A

Conventional Classifiers

A.1 Bayesian Classifiers

Bayesian classifiers are based on probability theory and give the theoretical basis for pattern classification.

Let ω be a random variable and take one of n states: $\omega_1, \ldots, \omega_n$, where ω_i indicates class i, and an m-dimensional feature vector $\mathbf{x}$ be a random variable vector. We assume that we know the a priori probabilities $P(\omega_i)$ and conditional densities $p(\mathbf{x} \,|\, \omega_i)$. Then when $\mathbf{x}$ is observed, the a posteriori probability of ω_i, $P(\omega_i \,|\, \mathbf{x})$, is calculated by the Bayes' rule:

$$P(\omega_i \,|\, \mathbf{x}) = \frac{p(\mathbf{x} \,|\, \omega_i)\, P(\omega_i)}{p(\mathbf{x})}, \tag{A.1}$$

where

$$p(\mathbf{x}) = \sum_{i=1}^{n} p(\mathbf{x} \,|\, \omega_i)\, P(\omega_i). \tag{A.2}$$

Assume that the cost c_{ij} is given when $\mathbf{x}$ is classified into class i although it is class j. Then the expected conditional cost in classifying $\mathbf{x}$ into class i, $C(\omega_i \,|\, \mathbf{x})$, is given by

$$C(\omega_i \,|\, \mathbf{x}) = \sum_{j=1}^{n} c_{ij}\, P(\omega_j \,|\, \mathbf{x}). \tag{A.3}$$

The conditional cost is minimized when $\mathbf{x}$ is classified into the class

$$\arg \min_{i=1,\ldots,n} C(\omega_i \,|\, \mathbf{x}). \tag{A.4}$$

This is called *Bayes' decision rule*.

In diagnosis problems, usually there are normal and abnormal classes. Misclassification of normal data into the abnormal class is less favorable than

misclassification of abnormal data into the normal class. In such a situation, we set a smaller cost to the former than the latter.

If we want to minimize the average probability of misclassification, we set the cost as follows:

$$c_{ij} = \begin{cases} 0 & \text{for} \quad i = j, \\ 1 & \text{for} \quad i \neq j, \end{cases} \qquad i, j = 1, \ldots, n. \tag{A.5}$$

Then, from (A.1) and (A.2) the conditional cost given by (A.3) becomes

$$\begin{aligned} C(\omega_i \,|\, \mathbf{x}) &= \sum_{\substack{j \neq i, \\ j = 1}}^{n} P(\omega_j \,|\, \mathbf{x}) \\ &= 1 - P(\omega_i \,|\, \mathbf{x}). \end{aligned} \tag{A.6}$$

Therefore, the Bayes decision rule given by (A.4) becomes

$$\begin{aligned} &\arg \max_{i = 1, \ldots, n} P(\omega_i \,|\, \mathbf{x}) \\ &= \arg \max_{i = 1, \ldots, n} p(\mathbf{x} \,|\, \omega_i)\, P(\omega_i). \end{aligned} \tag{A.7}$$

Now, we assume that the conditional densities $p(\mathbf{x} \,|\, \omega_i)$ are normal:

$$p(\mathbf{x} \,|\, \omega_i) = \frac{1}{\sqrt{(2\pi)^n \det(Q_i)}} \exp\left(-\frac{(\mathbf{x} - \mathbf{c}_i)^T\, Q_i^{-1}\, (\mathbf{x} - \mathbf{c}_i)}{2} \right), \tag{A.8}$$

where $\mathbf{c}_i$ is the mean vector and Q_i is the covariance matrix of the normal distribution for class i. If the a priori probabilities $P(\omega_i)$ are the same for $i = 1, \ldots, n$, $\mathbf{x}$ is classified into class i with the maximum $p(\mathbf{x} \,|\, \omega_i)$.

A.2 Nearest Neighbor Classifiers

Nearest neighbor classifiers use all the training data as templates for classification. In the simplest form, for a given input vector, the nearest neighbor classifier searches the nearest template and classifies the input vector into the class to which the template belongs. In the complex form the classifier treats k nearest neighbors. For a given input vector, the k nearest templates are searched and the input vector is classified into the class with the maximum number of templates. The classifier architecture is simple, but as the number of training data becomes larger, the classification time becomes longer. Therefore many methods for speeding up classification are studied [33, pp. 181–91], [202, pp. 191–201]. One uses the branch-and-bound method [90, pp. 360–2] and another edits the training data, i.e., selects or replaces the data with the suitable templates. It is proved theoretically that as the number of templates becomes larger, the expected error rate of the nearest neighbor classifier is bounded by twice that of the Bayesian classifier [97, pp. 159–75].

Usually the Euclidean distance is used to measure the distance between two data $\mathbf{x}$ and $\mathbf{y}$:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}, \tag{A.9}$$

but other distances, such as the Manhattan distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} |x_i - y_i| \tag{A.10}$$

are used. It is clear from the architecture that the recognition rate of the training data for the one-nearest neighbor classifier is 100 percent. But for the k-nearest neighbor classifier with $k > 1$, the recognition rate of the training data is not always 100 percent.

Because the distances such as the Euclidean and Manhattan distances are not invariant in scaling, classification performance varies according to the scaling of input ranges.

B

Matrices

B.1 Matrix Properties

In this section, we summarize the matrix properties used in this book. For more detailed explanation, see, e.g., [96].

Vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are *linearly independent* if

$$a_1 \mathbf{x}_1 + \cdots + a_n \mathbf{x}_n = 0 \tag{B.1}$$

holds only when $a_1 = \cdots = a_m = 0$. Otherwise, namely, at least one a_i is nonzero, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are *linearly dependent*.

Let A be an $m \times m$ matrix:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix}. \tag{B.2}$$

Then the *transpose* of A, denoted by A^T, is

$$A^T = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \cdots & \cdots & \cdots \\ a_{1m} & \cdots & a_{mm} \end{pmatrix}. \tag{B.3}$$

If A satisfies $A = A^T$, A is a *symmetric matrix*. If A satisfies $A^T A = A A^T = I$, A is an *orthogonal matrix*, where I is the $m \times m$ *unit matrix*:

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}. \tag{B.4}$$

If $m \times m$ matrices A and B satisfy $A B = I$, B is called the *inverse* of A and is denoted by A^{-1}. If A has the inverse, A is *regular* (or *nonsingular*). Otherwise, A is singular.

Lemma B.1. *Let matrices A, B, C, and D be $n \times n$, $n \times m$, $m \times n$, and $m \times m$ matrices, respectively and A, D, and $D + CA^{-1}B$ be nonsingular. Then the following relation holds:*

$$\left(A + BD^{-1}C\right)^{-1} = A^{-1} - A^{-1}B\left(D + CA^{-1}B\right)^{-1}CA^{-1}. \tag{B.5}$$

This is called the *matrix inversion lemma*.

Let $m = 1$, $B = \mathbf{b}$, $C = \mathbf{c}^T$, and $D = 1$. Then (B.5) reduces to

$$\left(A + \mathbf{b}\mathbf{c}^T\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{b}\mathbf{c}^T A^{-1}}{1 + \mathbf{c}^T A^{-1}\mathbf{b}}. \tag{B.6}$$

Using (B.6), LOO error rate estimation of linear equation based–machines such as LS support vector machines can be sped up [42, 55, 209, 278].

The *determinant* of an $m \times m$ matrix $A = \{a_{ij}\}$, $\det(A)$, is defined recursively by

$$\det(A) = \sum_{i=1}^{m} (-1)^{i+1} a_{1i} \det(A_{1i}), \tag{B.7}$$

where A_{1i} is the $(m-1) \times (m-1)$ matrix obtained by deleting the first row and the ith column from A. When $m = 1$, $\det(A) = a_{11}$.

If, for the $m \times m$ matrix A, a nonzero m-dimensional vector $\mathbf{x}$ exists for a constant λ:

$$A\,\mathbf{x} = \lambda\,\mathbf{x}, \tag{B.8}$$

λ is called an *eigenvalue* and $\mathbf{x}$ an *eigenvector*. Rearranging (B.8) gives

$$(A - \lambda\, I)\,\mathbf{x} = 0. \tag{B.9}$$

Thus, (B.9) has nonzero $\mathbf{x}$, when

$$\det(A - \lambda\, I) = 0, \tag{B.10}$$

which is called the *characteristic equation*.

Theorem B.2. *All the eigenvalues of a real symmetric matrix are real.*

Theorem B.3. *Eigenvectors associated with different eigenvalues for a real symmetric matrix are orthogonal.*

For an m-dimensional vector $\mathbf{x}$ and an $m \times m$ symmetric matrix A, $Q = \mathbf{x}^T A\,\mathbf{x}$ is called the *quadratic form*. If for any nonzero $\mathbf{x}$, $Q = \mathbf{x}^T\,A\,\mathbf{x} \geq 0$, Q is *positive semidefinite*. Matrix Q is *positive definite* if the strict inequality holds. Let L be an $m \times m$ orthogonal matrix. By $\mathbf{y} = L\,\mathbf{x}$, $\mathbf{x}$ is transformed into $\mathbf{y}$. This is the transformation from one orthonormal base into another orthonormal basis. The quadratic form Q is

$$\begin{aligned} Q &= \mathbf{x}^T A\,\mathbf{x} \\ &= \mathbf{y}^T L\,A\,L^T\,\mathbf{y}. \end{aligned} \tag{B.11}$$

Theorem B.4. *The characteristic equations for A and $L\,A\,L^T$ are the same.*

Theorem B.5. *If an $m \times m$ real symmetric matrix A is diagonalized by L:*

$$L\,A\,L^T = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_m \end{pmatrix}, \tag{B.12}$$

$\lambda_1, \ldots, \lambda_m$ are the eigenvalues of A and the ith row of L is the eigenvector associated with λ_i.

If all the eigenvalues of A are positive, A is *positive definite*. If all the eigenvalues are nonnegative, A is *positive semidefinite*.

B.2 Least Squares Methods and Singular Value Decomposition

Assume that we have M input-output pairs $\{(\mathbf{a}'_1, b_1), \ldots, (\mathbf{a}'_M, b_M)\}$ in the $(n-1)$-dimensional input space $\mathbf{x}'$ and one-dimensional output space y. Now using the least squares method, we determine the linear relation of the input-output pairs:

$$y = \mathbf{p}^T\,\mathbf{x}' + q, \tag{B.13}$$

where $\mathbf{p}$ is the $(n-1)$-dimensional vector, q is a scalar constant, and $M \geq n$.

Rewriting (B.13), we get

$$(\mathbf{x}'^T, 1) \begin{pmatrix} \mathbf{p} \\ q \end{pmatrix} = y. \tag{B.14}$$

Substituting $\mathbf{a}'_i$ and b_i into $\mathbf{x}'$ and y in (B.14), respectively, and replacing $(\mathbf{p}^T, q)^T$ with the n-dimensional parameter vector $\mathbf{x}$, we obtain

$$\mathbf{a}_i^T\,\mathbf{x} = b_i \qquad \text{for} \quad i = 1, \ldots, M, \tag{B.15}$$

where $\mathbf{a}_i = (\mathbf{a}_i'^T, 1)^T$.

We determine the parameter vector $\mathbf{x}$ so that the sum-of-squares error:

$$E = (A\,\mathbf{x} - \mathbf{b})^T (A\,\mathbf{x} - \mathbf{b}) \tag{B.16}$$

is minimized, where A is an $M \times n$ matrix and $\mathbf{b}$ is an M-dimensional vector:

$$A = \begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_M^T \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{pmatrix}. \tag{B.17}$$

Here, if the rank of A is smaller than n, there is no unique solution. In that situation, we determine $\mathbf{x}$ so that the Euclidean norm of $\mathbf{x}$ is minimized.

