Multicategory Classification by Support Vector Machines

Erin J. Bredensteiner
Department of Mathematics
University of Evansville
1800 Lincoln Avenue
Evansville, Indiana 47722
eb6@evansville.edu

Kristin P. Bennett
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12180
bennek@rpi.edu

Abstract

We examine the problem of how to discriminate between objects of three or more classes. Specifically, we investigate how two-class discrimination methods can be extended to the multiclass case. We show how the linear programming (LP) approaches based on the work of Mangasarian and quadratic programming (QP) approaches based on Vapnik's Support Vector Machines (SVM) can be combined to yield two new approaches to the multiclass problem. In LP multiclass discrimination, a single linear program is used to construct a piecewise linear classification function. In our proposed multiclass SVM method, a single quadratic program is used to construct a piecewise nonlinear classification function. Each piece of this function can take the form of a polynomial, radial basis function, or even a neural network. For the $k > 2$ class problems, the SVM method as originally proposed required the construction of a two-class SVM to separate each class from the remaining classes. Similarily, k two-class linear programs can be used for the multiclass problem. We performed an empirical study of the original LP method, the proposed k LP method, the proposed single QP method and the original k QP methods. We discuss the advantages and disadvantages of each approach.

1 Introduction

We investigate the problem of discriminating large real-world datasets with more than two classes. Given examples of points known to come from $k > 2$ classes, we construct a function to discriminate between the classes. The goal is to select a function that will efficiently and correctly classify future points. This classification technique can be used for data mining or pattern recognition. For example, the United States Postal Service is interested in an efficient yet accurate method of classifying zipcodes. Actual handwritten digits from zipcodes collected by the United States Postal Service are used in our study. Each digit is represented by a 16 by 16 pixel grayscale map, resulting in 256 attributes for each sample number. Given the enormous quantities of mail the Postal Service sorts each day, the accuracy and efficiency in evaluation are extremely important.

In this paper, we combine two independent but related research directions developed for solving the two-class linear discrimination problem. The first is the linear programming (LP) methods stemming from the Multisurface Method of Mangasarian [12, 13]. This method and it's later extension the Robust Linear Programming (RLP) approach [6] have been used in a highly successfully breast cancer diagnosis system [26]. The second direction is the quadratic programming (QP) methods based on Vapnik's Statistical Learning Theory [24, 25]. Statistical Learning Theory addresses mathematically the problem of how to best construct functions that generalize well on future points. The problem of constructing the best linear two-class discriminant can be posed as a convex quadratic program with linear constraints. The resulting linear discriminant is known as a Support Vector Machine (SVM) because it is a function of a subset of the training data known as *support vectors*. Specific implementations such as the Generalized Optimal Plane (GOP) method has proven to perform very well in practice [8]. Throughout this paper we will refer to the two different approaches as RLP and SVM.

The primary focus of this paper is how the the two research directions have differed in their approach to solving problems with $k > 2$ classes. The original SVM method for multiclass problems was to find k separate two-class discriminants [23]. Each discriminant is constructed by separating a single class from all the others. This process requires the solution of k quadratic programs. When applying all k classifiers to the original multicategory dataset, 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 directly construct k classification functions such that for each point the corresponding class function is maximized [5, 6]. The Multicategory Discrimination Method [5, 6] constructs a piecewise-linear discriminate for the k- class problem using a single linear program. We will call this method M-RLP since it is a direction extension of the RLP approach. We will show how these two different approaches can be combined two yield two new methods: k-RLP, and M-SVM.

In Section 2, we will provide background on the existing RLP and SVM

methods. While the k-class cases are quite different, the two-class linear discrimination methods for SVM and RLP are almost identical. They differ only in the regularization term used in the objective. We use the regularized form of RLP proposed in [3] which is equivalent to SVM except that a different norm is used for the regularization term. For two-class linear discrimination, RLP generalizes equally well and is more computationally efficient than SVM. RLP exploits the fact that state-of-the-art LP codes are far more efficient and reliable than QP codes.

The primary appeal of SVM is that they can be simply and elegantly applied to nonlinear discrimination. With only minor changes, SVM methods can construct a wide class of two-class nonlinear discriminants by solving a single QP [24]. The basic idea is that the points are mapped nonlinearly to a higher dimensional space. Then the dual SVM problem is used to construct a linear discriminant in the higher dimensional space that is nonlinear in the original attribute space. By using kernel functions in the dual SVM problem, SVM can efficiently and effectively construct many types of nonlinear discriminant functions including polynomial, radial basis function machine, and neural networks. The successful polynomial-time nonlinear methods based on LP use a multi-step approaches. The methods of Roy *et al* [20, 19, 18] use clustering in conjuction with LP to generate neural networks in polynomial time. Another approach is to recursively construct piecewise-linear discriminants using a series of LP's [13, 2, 15]. These approaches could also be used with SVM but we limit discussion to nonlinear discriminants constructed using the SVM kernel-type approaches.

After the introduction to the existing multiclass methods, M-RLP and k-SVM, we will show how same idea used in the M-RLP, can be adapted to construct multiclass SVM using a single quadratic program. We adapt a problem formulation similar to the two-class case. In the two-class case, initially the problem is to construct a linear discriminant. The data points are then transformed to a higher dimensional feature space. A linear discriminant is constructed in the higher dimension space. This results in a nonlinear classification function in the original feature space. In Section 3, for the $k > 2$ class case, we begin by constructing a piecewise-linear discriminant function. A regularization term is added to avoid overfitting. This method is then extended to piecewise-nonlinear classification functions in Section 4. The variables are mapped to a higher dimensional space. Then a piecewise-linear discriminant function is constructed in the new space. This results in a piecewise-nonlinear discriminant in the original space. In Section 5, we extend the method to piecewise inseparable datasets. We call the final approach the Multicategory Support Vector Machine (M-SVM). Depending on the choice of transformation, the pieces may be polynomials, radial basis functions, neural networks, etc. We concentrate our research on the polynomial classifier and leave the computational investigation other classification functions as future work. Figure 1 shows a piecewise-second-degree polynomial separating three classes in two dimensions.

M-SVM requires the solution of a very large quadratic program. When transforming the data points into a higher dimension feature space, the number

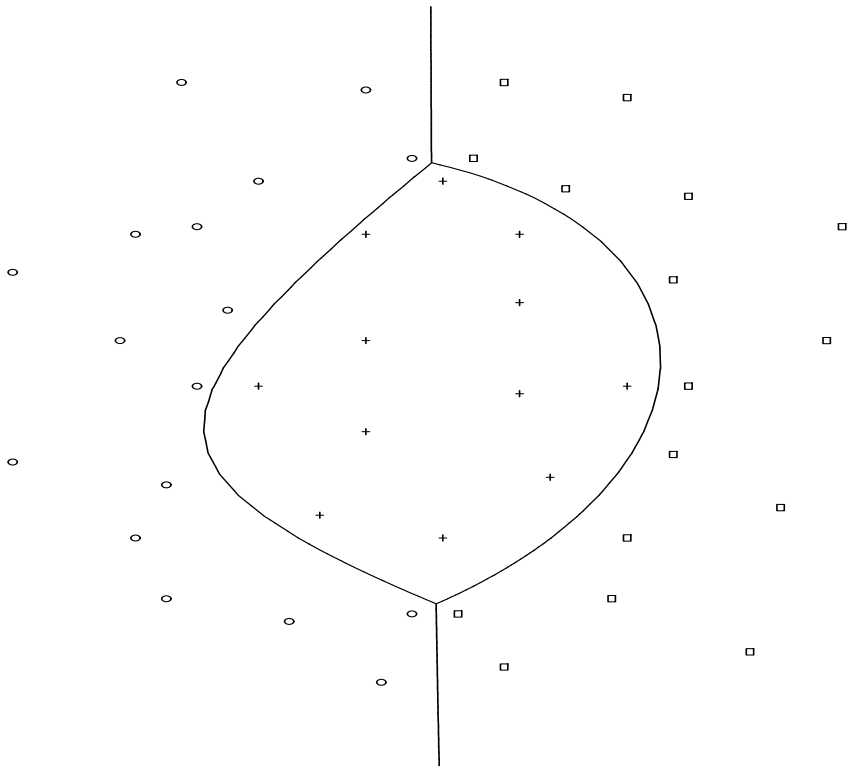

Figure 1: Piecewise-polynomial separation of three classes in two dimensions

of variables will grow exponentially. For example, a second degree polynomial classifier in two dimensions requires the original variables x_1 and x_2 as well as the variables $x_1^2,\ x_2^2$, and x_1x_2. In the primal problem, the problem size will explode as the degree of the polynomial increases. The dual problem, however, remains tractable. The number of dual variables is $k-1$ times the number of points regardless of what transformation is selected. In the dual problem, the transformation appears as an inner product in the high dimensional space. Inexpensive techniques exist for computing these inner products. Each dual variable corresponds to a point in the original feature space. A point with a corresponding positive dual variable is referred to as a support vector. The goal is to maintain a high accuracy while using a small number of support vectors. Minimizing the number of support vectors is important for generalization and also for reducing the computational time required to evaluate new examples. Section 6 contains computational results comparing the two LP approaches k-RLP and M-RLP; and the the two QP approaches k-SVM and M-SVM. The methods were compared in terms of generalization (testing set accuracy), number of support vectors, and computational time.

