NOTE **Communicated by John Platt**

Learning by Kernel Polarization

Yoram Baram
baram@cs.technion.ac.il
Department of Computer Science, Technion, Israel Institute of Technology, Haifa 3200, Israel

Kernels are key components of pattern recognition mechanisms. We propose a universal kernel optimality criterion, which is independent of the classifier to be used. Defining data polarization as a process by which points of different classes are driven to geometrically opposite locations in a confined domain, we propose selecting the kernel parameter values that polarize the data in the associated feature space. Conversely, the kernel is said to be polarized by the data. Kernel polarization gives rise to an unconstrained optimization problem. We show that complete kernel polarization yields consistent classification by kernel-sum classifiers. Tested on real-life data, polarized kernels demonstrate a clear advantage over the Euclidean distance in proximity classifiers. Embedded in a support vectors classifier, kernel polarization is found to yield about the same performance as exhaustive parameter search.

1 Introduction

Pattern recognition is central to biological and artificial intelligence. A pattern classification capability is normally acquired from knowledge of the class assignment of available data. Classical methods, such as nearest neighbor, classify new data points according to their Euclidean distance from class-labeled points (e.g., Fukunaga, 1990). More recently proposed criteria, such as margin maximization (Vapnik & Lerner, 1963), employ another brand of proximity measures, known as kernels. Kernels are at the heart of such successful methods as the Parzen density estimator (Parzen, 1962) and support vector machines (SVM; Vapnik, 1995). The choice of the kernel parameter values has been often resolved by exhaustive search, maximizing the empirical performance of the particular classifier used (Fukunaga, 1990; Scholkopf & Smola, 2002). A recent paper (Chapelle, Vapnik, Bousquet, & Mukhergee, 2002) suggests selecting the kernel parameter values that minimize a bound on the performance of the SVM method in the leave-one-out procedure. Other recent papers suggest optimizing measures of class separation, such as the kernel-target alignment (Cristianini, Shawe-Taylor, Elisseeff, & Kandola, 2001) and the scatter-matrix ratio (Wang & Chan, 2002). These approaches give rise to computationaly heavy optimization

problems (while the first has been limited to transduction problems, which are, in principle, considerably simpler than their more common induction versions, the second involves exhaustive search over sample matrix traces). A different approach (Platt, Burges, Swenson, Weare, & Zheng, 2002) is aimed at optimizing the linear metaparameters of an approximated kernel by minimizing the Frobenius norm.

In this note, we propose a kernel optimality criterion, which represents the fundamental structure of the classification problem at hand, regardless of the particular classifier to be used. This approach will facilitate a complete separation between kernel and classifier design. We propose that kernel parameter values be selected so as to drive points with different class labels to opposite locations in the abstract feature space defined by kernel geometry. Drawing from physics, we call such a process *kernel polarization*. Kernel polarization gives rise to an unconstrained optimization problem. Yet employing bounded kernels, the associated feature space geometry guarantees that the optimization problem is well posed. Independently derived, kernel polarity simplifies the measure proposed in Cristianini et al. (2001) by ridding the latter of its denominator, making the optimization problem considerably easier. Polarized kernels can be used in a variety of kernel sum classifiers, ranging from the Parzen classifier, through proximity classifiers, to support vector classifiers. We show that complete kernel polarization yields consistent classification (i.e., the expected risk vanishes as the amount of labeled data increases). This will motivate kernel polarization from a theoretical viewpoint. Testing the proposed methods on benchmark real-life data, we find that polarized kernels significantly outperform the Euclidean distance in proximity classifiers. In support vector classifiers, kernel polarization allows for eliminating the soft margin parameter, producing about the same results as exhaustive parameter search. While we employ the gaussian kernel throughout, other kernel functions may be used in a similar fashion.

2 Classification by Kernel Sums

We consider a set $X \subset R^n$, with each $x \in X$ belonging to one of two classes, labeled $y(x) \in Y = \{-1, +1\}$. Given a set of labeled points (the training set), $\{x^{(i)}, y^{(i)},\ i = 1, \ldots, M\}$, where $x^{(i)} \in X$ and $y^{(i)} \in Y$, and given a point $x \in X$, we would like to assign x with a label $y(x)$. We consider classifiers of the general form

$$f(x) = \operatorname{sign}\left\{\sum_{i=1}^{M} \alpha_i k_v\left(x^{(i)}, x\right) y^{(i)} + b\right\}, \tag{2.1}$$

where $\alpha_i,\ i = 1, \ldots, M$, and b are scalar parameters and $k_v(x^{(i)}, x)$ is a function from $X \times X$ into R, called a kernel, with v, a vector of kernel parameters.

