the fifth decile is underlined. We reached 77% of the Class A^1 population at the fifth decile. The testing gainschart combined with a model of expected profitability can be used to determine thresholds for scoring. In the scoring process, potential customers are selected based on the model and the selected threshold (Thomas, 1996; Hughes, 1996).

SVDT was also tested on two other database marketing problems with similar results. The largest data set attempted contained 33.6 megabytes of training data. The largest root decision was solved in 23 minutes by CPLEX on a Sun Ultra 1/140 with 688 megabytes of memory. The interested reader should consult (Bennett et al., 1998) for full details of these experiments.

Other versions of SVDT are possible. The original SVM quadratic program (1.37)

approach could be used in a TDIDT algorithm. The catch is that this approach frequently results in only one decision in the tree. Thus alternative SVM-based algorithms that consider and optimize all the decisions in the tree simultaneously have been proposed (Blue, 1998; Bennett and Blue, 1997). The trees found by SVDT on the database marketing problem look very much like the classifiers produced by the original MSM algorithm. So another possibility is to use trees with three-way splits at each decision, the region to the left of the margin, the region in the margin, and the region to the right of the margin. The more popular decision tree algorithms like CART and C4.5 work on attributes that are symbolic. What does the margin mean in the context of symbolic attributes? Can statistical learning theory help in algorithms for problems with symbolic attributes? These are open research topics.

19.4 Multicategory Classification

In this section we focus on the different approaches MPM and SVM have used to solve problems with $\Psi > 2$ classes. The original SVM method for multiclass problems was to find Ψ separate two-class discriminants (Cortes and Vapnik, 1995; Vapnik, 1995). Each discriminant is constructed by separating a single class from all the others. This process requires the solution of Ψ quadratic programs. We will denote this method Ψ-SVM. When applying all Ψ classifiers to the original multicategory data set, multiply classified points or unclassified points may occur. This ambiguity has been avoided by choosing the class of a point corresponding to the classification function that is maximized at that point. The LP approach has been to construct directly Ψ classification functions such that for each point the corresponding class function is maximized (Bennett and Mangasarian, 1993, 1994). The Multicategory Discrimination Method (Bennett and Mangasarian, 1993, 1994) constructs a piecewise-linear discriminant for the Ψ-class problem using a single linear program. We will call this method M-RLP since it is a direction extension of the RLP approach. The Ψ-SVM and M-RLP approaches can be combined to yield two new methods: Ψ-RLP, and M-SVM. We will provide a very brief description of this work. Full details on all the results of this section can be found in (Bredensteiner and Bennett, 1998; Bredensteiner, 1997).

To simplify the equations we introduce some notation. We wish to construct a discriminant function between the elements of the sets, $A^i, i = 1, \ldots, \Psi$, in the N-dimensional real space $\mathbb{R}^N$. Let $\mathbf{A}^i$ be an $\ell_i \times N$ matrix whose rows are the points in A^i. The j^{th} point in A^i and the j^{th} row of $\mathbf{A}^i$ are both denoted $\mathbf{A}^j_i$. Let $\mathbf{e}$ denote a vector of ones of the appropriate dimension. We can express a set of constraints such as $\mathbf{w} \cdot \mathbf{A}^i_j \geq \gamma + 1,\ j = 1, \ldots, \ell_i$ as $\mathbf{A}^i\mathbf{w} \geq (\gamma + 1)\mathbf{e}$.

In the linear case the original MPM and SVM methods both construct a piecewise-linear separator to discriminate between $\Psi > 2$ classes of $\ell^i,\ i = 1, \ldots, \Psi$, points. In Ψ-SVM (Vapnik, 1995; Cortes and Vapnik, 1995) a quadratic program is solved to construct a discriminant function to separate one class from the remaining $\Psi - 1$ classes. This process is repeated Ψ times. In the separable case, the

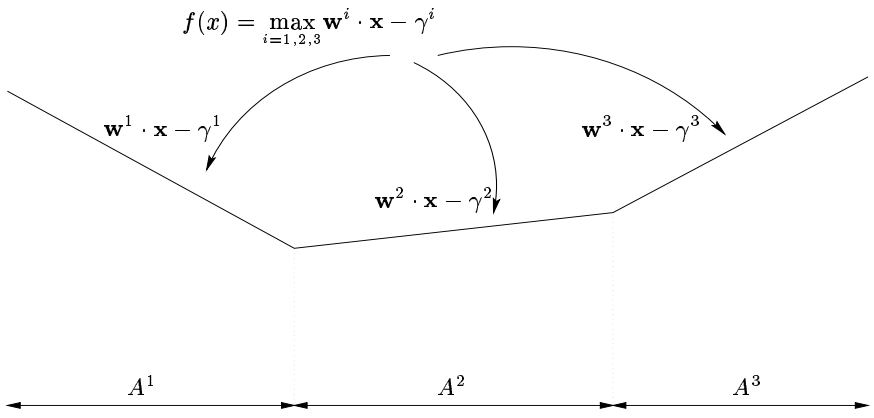

Figure 19.3 Piecewise-linear separation of sets A^1, A^2, and A^3 by the convex piecewise-linear function $f(\mathbf{x})$

Ψ-SVM

linear discriminant for each class must satisfy the following set of inequalities: Find $(\mathbf{w}^1, \gamma^1), \ldots, (\mathbf{w}^\Psi, \gamma^\Psi)$, such that

$$\begin{aligned} A^i\mathbf{w}^i - \gamma^i\mathbf{e} &\geq \mathbf{e} \\ -\mathbf{e} &\geq A^j\mathbf{w}^i - \gamma^i\mathbf{e}, \quad i,j = 1,\ldots,\Psi, \quad i \neq j. \end{aligned} \tag{19.7}$$

For the separable case, solving the Ψ one-class-from-the-rest SVM will yield $(\mathbf{w}^1, \gamma^1), \ldots, (\mathbf{w}^\Psi, \gamma^\Psi)$ if a solution exists.

To classify a new point $\mathbf{x}$, compute $f_i(\mathbf{x}) = \mathbf{w}^i \cdot \mathbf{x} - \gamma^i$. If $f_i(\mathbf{x}) > 0$ for only one i then clearly the point belongs to Class A^i. If more than one $f_i(\mathbf{x}) > 0$ or $f_i(\mathbf{x}) \leq 0$ for all $i = 1, \ldots, \Psi$ then the class is ambiguous. Thus the general rule is that the class of a point $\mathbf{x}$ is determined from $(\mathbf{w}^i, \gamma^i)$, $i = 1, \ldots, \Psi$, by finding i such that

$$f_i(\mathbf{x}) = \mathbf{w}^i \cdot \mathbf{x} - \gamma^i \tag{19.8}$$

is maximized. Figure 19.3 shows a piecewise-linear function $f(x) = \max_{i=1,2,3} f_i(\mathbf{x})$ on R that separates three sets. Note that while $(\mathbf{w}^i, \gamma^i)$, $i = 1 \ldots, \Psi$, are constructed in Ψ separate optimization problems, in the final classification function the problem is not separable into Ψ separate functions.

Note that either SVM (1.37) or RLP can be used to construct the Ψ two-class discriminants depending on the norm desired for capacity control. For clarity, we will call this method used with SVM (1.37), Ψ-SVM. We will denote this method used with RLP (19.6), Ψ-RLP. For both Ψ-SVM and Ψ-RLP to attain perfect training set accuracy using the function (19.8), the following inequalities must be feasible, i.e. there exist $(\mathbf{w}^1, \gamma^1), \ldots, (\mathbf{w}^\Psi, \gamma^\Psi)$ satisfying

$$\mathbf{A}^i\mathbf{w}^i - \gamma^i\mathbf{e} > \mathbf{A}^i\mathbf{w}^j - \gamma^j\mathbf{e}, \quad i,j = 1,\ldots,\Psi, \quad i \neq j \tag{19.9}$$

or equivalently

$$\mathbf{A}^i(\mathbf{w}^i - \mathbf{w}^j) - (\gamma^i - \gamma^j)\mathbf{e} \geq \mathbf{e}, \quad i,j = 1,\ldots,\Psi, \quad i \neq j \tag{19.10}$$

The M-RLP method[1] proposed and investigated in (Bennett and Mangasarian, 1993, 1994) can be used to find $(\mathbf{w}_i,\ \gamma_i),\ i = 1,\ldots,\Psi$ satisfying the inequalities (19.10). In the two-class case, M-RLP simplifies to the original RLP method:

M-RLP

$$\min_{\mathbf{w}^i,\gamma^i,\mathbf{z}^{ij}} \left\{ \sum_{i=1}^{\Psi} \sum_{\substack{j=1 \\ j\neq i}}^{\Psi} \frac{e^T\mathbf{z}^{ij}}{\ell^i} \;\middle|\; \begin{array}{c} \mathbf{z}^{ij} \geq -A^i(\mathbf{w}^i - \mathbf{w}^j) + (\gamma^i - \gamma^j)e + e,\ \mathbf{z}^{ij} \geq 0, \\ i,\ j = 1,\ldots,\Psi \quad i \neq j \end{array} \right\} \tag{19.11}$$

where $\mathbf{z}^{ij} \in R^{\ell_i \times 1}$. In M-RLP (19.11), if the optimal objective value is zero, then the data set is piecewise-linearly separable. If the data set is not piecewise-linearly separable, the positive values of the variables $\mathbf{z}_l^{ij}$ are proportional to the magnitude of the misclassified points from the plane $(\mathbf{w}^i - \mathbf{w}^j)\cdot\mathbf{x} = (\gamma^i - \gamma^j) + 1$. M-RLP (19.11) is a linear program. Like the original RLP (19.3), M-RLP does not include any terms for maximizing the margin. So we will now show how M-RLP and SVM can be combined by including margin maximization and generalized inner products into M-RLP.