The following notation will be used throughout this paper. Mathematically we can abstract the problem as follows: Given the elements of the sets, $\mathcal{A}^i, i = 1,\ldots,k$, in the n-dimensional real space R^n, construct a a discriminant function is determined which separates these points into distinct regions. Each region should contains points belonging to all or almost all of the same class. Let $\mathcal{A}^j$ be a set of points in the n-dimensional real space R^n with cardinality m_j. Let A^j be an $m_j \times n$ matrix whose rows are the points in $\mathcal{A}^j$. The i^{th} point in $\mathcal{A}^j$ and the i^{th} row of A^j are both denoted A_i^j. Let e denote a vector of ones of the appropriate dimension. The scalar 0 and a vector of zeros are both represented by 0. Thus, for $x \in R^n$, $x > 0$ implies that $x_i > 0$ for $i = 1,\ldots,n$. Similarly, $x \geq y$ implies that $x_i \geq y_i$ for $i = 1,\ldots,n$. The set of minimizers of $f(x)$ on the set $\mathcal{S}$ is denoted by $arg\min_{x\in\mathcal{S}} f(x)$. For a vector x in R^n, x_+ will denote the vector in R^n with components $(x_+)_i := max\{x_i, 0\},\ i = 1,\ldots,n$. The step function x_* will denote the vector in $[0,1]^n$ with components $(x_*)_i := 0$ if $(x)_i \leq 0$ and $(x_*)_i := 1$ if $(x)_i > 0,\ i = 1,\ldots,n$. For the vector x in R^n and the matrix A in $R^{n\times m}$, the transpose of x and A are denoted x^T and A^T respectively. The dot product of two vectors x and y will be denoted x^Ty and $(x \cdot y)$.

2 Background

This section contains a brief overview of the RLP and SVM methods for classification. First we will discuss the two-class problem using a linear classifier. Then SVM for two classes will be defined. Then RLP will be reviewed. Finally, the piecewise-linear function used for multicategory classification in M-RLP will be reviewed.

2.1 Two Class Linear Discrimination

Commonly, the method of discrimination for two classes of points involves determining a linear function that consists of a linear combination of the attributes of the given sets. In the simplest case, a linear function can be used to separate two sets as shown in Figure 2. This function is the separating plane $x^T w = \gamma$

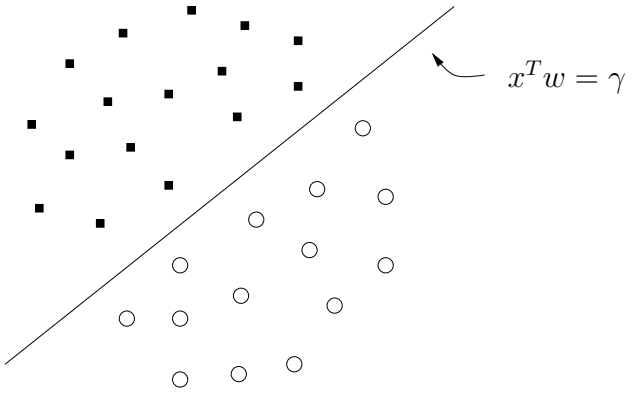

Figure 2: Two linearly separable sets and a separating plane

where w is the normal to the plane and γ is the distance from the origin. Let $\mathcal{A}^1$ and $\mathcal{A}^2$ be two sets of points in the n-dimensional real space R^n with cardinality m_1 and m_2 respectively. Let A^1 be an $m_1 \times n$ matrix whose rows are the points in $\mathcal{A}^1$. Let A^2 be an $m_2 \times n$ matrix whose rows are the points in $\mathcal{A}^2$. Let $x \in R^n$ be a point to be classified as follows:

$$\begin{array}{lll} x^T w - \gamma > 0 & \Rightarrow & \mathrm{x} \in \mathcal{A}^1 \\ x^T w - \gamma < 0 & \Rightarrow & \mathrm{x} \in \mathcal{A}^2 \end{array} \tag{1}$$

The two sets of points, $\mathcal{A}^1$ and $\mathcal{A}^2$, are linearly separable if

$$A^1 w > \gamma e \qquad \gamma e > A^2 w \tag{2}$$

where e is a vector of ones of the appropriate dimension. If the two classes are linear separable, there are infinitely many planes that separate the two classes. The goal is two choose the plane that will generalize best on future points.

Both Mangasarian [12] and Vapnik and Chervonenkis [25] concluded that the best plane in the separable case is the one that minimizes the distance of the closest vector in each class to the separating plane. For the separable case the formulations of Mangasarian's Multi-surface Method of Pattern Recognition [13] and those of Vapnik's Optimal Hyperplane [24, 25] are very similar [3]. We will concentrate on the Optimal Hyperplane problem since it the basis of SVM, and it is validated theoretically by Statistical Learning Theory [24]. According to Statistical Learning Theory, the Optimal Hyperplane can construct linear

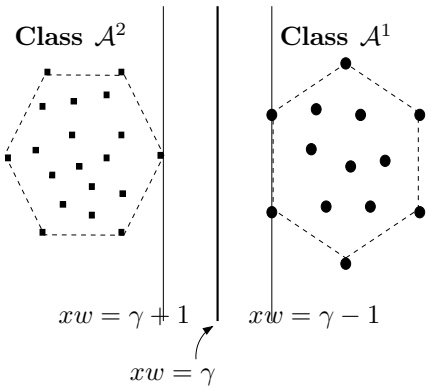

Figure 3: Two supporting planes and the resulting optimal separating plane

discriminants in very high dimensional spaces without overfitting. The reader should consult [24] for full details of Statistical Learning Theory not covered in this paper.

The problem in the canonical form of Vapnik [24] becomes to determine two parallel planes $xw = \gamma + 1$ and $xw = \gamma - 1$ such that

$$A^1 w - \gamma e - e \geq 0 \qquad -A^2 w + \gamma e - e \geq 0. \tag{3}$$

and the margin or distance between the two planes is maximized. The margin of seperation between the two supporting planes is $\frac{2}{\|w\|}$. An example of such a plane is shown in Figure 3. The problem of finding the maximum margin becomes[24]:

$$\begin{array}{ll} \min\limits_{w,\gamma} & \frac{1}{2} w^T w \\ s.t. & A^1 w - \gamma e - e \geq 0 \\ & -A^2 w + \gamma e - e \geq 0. \end{array} \tag{4}$$

In general it is not always possible for a single linear function to completely separate two given sets of points. Thus, it is important to find the linear function that discriminates best between the two sets according to some error minimization criterion. Bennett and Mangasarian [4] minimize the average magnitude of the misclassification errors in the construction of their following robust linear programming problem (RLP).

$$\begin{array}{ll} \min\limits_{w,\gamma,y,z} & \frac{1}{\delta_1} e^T y + \frac{1}{\delta_2} e^T z \\ subject\ to & y + A^1 w - \gamma e - e \geq 0 \\ & z - A^2 w + \gamma e - e \geq 0 \\ & y \geq 0, \quad z \geq 0 \end{array} \tag{5}$$

where $\delta_1 > 0$ and $\delta_2 > 0$ are the misclassification costs. To avoid the null solution $w = 0$, use $\delta_1 = \frac{1}{m_1}$ and $\delta_2 = \frac{1}{m_2}$ where m_1 and m_2 are the cardinalities of $\mathcal{A}^1$ and $\mathcal{A}^2$ respectively. The RLP method is very effective in practice.

The functions generated by RLP generalize well on many real-world problems. Additionally, the computational time is reasonably small because its solution involves only a single linear program. Note however that the RLP method no longer includes any notion of maximizing the margin. Statistical Learning Theory indicates that the maximizing the margin is essential for good generalization.

The SVM approach [8, 23] is a multiobjective quadratic program which minimizes the absolute misclassification errors, and maximizing the separation margin by minimizing $\|w\|^2$.

$$\begin{array}{ll} \min\limits_{w,y,z,\gamma} & (1-\lambda)(e^T y + e^T z) + \frac{\lambda}{2} w^T w \\ s.t. & A^1 w - \gamma e + y - e \geq 0 \\ & -A^2 w + \gamma e + z - e \geq 0 \\ & y \geq 0 \quad z \geq 0 \end{array} \tag{6}$$

where $0 < \lambda < 1$ is a fixed constant. Note that Problem 6 is equivalent to RLP with the addition of a regularization term $\frac{\lambda}{2} w^T w$, and $\delta_1 = \delta_2 = 1$.