For instance, the gaussian kernel

$$k_v(x^{(i)}, x) = \exp\left(-\frac{\| x^{(i)} - x \|^2}{v^2}\right) \tag{2.2}$$

may be viewed as a probability density function of x centered at $x^{(i)}$, or, since it is a monotone decreasing function of the Euclidean distance between its arguments, a proximity measure. But other kernels, some neither densities nor positive, have been proposed and used successfully (e.g., Scholkopf & Smola, 2002). Of course, we would like to have $f(x) = y(x)$. Many known classifiers can be put in the form of equation 2.1. Some are discussed next for the sake of later comparison.

2.1 Parzen Classifier. Assuming the existence of a probability measure P on $X \times Y$, it would be desirable to find a classifier $f(x)$ that makes the probability of an error as small as possible. The minimum possible value of $P[f(x) \neq y(x)]$ with respect to all $f(x)$ is called the *Bayes risk*. A classifier that achieves the Bayes risk is $y(x) = 1$ if $P_+ p_+(x) \geq P_- p_-(x)$, and $y(x) = -1$ otherwise, where $p_+(x)$ and $p_-(x)$ are the probability density functions and P_+ and P_- are the prior probabilities corresponding to the two classes. Substituting the densities by their Parzen estimates (Parzen, 1962), the classifier attains the form of equation 2.1 with $\alpha_i = 1/(P_+ m_+)$ for $i \mid y^{(i)} = +1$ and $\alpha_i = 1/(P_- m_-)$ for $i \mid y^{(i)} = -1$, $b = 0$, and $k(x^{(i)}, x)$ is a density kernel, such as equation 2.2. We call this classifier a *Parzen classifier*.

The Parzen classifier assumes a good approximation of the actual probability distributions of the classes. However, this cannot be guaranteed. Yet the kernel sum form of equation 2.1 is also admitted by other classifiers. Let us first consider one that does not possess this representation. The k-nearest neighbors classifier (e.g., Fukunaga, 1990) assigns a new point to the class of the majority of its k-nearest labeled neighbors, employing the Euclidean distance. While this classifier does not explicitly take into account the distribution of the data, its asymptotic error probability for any $k \geq 1$ is bounded above by twice the Bayes risk (Cover & Hart, 1967). Replacing the Euclidean distance by a monotone function of it in the k-nearest neighbors classifier will clearly have no advantage, and both measures will produce the same classification results. However, such replacement can be quite advantageous when we weight the neighbors by their proximity to x.

2.2 Weighted Neighbors. Employing the Euclidean distance d, a bounded proximity measure may be defined by $1/[1 + d(x^{(i)}, x)]$. As we shall see in the next section, kernel boundedness, which implies that the associated feature space is confined, is necessary for kernel polarization. However, unlike the kernel functions under consideration, the measure $1/[1 + d(x^{(i)}, x)]$ is nonparametric and consequently is not polarizable.

(Indeed, it will produce poor classification results, as will the unbounded measure $1/d(x^{(i)}, x)$.) Employing, instead, a parametric kernel such as equation 2.2 as a proximity measure and weighting all training data points by their proximity to x, we obtain a classifier in the form of equation 2.1 with $\alpha_i = 1,\ i = 1, \ldots, M$, and $b = 0$. It can be seen that the latter coincides with the Parzen classifier when the kernel is a density and when $m_+ = m_-$ and $P_+ = P_-$.

2.3 Nearest Mean. The kernel version of the nearest mean classifier (Scholkopf & Smola, 2002) has the form of equation 2.1 with $\alpha_i = 1/m_+$ for $i \mid y^{(i)} = +1$ and $\alpha_i = m_-$ for $i \mid y^{(i)} = -1$, and

$$b = 0.5\left(\frac{1}{m_-^2}\sum_{i,j|y^{(i)}=y^{(j)}=-1} k_v\big(x^{(i)}, x^{(j)}\big) - \frac{1}{m_+^2}\sum_{i,j|y^{(i)}=y^{(j)}=+1} k_v\big(x^{(i)}, x^{(j)}\big)\right). \tag{2.3}$$

2.4 A Neural Network. A neural network (e.g., Haykin, 1998) has the form of equation 2.1, where $\alpha_i,\ b \in R$, are weights, to be found from the data. The training points $x^{(i)}$ are often replaced by a smaller set of weights $w^{(i)},\ i = 1, \ldots, K$. These may represent a random sampling of the input space, (for example, some of the data points themselves, K being the random r-covering number (Baram, 1996)), while the α_i's and b are obtained from some learning algorithm, such as the perceptron (Rosenblatt, 1958). The main difficulty with such designs is that they merely seek a linear separation surface in feature space, without regard to its generalization property. The following approach brings us closer to resolving this issue.