Matrix A is decomposed into singular values [96]:

$$A = USV^T, \tag{B.18}$$

where U and V are $M \times M$ and $n \times n$ orthogonal matrices, respectively, and S is an $M \times n$ diagonal matrix given by

$$S = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ \hline & 0_{M-n,n} & \end{pmatrix}. \tag{B.19}$$

Here, σ_i are singular values and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, and $0_{M-n,n}$ is the $(M-n) \times n$ zero matrix.

It is known that the columns of U and V are the eigenvectors of AA^T and A^TA, respectively, and that the singular values correspond to the square roots of the eigenvalues of AA^T, which are the same as those of A^TA [60, pp. 434–5]. Thus when A is a symmetric square matrix, $U = V$ and $A = USU^T$. This is similar to the diagonalization of the square matrix given by Theorem B.5. The difference is that the singular values A are the absolute values of the eigenvalues of A. Thus, if A is a positive (semi)definite matrix, both decompositions are the same.

Rewriting (B.16), we get [96, p. 256]

$$\begin{aligned} E &= (A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - \mathbf{b}) \\ &= (USV^T\mathbf{x} - UU^T\mathbf{b})^T (A\mathbf{x} - \mathbf{b}) \\ &= (SV^T\mathbf{x} - U^T\mathbf{b})^T (SV^T\mathbf{x} - U^T\mathbf{b}) \\ &= \sum_{i=1}^{n} (\sigma_i \mathbf{v}_i^T \mathbf{x} - \mathbf{u}_i^T \mathbf{b})^2 + \sum_{i=n+1}^{M} (\mathbf{u}_i^T \mathbf{b})^2, \end{aligned} \tag{B.20}$$

where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_M)$ and $V = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$. Assuming that the rank of A is $r\,(\leq n)$, (B.20) is minimized when

$$\sigma_i \mathbf{v}_i^T \mathbf{x} = \mathbf{u}_i^T \mathbf{b} \qquad \text{for} \quad i = 1, \ldots, r, \tag{B.21}$$

$$\mathbf{v}_i^T \mathbf{x} = 0 \qquad \text{for} \quad i = r+1, \ldots, n. \tag{B.22}$$

Equation (B.22) is imposed to obtain the minimum Euclidean norm solution. From (B.21) and (B.22), we obtain

$$\mathbf{x} = VS^+U^T\mathbf{b} = A^+\mathbf{b}, \tag{B.23}$$

where S^+ is the $n \times M$ diagonal matrix given by

$$S^+ = \left(\begin{array}{ccc|c} \frac{1}{\sigma_1} & & 0 & \\ & \ddots & & 0 \\ & & \frac{1}{\sigma_r} & \\ \hline 0 & & 0 & \end{array}\right). \tag{B.24}$$

We call A^+ the pseudo-inverse of A. We must bear in mind that in calculating the pseudo-inverse, we replace the reciprocal of 0 with 0, not with infinity. This ensures the minimum norm solution.

From (B.18) and (B.23),

$$\begin{aligned} A^+ A &= V\, S^+ U^T\, U\, S\, V^T \\ &= V\, S^+\, S\, V^T \\ &= V \begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{n-r,r} & 0_{n-r} \end{pmatrix} V^T \\ &= \begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{n-r,r} & 0_{n-r} \end{pmatrix}, \end{aligned} \tag{B.25}$$

$$\begin{aligned} A\, A^+ &= U\, S\, S^+\, U^T \\ &= \begin{pmatrix} I_r & 0_{r,M-r} \\ 0_{M-r,r} & 0_{M-r} \end{pmatrix}, \end{aligned} \tag{B.26}$$

where I_r is the $r \times r$ unit matrix, 0_i is the $i \times i$ zero matrix, and $0_{i,j}$ is the $i \times j$ zero matrix. Therefore, if A is a square matrix with rank n, $A^+ A = A\, A^+ = I$. Namely, the pseudo-inverse of A coincides with the inverse of A, A^{-1}. If $M > n$ and the rank of A is n, $A^+ A = I$ but $A\, A^+ \neq I$. In this case A^+ is given by

$$A^+ = (A^T A)^{-1} A^T. \tag{B.27}$$

This is obtained by taking the derivative of (B.16) with respect to $\mathbf{x}$ and equating the result to zero.

When $M > n$ and the rank of A is smaller than n, $A^+ A \neq I$ and $A\, A^+ \neq I$.

Even when $A^T A$ is nonsingular, it is recommended to calculate the pseudo-inverse by singular value decomposition, not using (B.27). Because if $A^T A$ is near singular, $(A^T A)^{-1} A^T$ is vulnerable to the small singular values [195, pp. 59–70].

B.3 Covariance Matrices

Let $\mathbf{x}_1, \ldots, \mathbf{x}_M$ be M samples of the m-dimensional random variable X. Then the sample covariance matrix of X is given by

$$Q = \frac{1}{M} \sum_{i=1}^{M} (\mathbf{x}_i - \mathbf{c})\, (\mathbf{x}_i - \mathbf{c})^T, \tag{B.28}$$

where $\mathbf{c}$ is the mean vector:

$$\mathbf{c} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{x}_i. \tag{B.29}$$

To get the unbiased estimate of the covariance matrix, we replace M with $M - 1$ in (B.28), but in this book we use (B.28) as the sample covariance matrix.

Let

$$\mathbf{y}_i = \mathbf{x}_i - \mathbf{c}. \tag{B.30}$$

Then (B.28) becomes

$$Q = \frac{1}{M} \sum_{i=1}^{M} \mathbf{y}_i \, \mathbf{y}_i^T. \tag{B.31}$$

From (B.29) and (B.30), $\mathbf{y}_1, \ldots, \mathbf{y}_M$ are linearly dependent.

According to the definition, the covariance matrix Q is symmetric. Matrix Q is positive (semi)definite, as the following theorem shows.

Theorem B.6. *The covariance matrix Q given by (B.31) is positive definite if $\mathbf{y}_1, \ldots, \mathbf{y}_M$ have at least m linearly independent data. Matrix Q is positive semidefinite, if any m data from $\mathbf{y}_1, \ldots, \mathbf{y}_M$ are linearly dependent.*

Proof. Let $\mathbf{z}$ be an m-dimensional nonzero vector. From (B.31),

$$\begin{aligned} \mathbf{z}^T Q \, \mathbf{z} &= \mathbf{z}^T \left(\frac{1}{M} \sum_{i=1}^{M} \mathbf{y}_i \, \mathbf{y}_i^T \right) \mathbf{z} \\ &= \frac{1}{M} \sum_{i=1}^{M} \left(\mathbf{z}^T \, \mathbf{y}_i \right) \left(\mathbf{z}^T \, \mathbf{y}_i \right)^T \\ &= \frac{1}{M} \sum_{i=1}^{M} \left(\mathbf{z}^T \, \mathbf{y}_i \right)^2 \geq 0. \end{aligned} \tag{B.32}$$

Thus Q is positive semidefinite. If there are m linearly independent data in $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$, they span the m-dimensional space. Because any $\mathbf{z}$ is expressed by a linear combination of these data, the strict inequality holds for (B.32).

Because $\mathbf{y}_1, \ldots, \mathbf{y}_M$ are linearly dependent, at least $m + 1$ samples are necessary so that Q becomes positive definite. ■

Assuming that Q is positive definite, the following theorem holds.

Theorem B.7. *If Q is positive definite, the mean square weighted distance for $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ is m:*

$$\frac{1}{M} \sum_{i=1}^{M} \mathbf{y}_i^T \, Q^{-1} \, \mathbf{y}_i = m. \tag{B.33}$$

Proof. Let P be the orthogonal matrix that diagonalizes Q. Namely,

$$P\,Q\,P^T = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \tag{B.34}$$

where diag denotes the diagonal matrix, and $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of Q. From (B.34),

$$Q = P^T \mathrm{diag}(\lambda_1, \ldots, \lambda_m) P, \tag{B.35}$$

$$Q^{-1} = P^T \mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_m^{-1}) P. \tag{B.36}$$

Let

$$\tilde{\mathbf{y}}_i = P\,\mathbf{y}_i. \tag{B.37}$$

Then from (B.31) and (B.37), (B.34) becomes

$$\frac{1}{M} \sum_{i=1}^{M} \tilde{\mathbf{y}}_i \tilde{\mathbf{y}}_i^T = \mathrm{diag}(\lambda_1, \ldots, \lambda_m). \tag{B.38}$$

Thus for the diagonal elements of (B.38),

$$\frac{1}{M} \sum_{i=1}^{M} \tilde{y}_{ik}^2 = \lambda_k \qquad \text{for} \quad k = 1, \ldots, m, \tag{B.39}$$

where $\tilde{y}_{ik}$ is the kth element of $\tilde{\mathbf{y}}_i$. From (B.36) and (B.37), the left-hand side of (B.33) becomes

$$\begin{aligned} \frac{1}{M} \sum_{i=1}^{M} \mathbf{y}_i^T\, Q^{-1}\, \mathbf{y}_i &= \frac{1}{M} \sum_{i=1}^{M} \tilde{\mathbf{y}}_i^T\, \mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_m^{-1})\, \tilde{\mathbf{y}}_i \\ &= \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{m} \lambda_k^{-1}\, \tilde{y}_{ik}^2. \end{aligned} \tag{B.40}$$

Thus from (B.39) and (B.40), the theorem holds. ■

C

Quadratic Programming

Quadratic programming is the basis of support vector machines. Here we summarize some of the basic properties of quadratic programming.

C.1 Optimality Conditions

Consider the following optimization problem. Minimize

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\,\mathbf{x} + \mathbf{c}^T\,\mathbf{x} \tag{C.1}$$

subject to

$$g_i(\mathbf{x}) = \mathbf{a}_i^T\mathbf{x} + b_i \geq 0 \quad \text{for} \quad i = 1, \ldots, k, \tag{C.2}$$
$$h_i(\mathbf{x}) = \mathbf{d}_i^T\mathbf{x} + e_i = 0 \quad \text{for} \quad i = 1, \ldots, o, \tag{C.3}$$

where $\mathbf{x}$, $\mathbf{a}_i$, and $\mathbf{d}_i$ are m-dimensional vectors; Q is an $m \times m$ positive semidefinite matrix; and b_i and e_i are scalar constants. This problem is called the *quadratic programming problem*. Because of the linear equality and inequality constraints, $\mathbf{x}$ is in a closed convex domain.

We introduce the Lagrange multipliers:

$$L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) - \sum_{i=1}^{k} \alpha_i\, g_i(\mathbf{x}) + \sum_{i=1}^{o} \beta_i\, h_i(\mathbf{x}), \tag{C.4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$, $\alpha_i \geq 0$ for $i = 1, \ldots, k$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_o)^T$. Then the following theorem holds.

Theorem C.1. *The optimal solution* $(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ *exists if and only if the following conditions are satisfied:*

$$\frac{\partial L(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \mathbf{x}} = \mathbf{0}, \tag{C.5}$$

$$\alpha_i^* g_i(\mathbf{x}^*) = 0 \qquad for \quad i = 1, \ldots, k, \tag{C.6}$$

$$\alpha_i^* \geq 0 \qquad for \quad i = 1, \ldots, k, \tag{C.7}$$

$$h_i(\mathbf{x}^*) = 0 \qquad for \quad i = 1, \ldots, o. \tag{C.8}$$

These conditions are called the *Karush-Kuhn-Tucker (KKT) conditions* and the conditions given by (C.6) are called the *Karush-Kuhn-Tucker complementarity conditions*. If there is no confusion, the KKT complementarity conditions are abbreviated the *KKT conditions*.

The KKT complementarity condition means that if $\alpha_i^* > 0$, $g_i(\mathbf{x}^*) = 0$ (it is called *active*); and if $\alpha_i^* = 0$, $g_i(\mathbf{x}^*) \geq 0$ (it is called *inactive*).

