Journal of VLSI Signal Processing 45, 49–65, 2006

DOI: 10.1007/s11265-006-9771-8

Some Equivalences between Kernel Methods and Information Theoretic Methods

ROBERT JENSSEN AND TORBJØRN ELTOFT
Department of Physics and Technology, University of Tromsø, N–9037 Tromsø, Norway

DENIZ ERDOGMUS
Computer Science and Engineering Department, Oregon Graduate Institute, OHSU, Portland, OR 97006, USA

JOSE C. PRINCIPE
Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA

Abstract. In this paper, we discuss some equivalences between two recently introduced statistical learning schemes, namely Mercer kernel methods and information theoretic methods. We show that Parzen window-based estimators for some information theoretic cost functions are also cost functions in a corresponding Mercer kernel space. The Mercer kernel is directly related to the Parzen window. Furthermore, we analyze a classification rule based on an information theoretic criterion, and show that this corresponds to a linear classifier in the kernel space. By introducing a weighted Parzen window density estimator, we also formulate the support vector machine in this information theoretic perspective.

Keywords: Mercel kernel methods, information theoretic methods, Parzen window

1. Introduction

During the last decade, research on Mercer kernel-based learning algorithms has flourished [1–4]. These algorithms include for example the support vector machine (SVM) [5–9], kernel principal component analysis (KPCA) [10] and kernel Fisher discriminant analysis (KFDA) [11, 12]. The common property of these methods is that they are linear in nature, as they are being explicitly expressed in terms of inner-products. However, they may be applied to *non-linear* problems using the so-called "kernel-trick." The kernel trick refers to the technique of computing inner-products in a potentially infinite-dimensional *kernel feature space*, using so-called Mercer kernels. Mercer kernel-based methods have been applied successfully in several applications, e.g., pattern and object recognition [13], time series prediction [14] and DNA and protein analysis [15], to name a few.

The Mercer kernel-based methods rely on the assumption that the data becomes easier to handle after the transformation to the Mercer kernel feature space. In the case of the SVM, the assumption is that the data classes become linearly separable, and therefore a separating hyperplane can be created. In practice, one cannot know for certain that this assumption holds. In fact, one has to hope that the user chooses a kernel which turns out to properly separate the data.

Independent of the research on Mercer kernel-based learning algorithms another very powerful machine learning scheme has emerged. This is coined *information theoretic learning* [16, 17]. In information theoretic learning, the starting point is a data set that globally conveys information about a real-world event. The goal is to capture the information in the parameters of a learning machine, using some *information theoretic performance criterion*. A typical setup for information theoretic learning is shown in Fig. 1. The system output is given by $\mathbf{y}_i = \mathbf{g}(\mathbf{W})\mathbf{x_i}$, where $\mathbf{x}_i$ is the data pattern presented to the system at iteration i. The function $\mathbf{g}(\mathbf{W})$ represents a possibly non-linear data transformation, which depends on a parameter matrix $\mathbf{W}$. At each iteration, the criterion is evaluated and a correction term e_i generated, which is fed back to the system to guide the adjustment of the system parameters. The system may receive external input in the form of a desired response, in which case the system operates in a supervised learning mode.

The mean squared error criterion (MSE) has traditionally been the workhorse of adaptive systems training [18]. However, the great advantage of information theoretic criteria is that they are able to capture higher order statistical information in the data, as opposed to the MSE, which is a second order statistical criterion. This property is important, since recently many machine learning problems have been encountered where the MSE criterion is insufficient. Such problems include blind source separation and independent component analysis, blind equalization and deconvolution, subspace projections, dimensionality reduction, feature extraction, classification and clustering.

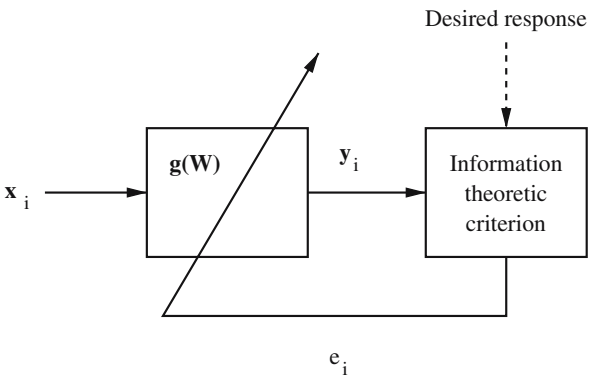


Figure 1. Information theoretic learning setup.

Information theoretic criteria are expressed as integrals over functions of probability densities. One possible approach to evaluate such criteria for an observed data set is to replace the densities by density estimators. Using parametric density estimators may be problematic, because they often require numerical integration procedures to be developed. Parzen windowing [19–23] has been proposed as an appropriate density estimation technique, since this method makes no parametric assumptions. Viola et al. [24] proposed to approximate Shannon-based measures using sample means, integrated with Parzen windowing [25]. Principe et al. [16] went a step further, by introducing a series of information theoretic quantities which can be estimated without the sample mean approximation [17, 26]. This is important, since the sample mean approximation may not hold very well for small sample sizes. The proposed measures were all based on the generalizations of the Shannon entropy derived by Renyi [27, 28], and include Renyi's quadratic entropy, the Cauchy–Schwarz (CS) pdf divergence measure, and the integrated squared error divergence measure. These will be discussed in more detail in Section 4. Since these measures all include quantities which are expressed as integrals over products and squares of densities, we will refer to them as quadratic information measures. Information theoretic learning based on the quadratic information measures, combined with Parzen windowing, has been applied with great success on several supervised and unsupervised learning problems [17, 29–36].

Information theoretic methods have the advantage over Mercer kernel-based methods that they are easier to interpret. Also, the information theoretic measures can be estimated using Parzen windowing. Parzen windowing is a well established density estimation technique, which has been studied since the 1960s. The strengths and weaknesses of the method are well understood. Moreover, techniques for determining a proper *data-driven* size for the Parzen window have been thoroughly studied [19–23].

In this paper, we will show some equivalences between these two learning schemes, which until now has been treated separately. Specifically, we show that Parzen window-based estimators for the quadratic information measures have a dual interpretation as Mercer kernel-based measures, where they are

expressed as functions of *mean values* in the Mercer kernel feature space. The Mercer kernel and the Parzen window are shown to be equivalent. This means that if the Parzen window size can be reliably determined, then the corresponding Mercer kernel size is simultaneously determined by the same procedure.

Furthermore, we analyze a classification rule based on the integrated squared error measure, and show that this corresponds to a *linear* classifier in the kernel feature space. The SVM is also a linear classifier in the kernel space. By introducing *weighted* Parzen window estimators, we show that the SVM can be related to the integrated squared error measure, hence pointing out an equivalence between this information theoretic approach and the SVM.

This paper is organized as follows. In Section 2, we review the idea behind Mercer kernel-based learning theory. In Section 3, we give a brief review of the SVM. We discuss the Parzen window-based estimators for the quadratic information measures in Section 4, and show the relationship to Mercer kernel feature space quantities. The information theoretic classification rule is analyzed in Section 5. Thereafter, we derive the connection between this classifier and the SVM in Section 6. We make our concluding remarks in Section 7.

2. Mercer Kernel-Based Learning Theory

Mercer kernel-based learning algorithms make use of the following idea: via a non-linear mapping

$$\begin{aligned} \Phi : R^d &\to \mathcal{F} \\ \mathbf{x} &\to \Phi(\mathbf{x}) \end{aligned} \tag{1}$$

the data $\mathbf{x}_1, \ldots, \mathbf{x}_N \in R^d$ is mapped into a potentially much higher dimensional feature space $\mathcal{F}$. For a given learning problem one now considers the same learning problem in $\mathcal{F}$ instead of in R^d, working with $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_N) \in \mathcal{F}$.

The learning algorithm itself is typically linear in nature, and can be expressed solely in terms of inner-product evaluations. This makes it possible to apply the algorithm in feature space without actually carrying out the data mapping. The key ingredient is a highly effective trick for computing inner products in the feature space using *kernel functions*. One therefore *implicitly* executes the linear algorithm in kernel feature space. This property is advantageous since execution of the learning algorithm in a very high dimensional space is avoided. Because of the non-linear data mapping, the linear operation in kernel feature space corresponds to a non-linear operation in the input space.

Consider a symmetric kernel function $k(\mathbf{x}, \mathbf{y})$. If $k : \mathcal{C} \times \mathcal{C} \to R$ is a continuous kernel of a positive integral operator in a Hilbert space $L_2(\mathcal{C})$ on a compact set $\mathcal{C} \in R^d$, i.e.,

$$\forall \psi \in L_2(\mathcal{C}) : \int_{\mathcal{C}} k(\mathbf{x}, \mathbf{y}) \psi(\mathbf{x}) \psi(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0. \tag{2}$$

Then there exists a space $\mathcal{F}$ and a mapping $\Phi : R^d \to \mathcal{F}$, such that by Mercer's theorem [37]

$$k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle = \sum_{i=1}^{N_{\mathcal{F}}} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{y}), \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product, the ϕ_i's are the eigenfunctions of the kernel and $N_{\mathcal{F}} \leq \infty$ [6, 14]. This operation is known as the "kernel-trick," and it implicitly computes an inner-product in the kernel feature space via $k(\mathbf{x}, \mathbf{y})$.