A linear programming version of (6) can be constructed by replacing the norm used to minimize the weights w [3]. Recall that the SVM objective minimizes the square of the 2-norm of w, $\|w\|^2 = w^T w$. The 1-norm of w, $\|w\|_1 = e^T |w|$, can be used instead. The absolute value function can be removed by introducing the variable s and the constraints $-s \leq w \leq s$. The SVM objective is then modified by substituting $e^T s$ for $\frac{w^T w}{2}$. At optimality, $s_i = |w_i|,\ i = 1, \ldots, k$. The resulting LP is:

$$\begin{array}{ll} \min\limits_{w,\alpha,\beta,y,z,s} & (1-\lambda)(\frac{1}{m_1} e^T y + \frac{1}{m_2} e^T z) + \lambda e^T s \\ s.t. & A^1 w - \gamma e + y \geq 0 \\ & -A^2 w + \gamma e + z \geq 0 \\ & -s \leq w \leq s \\ & y \geq 0 \quad z \geq 0 \quad s \geq 0. \end{array} \tag{7}$$

We will refer to this problem as RLP since $\lambda = 0$ yields the original RLP method.

As in the SVM method, the RLP method minimizes both the average distance of the misclassified points from the relaxed supporting planes and the maximum classification error. The main advantage of the RLP method over the SVM problem is that RLP is a linear program solvable using very robust algorithms such as the Simplex Method [17]. SVM requires the solution of quadratic program that is typically much more computationally costly for the same size problem. In [3], the RLP method was found to generalize as well as the linear SVM but with much less computational cost.

It is more efficient computationally to solve the dual RLP and SVM problems. The dual RLP problem is

$$\begin{array}{ll} \min\limits_{u,v} & e^T u + e^T v \\ s.t. & -\lambda e \leq u^T A^1 - v^T A^2 \leq \lambda e \\ & e^T u - e^T v = 0 \\ & 0 \leq u \leq (1-\lambda)\delta_1 \quad 0 \leq v \leq (1-\lambda)\delta_2 \end{array} \tag{8}$$

In this paper we use $\delta_1 = \frac{1}{m}$ and $\delta_2 = \frac{1}{k}$ but δ_1 and δ_2 may be any positive weights for the misclassification costs. The dual SVM problem and its extension to nonlinear discriminants is given in the next section.

2.2 Nonlinear Classifiers Using Support Vector Machines

The primary advantage of the SVM (6) over RLP (7) is that in its dual form it can be used to construct nonlinear discriminants. using polynomial separators, radial basis functions, neural networks, etc. The basic idea is to map the original problems to a higher dimensional space and then to construct a linear discriminant in a higher dimensional space that corresponds to a linear discriminant in the original space. So for example, to construct a quadratic discriminant for a two dimensional problems, the input attributes $[x_1, x_2]$ are mapped into $[x_1^2, x_1^2, \sqrt{2}x_1x_2, x_1, x_2]$ and a linear discriminant function is constructed in the new five-dimensional space. Two examples of possible polynomial classifiers are given in Figure 4. The dual SVM is applied to the mapped points. The regularization term in the primal objective helps avoid overfitting the higher dimensional space. The dual SVM provides a practical computational approach through the use of generalized inner products or kernels.

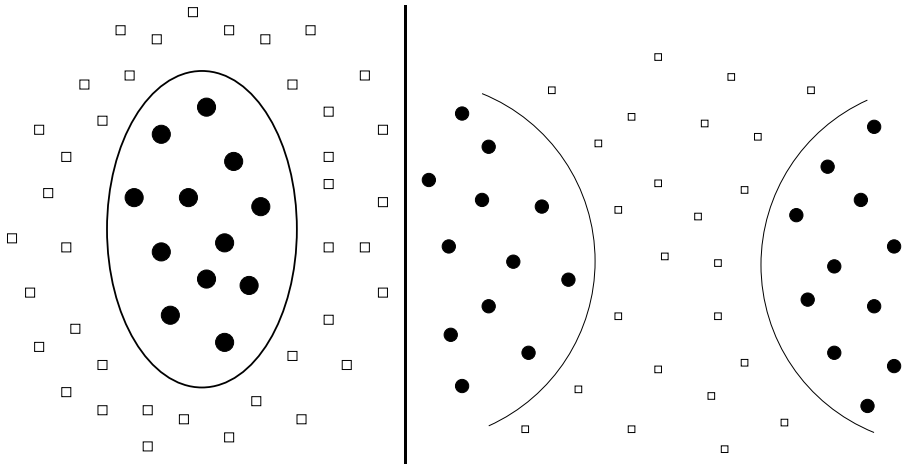

Figure 4: Two examples of second degree polynomial separations of two sets

The dual SVM is as follows: as follows:

$$\begin{array}{ll} \min\limits_{u,v} & \frac{1}{2\lambda}\left\|{A^1}^T u - {A^2}^T v\right\|^2 - e^T u - e^T v \\ s.t. & e^T u = e^T v \\ & (1-\lambda)e \geq u \geq 0 \quad (1-\lambda)e \geq v \geq 0. \end{array} \tag{9}$$

To formulate the nonlinear case it is convenient to rewrite the problem in summation notation. Let $\mathcal{A}$ be the set of all points $\mathcal{A}^1$ and $\mathcal{A}^2$. Define $M = m_1 +$

m_2 to be the total number of points. Let $\alpha^T = [\alpha_1, \alpha_2, \ldots, \alpha_M] = [\frac{1}{\lambda}u^T \ \frac{1}{\lambda}v^T]$. Let $t \in R^M$ be such that for $x_i \in \mathcal{A}$ $t_i = \begin{cases} 1 & x_i \in A^1 \\ -1 & x_i \in A^2 \end{cases}$.

To construct the nonlinear classification function, the original data points x are transformed to the higher dimension feature space by the function $\phi(x) : R^n \rightarrow R^{n'}$, $n' >> n$. The dot product of the original vectors $x_i^T x_j$ is replaced by the dot product of the transformed vectors $(\phi(x_i) \cdot \phi(x_j))$.

The first term of the objective function can then be written as the sum:

$$\frac{\lambda}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} t_i t_j \alpha_i \alpha_j (\phi(x_i) \cdot \phi(x_j)).$$

Using this notation and simplifying the problem becomes:

$$\begin{array}{ll} \min\limits_{\alpha} & \frac{1}{2}\sum\limits_{i=1}^{M}\sum\limits_{j=1}^{M} t_i t_j \alpha_i \alpha_j (\phi(x_i) \cdot \phi(x_j)) - \sum\limits_{i=1}^{M} \alpha_i \\ s.t. & \sum\limits_{i=1}^{M} \alpha_i t_i = 0 \\ & \frac{(1-\lambda)}{\lambda} e \geq \alpha \geq 0. \end{array} \tag{10}$$

In the support vector machine (SVM), Vapnik replaces the inner product $(\phi(x) \cdot \phi(x_i))$ with the inner product in the Hilbert space $K(x, x_i)$. This symmetric function $K(x, x_i)$ must satisfy Theorem 5.3 in [23]. This theorem ensures $K(x, x_i)$ is an inner product in some feature space. The choice of $K(x, x_i)$ determines the type of classifier that is constructed. Possible choices include polynomial classifiers as in Figure 4 ($K(x, x_i) = (x^T x_i + 1)^d$, where d is the degree of the polynomial), radial basis function machines ($K_\gamma(|x - x_i|) = \exp\{-\gamma|x - x_i|^2\}$ where $|x - x_i|$ is the distance between two vectors and γ is the width parameter), and two-layer neural networks ($K(x, x_i) = S[v(x^T x_i) + c]$ where $S(u)$ is a sigmoid function) [23]. Variants of SVM (10) have proven to be quite successful in paractice [21, 22, 7].

Note that the number of variables in Program (10) remains constant as $K(x, x_i)$ increases in dimensionality. Additionally, the objective function remains quadratic and thus the complexity of the problem does not increase. In fact, the size of the problem is dependent on the number of nonzero dual variables α_i. The points x_i corresponding to these variables are called the *support vectors*. According to Statistical Learning Theory, the best solution for a given misclassification error uses the minimum number of support vectors.

The final classification function with the generalized kernel function $K(x, x_i)$ is:

$$f(x) = \text{sign}\left(\sum_{\text{support vectors}} t_i \alpha_i K(x, x_i) - \gamma\right) \tag{11}$$

where $x \in \mathcal{A}^1$ if $f(x) = 1$, otherwise $x \in \mathcal{A}^2$.

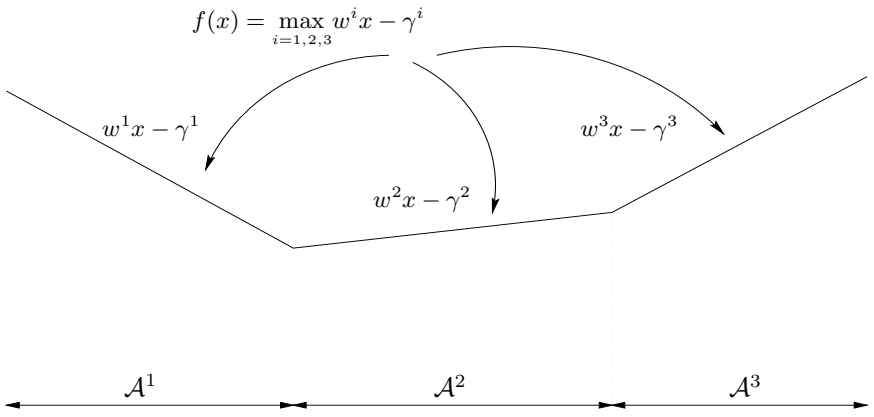

Figure 5: Piecewise-linear separation of sets $\mathcal{A}^1, \mathcal{A}^2$, and $\mathcal{A}^3$ by the convex piecewise-linear function $f(x)$.