C.2 Properties of Solutions

The optimal solution can be interpreted geometrically. From (C.4),

$$\frac{\partial L(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}} - \sum_{i=1}^{k} \alpha_i \frac{\partial g_i(\mathbf{x}^*)}{\partial \mathbf{x}} + \sum_{i=1}^{o} \beta_i \frac{\partial h_i(\mathbf{x}^*)}{\partial \mathbf{x}} = \mathbf{0}. \tag{C.9}$$

If some inequality constraints are inactive (i.e., the associated Lagrange multipliers are zero) for the optimal solution, we can discard the associated terms from (C.9). If they are active, they can be treated as the equality constraints. Thus, without loss of generality, we can assume that the inequality constraints are all inactive. Then the optimal solution must satisfy

$$-\frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}} = \sum_{i=1}^{o} \beta_i^* \frac{\partial h_i(\mathbf{x}^*)}{\partial \mathbf{x}}. \tag{C.10}$$

The negative gradient of $f(\mathbf{x})$, $-\partial f(\mathbf{x})/\partial \mathbf{x}$, points the direction in which $f(\mathbf{x})$ decreases the most. And at the optimal solution the negative gradient must be perpendicular to the equality constraint $h_i(\mathbf{x}) = 0$ or parallel to $\partial h_i(\mathbf{x}^*)/\partial \mathbf{x}$. Therefore, the negative gradient must be in the subspace spanned by $\partial g_i(\mathbf{x}^*)/\partial \mathbf{x}$ $(i = 1, \ldots, o)$, which is equivalent to (C.10).

If Q is positive definite, the solution is unique. And if Q is positive semidefinite, the solution may not be unique. But if $\mathbf{x}_o$ and $\mathbf{x}'_o$ are solutions, $\lambda \mathbf{x}_o + (1 - \lambda)\mathbf{x}'_o$, where $1 \geq \lambda \geq 0$, is also a solution.

For the optimal solution $(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, the following relation holds:

$$L(\mathbf{x}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \geq L(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \geq L(\mathbf{x}^*, \boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{C.11}$$

Namely, $L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is minimized with respect to $\mathbf{x}$ and maximized with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Thus the optimal point is a saddle point.

Example C.2. Consider the following problem. Minimize

$$f(\mathbf{x}) = \frac{1}{2}\,\mathbf{x}^T\,Q\,\mathbf{x} \tag{C.12}$$

subject to

$$2 \geq x_1 + x_2 \geq 1, \tag{C.13}$$

where $\mathbf{x} = (x_1\, x_2)^T$ and Q is positive definite:

$$Q = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}. \tag{C.14}$$

Because

$$L(\mathbf{x}, \boldsymbol{\alpha}) = \frac{1}{2}\,(x_1^2 + x_1 x_2 + x_2^2) - \alpha_1\,(x_1 + x_2 - 1) - \alpha_2\,(2 - x_1 - x_2), \tag{C.15}$$

the KKT conditions are given by

$$\frac{\partial L(\mathbf{x}, \boldsymbol{\alpha})}{\partial x_1} = x_1 + \frac{1}{2}\,x_2 - \alpha_1 + \alpha_2 = 0, \tag{C.16}$$

$$\frac{\partial L(\mathbf{x}, \boldsymbol{\alpha})}{\partial x_2} = \frac{1}{2}\,x_1 + x_2 - \alpha_1 + \alpha_2 = 0, \tag{C.17}$$

$$\alpha_1\,(x_1 + x_2 - 1) = 0, \quad \alpha_2\,(2 - x_1 - x_2) = 0, \tag{C.18}$$

$$\alpha_1 \geq 0, \quad \alpha_2 \geq 0. \tag{C.19}$$

Subtracting (C.17) from (C.16), we obtain $x_1 = x_2$. Therefore, from $f(\mathbf{x}) = 3\,x_1^2/2$ and the KKT conditions, the optimal solution satisfies

$$x_1 = x_2 = \frac{1}{2}, \quad \alpha_1 = \frac{3}{2}, \quad \alpha_2 = 0. \tag{C.20}$$

Thus the solution is unique (see Fig. C.1).

Let Q be positive semidefinite:

$$Q = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{C.21}$$

Because

$$L(\mathbf{x}, \boldsymbol{\alpha}) = \frac{1}{2}\,(x_1 + x_2)^2 - \alpha_1\,(x_1 + x_2 - 1) - \alpha_2\,(2 - x_1 - x_2), \tag{C.22}$$

the KKT conditions are given by

$$\frac{\partial L(\mathbf{x}, \boldsymbol{\alpha})}{\partial x_1} = \frac{\partial L(\mathbf{x}, \boldsymbol{\alpha})}{\partial x_2} = x_1 + x_2 - \alpha_1 + \alpha_2 = 0, \tag{C.23}$$

$$\alpha_1\,(x_1 + x_2 - 1) = 0, \quad \alpha_2\,(2 - x_1 - x_2) = 0, \tag{C.24}$$

$$\alpha_1 \geq 0, \quad \alpha_2 \geq 0. \tag{C.25}$$

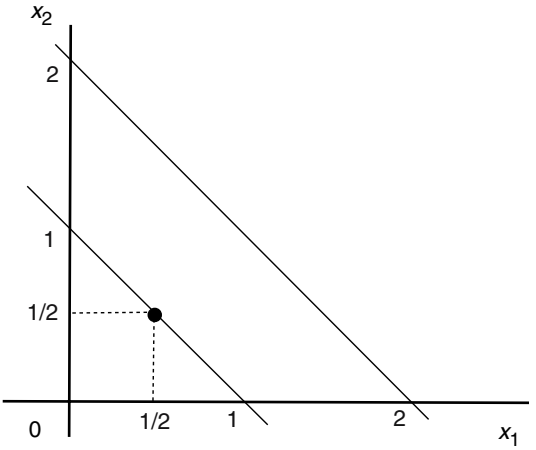

Fig. C.1. Unique solution with a positive definite matrix

So long as $x_1 + x_2$ is constant, the value of the objective function does not change. Thus the optimal solution satisfies $x_1 + x_2 = 1$. Therefore, from (C.23) and (C.24), the optimal solution satisfies

$$x_1 + x_2 = 1, \quad \alpha_1 = 1, \quad \alpha_2 = 0. \tag{C.26}$$

Thus the solution is nonunique (see Fig. C.2).

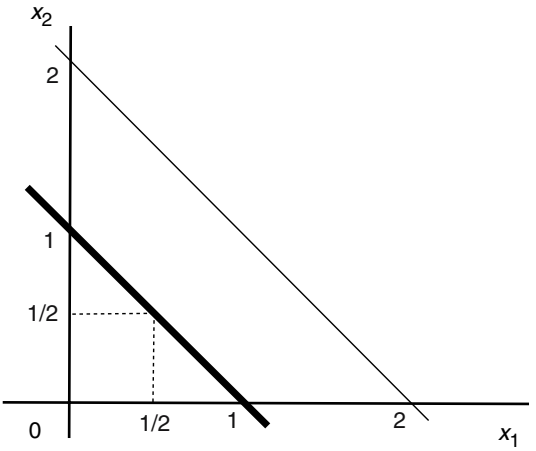

Fig. C.2. Nonunique solution with a positive semidefinite matrix

D

Positive Semidefinite Kernels and Reproducing Kernel Hilbert Space

Support vector machines are based on the theory of reproducing kernel Hilbert space. Here, we summarize some of the properties of positive semidefinite kernels and reproducing kernel Hilbert space based on [30].

D.1 Positive Semidefinite Kernels

Definition D.1. Let $H(\mathbf{x}, \mathbf{x}')$ be a real-valued symmetric function with $\mathbf{x}$ and $\mathbf{x}'$ being m-dimensional vectors. For any set of data $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ and $\mathbf{h}_M = (h_1, \ldots, h_M)^T$ with M being any natural number, if

$$\mathbf{h}_M^T H_M \mathbf{h}_M \geq 0 \tag{D.1}$$

is satisfied (i.e., H_M is a positive semidefinite matrix), we call $H(\mathbf{x}, \mathbf{x}')$ a *positive semidefinite kernel*, where

$$H_M = \begin{pmatrix} H(\mathbf{x}_1, \mathbf{x}_1) & \cdots & H(\mathbf{x}_1, \mathbf{x}_M) \\ \vdots & \ddots & \vdots \\ H(\mathbf{x}_M, \mathbf{x}_1) & \cdots & H(\mathbf{x}_M, \mathbf{x}_M) \end{pmatrix}. \tag{D.2}$$

If (D.1) is satisfied under the constraint

$$\sum_{i=1}^{M} h_i = 0, \tag{D.3}$$

$H(\mathbf{x}, \mathbf{x}')$ is called a *conditionally positive semidefinite kernel*.

From the definition it is obvious that if $H(\mathbf{x}, \mathbf{x}')$ is positive semidefinite, it is also conditionally positive semidefinite. In the following we discuss several properties of (conditionally) positive semidefinite kernels that are useful for constructing positive semidefinite kernels.

Theorem D.2. *If*

$$H(\mathbf{x}, \mathbf{x}') = a, \tag{D.4}$$

where $a > 0$, $H(\mathbf{x}, \mathbf{x}')$ *is positive semidefinite.*

Proof. Because for any natural number M,

$$H_M = (\sqrt{a}, \ldots, \sqrt{a})^T (\sqrt{a}, \ldots, \sqrt{a}), \tag{D.5}$$

$H(\mathbf{x}, \mathbf{x}')$ is positive semidefinite. ■

Theorem D.3. *If* $H_1(\mathbf{x}, \mathbf{x}')$ *and* $H_2(\mathbf{x}, \mathbf{x}')$ *are positive semidefinite kernels,*

$$H(\mathbf{x}, \mathbf{x}') = a_1 H_1(\mathbf{x}, \mathbf{x}') + a_2 H_2(\mathbf{x}, \mathbf{x}') \tag{D.6}$$

is also positive semidefinite, where a_1 *and* a_2 *are positive.*

Proof. Because for any M, h_i, and $\mathbf{x}_i$

$$\mathbf{h}_M^T \left(a_1 H_{1M} + a_2 H_{2M}\right) \mathbf{h}_M = a_1 \mathbf{h}_M^T H_{1M} \mathbf{h}_M + a_2 \mathbf{h}_M^T H_{2M} \mathbf{h}_M \geq 0, \tag{D.7}$$

$H(\mathbf{x}, \mathbf{x}')$ is positive semidefinite. ■

Theorem D.4. *If* $H(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})\, f(\mathbf{x}')$, *where* $f(\mathbf{x})$ *is an arbitrary scalar function,* $H(\mathbf{x}, \mathbf{x}')$ *is positive semidefinite.*

Proof. Because for any M, h_i, and $\mathbf{x}_i$

$$\sum_{i,j=1}^{M} h_i\, h_j\, f(\mathbf{x}_i)\, f(\mathbf{x}_j) = \left(\sum_{i=1}^{M} h_i\, f(\mathbf{x}_i) \right)^2 \geq 0, \tag{D.8}$$

$H(\mathbf{x}, \mathbf{x}')$ is positive semidefinite. ■

Theorem D.5. *If* $H_1(\mathbf{x}, \mathbf{x}')$ *and* $H_2(\mathbf{x}, \mathbf{x}')$ *are positive semidefinite,*

$$H(\mathbf{x}, \mathbf{x}') = H_1(\mathbf{x}, \mathbf{x}')\, H_2(\mathbf{x}, \mathbf{x}') \tag{D.9}$$

is also positive semidefinite.

Proof. It is sufficient to show that if $M \times M$ matrices $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ are positive semidefinite, $\{a_{ij}\, b_{ij}\}$ is also positive semidefinite.