Suppose that the underlying kernel satisfies Mercer's conditions (equivalently, the kernel is said to be positive semi-definite, e.g., Scholkopf & Smola, 2002). Then there exists a mapping ϕ into a space where k_v acts as a dot (inner) product, that is,

$$k_v\big(x^{(i)}, x\big) = \big\langle \phi\big(x^{(i)}\big), \phi(x)\big\rangle. \tag{2.4}$$

A kernel sum classifier can now be written as

$$y(x) = \text{sign}\{\langle w, \phi(x)\rangle + b\}, \tag{2.5}$$

where

$$w = \sum_{i=1}^{M} \alpha_i y^{(i)} \phi\big(x^{(i)}\big). \tag{2.6}$$

This means that the classifier constitutes a linear separation surface (a hyperplane) w in feature space. Let us place, symmetrically about this linear surface, two linear surfaces parallel to it. The width of themargin between these two surfaces is maximized by a support vector classifier.

2.5 A Support Vector. A support vector (SV) classifier (Vapnik, 1995) has the form of equation 2.1, where the coefficients α_i are found by solving the quadratic programming problem

$$\max_{\alpha} \left\{ \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \alpha_i \alpha_j y^{(i)} y^{(j)} k_v\left(x^{(i)}, x^{(j)}\right) \right\} \tag{2.7}$$

subject to

$$0 \le \alpha_i \le \frac{C}{M}, \quad i = 1, \ldots, M, \tag{2.8}$$

with C, the "soft margin" (or "error cost") parameter, to be determined, and

$$\sum_{i=1}^{M} \alpha_i y^{(i)} = 0, \quad i = 1, \ldots, M. \tag{2.9}$$

The additional parameter b is found from

$$b = \frac{1}{M} \sum_{j=1}^{M} \left(y^{(j)} - \sum_{i=1}^{M} \alpha_i k_v\left(x^{(i)}, x^{(j)}\right) y^{(i)} \right), \tag{2.10}$$

where the first sum is taken with respect to the support vectors—those data points $x^{(i)}$ for which the corresponding α_i's are nonzero. Margin maximization has been justified by generalization bounds (Cristianini & Shawe-Taylor, 2000).

The kernel parameters v and the soft margin parameter C are not solved by the SV machinery itself and require some external search process. The most widely used approach to this problem has been exhaustive search in conjunction with empirical error calculation. The leave-one-out error bound has been proposed for this purpose (Scholkopf & Smola, 2002). In practice, computationally less demanding cross-validation strategies have been used for error calculation in this context. A recent paper (Chapelle et al., 2002) suggests finding these parameters by minimizing a certain bound on the number of errors produced by the SV classifier in the leave-one-out procedure. In the next section we propose a method for optimizing the kernel parameters, which is independent of the particular classifier to be used.

3 Kernel Polarization

The mutual proximity of two points $x^{(i)}$, $x^{(j)}$ in X is represented by a kernel $k_v(x^{(i)}, x^{(j)})$, where v denotes the vector of kernel parameters. Separation of the two classes will become easier if there is a clear correspondence between the kernel proximity of points and the identity of their labels. In particular, this will be true if, by choice of the values of the kernel parameters, points of the same class come close together, while points of different classes go far apart, in a geometrically confined domain. In physics, such an action, applied to objects with opposite electric charges, has been called *polarization*. Given a set of labeled points (the training set), $\{x^{(i)}, y^{(i)},\ i = 1, \ldots, M\}$, where $x^{(i)} \in X$ and $y^{(i)} \in \{-1, +1\}$, we define its polarity as

$$K_v = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} k_v\big(x^{(i)}, x^{(j)}\big) y^{(i)} y^{(j)}. \tag{3.1}$$

Clearly, K_v will increase if points of the training set having the same label come closer together and points of different classes go farther apart, in the sense of the kernel being a proximity measure. How can we be assured that when this is true for points of the training set, it will also hold true, at least in some sense of approximation, for other points of X? This will be the case if, for instance, the kernel is a continuous monotone function of the Euclidean distance between its two arguments. Furthermore, the maximization problem will be well posed if the feature space is confined. Such properties are possessed by common kernel functions. For instance, the gaussian kernel, equation 2.2, is continuous, smooth, and monotone. Also, $k_v(x, x) = 1$ for all $x \in X$, and $k_v(x, x') > 0$ for all $x, x' \in X$. Since, the gaussian kernel can be shown to satisfy Mercer's condition of semi-positivity (Scholkopf & Smola, 2002), it is a dot product in feature space; hence, it is the cosine of the angle between the corresponding features. The feature space is, then, an orthant of the unity sphere; hence, it is quite confined (yet infinite dimensional; Scholkopf & Smola, 2002; Micchelli, 1986).