Indeed, it has been pointed out that the kernel trick can be used to develop non-linear generalizations to any algorithm that can be cast in terms of inner-products [4, 10]. For example, KPCA, KFDA and kernel K-means [10, 38, 39] are simply extensions of the corresponding linear algorithms by applying the kernel-trick on every inner-product evaluation.

A kernel which satisfies Eq. (2) is known as a Mercer kernel. The most widely used Mercer kernel is the radial-basis-function (RBF)

$$k(\mathbf{x}, \mathbf{y}) = \exp\left\{ -\frac{||\mathbf{x} - \mathbf{y}||^2}{2\sigma^2} \right\}, \tag{4}$$

where σ is a scale parameter which controls the width of the RBF. A RBF kernel corresponds to an infinite-dimensional Mercer kernel feature space, since the RBF has an infinite number of eigenfunctions.

3. The Support Vector Machine

The support vector machine is the most prominent Mercer kernel-based learning algorithm. It is a hyperplane classifier which is based on two crucial oper-

ations: (1) The kernel-trick, which makes the otherwise linear SVM algorithm non-linear. (2) The maximization of the hyperplane margin, which is a *regularizing* condition on the hyperplane solution. Basically, it limits the admissible separating hyperplanes to the one maximizing the margin. This regularization has a positive effect on the generalization capability of the classifier [6].

In the following, we give a brief review of the SVM theory. We formulate the problem directly in the Mercer kernel feature space. This Mercer kernel feature space is induced by some kernel function, which hopefully makes the feature space data linearly separable such that it can be separated by a hyperplane. Whether or not the data in fact is linearly separable, heavily depends on the user choosing a proper kernel.

Let ω_1 and ω_2 denote two data classes. We are given a training set consisting of $\{\mathbf{x}_i\},\ i = 1, \ldots, N_1$, from ω_1, and $\{\mathbf{x}_j\},\ j = 1, \ldots, N_2$, from ω_2. The task is to train a SVM classifier, such that it creates a maximum margin linear classifier in the kernel feature space. After training, the classification rule in feature space is

$$\mathbf{x}_0 \rightarrow \omega_1 : \quad \mathbf{w}^{*T}\Phi(\mathbf{x}_0) + b^* \geq 0, \tag{5}$$

otherwise, $\mathbf{x}_0 \rightarrow \omega_2$. Here, $\mathbf{x}_0$ is a new, previously unseen data point. Presumably, it has either been generated by the process generating the ω_1 data, or the process generating the ω_2 data.

Regularizing by maximizing the margin in feature space corresponds to *minimizing the squared norm* of the (canonical) separating hyperplane weight vector, that is $||\mathbf{w}^*||^2$, given the constraints

$$\begin{aligned} \mathbf{w}^{*T}\Phi(\mathbf{x}_i) + b^* &\geq +1, \quad \forall \mathbf{x}_i \in \omega_1 \\ \mathbf{w}^{*T}\Phi(\mathbf{x}_j) + b^* &\leq -1, \quad \forall \mathbf{x}_j \in \omega_2. \end{aligned} \tag{6}$$

This is a constrained optimization problem, which is dealt with by introducing Lagrange multipliers $\alpha_i \geq 0$, $\alpha_j \geq 0$, corresponding to the two classes, and a primal Lagrangian

$$\begin{aligned} L_P = \frac{1}{2}||\mathbf{w}^*||^2 - \sum_{i=1}^{N_1} \alpha_i[\mathbf{w}^{*T}\Phi(\mathbf{x}_i) + b^* - 1] \\ + \sum_{j=1}^{N_2} \alpha_j[\mathbf{w}^{*T}\Phi(\mathbf{x}_j) + b^* + 1]. \end{aligned} \tag{7}$$

The Lagrangian L_P has to be minimized with respect to the primal variables $\mathbf{w}^*$ and b^*, and maximized with respect to the *dual* variables α_i, α_j. Hence a saddle point must be found. At the saddle point, the derivatives of L_P with respect to the primal variables must vanish,

$$\frac{\partial}{\partial b^*}L_P = 0, \quad \frac{\partial}{\partial \mathbf{w}^*}L_P = 0, \tag{8}$$

which leads to

$$\sum_{i=1}^{N_1} \alpha_i = \sum_{j=1}^{N_2} \alpha_j = \Omega, \tag{9}$$

and

$$\mathbf{w}^* = \mathbf{m}_1^* - \mathbf{m}_2^*, \tag{10}$$

where

$$\mathbf{m}_1^* = \sum_{i=1}^{N_1} \alpha_i \Phi(\mathbf{x}_i), \quad \mathbf{m}_2^* = \sum_{j=1}^{N_2} \alpha_j \Phi(\mathbf{x}_j). \tag{11}$$

By substituting these constraints into Eq. (7), the dual Lagrangian

$$\begin{aligned} L_D = 2\Omega - \frac{1}{2}\Bigg\{ \sum_{i,i'=1}^{N_1,N_1} \alpha_i \alpha_{i'} k(\mathbf{x}_i, \mathbf{x}_{i'}) \\ - 2\sum_{i,j=1}^{N_1,N_2} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{j,j'=1}^{N_2,N_2} \alpha_j \alpha_{j'} k(\mathbf{x}_j, \mathbf{x}_{j'}) \Bigg\}, \end{aligned} \tag{12}$$

is obtained, where $k(\cdot, \cdot)$ denotes an inner product between any two training data points in the Mercer kernel feature space. L_D must be maximized with respect to the Lagrange multipliers. It can be seen that the solution vector $\mathbf{w}^*$ has an expansion in terms of the training patterns weighted by the Lagrange multipliers. The Karush–Kuhn–Tucker (KKT) conditions

$$\begin{aligned} \alpha_i[\mathbf{w}^{*T}\Phi(\mathbf{x}_i) + b^* - 1] = 0, \quad \forall i = 1, \ldots, N_1, \\ \alpha_j[\mathbf{w}^{*T}\Phi(\mathbf{x}_j) + b^* + 1] = 0, \quad \forall j = 1, \ldots, N_2, \end{aligned} \tag{13}$$

specify the *non-zero* Lagrange multipliers to be those training patterns which are situated on the margin in

feature space. Hence, $\mathbf{w}^*$ is a weighted combination of the patterns on the margin.

Let us determine the expression for b^* in the SVM theory. For those b^* corresponding to support vectors belonging to ω_1, we have $b_1^* = 1 - \mathbf{w}^{*T}\Phi(\mathbf{x}_i)$, where $\Phi(\mathbf{x}_i)$ is a support vector. By adding all b_1^* values corresponding to ω_1, we have (remember that only those α_i's corresponding to support vectors deviate from zero)

$$\begin{aligned} \sum_{i=1}^{N_1} \alpha_i b_1^* &= \sum_{i=1}^{N_1} \alpha_i - \mathbf{w}^{*^T} \sum_{i=1}^{N_1} \alpha_i \Phi(\mathbf{x}_i) \\ \Omega b_1^* &= \Omega - \mathbf{w}^{*T}\mathbf{m}_1^* \\ b_1^* &= 1 - \frac{1}{\Omega}||\mathbf{m}_1^*||^2 + \frac{1}{\Omega}\mathbf{m}_1^{*T}\mathbf{m}_2^*. \end{aligned} \tag{14}$$

Similarly, for those b^*corresponding to support vectors belonging to ω_2, we have $b_2^* = -1 - \mathbf{w}^{*T}\Phi(\mathbf{x}_j)$. Again, by adding all b_2^* corresponding to ω_2, we obtain

$$\begin{aligned} \sum_{j=1}^{N_2} \alpha_j b_2^* &= -\sum_{j=1}^{N_2} \alpha_j - \mathbf{w}^{*T} \sum_{j=1}^{N_2} \alpha_j \Phi(\mathbf{x}_j) \\ \Omega b_2^* &= -\Omega - \mathbf{w}^{*T}\mathbf{m}_2^* \\ b_2^* &= -1 - \frac{1}{\Omega}\mathbf{m}_1^{*T}\mathbf{m}_2^* + \frac{1}{\Omega}||\mathbf{m}_2^*||^2. \end{aligned} \tag{15}$$

Since $b_1^* = b_2^*$, we have $b^* = \frac{1}{2}[b_1^* + b_2^*]$, such that

$$b^* = \frac{1}{2\Omega}[||\mathbf{m}_2^*||^2 - ||\mathbf{m}_1^*||^2]. \tag{16}$$

4. Quadratic Information Measures and Parzen Windowing

In this section, we will review the quadratic information measures, and show how they may be estimated non-parametrically using the Parzen window technique for density estimation. For details on how these cost functions may be used in adaptive systems training, we refer to [16, 17]. We will also show how each of these measures can be expressed in terms of *mean values* in a Mercer kernel feature space.