2.3 Multicategory Discrimination

In multicategory classification a piecewise-linear separator is used to discriminate between $k > 2$ classes of m^i, $i = 1, \ldots, k$, points. We will examine two methods for accomplishing this. The first used in SVM [24] is two construct a discriminate function to separate one class from the remaining $k - 1$ classes. This is process is repeated k times. In the separable case, the linear discriminant for each class must satisfy the following set of inequalities. Find $(w^1, \gamma^1), \ldots, (w^k, \gamma^k)$, such that

$$A^i w^i - \gamma^i > A^j w^i - \gamma^i, \quad i, j = 1, \ldots, k, \quad i \neq j. \tag{12}$$

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

$$f_i(x) = x^T w^i - \gamma^i \tag{13}$$

is maximized. Figure 5 shows a piecewise-linear function $f(x) = \max_{i=1,2,3} f_i(x)$ on R that separates three sets.

Note either SVM (10) or RLP can be used to construct the k two-class discriminants. For clarity, we will call this method used with SVM (10), k-SVM. We will denote this method used with RLP (8), k-SVM. The advantage of k-SVM is that it can used for piecewise-nonlinear discriminants which k-RLP is limited to piecewise-linear discriminants. For both k-SVM and k-RLP to attain perfect training set accuracy, following inequalities must be satisfied:

$$A^i w^i - \gamma^i > A^j w^j - \gamma^j, \quad i, j = 1, \ldots, k, \quad i \neq j$$

This inequality can be used as a definition of piecewise-linear separability.

Definition 2.1 (Piecewise-linear Separability) *The sets of points* $\mathcal{A}^i,\ i = 1,\ldots,k$, *represented by the matrices* $A^i \in R^{m_i \times n},\ i = 1,\ldots,k$, *are piecewise-linearly separable if there exist* $w^i \in R^n$ *and* $\gamma^i \in R, i = 1,\ldots,k$, *such that*

$$A^i w^i - \gamma^i e > A^i w^j - \gamma^j e,\ i,\ j = 1,\ldots,k,\ i \neq j. \tag{14}$$

Equivalent to Definition 2.1, finding the piecewise-linear separator involves solving the equation $A^i w^i - \gamma^i e \geq A^i w^j - \gamma^j e + e,\ i,\ j = 1,\ldots,k,\ i \neq j$. This can be rewritten as $0 \geq -A^i(w^i - w^j) + (\gamma^i - \gamma^j)e + e,\ i,\ j = 1,\ldots,k,\ i \neq j$. Figure 6 shows an example of a piecewise-linear separator for three classes in two dimensions. The linear separating functions are represented by the quantities

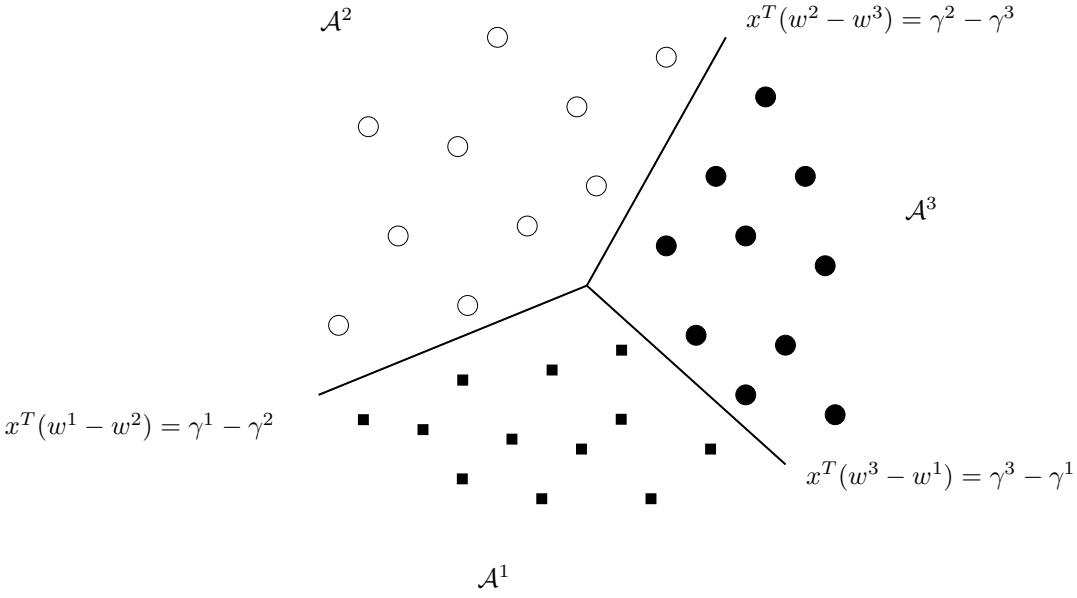

Figure 6: Three classes separated by a piecewise-linear function.

$(w^i - w^j, \gamma^i - \gamma^j),\ i,\ j = 1,\ldots,k,\ j \neq i$, where $w^i \in R^{n\times 1}$ and $\gamma^i \in R^1,\ i = 1,\ldots,k$.

The M-RLP method[1] proposed and investigated in [5, 6] can be used to find $(w_i,\ \gamma_i),\ i = 1,\ldots,k$ satisfying Definition 2.1.

$$\min_{w^i,\gamma^i,y^{ij}} \left\{ \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j\neq i}}^{k} \frac{e^T y^{ij}}{m^i} \;\middle|\; \begin{array}{c} y^{ij} \geq -A^i(w^i - w^j) + (\gamma^i - \gamma^j)e + e, y^{ij} \geq 0, \\ i \neq j,\ i,\ j = 1,\ldots,k \end{array} \right\} \tag{15}$$

where $y^{ij} \in R^{m_i \times 1}$. In M-RLP (15), if the optimal objective value is zero, then the dataset is piecewise-linearly separable. If the dataset is not piecewise-linearly separable, the positive values of the variables y_l^{ij} are proportional to the

[1]The method was originally called Multicategory Discrimination

magnitude of the misclassified points from the plane $x^T(w^i - w^j) = (\gamma^i - \gamma^j) + 1$. This program (15) is a generalization of the two-class RLP linear program (5) to the multicategory case. Like the original RLP (5) M-RLP does not include any terms for maximizing the margin and it does not directly permit the use of generalized inner products or kernels to allow extension to the nonlinear case. So in the next section we will show how M-RLP and SVM can be combined by including margin maximization and generalized inner products into M-RLP.

3 Formulation of M-SVM: Piecewise-linear Separable Case

We now propose to construct piecewise-linear and piecewise-nonlinear SVM using a single quadratic program. Analogous to the two class case we start by formulating the "optimal" piecewise-linear separator for the separable case. Assume that the k sets of points are piecewise-linearly separable, i.e., there exist $w^i \in R^n$ and $\gamma^i \in R, i = 1, \ldots, k$, such that

$$A^i w^i - \gamma^i e > A^i w^j - \gamma^j e,\ i,\ j = 1, \ldots, k,\ i \neq j. \tag{16}$$

The class of a point x is determined from $(w^i, \gamma^i),\ i = 1, \ldots, k$ by finding i such that

$$f_i(x) = x^T w^i - \gamma^i \tag{17}$$

is maximized.