Intuitively, the "optimal" $(\mathbf{w}^i, \gamma^i)$ should provide the largest margin of separation possible. So in an approach analogous to the two-class SVM approach, we add margin maximization terms to control capacity. The dashed lines in figure 19.4 represent the margins for each piece $(\mathbf{w}^i - \mathbf{w}^j, \gamma^i - \gamma^j)$ of the piecewise-linear separating function. The margin of separation between the classes i and j, i.e. the distance between

$$\begin{array}{l} A^i(\mathbf{w}^i - \mathbf{w}^j) \geq (\gamma^i - \gamma^j)e + e \quad \text{and} \\ A^j(\mathbf{w}^i - \mathbf{w}^j) \leq (\gamma^i - \gamma^j)e - e, \end{array} \tag{19.12}$$

is $\frac{2}{\|\mathbf{w}^i - \mathbf{w}^j\|}$. So, we would like to minimize $\|\mathbf{w}^i - \mathbf{w}^j\|$ for all $i,j = 1,\ldots,\Psi,\ i \neq j$. Also, we will add the regularization term $\frac{1}{2}\sum_{i=1}^{\Psi}\|\mathbf{w}^i\|^2$ to the objective. For the piecewise-linearly inseparable problem we get the following:

$$\begin{array}{ll} \min\limits_{\mathbf{w}^i,\gamma^i,\mathbf{z}^{ij}} & (1-\lambda)\sum\limits_{i=1}^{\Psi}\sum\limits_{\substack{j=1 \\ j\neq i}}^{\Psi} \frac{\mathbf{e}\cdot\mathbf{z}^{ij}}{\ell^i} + \frac{\lambda}{2}\left[\sum\limits_{i=1}^{\Psi}\sum\limits_{j=1}^{i-1}\|\mathbf{w}^i - \mathbf{w}^j\|^2 + \sum\limits_{i=1}^{\Psi}\|\mathbf{w}^i\|^2\right] \\ \text{subject to} & \mathbf{z}^{ij} + \mathbf{A}^i(\mathbf{w}^i - \mathbf{w}^j) - \mathbf{e}(\gamma^i - \gamma^j) - \mathbf{e} \geq 0 \\ & i,j = 1,\ldots,\Psi \quad i \neq j. \end{array} \tag{19.13}$$

where $\lambda \in (0,1)$. Note that the misclassification costs of $\frac{1}{\ell^i}$ could be any positive constant. In (Weston and Watkins, 1998) a Ψ-class formulation very similar to Problem (19.13) is proposed except the margin maximization term that minimizes $\|\mathbf{w}^i - \mathbf{w}^j\|^2$ is omitted.

1. The method was originally called Multicategory Discrimination.

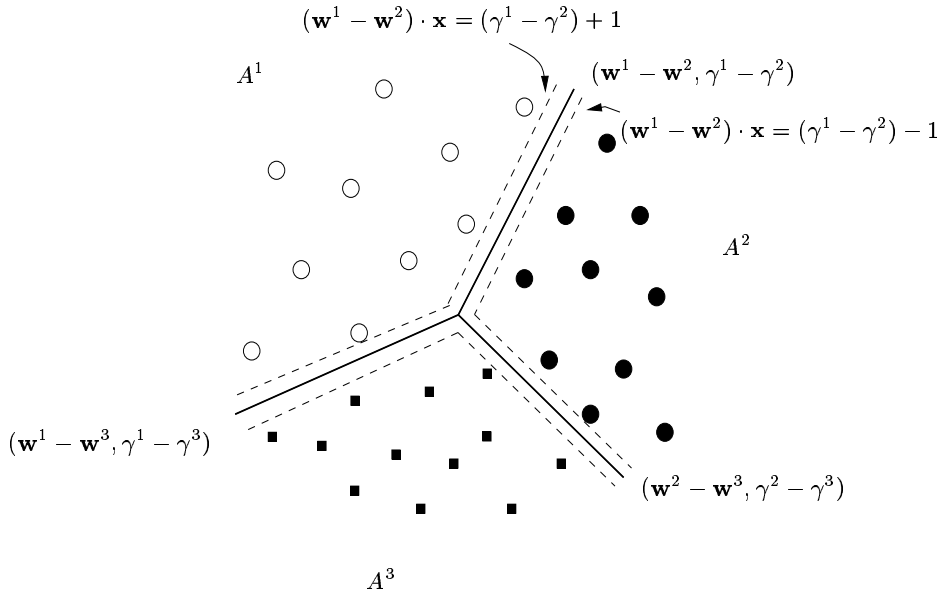

Figure 19.4 Piecewise-linear separator with margins for three classes

As in the two-class case, the dual of the problem can be formulated (see Bredensteiner and Bennett (1998)). Kernels can be easily incorporated into the dual formulation to allow piecewise-nonlinear discriminants. The notion of support vector exists in this formulation. There are $\Psi - 1$ Lagrangian multipliers, $\mathbf{u}$, for each point. The final M-SVM produces a piecewise-nonlinear classification that computes the class of a point $\mathbf{x}$ by finding $i = 1, \ldots, \Psi$ such that the classification function

$$f_i(\mathbf{x}) = \sum_{\substack{j=1 \\ j \neq i}}^{\Psi} \left[\sum_{\mathrm{SV} \in A^i} \mathbf{u}_p^{ij} K(\mathbf{x}, \mathbf{A}_p^{i\,T}) - \sum_{\mathrm{SV} \in A^j} \mathbf{u}_p^{ji} K(\mathbf{x}, \mathbf{A}_p^{j\,T}) \right] - \gamma^i \tag{19.14}$$

is maximized. Figure 19.5 illustrates the results of M-SVM on a three-class problem in two dimensions.

We summarize some of the computational results comparing M-SVM (19.13), M-RLP (19.11), Ψ-SVM using SVM (1.37), and Ψ-RLP using RLP (19.6). See (Bredensteiner and Bennett, 1998; Bredensteiner, 1997) for full details on the problem formulation and results. The quadratic programming problems for M-SVM and Ψ-SVM were solved using the nonlinear solver implemented in MINOS 5.4 (Murtagh and Saunders, 1993). This solver uses a reduced-gradient algorithm in conjunction with a quasi-Newton method. In M-SVM, Ψ-SVM, and M-RLP, the values for λ are .03, .05, and .03 respectively. Better solutions may result with different choices of λ. Additionally, it is not necessary for the same value of λ to be used for both methods. The kernel function for the piecewise-nonlinear M-SVM and Ψ-SVM methods is $K(x, x_i) = \left(\frac{x \cdot x_i}{n} + 1\right)^d$, where d is the degree of the desired polynomial.