4.1. Parzen Window Density Estimator

Parzen windowing is a well-known kernel-based density estimation method [19, 40]. Given a set of iid samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ drawn from the true density $f(\mathbf{x})$, the Parzen window estimator for this distribution is defined as

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\sum_{t=1}^{N} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_t). \tag{17}$$

Here, W_{σ^2} is the Parzen window, or kernel, and σ^2 controls the width of the kernel. The Parzen window must integrate to one, and is typically chosen to be a pdf itself, such as the Gaussian kernel. Hence,

$$W_{\sigma^2}(\mathbf{x}, \mathbf{x}_t) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \exp\left\{-\frac{||\mathbf{x} - \mathbf{x}_t||^2}{2\sigma^2}\right\}.$$

We will use the Gaussian kernel in the derivations that follows, but show in the Appendix that other choices may be used. Note also that the *width* of the Parzen window affects the density estimate much more than the actual form of the window function [22, 23].

It is easily shown that Eq. (17) is an asymptotically unbiased and consistent estimator, provided σ decays to zero at a certain rate as N tends to infinity [19]. In the finite sample case, the window width is usually chosen such that it minimizes the mean integrated squared error (MISE) between $\hat{f}(\mathbf{x})$ and the target density $f(\mathbf{x})$. It is easily shown that the MISE consists of a bias part and a variance part. Unfortunately, the bias part is minimized by minimizing the window width, while the variance is minimized by maximizing the window width. This is the inherent bias-variance trade-off associated with the Parzen window technique.

Finding a window width, or kernel size, which provides a good bias-variance trade-off has been thoroughly studied in the statistics literature [21–23]. Especially for data sets of low to moderate dimensionality, many reliable methods exist, such as for example least-squares cross-validation [23]. Another straight-forward and popular approach is to find the kernel size which minimizes the asymptotic MISE (AMISE). By assuming that the underlying density is Gaussian, this kernel size is given by

$$\sigma_{\text{AMISE}} = \sigma_X \left[\frac{4}{(2d+1)N}\right]^{\frac{1}{d+4}}, \tag{18}$$

where $\sigma_X^2 = d^{-1}\sum_i \Sigma_{\mathbf{X_{ii}}}$, and $\Sigma_{\mathbf{X_{ii}}}$ are the diagonal elements of the sample covariance matrix [21]. The main appeal of this approach is that is very easy to use. The obvious drawback is that it assumes that the underlying density is unimodal and Gaussian. Many other methods exist, each having specific properties.

For high-dimensional data sets, the Parzen window technique for density estimation is known to have severe limitations. The reason is that the usual bias-variance trade-off cannot be accomplished very well in higher dimensions without very large samples [21, 22]. This is known as the "curse-of-dimensionality."

Note however that this limitation may not apply directly when the Parzen window technique is used in clustering or classification, as discussed by Friedman [41]. He showed that in those applications, low variance is much more important than low bias, hence favoring a large kernel size.

4.2. Renyi Quadratic Entropy

The Renyi quadratic entropy associated with the pdf $f(\mathbf{x})$ is given by [27, 28]

$$H_{R_2}(f) = -\log \int f^2(\mathbf{x}) d\mathbf{x}. \tag{19}$$

We have available a sample from $f(\mathbf{x})$, namely $\{\mathbf{x}_t\},\ t = 1, \ldots, N$. Based on the sample, we estimate $f(\mathbf{x})$ by $\hat{f}(\mathbf{x})$, the Parzen window estimator. We obtain an estimate for the Renyi entropy using the *plug-in* a density estimator principle, by replacing $f(\mathbf{x})$ by $\hat{f}(\mathbf{x})$. However, since the logarithm is a monotonic function, we will focus on the quantity $V(f) = \int \hat{f}^{\,2}(\mathbf{x}) d\mathbf{x}$, thus given by[1]

$$\begin{aligned} V(f) &= \int \frac{1}{N}\sum_{t=1}^{N} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_t) \frac{1}{N}\sum_{t'=1}^{N} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_{t'}) d\mathbf{x} \\ &= \frac{1}{N^2}\sum_{t,t'=1}^{N,N} \int W_{\sigma^2}(\mathbf{x}, \mathbf{x}_t) W_{\sigma^2}(\mathbf{x}, \mathbf{x}_{t'}) d\mathbf{x}. \end{aligned} \tag{20}$$

Now a property of Gaussian functions is employed. By the convolution theorem for Gaussians, we have

$$\int W_{\sigma^2}(\mathbf{x}, \mathbf{x}_t) W_{\sigma^2}(\mathbf{x}, \mathbf{x}_{t'}) d\mathbf{x} = W_{2\sigma^2}(\mathbf{x}_t, \mathbf{x}_{t'}), \tag{21}$$

that is, the convolution of two Gaussians is a new Gaussian function having twice the (co)variance. Thus, we have

$$V(f) = \frac{1}{N^2}\sum_{t,t'=1}^{N,N} W_{2\sigma^2}(\mathbf{x}_t, \mathbf{x}_{t'}). \tag{22}$$

It can be seen that this estimation procedure involves no approximations, besides the pdf estimator itself. Eq. (22) was named the *information potential* [16], because of an analogy to a potential energy field.

The key point in the following discussion is to note that $W_{2\sigma^2}(\mathbf{x}_t, \mathbf{x}_{t'})$, for any $\mathbf{x}_t,\ \mathbf{x}_{t'}$, is a Gaussian kernel function, and hence it is also a *kernel function that satisfies Mercer's theorem*. Thus

$$W_{2\sigma^2}(\mathbf{x}_t, \mathbf{x}_{t'}) = k(\mathbf{x}_t, \mathbf{x}_{t'}) = \left\langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_{t'}) \right\rangle. \tag{23}$$

Hence, the Parzen window-based estimator for the information potential can be expressed in terms of inner products in a Mercer kernel space. In the following we make this connection explicit. We rewrite Eq. (20) as follows

$$\begin{aligned} V(f) &= \frac{1}{N^2}\sum_{t,t'=1}^{N,N} \left\langle \Phi(\mathbf{x}_t), \Phi(\mathbf{x}_{t'}) \right\rangle \\ &= \left\langle \frac{1}{N}\sum_{t=1}^{N} \Phi(\mathbf{x}_t), \frac{1}{N}\sum_{t'=1}^{N} \Phi(\mathbf{x}_{t'}) \right\rangle \\ &= \mathbf{m}^T\mathbf{m} \\ &= ||\mathbf{m}||^2, \end{aligned} \tag{24}$$

where $\mathbf{m}$ is the mean vector of the Φ-transformed data

$$\mathbf{m} = \frac{1}{N}\sum_{t=1}^{N} \Phi(\mathbf{x}_t). \tag{25}$$

That is, it turns out that the information potential may be expressed as the squared norm of the *mean vector* of the data in a Mercer kernel feature space. This connection was previously pointed out by [42] in a study relating orthogonal series density estimation to kernel principal component analysis.

4.3. Integrated Squared Error as a PDF Divergence

In order to measure the "distance" or divergence between two probability densities, $p(\mathbf{x})$ and $q(\mathbf{x})$, an integrated squared error (ISE) criterion may be used

$$\begin{aligned} ISE(p,q) &= \int [p(\mathbf{x}) - q(\mathbf{x})]^2 d\mathbf{x} \\ &= \int p^2(\mathbf{x}) d\mathbf{x} - 2\int p(\mathbf{x}) q(\mathbf{x}) d\mathbf{x} \\ &\quad + \int q^2(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{26}$$

It can be seen that the integrated squared error criterion is always non-negative, it is symmetric and it vanishes if and only if the two pdfs are identical. Such a measure is well-known in density estimation [43]. It has also been used for discrete distributions [44]. In [16], this measure was used primarily for computational simplicity.

We have available a sample from $p(\mathbf{x})$, namely $\{\mathbf{x}_i\}$, $i = 1, \ldots, N_1$, and a corresponding sample from $q(\mathbf{x})$, that is, $\{\mathbf{x}_j\}$, $j = 1, \ldots, N_2$.[2] We estimate the two pdfs by the Parzen window method

$$\hat{p}(\mathbf{x}) = \frac{1}{N_1} \sum_{i=1}^{N_1} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_i), \qquad \hat{q}(\mathbf{x}) = \frac{1}{N_2} \sum_{j=1}^{N_2} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_j). \tag{27}$$

These estimators are now used to estimate the *ISE* divergence. Note that we have for simplicity assumed that the same kernel size σ is appropriate for both estimators. This may not be the case in practice. The latter situation may easily be incorporated in the subsequent analysis. Now, performing a similar calculation as above, the *ISE* can be estimated non-parametrically as follows

$$\begin{aligned} \widehat{ISE}(p,q) &= \frac{1}{N_1^2} \sum_{i,i'=1}^{N_1,N_1} W_{2\sigma^2}(\mathbf{x}_i, \mathbf{x}_{i'}) - \frac{2}{N_1 N_2} \\ &\quad \times \sum_{i,j=1}^{N_1,N_2} W_{2\sigma^2}(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{N_2^2} \\ &\quad \times \sum_{j,j'=1}^{N_2,N_2} W_{2\sigma^2}(\mathbf{x}_j, \mathbf{x}_{j'}). \end{aligned}$$

In analogy to Eq. (22), the $\widehat{ISE}$ may also be expressed in terms of mean vectors in the Mercer kernel feature space. When we perform a similar calculation, we obtain

$$\begin{aligned} \widehat{ISE}(p,q) &= ||\mathbf{m}_1||^2 - 2\mathbf{m}_1^T \mathbf{m}_2 + ||\mathbf{m}_2||^2 \\ &= ||\mathbf{m}_1 - \mathbf{m}_2||^2, \end{aligned} \tag{28}$$

where $\mathbf{m}_1$ is the kernel feature space mean vector of the data points drawn from $p(\mathbf{x})$, and $\mathbf{m}_2$ is the kernel feature space mean vector of the data points drawn from $q(\mathbf{x})$. That is

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} \Phi(\mathbf{x}_i) \quad \mathbf{m}_1 = \frac{1}{N_2} \sum_{j=1}^{N_2} \Phi(\mathbf{x}_j). \tag{29}$$

Hence, the *ISE* divergence measure can also be seen to have a geometric interpretation in the kernel feature space. It measures the square of the norm of the difference vector between the two means $\mathbf{m}_1$ and $\mathbf{m}_2$.