For this piecewise-linearly separable problem, infinitely many (w^i, γ^i) exist that satisfy (16). Intuitively, the "optimal" (w^i, γ^i) provides the largest margin of classification. So in an approach analogous to the two class support vector machine (SVM) approach, we add regularization terms. The dashed lines in Figure 7 represent the margins for each piece $(w^i - w^j, \gamma^i - \gamma^j)$ of the piecewise-linear separating function. The margin of separation between the classes i and j, i.e. the distance between

$$\begin{gathered} A^i(w^i - w^j) \geq (\gamma^i - \gamma^j)e + e \\ \text{and} \\ A^j(w^i - w^j) \leq (\gamma^i - \gamma^j)e - e \end{gathered}$$

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

$$\begin{array}{ll} \min\limits_{w^i, \gamma^i} & \frac{1}{2}\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{i-1} \left\|w^i - w^j\right\|^2 + \frac{1}{2}\sum\limits_{i=1}^{k} \left\|w^i\right\|^2 \\ s.t. & A^i(w^i - w^j) - e(\gamma^i - \gamma^j) - e \geq 0 \\ & i, j = 1, \ldots, k \quad i \neq j. \end{array} \tag{18}$$

To simplify the notation for formulation of the piecewise-linear SVM, we rewrite this in matrix notation. See Appendix A for complete matrix definitions for general k. For the three class problem ($k = 3$) the following matrices are obtained:

Let

$$\bar{C} = \begin{bmatrix} I & -I & 0 \\ I & 0 & -I \\ 0 & I & -I \end{bmatrix}$$

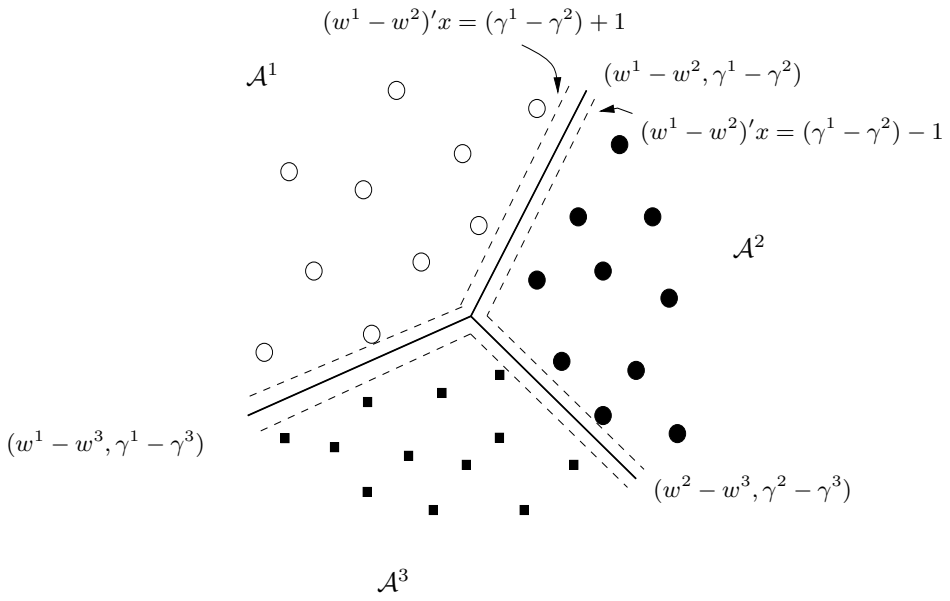

Figure 7: Piecewise-linear separator with margins for three classes.

where $I \in R^{n \times n}$ is the identity matrix.

Let

$$\bar{A} = \begin{bmatrix} A^1 & -A^1 & 0 \\ A^1 & 0 & -A^1 \\ -A^2 & A^2 & 0 \\ 0 & A^2 & -A^2 \\ -A^3 & 0 & A^3 \\ 0 & -A^3 & A^3 \end{bmatrix} \quad \bar{E} = \begin{bmatrix} -e^1 & e^1 & 0 \\ -e^1 & 0 & e^1 \\ e^2 & -e^2 & 0 \\ 0 & -e^2 & e^2 \\ e^3 & 0 & -e^3 \\ 0 & e^3 & -e^3 \end{bmatrix}$$

where $A^i \in R^{m_i \times n}, i = 1, \ldots, 3$, and $e^i \in R^{m_i \times 1}, i = 1, \ldots, 3$, is a vector of ones.

Using this notation for fixed $k > 2$ the program becomes:

$$\begin{array}{ll} \min\limits_{w,\gamma} & \frac{1}{2}\left\|\bar{C}w\right\|^2 + \frac{1}{2}\|w\|^2 \\ s.t. & \bar{A}w + \bar{E}\gamma - e \geq 0 \end{array} \tag{19}$$

where $w = [w^{1^T}, w^{2^T}, \ldots, w^{k^T}]^T$ and $\gamma = [\gamma^1, \gamma^2, \ldots, \gamma^k]^T$.

The dual of this problem can be written as:

$$\begin{array}{ll} \max\limits_{u,w,\gamma} & \frac{1}{2}\left\|\bar{C}w\right\|^2 + \frac{1}{2}\|w\|^2 - u^T(\bar{A}w + \bar{E}\gamma - e) \\ s.t. & (I + \bar{C}^T\bar{C})w = \bar{A}^T u \\ & -\bar{E}^T u = 0 \\ & u \geq 0. \end{array} \tag{20}$$

To eliminate the variables w and γ from this problem we will first show that the matrix $(I + \bar{C}^T\bar{C})$ is nonsingular.

Proposition 3.1 (Nonsingularity of $(I + \bar{C}^T\bar{C})$) *The inverse of matrix $(I + \bar{C}^T\bar{C})$ for $k > 2$ is*

$$(I_{kn} + \bar{C}^T\bar{C})^{-1} = \begin{bmatrix} \frac{2}{k+1}I_n & \frac{1}{k+1}I_n & \cdots & \frac{1}{k+1}I_n \\ \frac{1}{k+1}I_n & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{k+1}I_n \\ \frac{1}{k+1}I_n & \cdots & \frac{1}{k+1}I_n & \frac{2}{k+1}I_n \end{bmatrix} \tag{21}$$

where I_n indicates the $n \times n$ identity matrix.

Proof. To show that $(I + \bar{C}^T\bar{C})$ is nonsingular for some $k > 2$, we will calculate its inverse. The matrix $\bar{C}$ as defined in Appendix A has size $(n\sum_{i=2}^{k}(i-1) \times kn)$. Recall that n indicates the dimension of the feature space.

$$\bar{C}^T\bar{C} = \begin{bmatrix} (k-1)I_n & -I_n & \cdots & -I_n \\ -I_n & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -I_n \\ -I_n & \cdots & -I_n & (k-1)I_n \end{bmatrix}$$

has size $kn \times kn$.

Therefore

$$I_{kn} + \bar{C}^T\bar{C} = \begin{bmatrix} kI_n & -I_n & \cdots & -I_n \\ -I_n & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -I_n \\ -I_n & \cdots & -I_n & kI_n \end{bmatrix}.$$

Through simple calculations it can be shown that the inverse of this matrix is (21):

$$(I_{kn} + \bar{C}^T\bar{C})^{-1} = \begin{bmatrix} \frac{2}{k+1}I_n & \frac{1}{k+1}I_n & \cdots & \frac{1}{k+1}I_n \\ \frac{1}{k+1}I_n & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{k+1}I_n \\ \frac{1}{k+1}I_n & \cdots & \frac{1}{k+1}I_n & \frac{2}{k+1}I_n \end{bmatrix}.$$

□

Using Proposition 3.1 the following relationship results:

$$(I + \bar{C}^T\bar{C})^{-1}\bar{A}^T = \frac{1}{k+1}\bar{A}^T. \tag{22}$$

It follows from Problem (20) and equation (22) that

$$w = (I + \bar{C}^T\bar{C})^{-1}\bar{A}^T u = \frac{1}{k+1}\bar{A}^T u. \tag{23}$$

Using this relationship, we eliminate w from the dual problem. Additionally, γ is removed because $-\bar{E}^T u = 0$.

After some simplification the new dual problem becomes:

$$\begin{array}{ll} \max\limits_{u} & e^T u - \frac{1}{2(k+1)} u^T \bar{A}\bar{A}^T u \\ s.t. & \bar{E}^T u = 0 \\ & u \geq 0. \end{array} \tag{24}$$

To construct the multicategory support vector machine, it is convenient to write this problem in summation notation. Let the dual vector

$$u^T = [u^{12^T}, u^{13^T}, \ldots, u^{1k^T}, u^{21^T}, u^{23^T}, \ldots, u^{k(k-1)^T}]$$

where $u^{ij} \in R^{m_i \times 1}$. The resulting dual problem for piecewise-linear datasets is:

$$\begin{array}{ll} \max\limits_{u} & \sum\limits_{i=1}^{k}\sum\limits_{\substack{j=1\\ j\neq i}}^{k}\sum\limits_{l=1}^{m_i} u_l^{ij} - \frac{1}{2(k+1)}\sum\limits_{i=1}^{k}\sum\limits_{\substack{j=1\\ j\neq i}}^{k}\sum\limits_{\substack{l=1\\ l\neq i}}^{k}\left[\sum\limits_{p=1}^{m_i}\sum\limits_{q=1}^{m_i} u_p^{ij}u_q^{il}A_p^i {A_q^i}^T \right. \\ & \qquad \left. -2\sum\limits_{p=1}^{m_j}\sum\limits_{q=1}^{m_i} u_p^{ji}u_q^{il}A_p^j {A_q^i}^T + \sum\limits_{p=1}^{m_j}\sum\limits_{q=1}^{m_l} u_p^{ji}u_q^{li}A_p^j {A_q^l}^T\right] \\ s.t. & -\sum\limits_{\substack{j=1\\ j\neq i}}^{k}\sum\limits_{l=1}^{m_i} u_l^{ij} + \sum\limits_{\substack{j=1\\ j\neq i}}^{k}\sum\limits_{l=1}^{m_j} u_l^{ji} = 0 \ \text{ for } i = 1, \ldots, k \\ & u_l^{ij} \geq 0 \ \text{ for } i, j = 1, \ldots, k, \ i \neq j \ \text{ and } l = 1, \ldots, m_i \end{array} \tag{25}$$

where m_i is the number of points in class i.