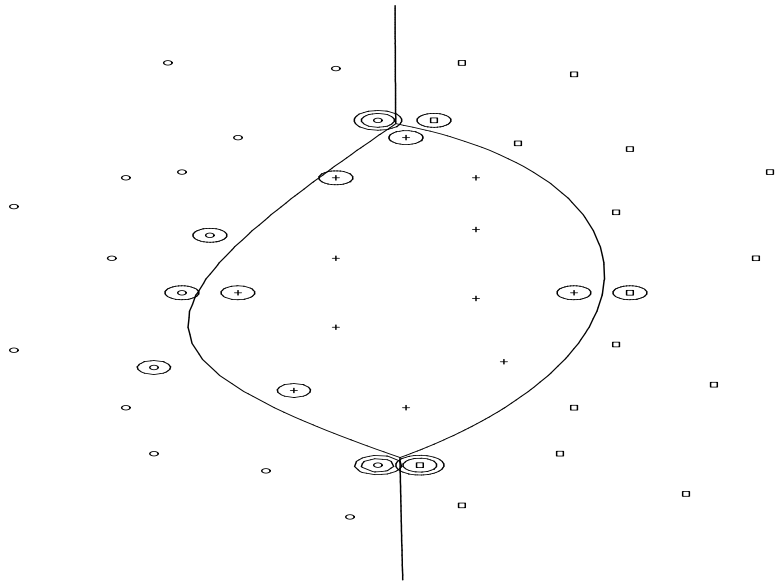

Figure 19.5 Piecewise-polynomial separation of three classes in two dimensions. Support vectors are indicated with large ovals.

We experimented with the United States Postal Service (USPS) Database (LeCun et al., 1989) containing zipcode samples from actual mail. This database is comprised of separate training and testing sets. There are 7291 samples in the training set and 2007 samples in the testing set. Each sample belongs to one of ten classes: the integers 0 through 9. The samples are represented by 256 features. Our experiment was conducted on two subsets of the USPS. Subsets were selected because the M-SVM for the complete formulation was too large for our solver without using decomposition techniques. These data contain handwriting samples of the integers 0 through 9. The objective of this data set is to interpret zipcodes quickly and effectively. This data set has separate training and testing sets, each of which consists of the 10 integer classes. We compiled two individual training subsets from the USPS training data. The first subset contains 1756 examples each belonging to the classes 3, 5, and 8. We call this set USPS-1 training data. The second subset contains 1961 examples each belonging to the classes 4, 6, and 7. We call this set USPS-2 training data. Similarly, two subsets are created from the testing data. In all of these data sets the data values are scaled by $\frac{1}{200}$. Testing set accuracies are reported for all four methods. The total numbers of unique support vectors in the resulting classification functions for the M-SVM and Ψ-SVM methods are given. Reference SVM accuracies on the full 10-class USPS benchmark are 95.8%, using a polynomial kernel, and 97.0%, incorporating prior knowledge by using a local kernel and Virtual SVs (Schölkopf et al., 1998d).

Data	Method	Degree				
		1	2	3	4	5
USPS-1	M-RLP	80.69	-	-	-	-
	Ψ-RLP	91.46	-	-	-	-
	M-SVM	91.26	91.87	92.28	92.07	92.28
		(415)	(327)	(312)	(305)	(317)
	Ψ-SVM	91.67	92.28	92.89	92.68	92.48
		(666)	(557)	(514)	(519)	(516)
USPS-2	M-RLP	80.66	-	-	-	-
	Ψ-RLP	96.13	-	-	-	-
	M-SVM	94.58	94.97	95.36	94.97	94.00
		(228)	(185)	(167)	(166)	(180)
	Ψ-SVM	96.13	96.52	96.13	95.16	94.58
		(383)	(313)	(303)	(294)	(289)

Table 19.2 Percent testing set accuracies and (total number of support vectors) for four multicategory discrimination methods

Table 19.2 contains results for the four methods on the USPS data subsets. Both of these data sets are piecewise-linearly separable. The solution that M-RLP has found for each of these data sets tests significantly worse than the other methods. This shows the importance of margin maximization, since M-RLP is the only method lacking capacity control. The Ψ-SVM method generalizes slightly better than M-SVM and is also more computationally efficient. The Ψ-RLP method reports accuracies similar to those of the Ψ-SVM method. Additionally, Ψ-RLP is solving many small linear program rather than one big linear program or quadratic programs, so the computational training time is significantly smaller than that of the other methods. Changing the parameter λ may further improve generalization. The M-SVM method consistently finds classification functions using fewer support vectors than those found by Ψ-SVM. With fewer support vectors, a sample can be classified more quickly since the dot-product of the sample with each support vector must be computed. Thus M-SVM would be a good method to choose when classification time is critical.

The results illustrate the value of combining SVM and MPM approaches. By incorporating margin maximization, the M-RLP method was greatly improved and two new methods Ψ-RLP and M-SVM were constructed. Overall, the one-class-from-the-rest approaches, Ψ-RLP and Ψ-SVM, are best both in terms of generalization and computational time on the problems we tested. Our computational experiments, however, were limited by the capacity of the solver used (MINOS). Decomposition methods such as the ones discussed in this book could be used to make the M-SVM method tractable for larger problems with more classes. Also, 1-norm capacity control and kernels could be added to the M-RLP formulation.

So the best multicategory formulation is still very much an open question both practically and theoretically.

19.5 Overall Risk Minimization and MPM

In this section we show how modeling techniques from mathematical programming can be used to help translate concepts from statistical learning theory into practical algorithms. As a concrete example, we examine the problem of overall risk minimization in transduction. Vapnik briefly presented this problem at the NIPS 1997 Support Vector Machine Workshop (see chapter 3) and it also can be found in chapter 10 of (Vapnik, 1979) and briefly in (Vapnik, 1995). Roughly, the transduction problem is: given a training set of labeled points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)$, estimate the value of a function $y = f(\mathbf{x})$, at a given unlabeled working set[2] $x_{\ell+1}, \ldots, x_m$. Vapnik distinguishes between this problem of transduction and the induction problem. In induction the goal is to estimate the function f at all possible points. Future testing points are classified using deduction. In transduction, the goal is to estimate the function value at a particular set of testing or working points. In induction, the structural risk is minimized. In transduction, the overall risk is minimized. According to overall risk minimization, by explicitly including the working set data in the problem formulation, we can expect better generalization on problems with insufficient data. We define the semi-supervised support vector machine problem (S^3VM) as: given a training set of points with known class, and a working set of data points with unknown class, construct a SVM to label the working set.

To formulate S^3VM, we start with either the 1-norm RLP or 2-norm SVM formulation, and then add two constraints for each point in the working set. One constraint calculates the misclassification error as if the point were in class A^1 and the other constraint calculates the misclassification error as if the point were in class A^2. The objective function calculates the minimum of the two possible misclassification errors. The final class of the points corresponds to the one that results in the smallest error. Specifically we define the S^3VM as:

Semi-supervised support vector machine

$$
\begin{array}{lll}
\min\limits_{\mathbf{w},\gamma,\xi,z} & (1-\lambda)\left[\sum\limits_{i=1}^{l_1}\delta_i\xi_i + \sum\limits_{j=1}^{l_2}\delta_i z_j + \sum\limits_{i=\ell+1}^{m}\delta_i min(\xi_i, z_i)\right] + \frac{\lambda}{2}\parallel \mathbf{w} \parallel & \\
\text{subject to} & \mathbf{w}\cdot\mathbf{x}_i - \gamma + \xi_i \geq 1 \quad \xi_i \geq 0 \quad i \in A^1 & \\
& -\mathbf{w}\cdot\mathbf{x}_j + \gamma + z_j \geq 1 \quad z_j \geq 0 \quad j \in A^2 & \\
& \mathbf{w}\cdot\mathbf{x}_s - \gamma + \xi_s \geq 1 \quad \xi_s \geq 0 \quad s \in \mathit{working\ set} & \\
& -\mathbf{w}\cdot\mathbf{x}_s + \gamma + z_s \geq 1 \quad z_s \geq 0 &
\end{array}
\tag{19.15}
$$

where $\delta_i > 0$ are fixed misclassification costs. For the experiments reported here we used $\delta_i = 1/m$ and $\lambda = .005$.

2. This set is refered to as the testing set in (Vapnik, 1995).

Integer programming can be used to solve this problem. The basic idea is to add a 0 or 1 decision variable, d_s, for each point $\mathbf{x}_s$ in the working set. This variable indicates the class of the point. If $d_s = 1$ then the point is in class A^1 and if $d_s = 0$ then the point is in class A^2. This results in the following mixed integer program (S^3VM-MIP):

S^3VMMixed Integer Program

$$\begin{aligned}
\min_{\mathbf{w},\gamma,\xi,z,d} \quad & (1-\lambda)\left[\sum_{i=1}^{l_1}\delta_i\xi_i + \sum_{j=1}^{l_2}\delta_i z_j + \sum_{i=\ell+1}^{m}\delta_i(\xi_i + z_i)\right] + \tfrac{\lambda}{2}\parallel \mathbf{w} \parallel_1 \\
\text{subject to} \quad & \mathbf{w}\cdot x_i - \gamma + \xi_i \geq 1 \quad \xi_i \geq 0 \quad i \in A^1 \\
& -\mathbf{w}\cdot x_j + \gamma + z_j \geq 1 \quad z_j \geq 0 \quad j \in A^2 \\
& \mathbf{w}\cdot x_s - \gamma + \xi_s + M(1-d_s) \geq 1 \quad \xi_s \geq 0 \quad s \in \textit{working set} \\
& -\mathbf{w}\cdot x_s + \gamma + z_s + Md_s \geq 1 \quad z_s \geq 0 \quad d_s = \{0,1\}
\end{aligned} \tag{19.16}$$

The constant $M > 0$ is chosen sufficiently large such that if $d_s = 0$ then $\xi_s = 0$ is feasible for any optimal $\mathbf{w}$ and γ. Likewise if $d_s = 1$ then $z_s = 0$.