Because A is positive semidefinite, A is expressed by $A = F^T F$, where F is an $M \times M$ matrix. Then $a_{ij} = \mathbf{f}_i^T \mathbf{f}_j$, where $\mathbf{f}_j$ is the jth column vector of F. Thus for arbitrary $h_1, \ldots, h_M$,

$$\sum_{i,j=1}^{M} h_i\, h_j\, \mathbf{f}_i^T\, \mathbf{f}_j\, b_{ij} = \sum_{i,j=1}^{M} (h_i\, \mathbf{f}_i)^T (h_j\, \mathbf{f}_j)\, b_{ij} \geq 0. \blacksquare \tag{D.10}$$

Example D.6. The linear kernel $H(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}$ is positive semidefinite because $H_M = (\mathbf{x}_1, \ldots, \mathbf{x}_M)^T (\mathbf{x}_1, \ldots, \mathbf{x}_M)$. Thus, from Theorems D.2 to D.5 the polynomial kernel given by $H(\mathbf{x}, \mathbf{x}') = (1+\mathbf{x}^T\mathbf{x}')^d$ is positive semidefinite.

Corollary D.7. *If $H(\mathbf{x}, \mathbf{x}')$ and $H'(\mathbf{y}, \mathbf{y}')$ are positive semidefinite kernels, where $\mathbf{x}$ and $\mathbf{y}$ may be of different dimensions, $H(\mathbf{x}, \mathbf{x}')\, H'(\mathbf{y}, \mathbf{y}')$ is also a positive semidefinite kernel.*

Corollary D.8. *Let $H(\mathbf{x}, \mathbf{x}')$ be positive semidefinite and satisfy*

$$|H(\mathbf{x}, \mathbf{x}')| \leq \rho, \tag{D.11}$$

where $\rho > 0$. Then if

$$f(y) = \sum_{i=1}^{\infty} a_i\, y^i \tag{D.12}$$

converges for $|y| \leq \rho$, where $a_i \geq 0$ for all integers i, the composed kernel $f(H(\mathbf{x}, \mathbf{x}'))$ is also positive semidefinite. ■

Proof. From Theorem D.5, $H^i(\mathbf{x}, \mathbf{x}')$ is positive semidefinite. Then from Theorem D.5,

$$\sum_{i=0}^{N} a_i\, H^i(\mathbf{x}, \mathbf{x}') \tag{D.13}$$

is positive semidefinite for all integers N. Therefore, so is $f(H(\mathbf{x}, \mathbf{x}'))$. ■

From Corollary D.8, especially for positive semidefinite kernel $H(\mathbf{x}, \mathbf{x}')$, $\exp(H(\mathbf{x}, \mathbf{x}'))$ is also positive semidefinite.

In the following we clarify the relations between positive and conditionally positive semidefinite kernels.

Lemma D.9. *Let*

$$H(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') + K(\mathbf{x}_0, \mathbf{x}_0) - K(\mathbf{x}, \mathbf{x}_0) - K(\mathbf{x}', \mathbf{x}_0). \tag{D.14}$$

Then $H(\mathbf{x}, \mathbf{x}')$ is positive semidefinite, if and only if $K(\mathbf{x}, \mathbf{x}')$ is conditionally positive semidefinite.

Proof. For $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ and $\mathbf{h}_M = (h_1, \ldots, h_M)^T$ with

$$\sum_{i=1}^{M} h_i = 0, \tag{D.15}$$

we have

$$\mathbf{h}_M^T H_M \mathbf{h}_M = \mathbf{h}_M^T K_M \mathbf{h}_M. \tag{D.16}$$

Thus, if $H(\mathbf{x}, \mathbf{x}')$ is positive semidefinite, $K(\mathbf{x}, \mathbf{x}')$ is conditionally positive semidefinite.

On the other hand, suppose that $K(\mathbf{x}, \mathbf{x}')$ is conditionally positive semidefinite. Then for $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ and $\mathbf{h}_M = (h_1, \ldots, h_M)^T$ with

$$h_0 = -\sum_{i=1}^{M} h_i, \tag{D.17}$$

we have

$$\begin{aligned} 0 &\leq \sum_{i,j=0}^{M} h_i\, h_j\, K(\mathbf{x}_i, \mathbf{x}_j) \\ &= \sum_{i,j=1}^{M} h_i\, h_j\, K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{M} h_i\, h_0\, K(\mathbf{x}_i, \mathbf{x}_0) + \sum_{j=1}^{M} h_0\, h_j\, K(\mathbf{x}_0, \mathbf{x}_j) \\ &\quad + h_0^2\, K(\mathbf{x}_0, \mathbf{x}_0) \\ &= \sum_{i,j=1}^{M} h_i\, h_j\, H(\mathbf{x}_i, \mathbf{x}_j). \end{aligned} \tag{D.18}$$

Therefore, $H(\mathbf{x}, \mathbf{x}')$ is positive semidefinite. ■

Theorem D.10. *Kernel $K(\mathbf{x}, \mathbf{x}')$ is conditionally positive semidefinite if and only if $\exp(\gamma\, K(\mathbf{x}, \mathbf{x}'))$ is positive semidefinite for any positive γ.*

Proof. If $\exp(\gamma\, K(\mathbf{x}, \mathbf{x}'))$ is positive semidefinite, $\exp(\gamma\, K(\mathbf{x}, \mathbf{x}')) - 1$ is conditionally positive semidefinite. So is the limit

$$K(\mathbf{x}, \mathbf{x}') = \lim_{\gamma \to +0} \frac{\exp(\gamma\, K(\mathbf{x}, \mathbf{x}')) - 1}{\gamma}. \tag{D.19}$$

Now let $K(\mathbf{x}, \mathbf{x}')$ be conditionally positive semidefinite and choose some $\mathbf{x}_0$ and $H(\mathbf{x}, \mathbf{x}')$ as in Lemma D.9. Then for positive γ

$$\gamma K(\mathbf{x}, \mathbf{x}') = \gamma H(\mathbf{x}, \mathbf{x}') - \gamma K(\mathbf{x}_0, \mathbf{x}_0) + \gamma K(\mathbf{x}, \mathbf{x}_0) + \gamma K(\mathbf{x}', \mathbf{x}_0). \tag{D.20}$$

Thus,

$$\begin{aligned} \exp(\gamma K(\mathbf{x}, \mathbf{x}')) = \exp(\gamma H(\mathbf{x}, \mathbf{x}'))\exp(-\gamma K(\mathbf{x}_0, \mathbf{x}_0)) \\ \times \exp(\gamma K(\mathbf{x}, \mathbf{x}_0))\exp(\gamma K(\mathbf{x}', \mathbf{x}_0)). \end{aligned} \tag{D.21}$$

From Theorems D.4 and D.5 and Corollary D.8, $\exp(\gamma K(\mathbf{x}, \mathbf{x}'))$ is positive semidefinite. ■

Example D.11. Kernel $H(\mathbf{x}, \mathbf{x}') = -\|\mathbf{x} - \mathbf{x}'\|^2$ is conditionally positive semidefinite because for $\sum_i^M h_i = 0$,

$$\begin{aligned}\mathbf{h}_M^T\, H_M\, \mathbf{h}_M &= -\sum_{i=1}^{M} h_i\, h_j\, \|\mathbf{x}_i - \mathbf{x}_j\|^2 \\ &= -\sum_{i,j=1}^{M} h_i\, h_j\, (\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i^T \mathbf{x}_j + \mathbf{x}_j^T \mathbf{x}_j) \\ &= 2\left(\sum_{i=1}^{M} h_i\, \mathbf{x}_i\right)^T \left(\sum_{i=1}^{M} h_i\, \mathbf{x}_i\right) \geq 0. \end{aligned} \tag{D.22}$$

Thus, $\exp(-\gamma\, \|\mathbf{x} - \mathbf{x}'\|^2)$ is positive semidefinite.

D.2 Reproducing Kernel Hilbert Space

Because a positive semidefinite kernel has the associated feature space called the *reproducing kernel Hilbert space (RKHS)*, support vector machines can determine the optimal hyperplane in that space using the kernel trick. In this section, we discuss reproducing kernel Hilbert spaces for positive and conditionally positive semidefinite kernels.

For the positive semidefinite kernels, the following theorem holds.

Theorem D.12. *Let X be the input space and $H(\mathbf{x}, \mathbf{x}')\, (\mathbf{x}, \mathbf{x}' \in X)$ be a positive semidefinite kernel. Let H_0 be the space spanned by the functions $\{H_{\mathbf{x}} \,|\, \mathbf{x} \in X\}$ where*

$$H_{\mathbf{x}}(\mathbf{x}') = H(\mathbf{x}, \mathbf{x}'). \tag{D.23}$$

Then there exist a Hilbert space H, which is a complete space of H_0, and the mapping from X to H such that

$$H(\mathbf{x}, \mathbf{x}') = \langle H_{\mathbf{x}}, H_{\mathbf{x}'} \rangle. \tag{D.24}$$

Here, instead of $\mathbf{x}^T \mathbf{x}'$, we use $\langle \mathbf{x}, \mathbf{x}' \rangle$ to denote the dot-product.

Proof. Let $H_{\mathbf{x}}(\mathbf{x}') = H(\mathbf{x}, \mathbf{x}')$ and H_0 be a linear subspace generated by the functions $\{H_{\mathbf{x}} \,|\, \mathbf{x} \in X\}$. Then for $f, g \in H_0$ expressed by

$$f = \sum_{\mathbf{x}_i \in X} c_i\, H_{\mathbf{x}_i}, \tag{D.25}$$

$$g = \sum_{\mathbf{x}'_j \in X} d_j\, H_{\mathbf{x}'_j}, \tag{D.26}$$

we define the dot-product as follows:

$$\begin{aligned}\langle f, g \rangle &= \sum_{\mathbf{x}'_j \in X} d_j\, f(\mathbf{x}'_j) \\ &= \sum_{\mathbf{x}_i, \mathbf{x}'_j \in X} c_i\, d_j\, H(\mathbf{x}_i, \mathbf{x}'_j) \\ &= \sum_{\mathbf{x}_i \in X} c_i\, g(\mathbf{x}_i). \end{aligned} \tag{D.27}$$

Now we show that (D.27) satisfies the properties of the dot-product. Clearly, (D.27) is symmetric and linear. Also, according to the assumption of $H(\mathbf{x}, \mathbf{x}')$ being positive semidefinite,

$$\langle f, f \rangle = \sum_{\mathbf{x}_i, \mathbf{x}_j \in X} c_i \, c_j \, H(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \tag{D.28}$$

is satisfied. Here, the strict equality holds if and only if f is identically zero. Thus, (D.27) is the dot-product. Hence, H_0 is a pre-Hilbert space and its completion H is a Hilbert space, which is called *RKHS associated with* $H_{\mathbf{x}}$.

From (D.27) the following reproducing property is readily obtained:

$$\langle f, H_{\mathbf{x}} \rangle = f(\mathbf{x}). \tag{D.29}$$

In particular,

$$\langle H_{\mathbf{x}}, H_{\mathbf{x}'} \rangle = H(\mathbf{x}, \mathbf{x}'). \blacksquare \tag{D.30}$$

For a conditionally positive semidefinite kernel, for $f \in H_0$ the following theorem holds.