Recall, for the piecewise-linear classification function, the class of a point x is determined by finding $i = 1, \ldots, k$, such that

$$f_i(x) = x^T w^i - \gamma^i \tag{26}$$

is maximized. From equation (23),

$$w = \begin{bmatrix} w^1 \\ w^2 \\ \vdots \\ w^k \end{bmatrix} = \frac{1}{k+1}\bar{A}^T u.$$

Solving for w^i in summation notation we get:

$$w^i = \sum_{\substack{j=1\\ j\neq i}}^{k}\sum_{p=1}^{m_i} u_p^{ij}{A_p^i}^T - \sum_{\substack{j=1\\ j\neq i}}^{k}\sum_{p=1}^{m_j} u_p^{ji}{A_p^j}^T.$$

Therefore,

$$f_i(x) = \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{p=1}^{m_i} u_p^{ij} x^T {A_p^i}^T - \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{p=1}^{m_j} u_p^{ji} x^T {A_p^j}^T - \gamma^i.$$

4 Formulation of M-SVM: Piecewise-nonlinearly Separable Case

Just like in the two-class case, M-SVM can be generalized to the piecewise-nonlinear functions. To construct the separating functions, $f_i(x)$, in a higher dimension feature space, the original data points x are transformed by some function $\phi(x) : R^n \to R^{n'}$ [23, 8]. The function $f_i(x)$ is now related to the sum of dot products of vectors in this higher dimension feature space:

$$f_i(x) = \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{p=1}^{m_i} u_p^{ij} (\phi(x) \cdot \phi({A_p^i}^T)) - \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{p=1}^{m_j} u_p^{ji} (\phi(x) \cdot \phi({A_p^j}^T)) - \gamma^i.$$

According to [23], any symmetric function $K(x, x_i) \in L_2$ that satisfies Mercer's Theorem [9] can replace the dot product $(\phi(x) \cdot \phi(x_i))$. Mercer's Theorem guarantees that any eigenvalue λ_j in the expansion $K(x, x_i) = \sum_{j=1}^{\infty} \lambda_j (\phi_j(x) \cdot \phi_j(x_i))$ is positive. This is a sufficient condition for a function $K(x, x_i)$ to define a dot product in the higher dimension feature space. Therefore we let $K(x, x_i) = (\phi(x) \cdot \phi(x_i))$.

Returning to dual Problem (25), the objective function contains the sum of dot products $A_p^j {A_q^i}^T$ of two points in the original feature space. To transform the points A_p^j to a higher dimension feature space we replace these dot products by $K({A_p^j}^T, {A_q^i}^T)$.

The resulting M-SVM for piecewise-linearly separable datasets is:

$$\begin{aligned}
\max_{u} \quad & \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{l=1}^{m_i} u_l^{ij} - \tfrac{1}{2(k+1)} \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{\substack{l=1 \\ l\neq i}}^{k} \Bigg[\sum_{p=1}^{m_i} \sum_{q=1}^{m_i} u_p^{ij} u_q^{il} K({A_p^i}^T, {A_q^i}^T) \\
& -2 \sum_{p=1}^{m_j} \sum_{q=1}^{m_i} u_p^{ji} u_q^{il} K({A_p^j}^T, {A_q^i}^T) + \sum_{p=1}^{m_j} \sum_{q=1}^{m_l} u_p^{ji} u_q^{li} K({A_p^j}^T, {A_q^l}^T) \Bigg] \qquad (27) \\
s.t. \quad & -\sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{l=1}^{m_i} u_l^{ij} + \sum_{\substack{j=1 \\ j\neq i}}^{k} \sum_{l=1}^{m_j} u_l^{ji} = 0 \quad \text{for } i = 1, \ldots, k \\
& u_l^{ij} \geq 0 \quad \text{for } i, j = 1, \ldots, k, \ i \neq j \quad \text{and } l = 1, \ldots, m_i.
\end{aligned}$$

The points A_l^i corresponding to nonzero dual variables $u_l^{ij}, j = 1, \ldots, k, j \neq i$ are referred to as support vectors. It is possible for A_l^i to correspond with more

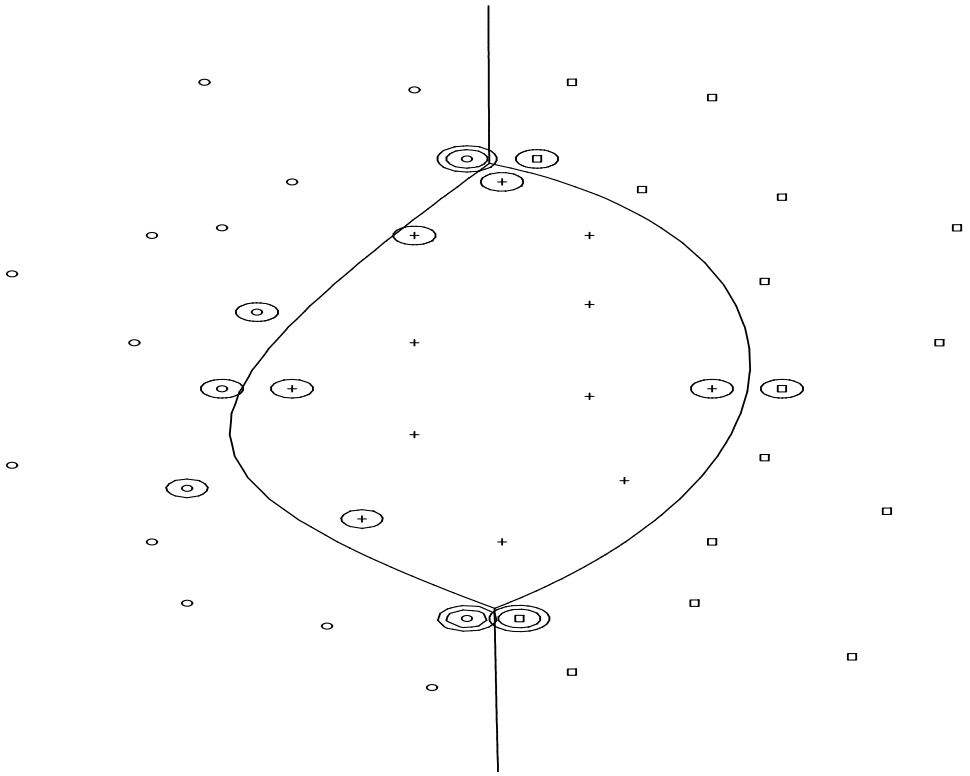

Figure 8: Piecewise-polynomial separation of three classes in two dimensions. Support vectors are indicated with circles.

than one nonzero variable $u_l^{ij}, j = 1, \ldots, k, j \neq i$. In Figure 8, support vectors are represented by a circle around the point. Some points have double circles which indicate that two dual variables $u_l^{ij} > 0, j = 1, \ldots, 3, j \neq i$. By the complementarity within the KKT conditions [14],

$$u_l^{ij} > 0 \Rightarrow A_l^i(w^i - w^j) = (\gamma^i - \gamma^j) + 1.$$

Consequently the support vectors are located "closest" to the separating function. In fact, the remainder of the points, those that are not support vectors, are not necessary in the construction of the separating function. The resulting nonlinear classification problem for a point x is to find $i = 1, \ldots, k$ such that the classification function

$$f_i(x) = \sum_{\substack{j=1 \\ j \neq i}}^{k} \left[\sum_{\substack{\text{support vectors} \\ \in A^i}} u_p^{ij} K(x, {A_p^i}^T) - \sum_{\substack{\text{support vectors} \\ \in A^j}} u_p^{ji} K(x, {A_p^j}^T) \right] - \gamma^i \tag{28}$$

is maximized.

5 Formulation of M-SVM: Piecewise Inseparable Case

The proceeding sections provided a formulation for the piecewise-linearly and piecewise-nonlinear separable cases. To construct a classification function for a piecewise-linearly *inseparable* dataset, we must first choose an error minimization criterion. The technique used in the preceeding sections of formulating the M-SVM for piecewise-linearly separable datasets can be combined with the 1-norm error criterion used in Problem (15) of Bennett and Mangasarian [6]. The result is the M-SVM for piecewise-linearly inseparable problems.

Using the same matrix notation as in Section 3, we add the terms $\frac{1}{2}\left\|\bar{C}w\right\|^2 + \frac{1}{2}\left\|w\right\|^2$ to the objective of Problem (15). The resulting primal problem is as follows:

$$\begin{array}{ll} \min\limits_{w,\gamma,y} & \lambda(\frac{1}{2}\left\|\bar{C}w\right\|^2 + \frac{1}{2}\left\|w\right\|^2) + (1-\lambda)e^T y \\ s.t. & \bar{A}w + \bar{E}\gamma - e + y \geq 0 \\ & y \geq 0 \end{array} \tag{29}$$

where $y = [y_{12}^T, y_{13}^T, \ldots, y_{1k}^T, y_{21}^T, \ldots, y_{k(k-1)}^T]^T$ and $0 < \lambda < 1$.