If the 1-norm is used, this problem can be exactly solved using CPLEX or other commercial integer programming codes (CPL, 1994). CPLEX uses a combination of branch-and-bound and branch-and-cut techniques to produce an enumeration tree. At each node of the tree a continuous relaxation of the integer program is solved using low-cost linear algebra. For problem (19.16) the effectiveness of the algorithm is dependent on the number of integer variables, i.e., the size of the working set, and the effectiveness of the algorithm at pruning the search space. Using the mathematical programming modeling language AMPL (Fourer et al., 1993), we were able to express the problem in approximately thirty lines of code plus a data file and solve it using CPLEX.[3] If the 2-norm is used for margin maximization, then the problem becomes a quadratic integer program. Methods exists for solving these problems but we did not have access to such a solver.

The S^3VM-MIP can be used to solve the transduction problem using overall risk minimization. Consider the simple problem given in figure 20 of (Vapnik, 1979). The results of RLP and SVM-MIP on this problem are shown in figure 19.6. The training set points are shown as transparent triangles and hexagons. The working set points are shown as filled circles. The left picture in figure 19.6 shows the solution found by RLP. Note that when the working set points are added, the resulting separation has a very small margin. The right picture shows the S^3VM-MIP solution constructed using the unlabeled working set. Note that a much larger and clearer separation margin is found. These computational solutions are virtually the same as the solution presented in (Vapnik, 1979).

We also tested S^3VM-MIP on ten real-world data sets from (Murphy and Aha, 1992). S^3VM-MIP tested better on nine of the ten data sets although not always significantly so. On no data set did S^3VM-MIP perform significantly worse. The

3. The AMPL code is available on request from the author at http://www.math.rpi.edu/~bennek.

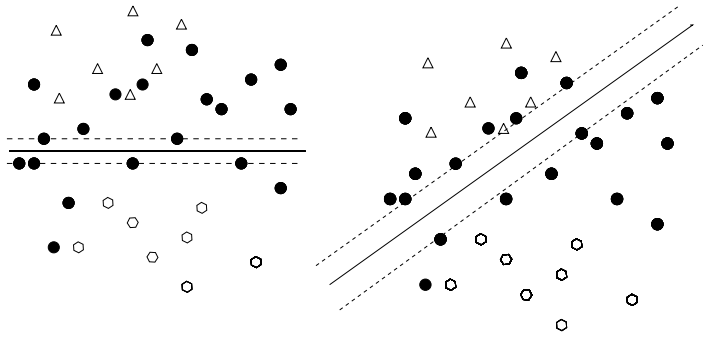

Figure 19.6 Left = solution found by RLP; Right = solution found by S^3*VM*-MIP

Data Set	Dim	Points	CV-size	RLP	S^3*VM*-MIP	p-value
Bright	14	2462	50*	0.02	**0.018**	0.343
Cancer	9	699	70	0.036	**0.034**	0.591
Cancer(Prognostic)	30	569	57	0.035	**0.033**	0.678
Dim	14	4192	50*	0.064	**0.054**	0.096
Heart	13	297	30	0.173	**0.160**	0.104
Housing	13	506	51	0.155	**0.151**	0.590
Ionosphere	34	351	35	0.109	**0.106**	0.59
Musk	166	476	48	**0.173**	**0.173**	0.999
Pima	8	769	50*	**0.220**	0.222	0.678
Sonar	60	208	21	0.281	**0.219**	0.045

Table 19.3 RLP vs S^3*VM*-MIP Average Testing Error

results are given in table 19.3. For each data set, we performed 10-fold cross-validation. For the three starred data sets, our integer programming solver failed due to excessive branching required within the CPLEX algorithm. On those data sets we randomly extracted 50-point working sets for each trial. The same parameters were used for each data set in both the RLP and S^3*VM*-MIP problems. While the p-values for the paired t-test of the testing set accuracies are not always small, this is not a surprise. Many algorithms have been applied successfully to these problems without incorporating working set information. Thus it was not clear *a priori* that S^3*VM* would improve generalization on these data sets. For the data sets where no improvement is possible, we would like S^3*VM*-MIP to not degrade the performance of RLP. Our results are consistant with the statistical learning theory results that incorporating working data improves generalization when insufficient training information is available. In every case, S^3*VM*-MIP either

improved or showed no significant difference in generalization compared to the baseline empirical risk minimization approach RLP. With additional constraints SVM-MIP can be adapted to clustering as well. Other problem formulations for S^3VM that incorporate kernels are being investigated.

19.6 Conclusions

We have shown that the past problem formulations for MPM from the work of Mangasarian share the same canonical form with SVM. These similarities allow MPM and SVM methods to be easily combined. We examined how the dissimilarities of the MPM and SVM approaches can be used to generate new methods for nonlinear discrimination using support vector decision trees and multicategory learning. In almost every case incorporating margin maximization into the MPM resulted in better generalization. We also showed how MPM models and tools allowed us to develop rapidly a practical approach to solving a transduction problem using overall risk minimization. This integer programming approach was able to solve moderately sized problem using commercial software. Our preliminary empirical results support the overall risk minimization theory and indicate that transduction is both a promising and practical research direction for both SVM and MPM. The review here has been solely limited to a few examples from a single family of MPM. There are many extensions of these methods such as those covered in (Bradley et al., 1998; Mangasarian, 1997) that can also be potentially combined with SVM. In addition, there are wide classes of totally unrelated MPM approaches, e.g. Glover (1990); Gochet et al. (1997), that also can be potentially synthesized with SVM. Omission of any method from this paper should not be used as an indication of the quality of the method. The primary weakness of the MPM approaches is that they have not been guided by statistical learning theory. In the problems investigated in this chapter, altering MPM methods to include principles of statistical learning theory almost always improved generalization. Many other optimization-based methods can potentially be improved by similar transformations.

Acknowledgments

Thanks to my collaborators in this work: Leonardo Auslender, Erin Bredensteiner, Ayhan Demiriz, and Donghui Wu and to Scott Vandenberg for his editorial comments. This work was supported by NSF grants IRI-9409427 and IRI-9702306.

References

H. D. I. Abarbanel. *Analysis of Observed Chaotic Data*. Springer Verlag, New York, 1996.

M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821 – 837, 1964.

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale–sensitive Dimensions, Uniform Convergence, and Learnability. *Journal of the ACM*, 44(4):615–631, 1997.

E. Amaldi and V. Kann. The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical Computer Science*, 147:181–210, 1995.

E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 1998. To appear.

M. Anthony. Probabilistic analysis of learning in artificial neural networks: The PAC model and its variants. *Neural Computing Surveys*, 1:1–47, 1997. http://www.icsi.berkeley.edu/~jagota/NCS.

M. Anthony and N. Biggs. *Computational Learning Theory*, volume 30 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1992.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337 – 404, 1950.

R. Ash. *Information Theory*. Interscience Publishers, New York, 1965.

P. L. Bartlett. Pattern classification in neural networks. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998a.

P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998b.

P. L. Bartlett, P. Long, and R. C. Williamson. Fat–Shattering and the Learnability of Real–Valued Functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

Y. Bengio, Y. LeCun, and D. Henderson. Globally trained handwritten word recognizer using spatial representation, convolutional neural networks and hidden

markov models. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 937–944, 1994.

K. P. Bennett. Decision tree construction via linear programming. In M. Evans, editor, *Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference*, pages 97–101, Utica, Illinois, 1992.

K. P. Bennett and J. A. Blue. A support vector machine approach to decision trees. In *Proceedings of IJCNN'98*, pages 2396 – 2401, Anchorage, Alaska, 1997.

K. P. Bennett and E. J. Bredensteiner. A parametric optimization method for machine learning. *INFORMS Journal on Computing*, 9(3):311–318, 1997.