4.4. Cauchy–Schwarz PDF Divergence

Based on the Cauchy–Schwarz inequality, [16] also used the following divergence measure between the pdfs $p(\mathbf{x})$ and $q(\mathbf{x})$

$$D_{CS}(p,q) = -\log \frac{\int p(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}}{\sqrt{\int p^2(\mathbf{x}) d\mathbf{x} \int q^2(\mathbf{x}) d\mathbf{x}}}, \tag{30}$$

which we refer to as the Cauchy–Schwarz pdf divergence. It is also always non-negative, it is symmetric and it vanishes if and only if the two densities are equal.

Since the logarithm is monotonic, we will focus on the quantity in the argument of the log in Eq. (30). The Parzen window-based estimator for this quantity was named the *information cut* (IC) in [45], because it was shown to be closely related to the graph theoretic *cut*. By a similar calculation as above, the IC can be expressed as

$$IC(p,q) = \frac{\frac{1}{N_1 N_2} \sum_{i,j=1}^{N_1,N_2} W_{2\sigma^2}(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{\frac{1}{N_1^2} \sum_{i,i'=1}^{N_1,N_1} W_{2\sigma^2}(\mathbf{x}_i, \mathbf{x}_{i'}) \; \frac{1}{N_2^2} \sum_{j,j'=1}^{N_2,N_2} W_{2\sigma^2}(\mathbf{x}_j, \mathbf{x}_{j'})}}. \tag{31}$$

Also, the information cut may be expressed in terms of mean vectors in the Mercer kernel feature space, as

$$IC = \frac{\mathbf{m}_1^T \mathbf{m}_2}{||\mathbf{m}_1||||\mathbf{m}_2||} = \cos \angle(\mathbf{m}_1, \mathbf{m}_2), \qquad (32)$$

Hence, it turns out that the information cut has a dual interpretation as a measure of the cosine of the angle between cluster mean vectors in the Mercer kernel feature space.

5. ISE-based Classification

In this section we will analyze a classification rule based on the ISE, which can be interpreted both in the input space and in the Mercer kernel space. In the next section, we will relate the SVM to such a classifier, hence providing another equivalence between kernel methods and information theoretic methods.

We have available the training data points $\{\mathbf{x}_i\},\ i = 1, \ldots, N_1$, drawn from $p(\mathbf{x})$, and a corresponding sample from $q(\mathbf{x})$, that is, $\{\mathbf{x}_j\},\ j = 1, \ldots, N_2$. Based on this training data set we wish to construct a classifier, which assigns a test data point $\mathbf{x}_0$ to one of the classes ω_1 or ω_2. Now, we define

$$\begin{aligned} \hat{p}'(\mathbf{x}) &= \frac{1}{N_1 + 1} \sum_{i=0}^{N_1} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_i), \\ \hat{q}'(\mathbf{x}) &= \frac{1}{N_2 + 1} \sum_{j=0}^{N_2} W_{\sigma^2}(\mathbf{x}, \mathbf{x}_j). \end{aligned} \qquad (33)$$

Hence, $\hat{p}'(\mathbf{x})$ is the Parzen estimator for $p(\mathbf{x})$, assuming $\mathbf{x}_0$ is included in the ω_1 data set. Likewise, $\hat{q}'(\mathbf{x})$ is the Parzen estimator for $q(\mathbf{x})$, assuming $\mathbf{x}_0$ is included in the ω_2 data set.

The proposed ISE-based strategy is to classify $\mathbf{x}_0$ according to the following rule:

$$\begin{aligned} \mathbf{x}_0 \to \omega_1 :\ & \int [\hat{p}'(\mathbf{x}) - \hat{q}(\mathbf{x})]^2 d\mathbf{x} \\ & \geq \int [\hat{p}(\mathbf{x}) - \hat{q}'(\mathbf{x})]^2 d\mathbf{x}, \end{aligned} \qquad (34)$$

otherwise, assign $\mathbf{x}_0$ to ω_2. In words; assign $\mathbf{x}_0$ to the class which, when having $\mathbf{x}_0$ appended to it, makes the estimated divergence between the classes the greatest.

We will now analyze this simple classification rule in terms of the Mercer kernel feature space. Let $\mathbf{m}_i',\ i = 1, 2$ be the Mercer kernel feature space mean vector of class ω_i, assuming $\Phi(\mathbf{x}_0)$ is assigned to that class. It is easily shown that

$$\begin{aligned} \mathbf{m}_1' &= \frac{N_1}{N_1 + 1} \mathbf{m}_1 + \frac{1}{N_1 + 1} \Phi(\mathbf{x}_0) \\ \mathbf{m}_2' &= \frac{N_2}{N_2 + 1} \mathbf{m}_2 + \frac{1}{N_2 + 1} \Phi(\mathbf{x}_0). \end{aligned} \qquad (35)$$

In the kernel feature space, the equivalent classification rule of Eq. (34) may be expressed as

$$\mathbf{x}_0 \to \omega_1 : \quad ||\mathbf{m}_1' - \mathbf{m}_2||^2 \geq ||\mathbf{m}_1 - \mathbf{m}_2'||^2. \qquad (36)$$

In what follows, we look at a special case. *Assume that $P(\omega_1) = P(\omega_2)$, that is the prior probabilities for the classes are equal. Let $P(\omega_1) = \frac{N_1}{N}$ and $P(\omega_2) = \frac{N_2}{N}$, which means that we assume that $N_1 = N_2$* In that case, we have

$$\begin{aligned} \mathbf{m}_1' &= \kappa_1 \mathbf{m}_1 + \kappa_2 \Phi(\mathbf{x}_0) \\ \mathbf{m}_2' &= \kappa_1 \mathbf{m}_2 + \kappa_2 \Phi(\mathbf{x}_0), \end{aligned} \qquad (37)$$

where $\kappa_1 = \frac{N_1}{N_1+1} = \frac{N_2}{N_2+1}$, and $\kappa_2 = \frac{1}{N_1+1} = \frac{1}{N_2+1}$.

For ease of notation, let $\Phi(\mathbf{x}_0) = \mathbf{y}$. The left-hand side of Eq. (36), becomes

$$\begin{aligned} ||\mathbf{m}_1' - \mathbf{m}_2||^2 &= \mathbf{m}_1'^T \mathbf{m}_1' - 2\mathbf{m}_1'^T \mathbf{m}_2 + \mathbf{m}_2^T \mathbf{m}_2 \\ &= \kappa_1^2 ||\mathbf{m}_1||^2 + 2\kappa_1 \kappa_2 \mathbf{m}_1^T \mathbf{y} + \kappa_2^2 ||\mathbf{y}||^2 \\ &\quad - 2\kappa_1 \mathbf{m}_1^T \mathbf{m}_2 - 2\kappa_2 \mathbf{m}_2^T \mathbf{y} + ||\mathbf{m}_2||^2. \end{aligned}$$

Similarly, the right-hand side of Eq. (36) becomes

$$\begin{aligned} ||\mathbf{m}_1 - \mathbf{m}_2'||^2 &= \mathbf{m}_1^T \mathbf{m}_1 - 2\mathbf{m}_1^T \mathbf{m}_2' + \mathbf{m}_2'^T \mathbf{m}_2 \\ &= ||\mathbf{m}_1||^2 - 2\kappa_1 \mathbf{m}_2^T \mathbf{m}_1 - 2\kappa_2 \mathbf{m}_1^T \mathbf{y} \\ &\quad + \kappa_1^2 ||\mathbf{m}_2||^2 + 2\kappa_1 \kappa_2 \mathbf{m}_2^T \mathbf{y} + \kappa_2^2 ||\mathbf{y}||^2. \end{aligned}$$

Using these results, the classification rule becomes

$$\begin{aligned} \mathbf{x}_0 \to \omega_1 :\quad & ||\mathbf{m}_1' - \mathbf{m}_2||^2 \geq ||\mathbf{m}_1 - \mathbf{m}_2'||^2 \\ \Leftrightarrow\quad & \mathbf{m}_1^T \mathbf{y} - \mathbf{m}_2^T \mathbf{y} - \frac{\kappa_1^2 - 1}{2\kappa_2[\kappa_1 + 1]} \\ & \left[||\mathbf{m}_2||^2 - |\mathbf{m}_1||^2\right] \geq 0 \\ \Leftrightarrow\quad & \mathbf{m}_1^T \mathbf{y} - \mathbf{m}_2^T \mathbf{y} + b \geq 0. \end{aligned} \qquad (38)$$

where $b = \frac{1}{2}\left[||\mathbf{m}_2||^2 - ||\mathbf{m}_1||^2\right]$, and the constant $\frac{\kappa_1^2-1}{\kappa_2[\kappa_1+1]} = -1$.