Solving for the dual, substituting $w = \frac{1}{k+1}\bar{A}^T u$, and simplifying produces the following problem:

$$\begin{array}{ll} \max\limits_{u} & u^T e - \frac{1}{2(k+1)} u^T \bar{A}\bar{A}^T u \\ s.t. & 0 \leq u \leq \frac{1-\lambda}{\lambda} e \\ & \bar{E}^T u = 0. \end{array} \tag{30}$$

As shown in Proposition 5.1, Problem (30) maximizes a concave quadratic objective over a bounded polyhedral set. Thus there exists a locally optimal solution that is globally optimal.

Proposition 5.1 (Concavity of objective) *The function* $u^T e - \frac{1}{2(k+1)} u^T \bar{A}\bar{A}^T u$ *is concave.*

Proof. The matrix $\bar{A}\bar{A}^T$ is always positive semi-definite and symmetric. Thus the Hessian matrix ($= -\frac{1}{(k+1)}\bar{A}\bar{A}^T$) is negative semi-definite. Therefore, the objective is a concave function. □

Problem (30) is identical to Problem (24) in the piecewise-linearly separable case except the dual variables are now bounded by $\frac{1-\lambda}{\lambda}$. Therefore, transforming the data points A_l^i will proceed identically as in Section 4. Using the function $K(x, x_i)$ to denote the dot product in some feature space, the final M-SVM results:

$$
\begin{aligned}
\max_{u} \quad & \sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\sum_{l=1}^{m_i} u_l^{ij} - \tfrac{1}{2(k+1)}\sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\sum_{\substack{l=1\\ l\neq i}}^{k}\left[\sum_{p=1}^{m_i}\sum_{q=1}^{m_i} u_p^{ij}u_q^{il}K({A_p^i}^T, {A_q^i}^T)\right. \\
& \left. -2\sum_{p=1}^{m_j}\sum_{q=1}^{m_i} u_p^{ji}u_q^{il}K({A_p^j}^T, {A_q^i}^T) + \sum_{p=1}^{m_j}\sum_{q=1}^{m_l} u_p^{ji}u_q^{li}K({A_p^j}^T, {A_q^l}^T)\right] \qquad (31)\\
s.t. \quad & -\sum_{\substack{j=1\\ j\neq i}}^{k}\sum_{l=1}^{m_i} u_l^{ij} + \sum_{\substack{j=1\\ j\neq i}}^{k}\sum_{l=1}^{m_j} u_l^{ji} = 0 \ \text{ for } i = 1,\ldots,k \\
& 0 \le u_l^{ij} \le \tfrac{1-\lambda}{\lambda} \ \text{ for } i,j = 1,\ldots,k,\ i\neq j \ \text{ and } l = 1,\ldots,m_i.
\end{aligned}
$$

As in Sections 3 and 4, the class of a point x is determined by finding the maximum function

$$
f_i(x) = \sum_{\substack{j=1\\ j\neq i}}^{k}\left[\sum_{\substack{\text{support vectors}\\ \in A^i}} u_p^{ij}K(x, {A_p^i}^T) - \sum_{\substack{\text{support vectors}\\ \in A^j}} u_p^{ji}K(x, {A_p^j}^T)\right] - \gamma^i \qquad (32)
$$

for $i = 1,\ldots,k$.

To determine the threshold values $\gamma^i, i = 1,\ldots,k$, we solve the primal problem (29) with w fixed, where $\bar{A}w$ is transformed to the higher dimension feature space. This problem is as follows:

$$
\begin{aligned}
\min_{\gamma,y} \quad & \sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\sum_{l=1}^{m_i} y_l^{ij} \\
s.t. \quad & -\gamma^i + \gamma^j + y_l^{ij} \ge \tfrac{1}{k+1}\sum_{\substack{l=1\\ l\neq i}}^{k}\sum_{q=1}^{m_i}\left[\sum_{r=1}^{m_i} K({A_q^i}^T, {A_r^i}^T)u_r^{il} - \sum_{r=1}^{m_l} K({A_q^i}^T, {A_r^l}^T)u_r^{li}\right] \\
& \qquad -\tfrac{1}{k+1}\sum_{\substack{l=1\\ l\neq j}}^{k}\sum_{q=1}^{m_i}\left[\sum_{r=1}^{m_j} K({A_q^i}^T, {A_r^j}^T)u_r^{jl} - \sum_{r=1}^{m_l} K({A_q^i}^T, {A_r^l}^T)u_r^{lj}\right] + 1 \\
& y_l^{ij} \ge 0, \quad i,j = 1,\ldots,k,\ i\neq j,\ l = 1,\ldots,m_i.
\end{aligned} \qquad (33)
$$

The right side of the constraints are constant. Thus Problem (33) is a linear program and is easily solved.

6 Computational Experiments

In this section, we present computational results comparing M-SVM (32), M-RLP (15), k-SVM using SVM (10), and k-RLP using RLP (8). Several experiments on real-world datasets are reported. A description of each of the

datasets follows this paragraph. Each of these methods was implemented using the MINOS 5.4 [17] solver. The quadratic programming problems for M-SVM and k-SVM were solved using the nonlinear solver implemented in Minos 5.4. This solver uses a reduced-gradient algorithm in conjunction with a quasi-Newton method. In M-SVM, k-SVM and M-RLP, the selected values for λ are given. 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 k-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.

Wine Recognition Data The Wine dataset [1] uses the chemical analysis of wine to determine the cultivar. There are 178 points with 13 features. This is a three class dataset distributed as follows: 59 points in class 1, 71 points in class 2, and 48 points in class 3. This dataset is available via anonymous file transfer protocol (ftp) from the UCI Repository of Machine Learning Databases and Domain Theories [16] at ftp://ftp.ics.uci.edu/pub/machine-learning-databases.

Glass Identification Database The Glass dataset [11] is used to identify the origin of a sample of glass through chemical analysis. This dataset is comprised of six classes of 214 points with 9 features. The distribution of points by class is as follows: 70 float processed building windows, 17 float processed vehicle windows, 76 non-float processed building windows, 13 containers, 9 tableware, and 29 headlamps. This dataset is available via anonymous file transfer protocol (ftp) from the UCI Repository of Machine Learning Databases and Domain Theories [16] at ftp://ftp.ics.uci.edu/pub/machine-learning-databases.

US Postal Service Database The USPS Database [10] contains 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.

Two experiments were performed. In the first, the datasets were normalized between -1 and 1. 10-fold cross validation was used to estimate generalization on future data. The second experiment was conducted on two subsets of the United States Postal Service (USPS) data. This data contains handwriting samples of the integers 0 through 9. The objective of this dataset is to quickly and effectively interpret zipcodes. This data has separate training and testing sets, each of which consist 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 datasets, 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 k-SVM methods are given.

Table 1 contains results for M-RLP, k-RLP, M-SVM, and k-SVM on the Wine and Glass datasets. As anticipated, adding the regularization term to

Data		Degree				
		1	2	3	4	5
Wine	M-RLP	91.01	-	-	-	-
	k-RLP	99.44	-	-	-	-
	M-SVM	97.19	97.19	97.75	96.63	96.63
		(378)	(291)	(258)	(239)	(228)
	k-SVM	99.44	98.88	99.44	99.44	99.44
		(537)	(424)	(405)	(394)	(413)
Glass	M-RLP	64.95	-	-	-	-
	k-RLP	54.67	-	-	-	-
	M-SVM	55.14	62.15	62.15	66.82	67.29
		(1759)	(1660)	(1595)	(1533)	(1476)
	k-SVM	43.46	55.61	64.95	70.56	72.43
		(1898)	(1854)	(1796)	(1769)	(1734)

Table 1: Percent testing set accuracies and (total number of support vectors) for M-SVM and k-SVM. $\lambda = .05$ for k-RLP, M-SVM, and k-SVM.

the degree one problem in M-SVM produced better testing generalization than M-RLP on the Wine dataset. The Wine dataset is piecewise-linearly separable. Therefore, the M-RLP method has infinitely many optimal solutions. However, the testing accuracy for M-SVM with degree one on the Glass data was much lower than the M-RLP accuracy. This may indicate that the choice of λ is too large. However, as the degree increases the accuracy of the M-SVM method improves and exceeds the M-RLP results. The k-SVM method generalized surprisingly well. The testing accuracies reported for k-SVM on the Wine dataset are higher than those of M-SVM. The linear k-RLP method performed just as well as the quadratic k-SVM program on the Wine dataset and better than the M-SVM and M-RLP methods. On the Glass data, as the degree increases, both methods, M-SVM and k-SVM, improve dramatically in testing accuracy. Using higher degree polynomials the M-SVM and k-SVM methods surpass the accuracies of M-RLP and k-RLP. This demonstrates the potential for polynomial and piecewise-polynomial classification functions over linear and piecewise-linear functions.