Theorem D.13. *Let* $H(\mathbf{x}, \mathbf{x}') \, (\mathbf{x}, \mathbf{x}' \in X)$ *be a conditionally positive semidefinite kernel. Then there exist a Hilbert space* H *and a mapping* $K_{\mathbf{x}}$ *from* X *to* H *such that*

$$H(\mathbf{x}, \mathbf{x}') - \frac{1}{2} H(\mathbf{x}, \mathbf{x}) - \frac{1}{2} H(\mathbf{x}', \mathbf{x}') = -\|K_{\mathbf{x}} - K_{\mathbf{x}'}\|^2. \tag{D.31}$$

Proof. For $\mathbf{x}_0$ we define

$$K(\mathbf{x}, \mathbf{x}') = \frac{1}{2} \left(H(\mathbf{x}, \mathbf{x}') + H(\mathbf{x}_0, \mathbf{x}_0) - H(\mathbf{x}, \mathbf{x}_0) - H(\mathbf{x}', \mathbf{x}_0) \right), \tag{D.32}$$

which is a positive semidefinite kernel from Lemma D.9. Let H be the associated RKHS for $K(\mathbf{x}, \mathbf{x}')$ and $K_{\mathbf{x}}(\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')$. Then

$$\begin{aligned} \|K_{\mathbf{x}} - K_{\mathbf{x}'}\|^2 &= K(\mathbf{x}, \mathbf{x}) + K(\mathbf{x}', \mathbf{x}') - 2K(\mathbf{x}, \mathbf{x}') \\ &= -H(\mathbf{x}, \mathbf{x}') + \frac{1}{2} H(\mathbf{x}, \mathbf{x}) + \frac{1}{2} H(\mathbf{x}', \mathbf{x}'). \end{aligned} \tag{D.33}$$

Thus the theorem holds. $\blacksquare$

References

1. S. Abe. *Neural Networks and Fuzzy Systems: Theory and Applications*. Kluwer Academic Publishers, Norwell, MA, 1997.
2. S. Abe. Dynamic cluster generation for a fuzzy classifier with ellipsoidal regions. *IEEE Transactions on Systems, Man, and Cybernetics—Part B*, 28(6):869–76, 1998.
3. S. Abe. *Pattern Classification: Neuro-Fuzzy Methods and Their Comparison*. Springer-Verlag, London, 2001.
4. S. Abe. Analysis of support vector machines. In H. Bourlard, T. Adali, S. Bengio, J. Larsen, and S. Douglas, editors, *Neural Networks for Signal Processing XII—Proceedings of the 2002 IEEE Signal Processing Society Workshop*, pages 89–98, 2002.
5. S. Abe. Analysis of multiclass support vector machines. In *Proceedings of International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA 2003)*, pages 385–96, Vienna, Austria, 2003.
6. S. Abe. On invariance of support vector machines. Presented at the Fourth International Symposium on Intelligent Data Engineering and Learning (IDEAL 2003) but not included in the proceedings (http://www2.kobe-u.ac.jp/˜abe/pdf/ideal2003.pdf), 2003.
7. S. Abe. Fuzzy LP-SVMs for multiclass problems. In *Proceedings of the Twelfth European Symposium on Artificial Neural Networks (ESANN 2004)*, pages 429–34, Bruges, Belgium, 2004.
8. S. Abe, Y. Hirokawa, and S. Ozawa. Steepest ascent training of support vector machines. In E. Damiani, L. C. Jain, R. J. Howlett, and N. Ichalkaranje, editors, *Knowledge-Based Intelligent Engineering Systems and Allied Technologies (KES 2002)*, volume 82 Frontiers in Artificial Intelligence and Applications, Part 2, pages 1301–5, IOS Press, Amsterdam, the Netherlands, 2002.
9. S. Abe and T. Inoue. Fast training of support vector machines by extracting boundary data. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Artificial Neural Networks (ICANN 2001)—Proceedings of International Conference, Vienna, Austria*, pages 308–13. Springer-Verlag, Berlin, Germany, 2001.
10. S. Abe and T. Inoue. Fuzzy support vector machines for multiclass problems. In *Proceedings of the Tenth European Symposium on Artificial Neural Networks (ESANN 2002)*, pages 113–8, Bruges, Belgium, 2002.

11. S. Abe and M.-S. Lan. A method for fuzzy rules extraction directly from numerical data and its application to pattern classification. *IEEE Transactions on Fuzzy Systems*, 3(1):18–28, 1995.
12. S. Abe and K. Sakaguchi. Generalization improvement of a fuzzy classifier with ellipsoidal regions. In *Proceedings of the Tenth IEEE International Conference on Fuzzy Systems*, volume 1, pages 207–10, Melbourne, Australia, 2001.
13. S. Abe and R. Thawonmas. A fuzzy classifier with ellipsoidal regions. *IEEE Transactions on Fuzzy Systems*, 5(3):358–68, 1997.
14. E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–41, 2000.
15. S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–9, 1999.
16. S. Amari and S. Wu. An information-geometrical method for improving the performance of support vector machine classifiers. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN '99)*, volume 1, pages 85–90, Edinburgh, UK, 1999.
17. D. Anguita, A. Boni, and S. Pace. Fast training of support vector machines for regression. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000)*, volume 5, pages 210–4, Como, Italy, 2000.
18. D. Anguita, A. Boni, and S. Ridella. Evaluating the generalization ability of support vector machines through the bootstrap. *Neural Processing Letters*, 11(1):51–8, 2000.
19. D. Anguita, S. Ridella, and D. Sterpi. A new method for multiclass support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2004)*, volume 1, pages 407–12, Budapest, Hungary, 2004.
20. C. Angulo, X. Parra, and A. Català. An [*sic*] unified framework for "all data at once" multi-class support vector machines. In *Proceedings of the Tenth European Symposium on Artificial Neural Networks (ESANN 2002)*, pages 161–6, Bruges, Belgium, 2002.
21. J. K. Anlauf and M. Biehl. The Adatron—An adaptive perceptron algorithm. *Europhysics Letters*, 10:687–92, 1989.
22. K. Baba, I. Enbutu, and M. Yoda. Explicit representation of knowledge acquired from plant historical data using neural network. In *Proceedings of 1990 IJCNN International Joint Conference on Neural Networks*, volume 3, pages 155–60, San Diego, 1990.
23. B. Baesens, S. Viaene, T. Van Gestel, J. A. K. Suykens, G. Dedene, B. De Moor, and J. Vanthienen. An empirical assessment of kernel type performance for least squares support vector machine classifiers. In *Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies (KES 2000)*, volume 1, pages 313–6, Brighton, UK, 2000.
24. T. Ban and S. Abe. Spatially chunking support vector clustering algorithm. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2004)*, volume 1, pages 413–8, Budapest, Hungary, 2004.
25. A. Barla, E. Franceschi, F. Odone, and A. Verri. Image kernels. In S.-W. Lee and A. Verri, editors, *Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002, Niagara Falls*, pages 83–96. Springer-Verlag, Berlin, Germany, 2002.

26. G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–404, 2000.
27. G. Baudat and F. Anouar. Kernel-based methods and function approximation. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 2, pages 1244–9, Washington, DC, 2001.
28. A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–37, 2001.
29. K. P. Bennett. Combining support vector and mathematical programming methods for classification. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 307–26. MIT Press, Cambridge, MA, 1999.
30. C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions.* Springer-Verlag, New York, 1984.
31. D. P. Bertsekas. *Nonlinear Programming, second edition.* Athena Scientific, Belmont, MA, 1999.
32. J. C. Bezdek, J. M. Keller, R. Krishnapuram, L. I. Kuncheva, and N. R. Pal. Will the real iris data please stand up? *IEEE Transactions on Fuzzy Systems*, 7(3):368–9, 1999.
33. J. C. Bezdek, J. Keller R. Krisnapuram, and N. R. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* Kluwer Academic Publishers, Norwell, MA, 1999.
34. J. Bi, K. P. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229–43, 2003.
35. C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, UK, 1995.
36. S. Borer and W. Gerstner. Support vector representation of multi-categorical data. In J. R. Dorronsoro, editor, *Artificial Neural Networks (ICANN 2002)—Proceedings of International Conference, Madrid, Spain*, pages 733–8. Springer-Verlag, Berlin, Germany, 2002.
37. P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, pages 82–90, Madison, 1998.
38. P. S. Bradley and O. L. Mangasarian. Massive data discrimination via linear support vector machines. *Optimization Methods and Software*, 13(1):1–10, 2000.
39. V. L. Brailovsky, O. Barzilay, and R. Shahave. On global, local, mixed and neighborhood kernels for support vector machines. *Pattern Recognition Letters*, 20(11–13):1183–90, 1999.
40. E. J. Bredensteiner and K. P. Bennett. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12(1–3):53–79, 1999.
41. M. Brown. Exploring the set of sparse, optimal classifiers. In *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR 2003)*, pages 178–84, Florence, Italy, 2003.
42. M. Brown, N. P. Costen, and S. Akamatsu. Efficient calculation of the complete optimal classification set. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, volume 2, pages 307–10, Cambridge, UK, 2004.

43. C. J. C. Burges. Simplified support vector decision rules. In L. Saitta, editor, *Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96)*, pages 71–7. Morgan Kaufmann, San Francisco, 1996.
44. C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–67, 1998.
45. C. J. C. Burges. Geometry and invariance in kernel based methods. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 89–116. MIT Press, Cambridge, MA, 1999.
46. C. J. C. Burges and D. J. Crisp. Uniqueness of the SVM solution. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 223–9. MIT Press, Cambridge, MA, 2000.
47. C. J. C. Burges and B. Schölkopf. Improving the accuracy and speed of support vector machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 375–81, 1997.
48. C. Campbell, T.-T. Frieß, and N. Cristianini. Maximal margin classification using the KA algorithm. In *Proceedings of the First International Symposium on Intelligent Data Engineering and Learning (IDEAL '98)*, pages 355–62, Hong Kong, China, 1998.
49. D. Caragea, D. Cook, and V. Honavar. Towards simple, easy-to-understand, yet accurate classifiers. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003)*, pages 497–500, Melbourne, FL, 2003.
50. G. L. Cash and M. Hatamian. Optical character recognition by the method of moments. *Computer Vision, Graphics, and Image Processing*, 39(3):291–310, 1987.
51. G. Cauwenberghs and T. Poggio. Incremental and decremental support vector machine learning. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 409–15. MIT Press, Cambridge, MA, 2000.
52. G. C. Cawley and N. L. C. Talbot. Manipulation of prior probabilities in support vector classification. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 4, pages 2433–8, Washington, DC, 2001.
53. G. C. Cawley and N. L. C. Talbot. Efficient formation of a basis in a kernel feature space. In *Proceedings of the Tenth European Symposium on Artificial Neural Networks (ESANN 2002)*, pages 1–6, Bruges, Belgium, 2002.
54. G. C. Cawley and N. L. C. Talbot. A greedy training algorithm for sparse least-squares support vector machines. In J. R. Dorronsoro, editor, *Artificial Neural Networks (ICANN 2002)—Proceedings of International Conference, Madrid, Spain*, pages 681–6. Springer-Verlag, Berlin, Germany, 2002.
55. G. C. Cawley and N. L. C. Talbot. Efficient model selection for kernel logistic regression. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, volume 2, pages 439–42, Cambridge, UK, 2004.
56. O. Chapelle and V. Vapnik. Model selection for support vector machines. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 230–6. MIT Press, Cambridge, MA, 2000.

57. J.-H. Chen. M-estimator based robust kernels for support vector machines. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, volume 1, pages 168–71, Cambridge, UK, 2004.
58. S. Chen, S. R. Gunn, and C. J. Harris. The relevance vector machine technique for channel equalization application. *IEEE Transactions on Neural Networks*, 12(6):1529–32, 2001.
59. S. Chen, S. R. Gunn, and C. J. Harris. Errata to "The relevance vector machine technique for channel equalization application." *IEEE Transactions on Neural Networks*, 13(4):1024, 2002.
60. V. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory, and Methods*. John Wiley & Sons, New York, 1998.
61. S. L. Chiu. Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy Systems*, 2:267–78, 1994.
62. C. S. Chu, I. W. Tsang, and J. T. Kwok. Scaling up support vector data description by using core-sets. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2004)*, volume 1, pages 425–31, Budapest, Hungary, 2004.
63. V. Chvátal. *Linear Programming*. W. H. Freeman and Company, New York, 1983.
64. C. Cortes, P. Haffner, and M. Mohri. Rational kernels. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 601–8. MIT Press, Cambridge, MA, 2003.
65. K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory (COLT 2000)*, pages 35–46, Palo Alto, CA, 2000.
66. K. Crammer and Y. Singer. Improved output coding for classification using continuous relaxation. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 437–43. MIT Press, Cambridge, MA, 2001.
67. K. Crammer and Y. Singer. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–91, 2003.
68. N. Cristianini and C. Campbell. Dynamically adapting kernels in support vector machines. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 204–10. MIT Press, Cambridge, MA, 1999.
69. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK, 2000.
70. R. S. Crowder. Predicting the Mackey-Glass time series with cascade-correlation learning. In *Proceedings of 1990 Connectionist Models Summer School*, pages 117–23, Carnegie Mellon University, 1990.
71. M. B. de Almeida, A. de Pádua Braga, and J. P. Braga. SVM-KM: Speeding SVMs learning with a priori cluster selection and k-means. In *Proceedings of the Sixth Brazilian Symposium on Neural Networks (SBRN 2000)*, volume 1, pages 162–7, Rio de Janeiro, Brazil, 2000.
72. N. de Freitas, M. Milo, P. Clarkson, M. Niranjan, and A. Gee. Sequential support vector machines. In *Neural Networks for Signal Processing IX—Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 31–40, 1999.