In fact, the above classification rule has a simple geometrical interpretation. *The point* $\mathbf{y}$ *is assigned to the class whose mean it is closest, and the class boundary in kernel feature space is a hyperplane given by a vector* $\mathbf{w}$. Let $\mathbf{w} = \mathbf{m}_1 - \mathbf{m}_2$, and let the midpoint between $\mathbf{m}_1$ and $\mathbf{m}_2$ be given by $\mathbf{v} = \frac{1}{2}(\mathbf{m}_1 + \mathbf{m}_2)$. Now the class of $\mathbf{y}$ is determined by examining whether the vector $(\mathbf{y} - \mathbf{v})$ encloses an angle smaller than $\frac{\pi}{2}$ with the vector $\mathbf{w}$ or not. If it does, $\mathbf{y}$ is closest to $\mathbf{m}_1$, and $\mathbf{y}$ is assigned to ω_1. Hence,

$$\begin{aligned} \mathbf{x}_0 \to \omega_1 : \quad & \mathbf{w}^T(\mathbf{y} - \mathbf{v}) \geq 0 \\ \Leftrightarrow \quad & \mathbf{w}^T\mathbf{y} + b \geq 0 \\ \Leftrightarrow \quad & \mathbf{m}_1^T\mathbf{y} - \mathbf{m}_2^T\mathbf{y} + b \geq 0, \end{aligned} \tag{39}$$

Figure 2 geometrically illustrates this simple classification rule, which we have derived using the ISE criterion as a starting point.

As explained above, in the Mercer kernel space, the value of the inner-product between the class mean values and the new data point determines which class it is assigned to. The threshold value, b, depends on the squared Euclidean norms of the mean values, which according to Eq. (24) are equivalent to the class information potentials, and hence the class entropies.

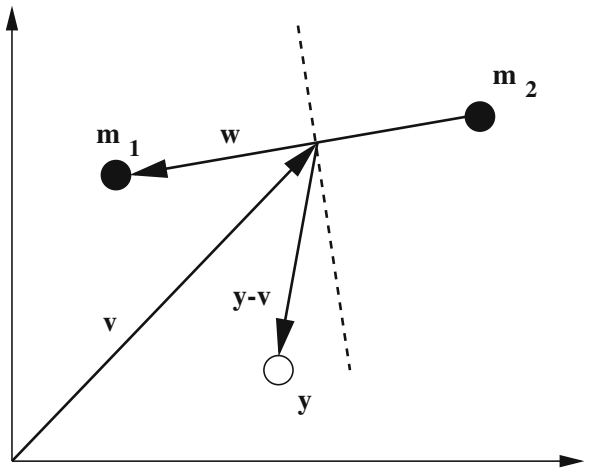


Figure 2. ISE-based geometric classification rule: Assign the point $\mathbf{y}$ to the class whose mean it is closest to. This can be done by looking at the inner-product between $(\mathbf{y} - \mathbf{v})$ and $\mathbf{w}$. It changes sign as the enclosed angle passes through $\frac{\pi}{2}$. The corresponding decision boundary is given by a *hyperplane* orthogonal to $\mathbf{w}$ (*dashed line*).

We now complete the circle, and analyze the Mercer kernel feature space classification rule in terms of Parzen estimators in the input space. Note that

$$\begin{aligned} \mathbf{m}_1^T\mathbf{y} = \mathbf{m}_1^T\Phi(\mathbf{x}_0) &= \frac{1}{N_1}\sum_{i=1}^{N_1}\Phi^T(\mathbf{x}_i)\Phi(\mathbf{x}_0) \\ &= \frac{1}{N_1}\sum_{i=1}^{N_1} W_{\sigma^2}(\mathbf{x}_0, \mathbf{x}_i) = \hat{p}(\mathbf{x}_0). \end{aligned} \tag{40}$$

Likewise

$$\begin{aligned} \mathbf{m}_2^T\mathbf{y} = \mathbf{m}_2^T\Phi(\mathbf{x}_0) &= \frac{1}{N_2}\sum_{j=1}^{N_2}\Phi^T(\mathbf{x}_j)\Phi(\mathbf{x}_0) \\ &= \frac{1}{N_2}\sum_{j=1}^{N_2} W_{\sigma^2}(\mathbf{x}_0, \mathbf{x}_j) = \hat{q}(\mathbf{x}_0). \end{aligned} \tag{41}$$

The classification rule hence becomes

$$\mathbf{x}_0 \to \omega_1 : \quad \hat{p}(\mathbf{x}_0) - \hat{q}(\mathbf{x}_0) + b \geq 0. \tag{42}$$

We remark that this classification rule depends both on the estimated densities at $\mathbf{x}_0$, and on the information potentials of the two classes. We have already shown that these information potentials are equivalent to Renyi's quadratic entropies for the classes.

In the case that the classes have the same value for the information potential (entropy), which means that the kernel feature space mean values have equal length from the origin, we have $b = 0$, and the current classification rule reduces to the well-known Bayes classification rule (for equal priors), where the class probability densities are estimated using Parzen windowing.

The same direct connection cannot be obtained based on the Cauchy–Schwarz divergence.

6. ISE-Based Classification and the SVM

In the previous section, we analyzed an information theoretic classification rule, which turned out to have a dual interpretation as a *hyperplane classifier* in a Mercer kernel feature space. We will now relate this classifier to the SVM, by introducing weighted

Parzen window estimators. The following discussion therefore provides an information theoretic perspective to the SVM.

The ISE classifier is entirely determined by the mean vectors $\mathbf{m}_1$ and $\mathbf{m}_2$ of the training data, since both $\mathbf{w}$ and b are determined by these vectors. We are therefore totally dependent on these mean vectors to truly represent the structure of the data. For example, the presence of outliers in the training set may affect the computation of $\mathbf{w}$ and b in such a way that the performance of the classifier is degraded. A possible approach to make the procedure more robust may be to allow the contribution of each training data point to the mean vectors to be weighted differently.

Let us therefore introduce the weighting components $\alpha_i \geq 0$ associated with ω_1, and $\alpha_j \geq 0$ associated with ω_2. The *weighted mean vectors* then become

$$\mathbf{m}_1 = \frac{1}{\Omega_1}\sum_{i=1}^{N_1}\alpha_i\Phi(\mathbf{x}_i), \quad \mathbf{m}_2 = \frac{1}{\Omega_2}\sum_{j=1}^{N_2}\alpha_j\Phi(\mathbf{x}_j). \tag{43}$$

By introducing such weighted mean vectors, we also need to introduce some criterion to determine proper weights. Such a criterion should be optimal with respect to classifier performance. The performance of a classifier is measured by its success rate on test data. Hence, the classifier should generalize well. In statistical learning theory, it has been shown that minimization of the squared norm of the hyperplane weight vector, while satisfying the classification constraints on the training data, improves generalization performance.

Based on the arguments above, we may relate the vector $\mathbf{w} = \mathbf{m}_1 - \mathbf{m}_2$ to the SVM weight vector $\mathbf{w}^* = \mathbf{m}_1^* - \mathbf{m}_2^*$. Recall that the SVM is exactly based on regularization by minimization of $||\mathbf{w}^*||^2$. The minimization is accompanied by the classification constraints, which ensures that the training data is classified correctly. These constraints say

$$\begin{aligned} \mathbf{w}^{*T}\Phi(\mathbf{x}_i) + b^* \geq +1, \quad \forall \mathbf{x}_i \in \omega_1 \\ \mathbf{w}^{*T}\Phi(\mathbf{x}_j) + b^* \leq -1, \quad \forall \mathbf{x}_j \in \omega_2. \end{aligned} \tag{44}$$

In fact, if Ω_1 and Ω_2 were equal, then $\mathbf{w}$ and $\mathbf{w}^*$ would only differ by a constant.