Given a training set, our objective is to maximize K_v with respect to the kernel parameters. Newton's iteration is

$$v(k+1) = v(k) + \left[\nabla_v^2 K_v\right]^{-1} \nabla_v K_v, \tag{3.2}$$

where $\nabla_v K_v$ and $\nabla_v^2 K_v$ are, respectively, the gradient and the Hessian of K_v with respect to v.

In the case of a gaussian kernel, we have one kernel parameter v. The first and the second derivatives of the kernel with respect to v

are:

$$\frac{\partial k_v(x^{(i)}, x^{(j)})}{\partial v} = \frac{2 \parallel x^{(i)} - x^{(j)} \parallel^2}{v^3} \exp\left(-\frac{\parallel x^{(i)}, x^{(j)} \parallel^2}{v^2}\right) \tag{3.3}$$

$$\frac{\partial^2 k_v(x^{(i)}, x^{(j)})}{\partial v^2} = \frac{-6 \parallel x^{(i)} - x^{(j)} \parallel^2}{v^4} \exp\left(-\frac{\parallel x^{(i)} - x^{(j)} \parallel^2}{v^2}\right) + \frac{4 \parallel x^{(i)} - x^{(j)} \parallel^4}{v^6} \exp\left(-\frac{\parallel x^{(i)} - x^{(j)} \parallel^2}{v^2}\right). \tag{3.4}$$

Substituting these expressions into equation 3.2 will complete the specification of the parameter optimization step.

How can we evaluate the performance of kernel sum classifiers? In learning theory, performance is often measured by risk functions (Scholkopf & Smola, 2002). In particular, given a classifier f and a cost function $c(x, y, f(x))$ (e.g., $c(x, y, f(x)) = 0.5 \mid y - f(x) \mid$), the expected risk is defined as

$$R[f] = E\{c(x, y, f(x))\} = \int_{X \times Y} c(x, y, f(x)) dP(x, y), \tag{3.5}$$

whereas the empirical risk is

$$R_{emp}[f] = \frac{1}{M} \sum_{i=1}^{M} c\left(x^{(i)}, y^{(i)}, f\left(x^{(i)}\right)\right). \tag{3.6}$$

Rooted in the VC theory (Vapnik & Chervonenkis, 1968, 1971), generalization is defined as uniform (with respect to a classifier family) convergence of the empirical risk to the actual risk, as the size M of the training set grows indefinitely. When the empirical risk is zero, generalization has a particularly strong meaning, since it guarantees, asymptotically, a faultless classifier. Since for a commonly used cost per error ($c = 0.5 \mid y(x) - f(x) \mid$), the expected risk coincides with the error probability, we shall call such convergence *consistency* (this definition, which is consistent with the one used in statistics, is different from the one used by Scholkopf & Smola, 2002, which implies convergence of the expected risk to the Bayes risk). When the kernel in the kernel sum classifier is positive definite and when the dimension of the corresponding feature space is finite, so is the VC dimension. Generalization is then implied by the VC bound. Unfortunately, the feature spaces associated with some useful kernels (e.g., the gaussian one) are infinite dimensional, making the VC bound unusable. Several alternative bounds on classifier performance have been proposed (e.g., Scholkopf & Smola, 2002).

We find the PAC-Bayesian margin bound particularly useful in the present context here. We shall assume that polarization is *complete*, in the sense that the polarity coefficient, equation 3.1, attains its absolute maximum value. Then we have the following result:

Lemma. *Under complete kernel polarization, the expected risk of a kernel sum classifier* $f(x) = sign\{\langle w, \phi(x)\rangle\}$ *with a positive definite kernel and a feature space of dimension* $n \in N \cup \infty$ *is, with probability* $1 - \delta$, *bounded as*

$$R[f] \leq \frac{2}{M}\left[-d \ln\left(1 - \sqrt{1 - \rho^2(f)}\right) + 2 \ln M - \ln \delta + 2\right], \tag{3.7}$$

where $d = \min(M, n)$ and

$$\rho(f) = \min_{i \in [M]} \frac{y^{(i)}\langle w, \phi(x^{(i)})\rangle}{\| \phi(x^{(i)}) \|},$$

where $\phi(x^{(i)})$ is the feature corresponding to $x^{(i)}$ and $[M] = \{1, 2, \ldots, M\}$.