Table 2 contains results for the four methods on the USPS data subsets. Similar observations as above can be made. Both of these datasets are piecewise-linearly separable. The solution that m-RLP has found for each of these datasets tests significantly lower than the other methods. The k-SVM method generalizes slightly better than M-SVM. The k-RLP method reports similar accuracies as the k-SVM method. Additionally, it is solving linear programs rather than quadratic programs, so the computational training time is significantly smaller than 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 of k-SVM. With fewer support vectors, a sam-

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

Table 2: Percent testing set accuracies and (total number of support vectors) for M-SVM and SVM. $\lambda = .05$ for k-SVM and $\lambda = .03$ for k-RLP and M-SVM.

	Degree				
	1	2	3	4	5
M-RLP	7922	-	-	-	-
k-RLP	341	-	-	-	-
M-SVM	20359	16978	11422	10402	12673
k-SVM	6883	5709	4819	4929	4932

Table 3: Total computational training time (in seconds) for M-RLP,k-RLP, M-SVM, and k-SVM on USPS-1.

ple can be classified more quickly since the dot-product of the sample with each support vector must be computed. Thus the M-SVM would be a good method to choose when classification time is critical.

CPU times for training all four methods on the USPS-1 dataset are reported in Table 3. The times for all the datasets are not listed because the programs were run using a batch system on clusters of machines so the timing was not reliable. However, the trends were clear. The k-RLP method is significantly faster than the other methods. In the M-SVM and k-SVM methods, as the degree increased the computational time would decrease and then after a certain degree is reached it would increase. The degree of the polynomial for which it starts to increase varies by dataset. Surprisingly, for the USPS datasets the k-SVM method was faster than the M-RLP method. This was not the case for the Wine and Glass datasets. The M-RLP method had faster training times than k-SVM for these datasets. The times reported are for IBM RS6000 model 590 workstations with 128 MB RAM.

7 Conclusions

We have examined four methods for the solution of multicategory discrimination problems based on the LP methods of Mangasarian and the QP methods for SVM of Vapnik. The two-class methods, RLP and SVM are differ only in the norm of the regularization term. In the past two different approaches had been used for the $k > 2$ class case. The method we called k-SVM, constructed k two-class discriminants using k quadratic programs. The resulting classifier was a piecewise-linear or piecewise nonlinear discriminant function depending on what kernel function was used in the SVM. The original multicategory RLP for k classes, constructed a piecewise-linear discriminant using a single linear program. We proposed two new hybrid approaches. Like the k-SVM method, k-RLP uses LP to construct k two-class discriminants. We also formulated a new approach, M-SVM. We began the formulation by adding regularization terms to M-RLP. Then like k-SVM with piecewise-nonlinear discriminants, the nonlinear pieces are found by mapping the original data points into a higher dimension feature space. This transformation appeared in the dual problem as an inner product of two points in the higher dimension space. A generalized inner product was used to make the problem tractable. The new M-SVM method requires the solution of a single quadratic program. We performed a computational study of the four methods on four datasets. In general we found that the k-SVM and k-RLP generalized. However, M-SVM used fewer support vectors – a counter-intuitive result since for the two-class class Statistical Learning Theory predicts that fewer support vector should result in better generalization. The theoretic justification of the better generalization of k-SVM and k-RLP and M-SVM and M-RLP is an open question. The k-RLP method provided accurate and efficient results on the piecewise-linear separable datasets. The k-SVM also tested surprisingly well but requires the solution of k quadratic programs. Thus providing solutions with smaller classification time. On the piecewise-linearly inseparable dataset, the polynomial and piecewise-polynomial classifiers provided an improvement over the M-RLP and k-RLP methods. On the other datasets, the k-RLP method found solutions that generalized best or nearly best in less computational time.

A Matrix Representations for Multicategory Support Vector Machines

This appendix contains the definitions of the matrices used for the general k-class SVM formulation (18):

$$\begin{array}{ll} \min\limits_{w,\gamma} & \frac{1}{2}\left\|\bar{C}w\right\|^2 + \frac{1}{2}\left\|w\right\|^2 \\ s.t. & \bar{A}w + \bar{E}\gamma - e \geq 0 \end{array}$$

Let

$$\bar{C} = \begin{bmatrix} I & -I & 0 & 0 & \cdots & 0 \\ I & 0 & -I & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & 0 & 0 & \ddots & 0 \\ I & 0 & \cdots & \cdots & 0 & -I \\ 0 & I & -I & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & 0 & \ddots & 0 \\ 0 & I & 0 & \cdots & 0 & -I \\ & & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & & & \\ 0 & \cdots & 0 & I & -I & 0 \\ 0 & \cdots & 0 & I & 0 & -I \\ 0 & \cdots & \cdots & 0 & I & -I \end{bmatrix} \tag{34}$$

where $I \in R^{n\times n}$ is the identity matrix. The matrix $\bar{C}$ has $n\sum_{i=2}^{k}(i-1)$ rows and kn columns.

Let

$$\bar{A} = \begin{bmatrix} A^1 & -A^1 & 0 & 0 & \cdots & 0 \\ A^1 & 0 & -A^1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & 0 & 0 & \ddots & \vdots \\ A^1 & 0 & \cdots & \cdots & 0 & -A^1 \\ -A^2 & A^2 & 0 & 0 & \cdots & 0 \\ 0 & A^2 & -A^2 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & A^2 & 0 & \cdots & \cdots & -A^2 \\ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \\ -A^k & 0 & \cdots & \cdots & 0 & A^k \\ 0 & -A^k & 0 & \cdots & 0 & A^k \\ \vdots & 0 & \ddots & 0 & \vdots & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ 0 & \cdots & \cdots & 0 & -A^k & A^k \end{bmatrix} \tag{35}$$

where $A^i \in R^{m_i \times n}$. The matrix $\bar{A}$ has $(k-1)\sum_{i=1}^{k} m_i$ rows and kn columns.

Let

$$\bar{E} = \begin{bmatrix} -e^1 & e^1 & 0 & 0 & \cdots & 0 \\ -e^1 & 0 & e^1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & 0 & 0 & \ddots & \vdots \\ -e^1 & 0 & \cdots & \cdots & 0 & e^1 \\ e^2 & -e^2 & 0 & 0 & \cdots & 0 \\ 0 & -e^2 & e^2 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & -e^2 & 0 & \cdots & \cdots & e^2 \\ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \\ e^k & 0 & \cdots & \cdots & 0 & -e^k \\ 0 & e^k & 0 & \cdots & 0 & -e^k \\ \vdots & 0 & \ddots & 0 & \vdots & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ 0 & \cdots & \cdots & 0 & e^k & -e^k \end{bmatrix}$$

where $e^i \in R^{m_i \times 1}$ is a vector of ones. The matrix $\bar{E}$ has $(k-1)\sum_{i=1}^{k} m_i$ rows and kn columns.

References

[1] S. Aeberhard, D. Coomans, and O. de Vel. Comparison of classifiers in high dimensional settings. Technical Report 92-02, Departments of Computer Science and of Mathematics and Statistics, James Cook University of North Queensland, 1992.

[2] 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.

[3] 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. To appear.

[4] K. P. Bennett and O. L. Mangasarian. Neural network training via linear programming. In P. M. Pardalos, editor, *Advances in Optimization and Parallel Computing*, pages 56–67, Amsterdam, 1992. North Holland.

[5] K. P. Bennett and O. L. Mangasarian. Multicategory discrimination via linear programming. *Optimization Methods and Software*, 3:27–39, 1994.

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

[7] 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.

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

[9] R. Courant and D. Hilbert. *Methods of Mathematical Physics*. J. Wiley, New York, 1953.

[10] Y. Le Cun, 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.

[11] I. W. Evett and E. J. Spiehler. Rule induction in forensic science. Technical report, Central Research Establishment, Home Office Forensic Science Service, Aldermaston, Reading, Berkshire RG7 4PN, 1987.

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

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

[14] O. L. Mangasarian. *Nonlinear Programming*. McGraw–Hill, New York, 1969.

[15] O. L. Mangasarian. Mathematical programming in machine learning. In G. DiPillo and F. Giannessi, editors, *Proceedings of Nonlinear Optimization and Applications Workshop*, pages 283–295, New York, 1996. Plenum Press.

[16] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases. [http://www.ics.uci.edu/˜mlearn/MLRepository.html]. Department of Information and Computer Science, University of California, Irvine, California, 1994.

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

[18] 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.

[19] 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.

[20] A. Roy and S. Mukhopadhyay. Pattern classification using linear programming. *ORSA Journal of Computing*, 3:66–80, 1990.

[21] B. Schölkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector 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, 1996. Springer Lecture Notes in Computer Science Vol. 1112.

[22] 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. AI Memo No. 1599; CBCL Paper No. 142, Massachusetts Institute of Technology, Cambridge, 1996.

[23] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[24] V. N. Vapnik. *The Nature of Statistical Learning Theory*. John Wiley & Sons, New York, 1996.

[25] V. N. Vapnik and A. Ja. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. In Russian.

[26] 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.