K. P. Bennett and E. J. Bredensteiner. Geometry in learning. In C. Gorini, E. Hart, W. Meyer, and T. Phillips, editors, *Geometry at Work*, Washington, D.C., 1998. Mathematical Association of America. Available http://www.math.rpi.edu/~bennek/geometry2.ps.

K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. Unpublished manuscript based on talk given at Machines That Learn Conference, Snowbird, 1998.

K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

K. P. Bennett and O. L. Mangasarian. Multicategory separation via linear programming. *Optimization Methods and Software*, 3:27–39, 1993.

K. P. Bennett and O. L. Mangasarian. Serial and parallel multicategory discrimination. *SIAM Journal on Optimization*, 4(4):722–734, 1994.

K. P. Bennett, D. H. Wu, and L. Auslender. On support vector decision trees for database marketing. R.P.I. Math Report No. 98-100, Rensselaer Polytechnic Institute, Troy, NY, 1998.

L. Bernhardt. Zur Klassifizierung vieler Musterklassen mit wenigen Merkmalen. In H. Kazmierczak, editor, *5. DAGM Symposium: Mustererkennung 1983*, pages 255 – 260, Berlin, 1983. VDE-Verlag.

D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1995.

M. Bierlaire, Ph. Toint, and D. Tuyttens. On iterative algorithms for linear least squares problems with bound constraints. *Linear Alebra Appl.*, pages 111–143, 1991.

C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

V. Blanz, B. Schölkopf, H. Bülthoff, C. Burges, V. Vapnik, and T. Vetter. Comparison of view-based object recognition algorithms using realistic 3D models. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Artificial Neural Networks — ICANN'96*, pages 251 – 256, Berlin, 1996. Springer Lecture Notes in Computer Science, Vol. 1112.

J. A. Blue. *A Hybrid of Tabu Search and Local Descent Algorithms with Applications in Artificial Intelligence.* PhD thesis, Rensselaer Polytechnic Institute, 1998.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

W. M. Boothby. *An introduction to differentiable manifolds and Riemannian geometry.* Academic Press, 2nd edition, 1986.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Müller, E. Säckinger, P. Simard, and V. Vapnik. Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the 12th International Conference on Pattern Recognition and Neural Networks, Jerusalem*, pages 77 – 87. IEEE Computer Society Press, 1994.

L. Bottou and V. N. Vapnik. Local learning algorithms. *Neural Computation*, 4(6): 888–900, 1992.

P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian. Data mining: Overview and optimization opportunities. Technical Report Mathematical Programming Technical Report 98-01, University of Wisconsin-Madison, 1998. Submitted for publication.

P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. Technical Report Mathematical Programming Technical Report 98-03, University of Wisconsin-Madison, 1998a. To appear in ICML-98.

P. S. Bradley and O. L. Mangasarian. Massive data discrimination via linear support vector machines. Technical Report Mathematical Programming Technical Report 98-05, University of Wisconsin-Madison, 1998b. Submitted for publication.

P. S. Bradley, O. L. Mangasarian, and W. N. Street. Feature selection via mathematical programming. Technical Report 95-21, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, 1995. To appear in *INFORMS Journal on Computing* 10, 1998.

E. J. Bredensteiner. *Optimization Methods in Data Mining and Machine Learning.* PhD thesis, Rensselaer Polytechnic Institute, 1997.

E. J. Bredensteiner and K. P. Bennett. Feature minimization within decision trees. *Computational Optimization and Applications*, 10:110–126, 1997.

E. J. Bredensteiner and K. P. Bennett. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 1998. To appear.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

L. Breiman. Bagging predictors. Technical Report 421, Department of Statistics, UC Berkeley, 1994. ftp://ftp.stat.berkeley.edu/pub/tech-reports/421.ps.Z.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International, California, 1984.

R. Brown, P. Bryant, and H. D. I. Abarbanel. Computing the lyapunov spectrum of a dynamical system from observed time-series. *Phys. Rev. Lett.*, 43(6):2787–2806, 1991.

J. R. Bunch and L. Kaufman. Some stable methods for calculating inertia and solving symmetric linear systems. *Mathematics of Computation*, 31:163–179, 1977.

C. J. C. Burges. Simplified support vector decision rules. In L. Saitta, editor, *Proceedings, 13th Intl. Conf. on Machine Learning*, pages 71–77, San Mateo, CA, 1996. Morgan Kaufmann.

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):1–47, 1998.

C. J. C. Burges and B. Schölkopf. Improving the accuracy and speed of support vector learning machines. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 375–381, Cambridge, MA, 1997. MIT Press.

C. J. C. Burges and V. Vapnik. A new method for constructing artificial neural networks: Interim technical report, ONR contract N00014-94-c-0186. Technical report, AT&T Bell Laboratories, 1995.

B. Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Annales de l'Institut Fourier*, 35(3):79–118, 1985.

B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators*. Cambridge University Press, Cambridge, UK, 1990.

Y. Censor. Row-action methods for huge and sparse systems and their applications. *SIAM Review*, 23(4):444–467, 1981.

Y. Censor and A. Lent. An iterative row-action method for interval convex programming. *J. Optimization Theory and Applications*, 34(3):321–353, 1981.

S. Chen. *Basis Pursuit*. PhD thesis, Department of Statistics, Stanford University, 1995.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. Technical Report 479, Department of Statistics, Stanford University, 1995.

C. R. Chester. *Techniques in Partial Differential Equations*. McGraw Hill, 1971.

E. T. Copson. *Metric Spaces*. Cambridge University Press, 1968.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273 – 297, 1995.

R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume 1. Interscience Publishers, Inc, New York, 1953.

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Elect. Comp.*, 14:326–334, 1965.

D. Cox and F. O'Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.*, 18:1676–1695, 1990.

CPLEX Optimization Incorporated, Incline Village, Nevada. *Using the CPLEX Callable Library*, 1994.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31:377–403, 1979.

N. Cristianini, J. Shawe-Taylor, and P. Sykacek. Bayesian classifiers are large margin hyperplanes in a hilbert space. In J. Shavlik, editor, *Machine Learning: Proceedings of the Fifteenth International Conference*, San Francisco, CA, 1998. Morgan Kaufmann.

K. I. Diamantaras and S. Y. Kung. *Principal Component Neural Networks.* Wiley, New York, 1996.

K. Dodson and T. Poston. *Tensor Geometry.* Springer-Verlag, 2nd edition, 1991.

H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, Cambridge, MA, 1997. MIT Press.

R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, 1973.

S. Dumais. Using SVMs for text categorization. *IEEE Intelligent Systems*, 13(4), 1998. In: M.A. Hearst, B. Schölkopf, S. Dumais, E. Osuna, and J. Platt: Trends and Controversies — Support Vector Machines.

N. Dunford and J. T. Schwartz. *Linear Operators Part II: Spectral Theory, Self Adjoint Operators in Hilbert Space.* Number VII in Pure and Applied Mathematics. John Wiley & Sons, New York, 1963.

J. P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Modern Phys.*, 57(3):617–656, 1985.

K. Efetov. *Supersymmetry in Disorder and Chaos.* Cambridge University Press, Cambridge, 1997.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82: 247–261, 1989.

L. Elden and L. Wittmeyer-Koch. *Numerical Analysis: An Introduction.* Academic Press, Cambrigde, 1990.

R. Fourer, D. Gay, and B. Kernighan. *AMPL A Modeling Language for Mathematical Programming.* Boyd and Frazer, Danvers, Massachusetts, 1993.

J. H. Friedman. Another approach to polychotomous classification. Technical re-

port, Department of Statistics and Stanford Linear Accelerator Center, Stanford University, 1996.

E. Gardner. The space of interactions in neural networks. *Journal of Physics A*, 21:257–70, 1988.

P. E. Gill, W. Murray, and M. A. Saunders. Snopt: An sqp algorithm for large-scale constrained optimization. Technical Report NA-97-2, Dept. of Mathematics, U.C. San Diego, 1997.

P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, 1981.

D. Girard. Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. *Ann. Statist.*, 19:1950–1963, 1991.

D. Girard. Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression. *Ann. Statist.*, 126:315–334, 1998.

F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. A.I. Memo No. 1430, MIT, 1993.

F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

F. Glover. Improved linear programming models for discriminant analysis. *Decision Sciences*, 21:771–785, 1990.

W. Gochet, A. Stam, V. Srinivasan, and S. Chen. Multigroup discriminant analysis using linear programming. *Operations Research*, 45(2):213–559, 1997.