73. T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–86, 1995.
74. T. Downs, K. E. Gates, and A. Masters. Exact simplification of support vector solutions. *Journal of Machine Learning Research*, 2:293–7, 2001.
75. P. M. L. Drezet and R. F. Harrison. A new method for sparsity control in support vector classification and regression. *Pattern Recognition*, 34(1):111–25, 2001.
76. K. Duan, S. S. Keerthi, and A. N. Poo. An empirical evaluation of simple performance measures for tuning SVM hyperparameters. In *Proceedings of the Eighth International Conference on Neural Information Processing (ICONIP-2001)*, Paper ID# 159, Shanghai, China, 2001.
77. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.
78. B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Press, Boca Raton, FL, 1993.
79. T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio. Image representations for object detection using kernel classifiers. In *Proceedings of Asian Conference on Computer Vision (ACCV 2000)*, pages 687–92, Taipei, Taiwan, 2000.
80. T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.
81. J. Feng and P. Williams. The generalization error of the symmetric and scaled support vector machines. *IEEE Transactions on Neural Networks*, 12(5):1255–60, 2001.
82. R. Fernández. Behavior of the weights of a support vector machine as a function of the regularization parameter C. In *Proceedings of the Eighth International Conference on Artificial Neural Networks (ICANN '98)*, volume 2, pages 917–22, Skövde, Sweden, 1998.
83. S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–64, 2001.
84. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–88, 1936.
85. P. Frasconi, A. Passerini, and A. Vullo. A two-stage SVM architecture for predicting the disulfide bonding state of cysteines. In H. Bourlard, T. Adali, S. Bengio, J. Larsen, and S. Douglas, editors, *Neural Networks for Signal Processing XII—Proceedings of the 2002 IEEE Signal Processing Society Workshop*, pages 25–34, 2002.
86. Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–96, 1999.
87. F. Friedrichs and C. Igel. Evolutionary tuning of multiple SVM parameters. In *Proceedings of the Twelfth European Symposium on Artificial Neural Networks (ESANN 2004)*, pages 519–24, Bruges, Belgium, 2004.
88. T.-T. Frieß, N. Cristianini, and C. Campbell. The kernel-Adatron algorithm: A fast and simple learning procedure for support vector machines. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, pages 188–96, Madison, 1998.
89. X. Fu, C.-J. Ong, S. Keerthi, G. G. Hung, and L. Goh. Extracting the knowledge embedded in support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2004)*, volume 1, pages 291–6, Budapest, Hungary, 2004.

90. K. Fukunaga. *Introduction to Statistical Pattern Recognition, second edition.* Academic Press, San Diego, 1990.
91. G. Fumera and F. Roli. Support vector machines with embedded reject option. In S.-W. Lee and A. Verri, editors, *Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002, Niagara Falls*, pages 68–82. Springer-Verlag, Berlin, Germany, 2002.
92. A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)*, pages 148–55, Madison, 1998.
93. C. Gentile. A new approximate maximal margin classification algorithm. *Journal of Machine Learning Research*, 2:213–42, 2001.
94. T. Van Gestel, J. A. K. Suykens, J. De Brabanter, B. De Moor, and J. Vandewalle. Least squares support vector machine regression for discriminant analysis. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 4, pages 2445–50, Washington, DC, 2001.
95. M. Girolami. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, 13(3):780–4, 2002.
96. G. H. Golub and C. F. Van Loan. *Matrix Computations, third edition.* The Johns Hopkins University Press, Baltimore, 1996.
97. E. Gose, R. Johnsonbaugh, and S. Jost. *Pattern Recognition and Image Analysis.* Prentice Hall, Upper Saddle River, NJ, 1996.
98. T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K.-R. Müller, K. Obermayer, and R. Williamson. Classification on proximity data with LP-machines. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN '99)*, volume 1, pages 304–9, Edinburgh, UK, 1999.
99. Y. Grandvalet and S. Canu. Adaptive scaling for feature selection in SVMs. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 569–76. MIT Press, Cambridge, MA, 2003.
100. Y. Guermeur, A. Elisseeff, and H. Paugam-Moisy. A new multi-class SVM based on a uniform convergence result. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000)*, volume 4, pages 183–8, Como, Italy, 2000.
101. S. R. Gunn. Support vector machines for classification and regression. Technical Report ISIS-1-98, Department of Electronics and Computer Science, University of Southampton, 1998.
102. S. R. Gunn and M. Brown. SUPANOVA: A sparse, transparent modelling approach. In *Neural Networks for Signal Processing IX—Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 21–30, 1999.
103. G. Guo, S. Z. Li, and K. L. Chan. Support vector machines for face recognition. *Image and Vision Computing*, 19(9–10):631–8, 2001.
104. I. Guyon and D. G. Stork. Linear discriminant and support vector classifiers. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 147–69. MIT Press, Cambridge, MA, 2000.
105. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002.

106. A. Hashizume, J. Motoike, and R. Yabe. Fully automated blood cell differential system and its application. In *Proceedings of the IUPAC Third International Congress on Automation and New Technology in the Clinical Laboratory*, pages 297–302, Kobe, Japan, 1988.
107. T. Hastie and R. Tibshirani. Classification by pairwise coupling. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 507–13. MIT Press, Cambridge, MA, 1998.
108. S. Haykin. *Neural Networks: A Comprehensive Foundation, second edition.* Prentice Hall, Upper Saddle River, NJ, 1999.
109. R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms.* MIT Press, Cambridge, MA, 2002.
110. R. Herbrich and J. Weston. Adaptive margin support vector machines for classification. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN '99)*, volume 2, pages 880–5, Edinburgh, UK, 1999.
111. Y. Hirokawa and S. Abe. Training of support vector regressors based on the steepest ascent method. In *Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP '02)*, volume 2, pages 552–5, Singapore, 2002.
112. Y. Hirokawa and S. Abe. Steepest ascent training of support vector regressors. *IEEJ Transactions of Electronics, Information and Systems*, 124(10):2066–73, 2004 (in Japanese).
113. K. Hotta. Support vector machine with local summation kernel for robust face recognition. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, volume 3, pages 482–5, Cambridge, UK, 2004.
114. C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–25, 2002.
115. T. M. Huang and V. Kecman. Bias term b in SVMs again. In *Proceedings of the Twelfth European Symposium on Artificial Neural Networks (ESANN 2004)*, pages 441–8, Bruges, Belgium, 2004.
116. T. Inoue and S. Abe. Fuzzy support vector machines for pattern classification. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 2, pages 1449–54, Washington, DC, 2001.
117. T. Inoue and S. Abe. Improvement of generalization ability of multiclass support vector machines by introducing fuzzy logic and Bayes theory. *Transactions of the Institute of Systems, Control and Information Engineers*, 15(12):643–51, 2002 (in Japanese).
118. K. Ito and R. Nakano. Optimizing support vector regression hyperparameters based on cross-validation. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2003)*, volume 3, pages 2077–82, Portland, OR, 2003.
119. J.-S. R. Jang. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3):665–85, 1993.
120. Jayadeva, A. K. Deb, and S. Chandra. Binary classification by SVM based tree type neural networks. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02)*, volume 3, pages 2773–8, Honolulu, 2002.
121. J.-T. Jeng and C.-C. Chuang. A novel approach for the hyperparameters of support vector regression. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02)*, volume 1, pages 642–7, Honolulu, 2002.

122. T. Joachims. Estimating the generalization performance of an SVM efficiently. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)*, pages 431–8, Stanford, CA, 2000.
123. T. Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell, MA, 2002.
124. E. M. Jordaan and G. F. Smits. Robust outlier detection using SVM for regression. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2004)*, volume 3, pages 2017–22, Budapest, Hungary, 2004.
125. A. Juneja and C. Espy-Wilson. Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2003)*, volume 1, pages 675–9, Portland, OR, 2003.
126. K. Kaieda and S. Abe. A kernel fuzzy classifier with ellipsoidal regions. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2003)*, volume 3, pages 2043–8, Portland, OR, 2003.
127. K. Kaieda and S. Abe. KPCA-based training of a kernel fuzzy classifier with ellipsoidal regions. *International Journal of Approximate Reasoning*, 37(3):145–253, 2004.
128. V. Kecman, T. Arthanari, and I. Hadzic. LP and QP based learning from empirical data. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 4, pages 2451–5, Washington, DC, 2001.
129. V. Kecman and I. Hadzic. Support vectors selection by linear programming. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000)*, volume 5, pages 193–8, Como, Italy, 2000.
130. V. Kecman, M. Vogt, and T. M. Huang. On the equality of kernel AdaTron and sequential minimal optimization in classification and regression tasks and alike algorithms for kernel machines. In *Proceedings of the Eleventh European Symposium on Artificial Neural Networks (ESANN 2003)*, pages 215–22, Bruges, Belgium, 2003.
131. S. S. Keerthi and E. G. Gilbert. Convergence of a generalized SMO algorithm for SVM classifier design. *Machine Learning*, 46:351–60, 2002.
132. S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. A fast iterative nearest point algorithm for support vector machine classifier design. *IEEE Transactions on Neural Networks*, 11(1):124–36, 2000.
133. S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13:637–49, 2001.
134. B. Kijsirikul and N. Ussivakul. Multiclass support vector machines using adaptive directed acyclic graph. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02)*, volume 1, pages 980–5, Honolulu, 2002.
135. T. Kikuchi and S. Abe. Error correcting output codes vs. fuzzy support vector machines. In *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR 2003)*, pages 192–6, Florence, Italy, 2003.
136. T. Kikuchi and S. Abe. Error correcting output codes vs. fuzzy support vector machines. *Pattern Recognition Letters* (to appear).
137. Y. Koshiba. Acceleration of training of support vector machines. Master's thesis, Graduate School of Science and Technology, Kobe University, Japan, 2004 (in Japanese).