Proof. Complete kernel polarization guarantees that the two classes in the training set reduce to a point each in feature space, and these two points are different. This means that the empirical risk of a kernel sum classifier on the training set is zero. The proof now follows directly from the PAC-Bayesian margin bound (Herbrich, 2000; Scholkopf & Smola, 2002).

Consistency of kernel sum classifiers now follows for positive unity kernels, that is, kernels satisfying $k(x, x') \geq 0$ and $k(x, x) = 1$.

Theorem. *For a positive unity kernel under complete polarization, a kernel sum classifier is consistent.*

Proof. We first note that under a positive unity kernel, the feature space is the positive orthant of the unity sphere. Under complete polarization, the two classes are concentrated at two points in feature space, separated by an angle $\pi/2$. Consider a homogeneous separating hyperplane in feature space, yielding a zero empirical risk and the highest possible expected risk. There may be, at most, two such hyperplanes, each passing arbitrarily close to one of the two polarized classes (one or both may possess the highest expected risk). Let us denote the normal to one of these hyperplanes w^* and the associated classifier f^*. Without loss of generality, suppose that this hyperplane passes arbitrarily close to the point corresponding to the polarized class labeled $y = +1$. Then w^* forms an angle arbitrarily close to π, with the vector associated with the point corresponding to the polarized class labeled $y = -1$. Clearly, for any other hyperplane w with zero empirical

risk (i.e., a hyperplane passing between the two points) and its associated classifier f, we have

$$R[f] \leq R[f^*]. \tag{3.8}$$

We now have $\min_{i\in[M]}\{y^{(i)}\langle w^*, \phi(x^{(i)})\rangle\} = \cos\pi = -1$ and $\| \phi(x^{(i)}) \| = k_v(x^{(i)}, x^{(i)}) = 1$; hence, $\rho^2(f^*) = 1$. It follows from equation 3.7 that

$$R[f^*] \leq \frac{2}{M}(2\ln M - \ln\delta + 2). \tag{3.9}$$

Hence,

$$R[f] \leq \frac{2}{M}(2\ln M - \ln\delta + 2). \tag{3.10}$$

It follows that $R[f] \to 0$ as $M \to \infty$, as asserted.

The theorem implies that a kernel sum classifier with a polarized gaussian kernel is consistent. Of course, this is also true for other kernels with similar properties. Even if complete polarization is not achieved, this result certainly motivates kernel polarization, which is further justified in the following section by numerical examples.

4 Examples

We have applied kernel polarization to the six smallest databases from UCI (2003) ranging in size from 215 to 1066 and in dimension from 5 to 20. Each database was divided into four parts, three parts serving as a training set and one part as a test set, in a fourfold cross-validation fashion. For each of the cases, we have polarized a gaussian kernel with respect to the training data. The polarized kernel was then employed by the weighted-neighbors (PKWN), the nearest-mean (PKNM), and the support vectors (PKSV) classifiers described before. We have also applied the Parzen classifier (PC) and the Euclidean versions of the nearest neighbor (NN), the nearest mean (NM), and the weighted neighbors (WN) classifiers to the same data. Performance results for the SV classifier employing exhaustive search for the kernel and the soft margin parameter values were obtained from R. El-Yaniv and E. Yom-Tov (personal communication, 2003). The results, in terms of the average percentile number of errors for the four folds and the standard error of the mean deviation (STDM) are summarized in Table 1. The average performance of each of the classifiers is given in the last row of the table, where the STDM value is calculated with respect to the results for the different databases.

Table 1: Classification Results (Error Percentage Mean+STDM) for the Examples.