H. Goldstein. *Classical Mechanics*. Addison-Wesley, Reading, MA, 1986.

M. Golea, P. L. Bartlett, W. S. Lee, and L. Mason. Generalization in decision trees and DNF: Does size matter? In *Advances in Neural Information Processing Systems 10*, 1998.

G. Golub and U. von Matt. Generalized cross-validation for large-scale problems. *J. Comput. Graph. Statist.*, 6:1–34, 1997.

J. Gong, G. Wahba, D. Johnson, and J. Tribbia. Adaptive tuning of numerical weather prediction models: simultaneous estimation of weighting, smoothing and physical parameters. *Monthly Weather Review*, 125:210–231, 1998.

Y. Gordon, H. König, and C. Schütt. Geometric and probabilistic estimates for entropy and approximation numbers of operators. *Journal of Approximation Theory*, 49:219–239, 1987.

T. Graepel and K. Obermayer. Fuzzy topographic kernel clustering. In W. Brauer, editor, *Proceedings of the 5th GI Workshop Fuzzy Neuro Systems '98*, pages 90 – 97, 1998.

R. E. Greene. *Isometric Embeddings of Riemannian and Pseudo-Riemannian*

Manifolds. American Mathematical Society, 1970.

C. Gu and G. Wahba. Semiparametric analysis of variance with tensor product thin plate splines. *J. Royal Statistical Soc. Ser. B*, 55:353–368, 1993.

L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in banach spaces. In *Proceedings of Algorithm Learning Theory, ALT-97*, 1997. Also: NECI Technical Report, 1997.

I. Guyon, B. Boser, and V. Vapnik. Automatic capacity tuning of very large VC-dimension classifiers. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 147–155. Morgan Kaufmann, San Mateo, CA, 1993.

M. Hamermesh. *Group theory and its applications to physical problems*. Addison Wesley, Reading, MA, 2 edition, 1962. Reprint by Dover, New York, NY.

D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. In *J. Environ. Economics & Management*, volume 5, pages 81–102, 1978. Original source of the Boston Housing data, actually from ftp://ftp.ics.uci.com/pub/machine-learning-databases/housing.

T. Hastie and W. Stuetzle. Principal curves. *JASA*, 84:502 – 516, 1989.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1990.

S. Haykin. *Neural Networks : A Comprehensive Foundation*. Macmillan, New York, 1994.

S. Haykin, S. Puthusserypady, and P. Yee. Reconstruction of underlying dynamics of an observed chaotic process. Technical Report 353, Comm. Res. Lab., McMaster University, 1997.

C. Hildreth. A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4:79–85, 1957.

T. K. Ho and E. Kleinberg. Building projectable classifiers for arbitrary complexity. In *Proceedings of the 12th International Conference on Pattern Recognition, Vienna*, pages 880–885, 1996.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441 and 498–520, 1933.

P. J. Huber. Robust statistics: a review. *Ann. Statist.*, 43:1041, 1972.

P. J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

A. M. Hughes. *The Complete Database Marketer*. Irwin Prof. Publishing, Chicago, 1996.

M. Hutchinson. A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines. *Commun. Statist.-Simula.*, 18:1059–1076, 1989.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. Technical Report 23, LS VIII, University of Dortmund,

1997.

T. Joachims. Text categorization with support vector machines. In *European Conference on Machine Learning (ECML)*, 1998.

I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

J. Karhunen and J. Joutsensalo. Generalizations of principal component analysis, optimization problems, and neural networks. *Neural Networks*, 8(4):549–562, 1995.

K. Karhunen. Zur Spektraltheorie stochastischer Prozesse. *Ann. Acad. Sci. Fenn.*, 34, 1946.

W. Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, Dept. of Mathematics, Univ. of Chicago, 1939.

M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2):115–141, 1994.

M. Kennel, R. Brown, and H. D. I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A.*, 45:3403–3411, 1992.

G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41:495–502, 1970.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

M. Kirby and L. Sirovich. Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

J. Kohlmorgen, K.-R. Müller, and K. Pawelzik. Analysis of drifting dynamics with neural network hidden markov models. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems 10*, Cambridge, MA, 1998. MIT Press.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59 – 69, 1982.

A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Prentice-Hall, Inc., 1970.

H. König. *Eigenvalue Distribution of Compact Operators*. Birkhäuser, Basel, 1986.

U. Kreßel. The impact of the learning–set size in handwritten–digit recognition. In T. Kohonen et al., editor, *Artificial Neural Networks — ICANN'91*, pages 1685 – 1689, Amsterdam, 1991. North–Holland.

U. Kreßel. Polynomial classifiers and support vector machines. In W. Gerstner et al., editor, *Artificial Neural Networks — ICANN'97*, pages 397 – 402, Berlin, 1997. Springer Lecture Notes in Computer Science, Vol. 1327.

U. Kreßel and J. Schürmann. Pattern classification techniques based on function

approximation. In H. Bunke and P.S.P. Wang, editors, *Handbook on Optical Character Recognition and Document Analysis*, pages 49 – 78. World Scientific Publishing Company, Singapore, 1997.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics*, pages 481–492, Berkeley, 1951. University of California Press.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. J. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541 – 551, 1989.

Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Müller, E. Säckinger, P. Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. In F. Fogelman-Soulié and P. Gallinari, editors, *Proceedings ICANN'95 — International Conference on Artificial Neural Networks*, volume II, pages 53 – 60, Nanterre, France, 1995. EC2. The MNIST benchmark data is available from http://www.research.att.com/~yann/ocr/mnist/.

W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 1998. to appear.

K. C. Li. Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing. *Ann. Statist.*, 14:1101–1112, 1986.

W. Liebert, K. Pawelzik, and H. G. Schuster. Optimal embeddings of chaotic attractors from topological considerations. *Europhys. Lett.*, 14:521 – 526, 1991.

B. Lillekjendlie, D. Kugiumtzis, and N. Christophersen. Chaotic time series: Part ii. system identification and prediction. *Modeling, Identification and Control*, 15 (4):225–243, 1994.

N. Littlestone and M. Warmuth. Relating data compression and learnability. Technical report, University of California Santa Cruz, 1986.

E. N. Lorenz. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20:130–141, 1963.

G. Loy and P. L. Bartlett. Generalization and the size of the weights: an experimental study. In *Proceedings of the Eighth Australian Conference on Neural Networks*, pages 60–64, 1997.

D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4:720–736, 1992a.

D. J. C. MacKay. A practical Bayesian framework for backprop networks. *Neural Computation*, 4:448–472, 1992b.

M. C. Mackey and L. Glass. Oscillation and chaos in physiological control systems. *Science*, 197:287–289, 1977.

S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

O. L. Mangasarian. Multi-surface method of pattern separation. *IEEE Transactions on Information Theory*, IT-14:801–807, 1968.

O. L. Mangasarian. Misclassification minimization. *J. Global Optimization*, 5:309–323, 1994.

O. L. Mangasarian. Mathematical programming in data mining. *Data Mining and Knowledge Discovery*, 42(1):183–201, 1997.

O. L. Mangasarian and R. Meyer. Nonlinear perturbations of linear programs. *SIAM Journal on Control and Optimization*, 17(6):745–752, 1979.

O. L. Mangasarian, R. Setiono, and W. H. Wolberg. Pattern recognition via linear programming: Theory and application to medical diagnosis. In *Proceedings of the Workshop on Large-Scale Numerical Optimization, 1989*, pages 22–31, Philadelphia, Pennsylvania, 1990. SIAM.

O. L. Mangasiarian. Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13:444–452, 1965.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London*, A 209:415–446, 1909.

C. J. Merz and P. M. Murphy. UCI repository of machine learning databases, 1998. [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using a support vector machine. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, *Neural Networks for Signal Processing VII — Proceedings of the 1997 IEEE Workshop*, pages 511 – 520, New York, 1997. IEEE.

B. Müller and J. Reinhardt. *Neural Networks: An Introduction*. Springer Verlag, 1990.

K.-R. Müller, J. Kohlmorgen, and K. Pawelzik. Analysis of switching dynamics with competing neural networks. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E78–A(10):1306–1315, 1995.

K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Artificial Neural Networks — ICANN'97*, pages 999 – 1004, Berlin, 1997. Springer Lecture Notes in Computer Science, Vol. 1327.

P.M. Murphy and D.W. Aha. UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, California, 1992.

B. A. Murtagh and M. A. Saunders. MINOS 5.4 user's guide. Technical Report SOL 83.20, Stanford University, 1993.

K. G. Murthy. *Linear Programming.* John Wiley & Sons, New York, New York, 1983.

S. Murthy, S. Kasif, and S. Salzberg. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2:1–32, 1994.

I. Nagayama and N. Akamatsu. Approximation of chaotic behavior by using neural network. *IEICE Trans. Inf. & Syst.*, E77-D(4), 1994.