138. Y. Koshiba and S. Abe. Comparison of L1 and L2 support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2003)*, volume 3, pages 2054–9, Portland, OR, 2003.
139. Z. Kou, J. Xu, X. Zhang, and L. Ji. An improved support vector machine using class-median vectors. In *Proceedings of the Eighth International Conference on Neural Information Processing (ICONIP-2001)*, Paper ID# 60, Shanghai, China, 2001.
140. U. H.-G. Kreßel. Pairwise classification and support vector machines. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 255–68. MIT Press, Cambridge, MA, 1999.
141. M.-S. Lan, H. Takenaga, and S. Abe. Character recognition using fuzzy rules extracted from data. In *Proceedings of the Third IEEE International Conference on Fuzzy Systems*, volume 1, pages 415–20, Orlando, 1994.
142. G. Lebrun, C. Charrier, and H. Cardot. SVM training time reduction using vector quantization. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, volume 1, pages 160–3, Cambridge, UK, 2004.
143. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–324, 1998.
144. K. K. Lee, S. R. Gunn, C. J. Harris, and P. A. S. Reed. Classification of imbalanced data with transparent kernels. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 4, pages 2410–5, Washington, DC, 2001.
145. C. Leslie, E. Eskin, J. Weston, and W. S. Noble. Mismatch string kernels for SVM protein classification. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1441–8. MIT Press, Cambridge, MA, 2003.
146. H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 97–104. MIT Press, Cambridge, MA, 2004.
147. Y. Li and P. M. Long. The relaxed online maximum margin algorithm. *Machine Learning*, 46(1–3):361–87, 2002.
148. Z. Li and S. Tang. Face recognition using improved pairwise coupling support vector machines. In *Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP '02)*, #1288, Singapore, 2002.
149. S.-P. Liao, H.-T. Lin, and C.-J. Lin. A note on the decomposition methods for support vector regression. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 2, pages 1474–9, Washington, DC, 2001.
150. C. Ap. M. Lima, A. L. V. Coelho, and F. J. Von Zuben. Ensembles of support vector machines for regression problems. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02)*, volume 3, pages 2381–6, Honolulu, 2002.
151. C.-F. Lin and S.-D. Wang. Fuzzy support vector machines. *IEEE Transactions on Neural Networks*, 13(2):464–71, 2002.
152. C.-J. Lin. On the convergence of the decomposition method for support vector machines. *IEEE Transactions on Neural Networks*, 12(6):1288–98, 2001.

153. C.-J. Lin. Asymptotic convergence of an SMO algorithm without any assumptions. *IEEE Transactions on Neural Networks*, 13(1):248–50, 2002.
154. C.-J. Lin. Errata to"A Comparison of Methods for Multiclass Support Vector Machines." *IEEE Transactions on Neural Networks*, 13(4):1026–7, 2002.
155. H. Lodhi, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 563–9, 2001.
156. B.-L. Lu, K.-A. Wang, M. Utiyama, and H. Isahara. A part-versus-part method for massively parallel training of support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2004)*, volume 1, pages 735–40, Budapest, Hungary, 2004.
157. L. Lukas, A. Devos, J. A. K. Suykens, L. Vanhamme, S. Van Huffel, A. R. Tate, C. Majós, and C. Arús. The use of LS-SVM in the classification of brain tumors based on magnetic resonance spectroscopy signals. In *Proceedings of the Tenth European Symposium on Artificial Neural Networks (ESANN 2002)*, pages 131–6, Bruges, Belgium, 2002.
158. J. Ma and S. Perkins. Time-series novelty detection using one-class support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2003)*, volume 3, pages 1741–5, Portland, OR, 2003.
159. E. Maeda and H. Murase. Kernel based nonlinear subspace method for pattern recognition. *Transactions of the Institute of Electronics, Information and Communication Engineers D-II*, J82D-II(4):600–12, 1999 (in Japanese).
160. O. L. Mangasarian and D. R. Musicant. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10(5):1032–7, 1999.
161. D. Martinez and G. Millerioux. Support vector committee machines. In *Proceedings of the Eleventh European Symposium on Artificial Neural Networks (ESANN 2000)*, pages 43–8, Bruges, Belgium, 2000.
162. F. Masulli and G. Valentini. Comparing decomposition methods for classification. In *Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies (KES 2000)*, volume 2, pages 788–91, Brighton, UK, 2000.
163. D. Mattera, F. Palmieri, and S. Haykin. An explicit algorithm for training support vector machines. *IEEE Signal Processing Letters*, 6(9):243–5, 1999.
164. D. Mattera, F. Palmieri, and S. Haykin. Simple and robust methods for support vector expansions. *IEEE Transactions on Neural Networks*, 10(5):1038–47, 1999.
165. E. Mayoraz and E. Alpaydin. Support vector machines for multi-class classification. In J. Mira and J. V. Sanchez-Andres, editors, *Engineering Applications of Bio-Inspired Artificial Neural Networks (IWANN'99)—Proceedings of International Work—Conference on Artificial and Natural Neural Networks, Alicante, Spain*, volume 2, pages 833–42, 1999.
166. S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX—Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–8, 1999.
167. S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. Kernel PCA and de-noising in feature spaces. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 536–42. MIT Press, Cambridge, MA, 1999.

168. S. Miyamoto and D. Suizu. Fuzzy c-means clustering using transformations into high dimensional spaces. In *Proceedings of the First International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '02)*, volume 2, pages 656–60, Singapore, 2002.
169. M. Moreira and E. Mayoraz. Improved pairwise coupling classification with correcting classifiers. In *Proceedings of the Tenth European Conference on Machine Learning (ECML-98)*, pages 160–71, Chemnitz, Germany, 1998.
170. K. Morikawa. Pattern classification and function approximation by kernel least squares. Bachelor's thesis, Electrical and Electronics Engineering, Kobe University, Japan, 2004 (in Japanese).
171. S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *Neural Networks for Signal Processing VII—Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 511–20, 1997.
172. S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. P. Mesirov, and T. Poggio. Support vector machine classification of microarray data. Technical Report AI Memo 1677, Massachusetts Institute of Technology, 1999.
173. K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.
174. K.-R. Müller, A. J. 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)—Proceedings of the Seventh International Conference, Lausanne, Switzerland*, pages 999–1004. Springer-Verlag, Berlin, Germany, 1997.
175. C. Nakajima, M. Pontil, and T. Poggio. People recognition and pose estimation in image sequences. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2000)*, volume IV, pages 189–94, Como, Italy, 2000.
176. H. Nakayama and T. Asada. Support vector machines using multi objective programming and goal programming. In *Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP '02)*, volume 2, pages 1053–7, Singapore, 2002.
177. A. Navia-Vázquez, F. Pérez-Cruz, A. Artés-Rodríguez, and A. R. Figueiras-Vidal. Weighted least squares training of support vector classifiers leading to compact and adaptive schemes. *IEEE Transactions on Neural Networks*, 12(5):1047–59, 2001.
178. T. Nishikawa and S. Abe. Maximizing margins of multilayer neural networks. In *Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP '02)*, volume 1, pages 322–6, Singapore, 2002.
179. H. Núñez, C. Angulo, and Català. Rule extraction from support vector machines. In *Proceedings of the Tenth European Symposium on Artificial Neural Networks (ESANN 2002)*, pages 107–12, Bruges, Belgium, 2002.
180. C. S. Ong and A. J. Smola. Machine learning using hyperkernels. In T. Fawcett and N. Mishra, editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), Washington, DC*, pages 568–75. AAAI Press, Menlo Park, CA, 2003.
181. C. S. Ong, A. J. Smola, and R. C. Williamson. Hyperkernels. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 495–502. MIT Press, Cambridge, MA, 2003.

182. E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Neural Networks for Signal Processing VII—Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 276–85, 1997.
183. S. K. Pal and S. Mitra. *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*. John Wiley & Sons, New York, 1999.
184. C. H. Park and H. Park. Efficient nonlinear dimension reduction for clustered data using kernel functions. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003)*, pages 243–50, Melbourne, FL, 2003.
185. J. P. Pedroso and N. Murata. Optimisation on support vector machines. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000)*, volume 6, pages 399–404, Como, Italy, 2000.
186. F. Pérez-Cruz and A. Artés-Rodríguez. Puncturing multi-class support vector machines. In J. R. Dorronsoro, editor, *Artificial Neural Networks (ICANN 2002)—Proceedings of International Conference, Madrid, Spain*, pages 751–6. Springer-Verlag, Berlin, Germany, 2002.
187. F. Pérez-Cruz, G. Camps-Valls, E. Soria-Olivas, J. J. Pérez-Ruixo, A. R. Figueiras-Vidal, and A. Artés-Rodríguez. Multi-dimensional function approximation and regression estimation. In J. R. Dorronsoro, editor, *Artificial Neural Networks (ICANN 2002)—Proceedings of International Conference, Madrid, Spain*, pages 757–62. Springer-Verlag, Berlin, Germany, 2002.
188. S. Perkins, K. Lacker, and J. Theiler. Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research*, 3:1333–56, 2003.
189. T. Phetkaew, B. Kijsirikul, and W. Rivepiboon. Reordering adaptive directed acyclic graphs: An improved algorithm for multiclass support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2003)*, volume 2, pages 1605–10, Portland, OR, 2003.
190. J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 185–208. MIT Press, Cambridge, MA, 1999.
191. J. C. Platt. Probabilities for SV machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–73. MIT Press, Cambridge, MA, 2000.
192. J. C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 547–53. MIT Press, Cambridge, MA, 2000.
193. M. Pontil and A. Verri. Properties of support vector machines. *Neural Computation*, 10(4):955–74, 1998.
194. M. Pontil and A. Verri. Support vector machines for 3-D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–46, 1998.
195. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing, second edition*. Cambridge University Press, Cambridge, UK, 1992.
196. T. Raicharoen and C. Lursinsap. Critical support vector machine without kernel function. In *Proceedings of the Ninth International Conference on Neural*

Information Processing (ICONIP '02), volume 5, pages 2532–6, Singapore, 2002.

197. A. Rakotomamonjy. Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3:1357–70, 2003.
198. L. Ralaivola and F. d'Alché-Buc. Incremental support vector machine learning: A local approach. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Artificial Neural Networks (ICANN 2001)—Proceedings of International Conference, Vienna, Austria*, pages 322–30. Springer-Verlag, Berlin, Germany, 2001.
199. G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for AdaBoost. *Machine Learning*, 42(3):287–320, 2001.
200. G. Rätsch, A. J. Smola, and S. Mika. Adapting codes and embeddings for polychotomies. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 529–36. MIT Press, Cambridge, MA, 2003.
201. R. M. Rifkin, M. Pontil, and A. Verri. A note on support vector machine degeneracy. In O. Watanabe and T. Yokomori, editors, *Proceedings of the Tenth International Conference on Algorithmic Learning Theory (ALT '99), Tokyo, Japan*, pages 252–63. Springer-Verlag, Berlin, Germany, 1999.
202. B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK, 1996.
203. D. Roobaert. DirectSVM: A fast and simple support vector machine perceptron. In *Neural Networks for Signal Processing X—Proceedings of the 2000 IEEE Signal Processing Society Workshop*, volume 1, pages 356–65, 2000.
204. R. Rosipal, M. Girolami, and L. J. Trejo. Kernel PCA feature extraction of event-related potentials for human signal detection performance. In H. Malmgren, M. Borga, and L. Niklasson, editors, *Artificial Neural Networks in Medicine and Biology—Proceedings of the ANNIMAB-1 Conference, Göteborg, Sweden*, pages 321–6. Springer-Verlag, Berlin, Germany, 2000.
205. R. Rosipal, M. Girolami, L. J. Trejo, and A. Cichocki. Kernel PCA for feature extraction and de-noising in nonlinear regression. *Neural Computing & Applications*, 10(3):231–43, 2001.
206. A. Ruiz and P. E. López de Teruel. Nonlinear kernel-based statistical pattern analysis. *IEEE Transactions on Neural Networks*, 12(1):16–32, 2001.
207. M. Rychetsky, S. Ortmann, M. Ullmann, and M. Glesner. Accelerated training of support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '99)*, volume 2, pages 998–1003, Washington, DC, 1999.
208. K. Saadi, G. C. Cawley, and L. C. Talbot. Fast exact leave-one-out cross-validation of least-squares support vector machines. In *Proceedings of the Tenth European Symposium on Artificial Neural Networks (ESANN 2002)*, pages 149–54, Bruges, Belgium, 2002.
209. K. Saadi, N. L. C. Talbot, and G. C. Cawley. Optimally regularised kernel fisher discriminant analysis. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, volume 2, pages 427–30, Cambridge, UK, 2004.
210. 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, Royal Holloway, University of London, London, 1998.