Let us take a closer look at the information potentials associated with the weighted mean vectors. We have

$$||\mathbf{m}_1||^2 = \frac{1}{\Omega_1^2}\sum_{i,i'=1}^{N_1,N_1}\alpha_i\alpha_{i'}k(\mathbf{x}_i,\mathbf{x}_{i'}) = \int \hat{p}^{\,2}(\mathbf{x})d\mathbf{x}. \tag{45}$$

Thus, the weighted mean vector $\mathbf{m}_1$ is associated with

$$\hat{p}(\mathbf{x}) = \frac{1}{\Omega_1}\sum_{i=1}^{N_1}\alpha_i W_{\sigma^2}(\mathbf{x},\mathbf{x}_i), \tag{46}$$

a weighted Parzen window estimator in the input space. We likewise have

$$\hat{q}(\mathbf{x}) = \frac{1}{\Omega_2}\sum_{j=1}^{N_2}\alpha_j W_{\sigma^2}(\mathbf{x},\mathbf{x}_j). \tag{47}$$

The kernels which constitute these Parzen window estimators are no longer equally important. Recall that the ISE classification rule based on the density estimators is

$$\mathbf{x}_0 \rightarrow \omega_1 : \quad \hat{p}(\mathbf{x}_0) - \hat{q}(\mathbf{x}_0) + b \geq 0, \tag{48}$$

with $b = \frac{1}{2}\left[||\mathbf{m}_2||^2 - ||\mathbf{m}_1||^2\right]$. In order to derive this classification rule using the traditional Parzen window estimators, we assumed that $N_1 = N_2$. Using the weighted Parzen window estimators instead, it is easily found that the corresponding assumption becomes $\Omega_1 = \Omega_2 = \Omega$ (see Eq. (37) and the related discussion in Section 5). Therefore,

$$\mathbf{m}_1 = \frac{1}{\Omega}\mathbf{m}_1^*, \quad \mathbf{m}_2 = \frac{1}{\Omega}\mathbf{m}_2^*, \tag{49}$$

and consequently

$$\mathbf{w} = \frac{1}{\Omega}\mathbf{w}^*. \tag{50}$$

Now, using the weighted Parzen window estimators we may express the SVM optimization problem in an *information theoretic framework* as follows

$$\min_{\alpha_i,\,\alpha_j} ||\mathbf{w}^*||^2 = \min_{\alpha_i,\,\alpha_j} \Omega^2||\mathbf{w}||^2 = \min_{\alpha_i,\,\alpha_j} \Omega^2||\mathbf{m}_1 - \mathbf{m}_2||^2. \tag{51}$$

Since $||\mathbf{m}_1 - \mathbf{m}_2||^2$ is the Mercer kernel feature space equivalent to the ISE pdf divergence, we have

$$\min_{\alpha_i,\, \alpha_j} \Omega^2||\mathbf{m}_1 - \mathbf{m}_2||^2 = \min_{\alpha_i,\, \alpha_j} \Omega^2 \int [\hat{p}(\mathbf{x}) - \hat{q}(\mathbf{x})]^2 d\mathbf{x}. \tag{52}$$

The optimization is subject to classification constraints, expressed as

$$\begin{aligned} 1) \quad & \mathbf{w}^{*T}\Phi(\mathbf{x}_i) + b^* \geq 1 \\ \Leftrightarrow \quad & \Omega\mathbf{w}^T\Phi(\mathbf{x}_i) + \Omega b \geq 1 \\ \Leftrightarrow \quad & \mathbf{w}^T\Phi(\mathbf{x}_i) + b \geq \frac{1}{\Omega} \\ \Leftrightarrow \quad & \hat{p}(\mathbf{x}_i) - \hat{q}(\mathbf{x}_i) + b \geq \frac{1}{\Omega}, \end{aligned} \tag{53}$$

for $i = 1, \ldots, N_1$.

$$\begin{aligned} 2) \quad & \mathbf{w}^{*T}\Phi(\mathbf{x}_j) + b^* \leq -1 \\ \Leftrightarrow \quad & \Omega\mathbf{w}^T\Phi(\mathbf{x}_j) + \Omega b \leq -1 \\ \Leftrightarrow \quad & \mathbf{w}^T\Phi(\mathbf{x}_j) + b \leq -\frac{1}{\Omega} \\ \Leftrightarrow \quad & \hat{p}(\mathbf{x}_j) - \hat{q}(\mathbf{x}_j) + b \leq -\frac{1}{\Omega}, \end{aligned} \tag{54}$$

for $j = 1, \ldots, N_2$.

Likewise, the SVM classification rule, using the weighted Parzen window estimators, becomes

$$\begin{aligned} \mathbf{x}_0 \rightarrow \omega_1 : \quad & \mathbf{w}^{*T}\Phi(\mathbf{x}_0) + b^* \geq 0 \\ \Leftrightarrow \quad & \Omega\mathbf{w}^T\Phi(\mathbf{x}_0) + \Omega b \geq 0 \\ \Leftrightarrow \quad & \mathbf{w}^T\Phi(\mathbf{x}_0) + b \geq 0 \\ \Leftrightarrow \quad & \hat{p}(\mathbf{x}_0) - \hat{q}(\mathbf{x}_0) + b \geq 0. \end{aligned} \tag{55}$$

The weighted Parzen window estimators $\hat{p}(\mathbf{x})$ and $\hat{q}(\mathrm{x})$, as defined above, are *bona fide* density estimators. That is, they are always non-negative and integrate to one. However, since the weights are determined by minimizing the ISE pdf divergence, which puts emphasis on the points close to the class boundary trying to maximize the overlap between the class pdfs, we do not regard them as proper estimators for the pdfs that generated the data. From SVM theory, we know that in the Mercer kernel feature space, the only non-zero weighting components are those which correspond to data patterns on the margin. In the input space, it seems that the corresponding non-zero weighting components will be associated with data patterns near the class boundary. We therefore interpret the minimization of the ISE pdf divergence as a *sparseness criterion*, which tunes the classifier to those patterns which are near the boundary. The other data patterns should be much easier to classify correctly, and are not given any weight in the design of the classifier. The performance of the classifier is secured by the classification constraints. Note that weighted Parzen window estimators have been previously proposed for improved Parzen window-based Bayes classification [46].

In summary, we have found that one may view the SVM theory in feature space in terms of weighted Parzen density estimation in the input space, where regularization is obtained by minimizing the integrated squared error criterion. Hence, in an information theoretic framework, the support vector machine is formulated by introducing the weights $\alpha_i \geq 0$, $\alpha_j \geq 0$, and estimating the class densities according to

$$\begin{aligned} \hat{p}(\mathbf{x}) &= \frac{1}{\Omega}\sum_{i=1}^{N_1} \alpha_i W_{\sigma^2}(\mathbf{x}, \mathbf{x}_i), \\ \hat{q}(\mathbf{x}) &= \frac{1}{\Omega}\sum_{j=1}^{N_2} \alpha_j W_{\sigma^2}(\mathbf{x}, \mathbf{x}_j). \end{aligned} \tag{56}$$

The weights, and hence $\hat{p}(\mathbf{x})$ and $\hat{q}(\mathbf{x})$, are learned by enforcing a regularization criterion

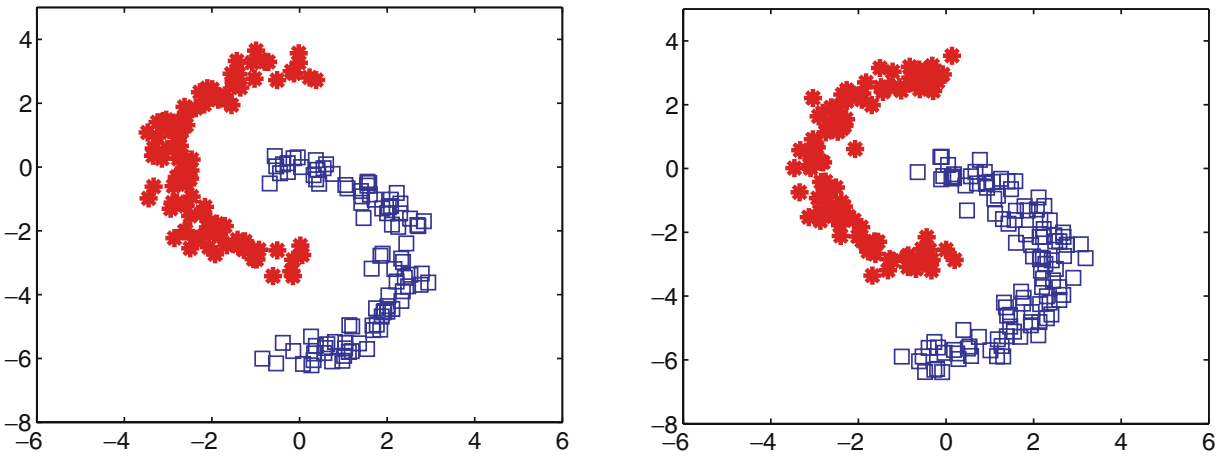
$$\min_{\alpha_i,\, \alpha_j} \Omega^2 \int [\, \hat{p}(\mathbf{x}) - \hat{q}(\mathbf{x})]^2 d\mathbf{x}, \tag{57}$$

subject to the classification constraints,

$$\begin{aligned} \hat{p}(\mathbf{x}_i) - \hat{q}(\mathbf{x}_i) + b \geq +\frac{1}{\Omega}, \quad \forall \mathbf{x}_i \in \omega_1, \\ \hat{p}(\mathbf{x}_j) - \hat{q}(\mathbf{x}_j) + b \leq -\frac{1}{\Omega}, \quad \forall \mathbf{x}_j \in \omega_2. \end{aligned} \tag{58}$$

In terms of the weighted Parzen window estimators, the classification rule then becomes $\mathbf{x}_0 \rightarrow \omega_1 : \hat{p}(\mathbf{x}_0) - \hat{q}(\mathbf{x}_0) + b \geq 0$.

The SVM may therefore be interpreted as an enhanced, more complex, version of the ISE-based classification rule analyzed in Section 5.



(a) Training data (b) Test data classified

Figure 3. Classification of simple artificially created data set using both the ISE classifier and the SVM.

6.1. Experimental Illustration

A simple artificial data set is constructed, in order to illustrate that the SVM, which is based on regularization using weighted Parzen window estimators, seems indeed to be more robust to outliers than the ISE classifier, which is based on traditional Parzen window estimators.