J. Nash. The embedding problem for riemannian manifolds. *Annals of Mathematics*, 63:20 – 63, 1956.

R. Neal. Priors for infinite networks. Technical Report CRG-TR-94-1, Dept. of Computer Science, University of Toronto, 1994.

R. Neal. *Bayesian Learning in Neural Networks.* Springer Verlag, 1996.

N. J. Nilsson. *Learning machines: Foundations of Trainable Pattern Classifying Systems.* McGraw-Hill, 1965.

E. Oja. A simplified neuron model as a principal component analyzer. *J. Math. Biology*, 15:267 – 273, 1982.

P. J. Olver. *Applications of Lie Groups to Differential Equations.* Springer-Verlag, 1986.

M. Opper. Learning in neural networks: Solvable dynamics. *Europhysics Letters*, 8 (4):389–392, 1989.

M. Opper and W. Kinzel. Physics of generalization. In E. Domany J.L. van Hemmen and K. Schulten, editors, *Physics of Neural Networks III.* Springer Verlag, New York, 1996.

M. Opper, P. Kuhlmann, and A. Mietzner. Convexity, internal representations and the statistical mechanics of neural networks. *Europhysics Letters*, 37(1):31–36, 1997.

M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Proc. Computer Vision and Pattern Recognition*, pages 193–199, Puerto Rico, 1997.

E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, *Neural Networks for Signal Processing VII — Proceedings of the 1997 IEEE Workshop*, pages 276 – 285, New York, 1997a. IEEE.

E. Osuna, R. Freund, and F. Girosi. Support vector machines: Training and applications. AI Memo 1602, Massachusetts Institute of Technology, 1997b.

E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proc. Computer Vision and Pattern Recognition '97*, pages 130–136, 1997c.

N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a

time series. *Phys. Rev. Lett.*, 45:712–716, 1980.

E. Parzen. An approach to time series analysis. *Ann. Math. Statist.*, 32:951–989, 1962.

E. Parzen. Statistical inference on time series by rkhs methods. In R. Pyke, editor, *Proceedings 12th Biennial Seminar*, Montreal, 1970. Canadian Mathematical Congress. 1-37.

K. Pawelzik, J. Kohlmorgen, and K.-R. Müller. Annealed competition of experts for a segmentation and classification of switching dynamics. *Neural Computation*, 8 (2):342–358, 1996a.

K. Pawelzik, K.-R. Müller, and J. Kohlmorgen. Prediction of mixtures. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Artificial Neural Networks — ICANN'96*, pages 127–133, Berlin, 1996b. Springer Lecture Notes in Computer Science, Vol. 1112.

K. Pearson. On lines and planes of closest fit to points in space. *Philosophical Magazine*, 2 (sixth series):559–572, 1901.

J. C. Platt. A resource-allocating network for function interpolation. *Neural Computation*, 3(2):213–225, 1991.

J. C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research, 1998.

T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics*, 19: 201–209, 1975.

T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), 1990a.

T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990b.

M. Pontil and A. Verri. Properties of support vector machines. *Neural Computation*, 10:955 – 974, 1997.

M. J. D. Powell. Radial basis functions for multivariable interpolation: A review. In *Algorithms for Approximation, J.C. Mason and M.G. Cox (Eds.)*, pages 143–167. Oxford Clarendon Press, 1987.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing (2nd ed.)*. Cambridge University Press, Cambridge, 1992.

J. C. Principe and J. M. Kuo. Dynamic modeling of chaotic time series with neural networks. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Precessing Systems 7*, San Mateo, CA, 1995. Morgan Kaufmann Publishers.

R.T. Prosser. The ε–Entropy and ε–Capacity of Certain Time–Varying Channels. *Journal of Mathematical Analysis and Applications*, 16:553–573, 1966.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

C. Rasmussen. *Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression.* PhD thesis, Department of Computer Science, University of Toronto, 1996. ftp://ftp.cs.toronto.edu/pub/carl/thesis.ps.gz.

H. J. Ritter, T. M. Martinetz, and K. J. Schulten. *Neuronale Netze: Eine Einführung in die Neuroinformatik selbstorganisierender Abbildungen.* Addison-Wesley, Munich, Germany, 1990.

A. Roy, S. Govil, and R. Miranda. An algorithm to generate radial basis function (RBF)-like nets for classification problems. *Neural Networks*, 8(2):179–202, 1995.

A. Roy, L. S. Kim, and S. Mukhopadhyay. A polynomial time algorithm for the construction and training of a class of multilayer perceptrons. *Neural Networks*, 6:535–545, 1993.

A. Roy and S. Mukhopadhyay. Iterative generation of higher-order nets in polynomial time using linear programming. *IEEE Transactions on Neural Networks*, 8 (2):402–412, 1997.

M. A. Saunders S. S. Chen, D. L. Donoho. Atomic decomposition by basis pursuit. Technical Report Dept. of Statistics Technical Report, Stanford University, 1996.

S. Saitoh. *Theory of Reproducing Kernels and its Applications.* Longman Scientific & Technical, Harlow, England, 1988.

T. D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward network. *Neural Networks*, 2:459–473, 1989.

T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *J. Stat. Phys.*, 65:579–616, 1991.

C. Saunders, M. O. Stitson, J. Weston, L. Bottou, B. Schölkopf, and A. Smola. Support vector machine - reference manual. Technical Report CSD-TR-98-03, Department of Computer Science, Royal Holloway, University of London, Egham, TW20 0EX, UK, 1998. TR available as http://www.dcs.rhbnc.ac.uk/research/compint/areas/comp_learn/sv/pub/report98-03.ps; SVM available at http://svm.dcs.rhbnc.ac.uk/.

R. J. Schalkoff. *Digital Image Processing and Computer Vision.* John Wiley and Sons, Inc., 1989.

R. Schapire, Y. Freund, P. Bartlett, and W. Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 1998. (To appear. An earlier version appeared in: D.H. Fisher, Jr. (ed.), Proceedings ICML97, Morgan Kaufmann.).

M. Schmidt and H. Gish. Speaker identification via support vector classifiers. In *Proc. ICASSP '96*, pages 105–108, Atlanta, GA, 1996.

I. Schoenberg. Positive definite functions on spheres. *Duke Math. J.*, 9:96–108, 1942.

B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In U. M. Fayyad and R. Uthurusamy, editors, *Proceedings, First International Conference on Knowledge Discovery & Data Mining.* AAAI Press, Menlo Park,

CA, 1995.

B. Schölkopf. *Support Vector Learning*. R. Oldenbourg Verlag, Munich, 1997.

B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson. Support vector regression with automatic accuracy control. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, Berlin, 1998a. Springer Verlag. In press.

B. Schölkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Artificial Neural Networks — ICANN'96*, pages 47 – 52, Berlin, 1996a. Springer Lecture Notes in Computer Science, Vol. 1112.

B. Schölkopf, P. Knirsch, A. Smola, and C. Burges. Fast approximation of support vector kernel expansions, and an interpretation of clustering as approximation in feature spaces. In *20. DAGM Symposium Mustererkennung*, Lecture Notes in Computer Science, Berlin, 1998b. Springer. To appear.

B. Schölkopf, S. Mika, A. Smola, G. Rätsch, and K.-R. Müller. Kernel PCA pattern reconstruction *via* approximate pre-images. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, Berlin, 1998c. Springer Verlag. In press.

B. Schölkopf, P. Simard, A. Smola, and V. Vapnik. Prior knowledge in support vector kernels. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems 10*, Cambridge, MA, 1998d. MIT Press.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. Technical Report 44, Max-Planck-Institut für biologische Kybernetik, 1996b.

B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Artificial Neural Networks — ICANN'97*, pages 583 – 588, Berlin, 1997a. Springer Lecture Notes in Computer Science, Vol. 1327.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299 – 1319, 1998e.

B. Schölkopf, A. Smola, K.-R. Müller, C. Burges, and V. Vapnik. Support vector methods in learning and feature extraction. In T. Downs, M. Frean, and M. Gallagher, editors, *Proceedings of the Ninth Australian Conference on Neural Networks*, pages 72 – 78, Brisbane, Australia, 1998f. University of Queensland.

B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans. Sign. Processing*, 45:2758 – 2765, 1997b.

J. C. Schouten, F. Takens, and C. M. van den Bleek. Estimation of the dimension of a noisy attractor. *Physical Review E*, 50(3):1851–1860, 1994.