211. B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson. Support vector regression with automatic accuracy control. In *Proceedings of the Eighth International Conference on Artificial Neural Networks (ICANN '98)*, volume 2, pages 111–6, Skövde, Sweden, 1998.
212. B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, MA, 1999.
213. B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Kernel-dependent support vector error bounds. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN '99)*, volume 1, pages 103–8, Edinburgh, UK, 1999.
214. B. Schölkopf, P. Simard, A. Smola, and V. Vapnik. Prior knowledge in support vector kernels. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 640–6. MIT Press, Cambridge, MA, 1998.
215. B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.
216. B. Schölkopf, A. J. Smola, and K.-R. Müller. Kernel principal component analysis. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 327–52. MIT Press, Cambridge, MA, 1999.
217. A. Schwaighofer and V. Tresp. The Bayesian committee support vector machine. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Artificial Neural Networks (ICANN 2001)—Proceedings of International Conference, Vienna, Austria*, pages 411–20. Springer-Verlag, Berlin, Germany, 2001.
218. F. Schwenker. Hierarchical support vector machines for multi-class pattern recognition. In *Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies (KES 2000)*, volume 2, pages 561–5, Brighton, UK 2000.
219. F. Schwenker. Solving multi-class pattern recognition problems with tree structured support vector machines. In B. Radig and S. Florczyk, editors, *Pattern Recognition 2001*, pages 283–90. Springer-Verlag, Berlin, Germany, 2001.
220. M. Seeger. Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 603–9. MIT Press, Cambridge, MA, 2000.
221. X. Shao and V. Cherkassky. Multi-resolution support vector machine. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '99)*, volume 2, pages 1065–70, Washington, DC, 1999.
222. A. Shilton, M. Palaniswami, D. Ralph, and A. C. Tsoi. Incremental training of support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, Washington, DC, 2001.
223. H. Shimodaira, K. Noma, M. Nakai, and S. Sagayama. Dynamic time-alignment kernel in support vector machine. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, volume 2, pages 921–8, MIT Press, Cambridge, MA, 2002.
224. T. Shimozaki, T. Takigawa, and S. Abe. A fuzzy classifier with polyhedral regions. *Transactions of the Institute of Systems, Control and Information Engineers*, 14(7):365–72, 2001 (in Japanese).
225. H. Shin and S. Cho. How many neighbors to consider in pattern pre-selection for support vector classifiers? In *Proceedings of International Joint Conference*

on Neural Networks (IJCNN 2003), volume 1, pages 565–70, Portland, OR, 2003.
226. P. K. Simpson. Fuzzy min-max neural networks—Part 1: Classification. *IEEE Transactions on Neural Networks*, 3(5):776–86, 1992.
227. N. Smith and M. Gales. Speech recognition using SVMs. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, volume 2, pages 1197–204, MIT Press, Cambridge, MA, 2002.
228. G. F. Smits and E. M. Jordaan. Improved SVM regression using mixtures of kernels. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02)*, volume 3, pages 2785–90, Honolulu, 2002.
229. A. Smola, B. Schölkopf, and G. Rätsch. Linear programs for automatic accuracy control in regression. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN '99)*, volume 2, pages 575–80, Edinburgh, UK, 1999.
230. A. J. Smola, O. L. Mangasarian, and B. Schölkopf. Sparse kernel feature analysis. Technical Report 99-04, University of Wisconsin, Data Mining Institute, Madison, 1999.
231. A. J. Smola, N. Murata, B. Schölkopf, and K.-R. Müller. Asymptotically optimal choice of ε-loss for support vector machines. In *Proceedings of the Eighth International Conference on Artificial Neural Networks (ICANN '98)*, volume 1, pages 105–10, Skövde, Sweden, 1998.
232. S. Sohn and C. H. Dagli. Advantages of using fuzzy class memberships in self-organizing map and support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 3, pages 1886–90, Washington, DC, 2001.
233. S.-Y. Sun, C. L. Tseng, Y. H. Chen, S. C. Chuang, and H. C. Fu. Cluster-based support vector machines in text-independent speaker identification. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2004)*, volume 1, pages 729–34, Budapest, Hungary, 2004.
234. J. A. K. Suykens. Least squares support vector machines for classification and nonlinear modelling. *Neural Network World*, 10(1–2):29–47, 2000.
235. J. A. K. Suykens, L. Lukas, and J. Vandewalle. Sparse least squares support vector machine classifiers. In *Proceedings of the Eighth European Symposium on Artificial Neural Networks (ESANN 2000)*, pages 37–42, Bruges, Belgium, 2000.
236. J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Publishing, Singapore, 2002.
237. J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.
238. J. A. K. Suykens and J. Vandewalle. Multiclass least squares support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '99)*, volume 2, pages 900–3, Washington, DC, 1999.
239. J. A. K. Suykens and J. Vandewalle. Training multilayer perceptron classifiers based on a modified support vector method. *IEEE Transactions on Neural Networks*, 10(4):907–11, 1999.
240. F. Takahashi and S. Abe. Decision-tree-based multiclass support vector machines. In *Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP '02)*, volume 3, pages 1418–22, Singapore, 2002.

241. F. Takahashi and S. Abe. Optimizing directed acyclic graph support vector machines. In *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR 2003)*, pages 166–70, Florence, Italy, 2003.
242. F. Takahashi and S. Abe. Optimal structure of decision-tree-based pairwise support vector machines. *Transactions of the Institute of Systems, Control and Information Engineers*, 17(3):122–30, 2004 (in Japanese).
243. T. Takahashi and T. Kurita. Robust de-noising by kernel PCA. In J. R. Dorronsoro, editor, *Artificial Neural Networks (ICANN 2002)—Proceedings of International Conference, Madrid, Spain*, pages 739–44. Springer-Verlag, Berlin, Germany, 2002.
244. H. Takenaga, S. Abe, M. Takatoo, M. Kayama, T. Kitamura, and Y. Okuyama. Input layer optimization of neural networks by sensitivity analysis and its application to recognition of numerals. *Electrical Engineering in Japan*, 111(4):130–8, 1991.
245. T. Takigawa and S. Abe. High speed training of a fuzzy classifier with polyhedral regions. *Transactions of the Institute of Systems, Control and Information Engineers*, 15(12):673–80, 2002 (in Japanese).
246. D. M. J. Tax and R. P. W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20(11–13):1191–9, 1999.
247. D. M. J. Tax and R. P. W. Duin. Outliers and data descriptions. In *Proceedings of the Seventh Annual Conference of the Advanced School for Computing and Imaging*, pages 234–41, Heijen, the Netherlands, 2001.
248. D. M. J. Tax and P. Juszczak. Kernel whitening for one-class classification. In S.-W. Lee and A. Verri, editors, *Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002, Niagara Falls*, pages 40–52. Springer-Verlag, Berlin, Germany, 2002.
249. T. B. Trafalis and H. Ince. Benders decomposition technique for support vector regression. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02)*, volume 3, pages 2767–72, Honolulu, 2002.
250. D. Tsujinishi and S. Abe. Fuzzy least squares support vector machines. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2003)*, volume 2, pages 1599–604, Portland, OR, 2003.
251. D. Tsujinishi and S. Abe. Fuzzy least squares support vector machines for multiclass problems. *Neural Networks*, 16(5–6):785–92, 2003.
252. D. Tsujinishi, Y. Koshiba, and S. Abe. Why pairwise is better than one-against-all or all-at-once. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2004)*, volume 1, pages 693–8, Budapest, Hungary, 2004.
253. V. Uebele, S. Abe, and M.-S. Lan. A neural-network-based fuzzy classifier. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(2):353–61, 1995.
254. R. J. Vanderbei. LOQO: An interior point code for quadratic programming. Technical Report SOR-94-15, Princeton University, 1998.
255. R. J. Vanderbei. *Linear Programming: Foundations and Extensions, second edition.* Kluwer Academic Publishers, Norwell, MA, 2001.
256. V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.
257. V. N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, New York, 1998.

258. V. Vapnik and O. Chapelle. Bounds on error expectation for SVM. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 261–80. MIT Press, Cambridge, MA, 2000.
259. K. Veropoulos. Machine learning approaches to medical decision making. Ph.D thesis, Department of Computer Science, University of Bristol, UK, 2001.
260. K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), Workshop ML3*, pages 55–60, 1999.
261. S. V. N. Vishwanathan and M. N. Murty. SSVM: A simple SVM algorithm. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02)*, volume 2, pages 2393–8, Honolulu, 2002.
262. M. Vogt. SMO algorithms for support vector machines without bias term. Technical report, Institute of Automatic Control, TU Darmstadt, Germany, 2002.
263. V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. In I. Bratko and S. Dzeroski, editors, *Machine Learning, Proceedings of the Sixteenth International Conference (ICML '99)*, pages 444–53. Morgan Kaufmann, San Francisco, 1999.
264. S. M. Weiss and I. Kapouleas. An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 781–7, Detroit, 1989.
265. J. Weston. Leave-one-out support vector machines. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*, volume 2, pages 727–33, Stockholm, Sweden, 1999.
266. J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–61, 2003.
267. J. Weston and R. Herbrich. Adaptive margin support vector machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 281–95. MIT Press, Cambridge, MA, 2000.
268. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 668–74. MIT Press, Cambridge, MA, 2001.
269. J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, London, 1998.
270. J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the Seventh European Symposium on Artificial Neural Networks (ESANN 1999)*, pages 219–24, Bruges, Belgium, 1999.
271. T. Windeatt and F. Roli, editors. *Multiple Classifier Systems—Proceedings of the fourth International Workshop, MCS 2003, Guildford, UK.* Springer-Verlag, Berlin, Germany, 2003.
272. S. J. Wright. *Primal-Dual Interior-Point Methods.* Society for Industrial and Applied Mathematics, Philadelphia, 1997.
273. R. Xiao, J. Wang, and F. Zhang. An approach to incremental SVM learning algorithm. In *Proceedings of the Twelfth IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2000)*, pages 268–73, Vancouver, BC, Canada, 2000.

274. J. Xu, X. Zhang, and Y. Li. Large margin kernel pocket algorithm. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 2, pages 1480–5, Washington, DC, 2001.
275. P. Xu and A. K. Chan. Support vector machines for multi-class signal classification with unbalanced samples. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2003)*, volume 2, pages 1116–9, Portland, OR, 2003.
276. J. Yang, V. Estivill-Castro, and S. K. Chalup. Support vector clustering through proximity graph modelling. In *Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP '02)*, volume 2, pages 898–903, Singapore, 2002.
277. M.-H. Yang and N. Ahuja. A geometric approach to train support vector machines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 430–7, 2000.
278. Z. Ying and K. C. Keong. Fast leave-one-out evaluation and improvement on inference for LS-SVMs. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, volume 3, pages 494–7, Cambridge, UK, 2004.
279. S. Young and T. Downs. CARVE—A constructive algorithm for real-valued examples. *IEEE Transactions on Neural Networks*, 9(6):1180–90, 1998.
280. C. Yuan and D. Casasent. Support vector machines for class representation and discrimination. In *Proceedings of International Joint Conference on Neural Networks (IJCNN 2003)*, volume 2, pages 1611–6, Portland, OR, 2003.
281. P. Zhang and J. Peng. SVM vs regularized least squares classification. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, volume 1, pages 176–9, Cambridge, UK, 2004.
282. P. Zhang, J. Peng, and C. Domeniconi. Dimensionality reduction using kernel pooled local discriminant information. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003)*, pages 701–4, Melbourne, FL, 2003.
283. W. Zhang and I. King. Locating support vectors via β-skeleton technique. In *Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP '02)*, volume 3, pages 1423–7, Singapore, 2002.
284. X. Zhang. Using class-center vectors to build support vector machines. In *Neural Networks for Signal Processing IX—Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 3–11, 1999.
285. W. Zhou, L. Zhang, and L. Jiao. Linear programming support vector machines. *Pattern Recognition*, 35(12):2927–36, 2002.

Index