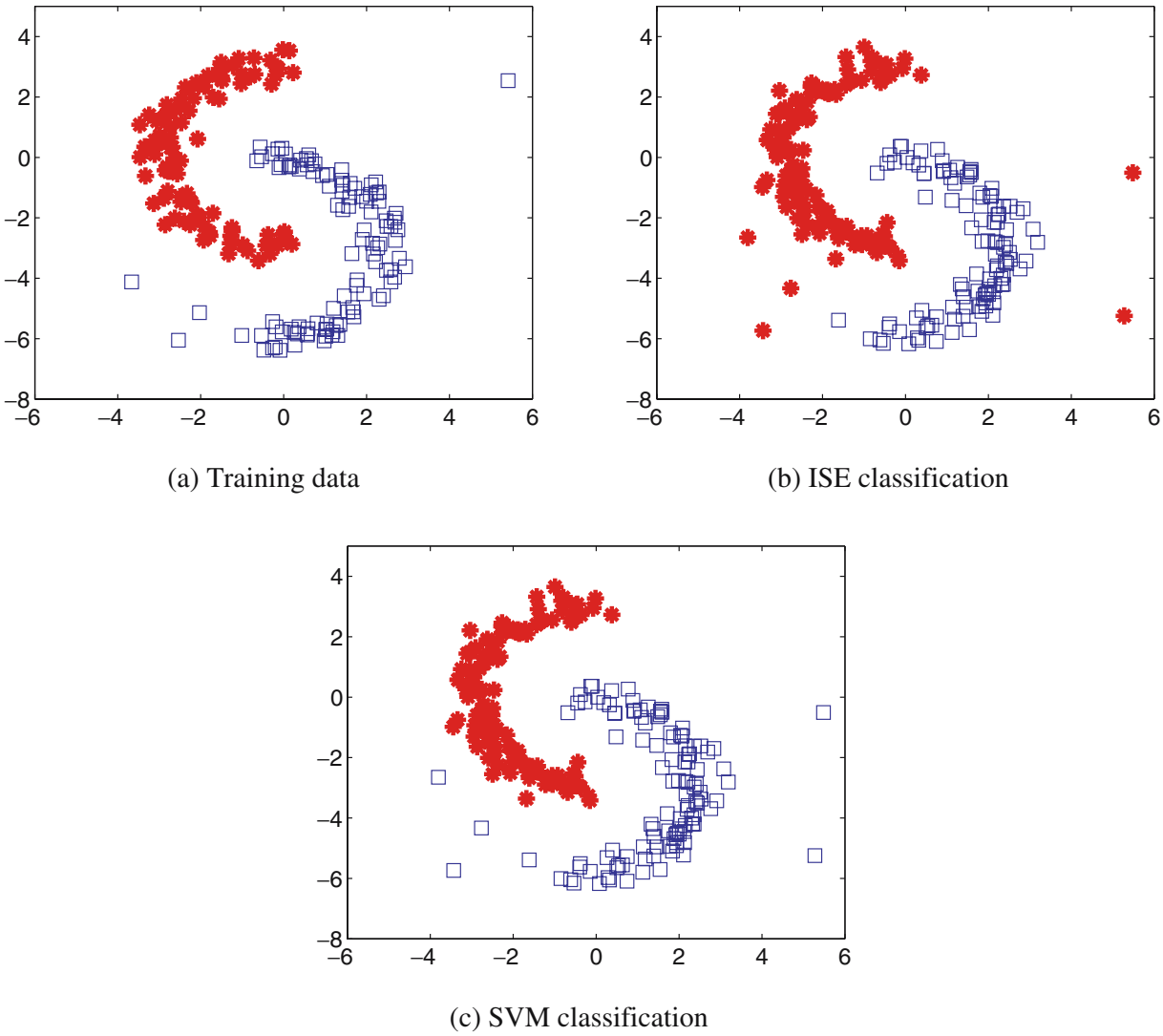
Figure 3a shows a training data set, with the known class labels indicated. It consists of two half-moon shaped classes, which are non-linearly separable. A test data set, drawn from the same distributions as the training data set, is available. The task is to correctly classify the test data set. The MISE rule, Eq. (18), is employed to determine a proper Parzen window size, hence determining the Mercer kernel size also. The resulting kernel size is $\sigma = 0.7$. Using this kernel size, an ISE-based classification is performed. The result is shown in Fig. 3b. By visual inspection, the result seems reasonable. A SVM is also trained using the same kernel size. In this case, the SVM classification and the ISE classification are identical.

Next, a few data points are added to the training data set. These data points may be considered outliers. The resulting data set is shown in Fig. 4a, with the known class labels indicated. The ISE-based classification is shown in Fig. 4b. The outliers turn out not to be assigned to the correct class. This may be a result of the fact that all data points are weighted equally in the computation of the Mercer kernel space mean vectors. On the other hand, the SVM obtains the result shown in Fig. 4c. This result reflects the structure of the training data to a higher degree. The improved performance in this case may be attributed to the weighting property of the SVM.

We have also conducted experiments on real data sets, which are not shown here. These experiments indicate that in many cases the ISE-based classification may perform quite well, but that in some cases the SVM regularization has a positive impact on the classifier performance. Since the purpose of this paper is to establish theoretical equivalences between Mercer kernel methods and information theoretic methods, we will not try to further analyze the classification improvements that the SVM regularization has over the simpler, but closely related, ISE-classifier.

7. Conclusions

We have shown that Parzen window-based estimators for the quadratic information measures are equivalent to Mercer kernel feature space measures, which can be expressed as functions of *mean values* in the Mercer kernel feature space. The Mercer kernel and the Parzen window are shown to be equivalent. This implies that Parzen window size selection procedures known from statistics can potentially be incorporated into Mercer kernel-based methods in order to determine a proper data-driven Mercer kernel size.



(a) Training data

(b) ISE classification

(c) SVM classification

Figure 4. Classification of data set with outliers using both the ISE classifier and the SVM.

This also means that the problems associated with applying the Parzen window technique on high dimensional data sets may also be problematic for some Mercer kernel-based methods. Note that these equivalences cannot be obtained using Parzen window-based estimators for the Shannon measures.

We have analyzed a classification rule based on the ISE measure, combined with Parzen windowing. The resulting classifier was shown to have a dual interpretation as a hyperplane classifier in a Mercer kernel feature space. By introducing weighted Parzen window estimators, we formulated the SVM classifier as a closely related enhancement of the ISE classifier. Thus, the ISE-classification rule is to some extent equivalent to the SVM classifier.

In our information theoretic framework, the SVM weights, which are related to weighted Parzen window density estimators, are determined by minimizing the ISE between the class densities. In future work, perhaps some other criteria could be used to learn proper weights. Preferably, such alternative methods should be easier to implement than the SVM optimization. In particular, we will investigate whether the Cauchy–Schwarz pdf divergence measure could be more advantageous in some respect than the integrated squared error criterion for this purpose.

Acknowledgments

This work was partially supported by NSF grant ECS-0300340. Deniz Erdogmus was with the Computational NeuroEngineering Laboratory during this work. Robert Jenssen would like to express gratitude to the University of Tromsø, for granting a research scholarship for the academic year 2002/2003, and a two-month research scholarship in the spring of 2004, for visiting the Computational NeuroEngineering Laboratory at the University of Florida.

Appendix: Using Non-Gaussian Mercer Kernels

In this Appendix, we will examine Parzen window-based estimator of $\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}$, using non-Gaussian Mercer kernels.

First, note that

$$\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x} = E_p\{q(\mathbf{x})\}, \tag{59}$$

where $E_p\{\cdot\}$ denotes expectation with respect to the density $p(\mathbf{x})$.

The expectation operator may be *approximated* based on the available samples, as follows

$$E_p\{q(\mathbf{x})\} \approx \frac{1}{N_1}\sum_{i=1}^{N_1} q(\mathbf{x}_i). \tag{60}$$

Assume now that

$$\hat{q}(\mathbf{x}) = \frac{1}{N_2}\sum_{j=1}^{N_2} k(\mathbf{x}, \mathbf{x}_j), \tag{61}$$

where $k(\mathbf{x}, \mathbf{x}_j)$ is a non-Gaussian Mercer/Parzen kernel. Eq. (59) can now be approximated by

$$\begin{aligned}\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x} &\approx \frac{1}{N_1}\sum_{i=1}^{N_1}\hat{q}(\mathbf{x}_i) \\ &= \frac{1}{N_1}\sum_{i=1}^{N_1}\frac{1}{N_2}\sum_{j=1}^{N_2} k(\mathbf{x}_i, \mathbf{x}_j) \\ &= \frac{1}{N_1 N_2}\sum_{i,j=1}^{N_1,N_2} k(\mathbf{x}_i, \mathbf{x}_j).\end{aligned} \tag{62}$$

Hence, the same result is obtained as in the case where Gaussian Parzen kernels were used. However, in this case, it required an additional approximation with regard to the expectation operator. Of course, the same reasoning can be used to approximate the quantities $\int p^2(\mathbf{x})d\mathbf{x}$ and $\int q^2(\mathbf{x})d\mathbf{x}$.

Notes

1. $\sum_{t,t'=1}^{N,N}$ equals the double summation $\sum_{t=1}^{N}\sum_{t'=1}^{N}$.
2. In the following, the index i always points to data drawn from $p(\mathbf{x})$, j always points to data drawn from $q(\mathbf{x})$, and t always points to data drawn from $f(\mathbf{x})$.