J. Schürmann. *Pattern Classification: a unified view of statistical and neural approaches*. Wiley, New York, 1996.

D.W Scott. *Multivariate Density Estimation*. Wiley-Interscience, New York, 1992.

H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992.

J. Shawe-Taylor, P. Bartlett, R. Williamson, and M. Anthony. A framework for structural risk minimization. In *COLT*, 1996.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 1998. To appear. Also: NeuroCOLT Technical Report NC-TR-96-053, 1996, ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports.

J. Shawe-Taylor and N. Cristianini. Data-dependent structural risk minimisation for perceptron decision trees. In *Advances in Neural Information Processing Systems 10*, 1998.

P. Simard, Y. LeCun, and J. Denker. Efficient pattern recognition using a new transformation distance. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 50–58, San Mateo, CA, 1993. Morgan Kaufmann.

A. Skorokhod and M. Yadrenko. On absolute continuity of measures corresponding to homogeneous Gaussian fields. *Theory of Probability and its Applications*, XVIII:27–40, 1973.

F. W. Smith. Pattern classifier design by linear programming. *IEEE Transactions on Computers*, C-17:367–372, 1968.

A. Smola and B. Schölkopf. From regularization operators to support vector kernels. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems 10*, Cambridge, MA, 1998a. MIT Press.

A. Smola and B. Schölkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 1998b. In press.

A. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998a.

A. Smola, B. Schölkopf, and K.-R. Müller. Convex cost functions for support vector regression. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, Berlin, 1998b. Springer Verlag. In press.

A. Smola, B. Schölkopf, and K.-R. Müller. General cost functions for support vector regression. In T. Downs, M. Frean, and M. Gallagher, editors, *Proc. of the Ninth Australian Conf. on Neural Networks*, pages 79 – 83, Brisbane, Australia, 1998c. University of Queensland.

J. Stewart. Positive definite funcions and generalizations, an historical survey. *Rocky Mountain Journal of Mathematics*, 6(3):409–434, 1978.

M. Stitson, A. Gammerman, V. Vapnik, V. Vovk, C. Watkins, and J. Weston. Support vector regression with ANOVA decomposition kernels. Technical Report CSD-TR-97-22, Royal Holloway, University of London, 1997.

F. Takens. Detecting strange attractors in fluid turbulence. In D. Rand and L.S. Young, editors, *Dynamical Systems and Turbulence*, pages 366–381. Springer-Verlag, Berlin, 1981.

M. Talagrand. The Glivenko–Cantelli problem, ten years later. *Journal of Theoretical Probability*, 9(2):371–384, 1996.

W. Thomas. Database marketing: Dual approach outdoes response modeling. *Database Marketing News*, page 26, 1996.

R. Vanderbei. Loqo: An interior point code for quadratic programming. Technical Report SOR 94-15, Princeton University, 1994.

R. J. Vanderbei. LOQO user's manual – version 3.10. Technical Report SOR-97-08, Princeton University, Statistics and Operations Research, 1997. Code available at http://www.princeton.edu/~rvdb/.

V. Vapnik. *Estimation of Dependences Based on Empirical Data [in Russian].* Nauka, Moscow, 1979. (English translation: Springer Verlag, New York, 1982).

V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.

V. Vapnik. Structure of statistical learning theory. In A. Gammerman, editor, *Computational and Probabalistic Reasoning*, chapter 1. Wiley, Chichester, 1996.

V. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998. forthcoming.

V. Vapnik and A. Chervonenkis. A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964.

V. Vapnik and A. Chervonenkis. Uniform convergence of frequencies of occurence of events to their probabilities. *Dokl. Akad. Nauk SSSR*, 181:915 – 918, 1968.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2): 264–280, 1971.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian].* Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin, 1979).

V. Vapnik and A. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications*, 26(3):532–553, 1981.

V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 281–287, Cambridge, MA, 1997. MIT Press.

V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method.

Automation and Remote Control, 24, 1963.

V. Vapnik, E. Levin, and Y. Le Cun. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6(5):851–876, 1994.

N. Ya. Vilenkin. *Special Functions and the Theory of Group Representations*, volume 22 of *Translations of Mathematical Monographs*. American Mathematical Society Press, Providence, NY, 1968.

M. Villalobos and G. Wahba. Inequality constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Am. Statist. Assoc.*, 82:239–248, 1987.

G. Wahba. Convergence rates of certain approximate solutions to Fredholm integral equations of the first kind. *Journal of Approximation Theory*, 7:167 – 185, 1973.

G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Stat. Soc. Ser. B*, 40:364–372, 1978.

G. Wahba. Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Stat. Comput.*, 2:5–16, 1981.

G. Wahba. Constrained regularization for ill posed linear operator equations, with applications in meteorology and medicine. In S. Gupta and J. Berger, editors, *Statistical Decision Theory and Related Topics, III, Vol.2*, pages 383–418. Academic Press, 1982a.

G. Wahba. Erratum: Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Stat. Comput.*, 3:385–386, 1982b.

G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, 13:1378–1402, 1985a.

G. Wahba. Multivariate thin plate spline smoothing with positivity and other linear inequality constraints. In E. Wegman and D. dePriest, editors, *Statistical Image Processing and Graphics*, pages 275–290. Marcel Dekker, 1985b.

G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

G. Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol XII*, pages 95–112. Addison-Wesley, 1992.

G. Wahba, D. Johnson, F. Gao, and J. Gong. Adaptive tuning of numerical weather prediction models: randomized GCV in three and four dimensional data assimilation. *Mon. Wea. Rev.*, 123:3358–3369, 1995a.

G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Structured machine learning for 'soft' classification with smoothing spline ANOVA and stacked tuning, testing and evaluation. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 415–422. Morgan Kauffman, 1994.

G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, 23:1865–1895, 1995b.

T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65:499–556, 1993.

A. S. Weigend and N. A. Gershenfeld (Eds.). *Time Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley, 1994. Santa Fe Institute Studies in the Sciences of Complexity.

P. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* PhD thesis, Harvard, 1974.

J. Werner. *Optimization - Theory and Applications.* Vieweg, 1984.

J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, Egham, TW20 0EX, UK, 1998.

H. Widom. Asymptotic behaviour of eigenvalues of certain integral operators. *Archive for Rational Mechanics and Analysis*, 17:215–229, 1964.

R. A. Wilkinson, J. Geist, S. Janet, P. J. Grother, C. J. C. Burges, R. Creecy, B. Hammond, J. J. Hull, N. J. Larsen, T. P. Vogl, and C. L. Wilson. The first census optical character recognition system conference. Technical Report NISTIR 4912, National Institute of Standards and Technology (NIST), Gaithersburg, 1992.

C. K. I. Williams. Computation with infinite networks. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, Cambridge, MA, 1997. MIT Press.

C. K. I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan, editor, *Learning and Inference in Graphical Models.* Kluwer, 1998. To appear. Also: Technical Report NCRG/97/012, Aston University.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Technical Report NC-TR-98-019, Royal Holloway College, University of London, UK, 1998a.

R. C. Williamson, A. J. Smola, and B. Schölkopf. A Maximum Margin Miscellany. Typescript, 1998b.

W. H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences,U.S.A.*, 87:9193–9196, 1990.

A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano. Determining lyapunov exponents from a time series. *Physica D*, 16:285–317, 1985.

D. Xiang. *Model Fitting and Testing for Non-Gaussian Data with a Large Data Set.* PhD thesis, Technical Report 957, University of Wisconsin-Madison, Madison

WI, 1996.

D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, 6:675–692, 1996.

D. Xiang and G. Wahba. Approximate smoothing spline methods for large data sets in the binary case. Technical Report 982, Department of Statistics, University of Wisconsin, Madison WI, 1997. To appear in the Proceedings of the 1997 ASA Joint Statistical Meetings, Biometrics Section, pp 94-98 (1998).

P. V. Yee. *Regularized Radial Basis Function Netowrks: Theory and Applications to Probability Estimation, Classification, and Time Series Prediction*. PhD thesis, Dept. of ECE, McMaster University, Hamilton, Canada, 1998.

E. C. Zachmanoglou and Dale W. Thoe. *Introduction to Partial Differential Equations with Applications*. Dover, Mineola, N.Y., 1986.

X. Zhang and J. Hutchinson. Simple architectures on fast machines: practical issues in nonlinear time series prediction. In A. S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*. Santa Fe Institute, Addison-Wesley, 1994.

G. Zoutendijk. *Methods of Feasible Directions: a Study in Linear and Non-line ar Programming*. Elsevier, 1970.