	NN	PC	WN	PKWN	NM	PKNM	SV	PKSV
Thyroid	3.3 ± 0.3	22.3 ± 0.8	30.2 ± 1.9	8.4 ± 0.41	15.2 ± 0.9	9.3 ± 0.9	5.2 ± 3.2	5.1 ± 0.1
Heart	23.7 ± 0.3	16.7 ± 3.0	41.0 ± 6.1	23.0 ± 0.6	18.4 ± 0.9	15.6 ± 0.5	15.3 ± 6.4	16.3 ± 2.0
Cancer	30.7 ± 2.1	27.4 ± 0.1	29.2 ± 0.4	28.9 ± 0.9	30.4 ± 0.5	29.9 ± 0.7	29.7 ± 2.8	27.1 ± 0.7
Diabetes	28.6 ± 0	24.9 ± 1.8	34.9 ± 2.1	25.9 ± 0.7	26.9 ± 1.2	27.3 ± 1.6	22.7 ± 2.3	22.9 ± 0.4
German	30.3 ± 1.3	26.7 ± 0.6	30.0 ± 1.4	29.4 ± 0.3	28.5 ± 0.8	28.2 ± 0.6	23.5 ± 0.8	23.7 ± 0.8
Flare	44.4 ± 0.3	44.0 ± 2.2	35.3 ± 0.5	34.2 ± 1.4	37.2 ± 1.4	33.4 ± 3.5	32.4 ± 1.8	34.0 ± 0.8
Mean	26.8 ± 5.5	27.0 ± 3.7	33.4 ± 1.9	24.9 ± 3.7	26.1 ± 3.3	24.1 ± 3.9	21.5 ± 4.1	21.6 ± 4.0

It can be seen that the kernel versions of the weighted neighbors and the nearest mean (PKWN and PKNM) classifiers produced better results than their Euclidean counterparts (WN and NM) in all cases. Furthermore, these results were better than those achieved by the NN and the PC classifiers in most cases. These relative advantages can also be seen from the average performance results, given in the last row of the table. The worst results were achieved by the Euclidean WN classifier. Switching to polarized kernels, the average performance of the WN classifier improved by about 25%, exceeding those of the NN and the PC classifiers by about 10% on average.

The performances of the PKWN and the PKNM classifiers are nearly the same. As for the SV classifier, the use of polarized kernels proves to be quite beneficial, as the classification results are nearly the same as those obtained by exhaustive parameter search. Moreover, since polarization implies that hard classification would be nearly as effective as soft classification, polarization involves the kernel parameter (v) alone, and the soft margin parameter is eliminated from our (PKSV) design.

5 Conclusion

We have introduced the concept of kernel polarization as a means for selecting kernel parameter values. We have shown that complete kernel polarization yields consistency in kernel sum classification. We have examined the use of polarized kernels in proximity classifiers on real-life data and found them to perform better than their Euclidean distance counterparts. Employed by a support vectors classifier, a polarized kernel was found to produce about the same classification results as a kernel optimized by exhaustive search.

Acknowledgments

I thank Ran El-Yaniv for his insightful comments and his help with the numerical examples.

References

Baram, Y. (1996). Classification by balanced representation. *Neurocomputing*, *13*, 347–357.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukhergee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, *46*, 131–159.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. on Information Theory*, *IT-13*, 21–27.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., & Kandola, J. (2001). On kernel-target alignment. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*, *14* (pp. 367–373). Cambridge, MA: MIT Press.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Orlando, FL: Academic Press.

Haykin, S. (1998). *Neural networks, a comprehensive foundation* (2nd ed.). New York: Macmillan.

Herbrich, R. (2000). *Learning linear classifiers*. Unpublished doctoral dissertation, TU Berlin.

Micchelli, C. A. (1986). Algebraic aspects of interpolation. In *Proceedings of Symposia in Applied Mathematics*, *36* (pp. 81–102). Providence, RI: American Mathematical Society.

Parzen, E. (1962). On the estimation of probability density function and the mode. *Ann. Math. Stat.*, *33*, 1065–1076.

Platt, J. C., Burges, C. J. C., Swenson, S., Weare, C., & Zheng, A. (2002). Learning a gaussian process prior for automatically generating music playlists. In S. Becker, S. Thrun, & K. Ober mayer (Eds.), *Advances in neural information processing systems*, *15* (pp. 1425–1432). Cambridge, MA: MIT Press.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Scholkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

UCI. (2003). Machine Learning Repository. Available online: www.ics.uci.edu/~mlearn/MLRepository.html.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Vapnik, V., & Chervonenkis, A. (1968). Uniform convergence of frequencies of occurrence of events to their probabilities. *Dokl. Akad. Nauk SSSR*, *181*, 915–918.

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

Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, *24*, 774–780.

Wang, L., & Chan, K. L. (2002). *Learning kernel parameters by using class separability measure*. Paper presented at the NIPS kernel workshop, Whistler, Canada.

Received December 18, 2003; accepted November 29, 2004.