References

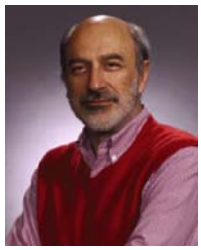
1. J. Shawe-Taylor and N. Cristianini, "Kernel Methods for Pattern Analysis," Cambridge University Press, 2004.
2. K.R. Müller, S. Mika, G. Rätsch, K. Tsuda and B. Schölkopf, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, 2001, pp. 181–201.
3. F. Perez-Cruz and O. Bousquet, "Kernel Methods and Their Potential Use in Signal Processing," *IEEE Signal Process. Mag.*, 2004, pp. 57–65, May.
4. B. Schölkopf and A.J. Smola, "*Learning with Kernels*," MIT, Cambridge, 2002.
5. C. Cortes and V.N. Vapnik, "Support Vector Networks," *Mach. Learn.*, vol. 20, 1995, pp. 273–297.
6. V.N. Vapnik, "*The Nature of Statistical Learning Theory*," Springer, Berlin Heidelberg New York, 1995.
7. N. Cristianini and J. Shawe-Taylor, "*An Introduction to Support Vector Machines*," Cambridge University Press, Cambridge, 2000.
8. C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Knowledge Discovery and Data Mining*, vol. 2, no. 2, 1998, pp. 121–167.
9. T. Hastie, S. Rosset, R. Tibshirani and J. Zhu, "The Entire Regularization Path for the Support Vector Machine," *J. Mach. Learn. Res.*, vol. 5, 2004, pp. 1391–1415.
10. B. Schölkopf, A.J. Smola and K.R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Comput.*, vol. 10, 1998, pp. 1299–1319.
11. S. Mika, G. Rätsch, J. Weston, B. Schölkopf and K.R. Müller, "Fisher Discriminant Analysis with Kernels," in *Proceedings of IEEE International Workshop on Neural Networks for Signal Processing*, Madison, USA, August 23–25, 1999, pp. 41–48.

12. V. Roth and V. Steinhage, "Nonlinear Discriminant Analysis using Kernel Functions," in *Advances in Neural Information Processing Systems 12*, MIT, Cambridge, 2000, pp. 568–574.
13. Y.A. LeCun, L.D. Jackel, L. Bottou, A. Brunot, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, U.A. Müller, E. Säckinger, P.Y. Simard and V.N. Vapnik, "Learning Algorithms for Classification: A Comparison on Handwritten Digit Reconstruction," *Neural Netw.*, 1995, pp. 261–276.
14. K.R. Müller, A.J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen and V.N. Vapnik, "Predicting Time Series with Support Vector Machines," in *Proceedings of International Conference on Artificial Neural Networks—Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 1997, vol. 1327, pp. 999–1004.
15. A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer and K.R. Müller, "Engineering Support Vector Machine Kernels that Recognize Translation Invariant Sites in DNA," *Bioinformatics*, vol. 16, 2000, pp. 906–914.
16. J. Principe, D. Xu and J. Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, S. Haykin (Ed.), Wiley, New York, 2000, vol. I, Chapter 7.
17. J.C. Principe, D. Xu, Q. Zhao and J.W. Fisher, "Learning From Examples with Information Theoretic Criteria," *J. VLSI Signal Process.*, vol. 26, no. 1, 2000, pp. 61–77.
18. S. Haykin, (Ed.), "*Unsupervised Adaptive Filtering: Volume 1, Blind Source Separation*, Wiley, New York, 2000.
19. E. Parzen, "On the Estimation of a Probability Density Function and the Mode," *Ann. Math. Stat.*, vol. 32, 1962, pp. 1065–1076.
20. L. Devroye, "On Random Variate Generation when only Moments or Fourier Coefficients are known," *Math. Comput. Simul.*, vol. 31, 1989, pp. 71–89.
21. B.W. Silverman, "*Density Estimation for Statistics and Data Analysis*," Chapman & Hall, London, 1986.
22. D.W. Scott, "*Multivariate Density Estimation*, " Wiley, New York, 1992.
23. M.P. Wand and M.C. Jones, "*Kernel Smooting*, " Chapman & Hall, London, 1995.
24. P.A. Viola, N.N. Schraudolph and T.J. Sejnowski, "Empirical Entropy Manipulation for Real-World Problems," in *Advances in Neural Information Processing Systems, 8*, MIT, Cambridge, 1995, pp. 851–857.
25. P. Viola and W.M. Wells, "Alignment by Maximization of Mutual Information," *Int. J. Comput. Vis.*, vol. 24, no. 2, 1997, pp. 137–154.
26. D. Xu, "*Energy, Entropy and Information Potential for Neural Computation*, Ph.D. thesis, University of Florida, Gainesville, FL, USA, 1999.
27. A. Renyi, "Some Fundamental Questions of Information Theory," *Selected Papers of Alfred Renyi, Akademiai Kiado, Budapest*, vol. 2, 1976, pp. 526–552.
28. A. Renyi, "On Measures of Entropy and Information," *Selected Papers of Alfred Renyi, Akademiai Kiado, Budapest*, vol. 2, 1976, pp. 565–580.
29. M. Lazaro, I. Santamaria, D. Erdogmus, K.E. Hild II, C. Pantaleon and J.C. Principe, "Stochastic Blind Equalization Based on PDF Fitting using Parzen Estimator," *IEEE Trans. Signal Process.*, vol. 53, no. 2, 2005, pp. 696–704.
30. D. Erdogmus, K.E. Hild, Y.N. Rao and J.C. Principe, "Minimax Mutual Information Approach for Independent Component Analysis," *Neural Comput.*, vol. 16, 2004, pp. 1235–1252.
31. D. Erdogmus, K.E. Hild, J.C. Principe, M. Lazaro and I. Santamaria, "Adaptive Blind Deconvolution of Linear Channels using Renyi's Entropy with Parzen Window Estimation," *IEEE Trans. Signal Process.*, vol. 52, no. 6, 2004, pp. 1489–1498.
32. D. Erdogmus and J.C. Principe, "Convergence Properties and Data Efficiency of the Minimum Error-Entropy Criterion in Adaline Training," *IEEE Trans. Signal Process.*, vol. 51, no. 7, 2003, pp. 1966–1978.
33. D. Erdogmus, K.E. Hild and J.C. Principe, "Blind Source Separation using Renyi's α-Marginal Entropies," *Neurocomputing*, vol. 49, 2002, pp. 25–38.
34. I. Santamaria, D. Erdogmus and J.C. Principe, "Entropy Minimization for Supervised Digital Communications Channel Equalization," *IEEE Trans. Signal Process.*, vol. 50, no. 5, 2002, pp. 1184–1192.
35. D. Erdogmus and J.C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, 2002, pp. 1035–1044.
36. D. Erdogmus and J.C. Principe, "An Error-Entropy Minimization Algorithm for Supervised Training of Nonlinear Adaptive Systems," *IEEE Trans. Signal Process.*, vol. 50, no. 7, 2002, pp. 1780–1786.
37. J. Mercer, "Functions of Positive and Negative Type and their Connection with the Theory of Integral Equations," *Philos. Trans. Roy. Soc. London*, vol. A, 1909, pp. 415–446.
38. M. Girolami, "Mercer Kernel-Based Clustering in Feature Space," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, 2002, pp. 780–784.
39. I.S. Dhillon, Y. Guan and B. Kulis, "Kernel K-means, Spectral Clustering and Normalized Cuts," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, USA, August 22–25, 2004, pp. 551–556.
40. L. Devroye and G. Lugosi, "*Combinatorial Methods in Density Estimation*," Springer, Berlin Heidelberg New York, 2001.
41. J.H. Friedman, "On Bias, Variance, 0/1 Loss, and the Curse-Of-Dimensionality," *Data Mining and Knowledge Discovery*, vol. 1, no. 1, 1997, pp. 55–77.
42. M. Girolami, "Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem," *Neural Comput.*, vol. 14, no. 3, 2002, pp. 669–688.
43. D.W. Scott, "Parametric Statistical Modeling by Integrated Squared Error," *Technometrics*, vol. 43, 2001, pp. 274–285.
44. J.N. Kapur, "*Measures of Information and their Applications*," Wiley, New York, 1994.
45. R. Jenssen, J.C. Principe and T. Eltoft, "Information Cut and Information Forces for Clustering," in *Proceedings of IEEE International Workshop on Neural Networks for Signal Processing*, Toulouse, France, September 17–19, 2003, pp. 459–468.
46. M. Di Marzio and C.C. Taylor, "Kernel Density Classification and Boosting: An L_2 Analysis," *Stat. Comput.*, vol. 15, no. 2, 2005, pp. 113–123.

Robert Jenssen received the M.S. and Ph.D. in Electrical Engineering (EE), in 2001 and 2005, respectively, from the University of Tromsø, Norway, focusing on an information theoretic approach to machine learning, including kernel methods, spectral clustering and independent component analysis. Jenssen spent the academic year 2002/2003 and March/April 2004 at the University of Florida, as a visitor at the Computational NeuroEngineering Laboratory. Jenssen is holding an Associate Professor position in EE at the University of Tromsø. Dr. Jenssen received "Honorable Mention for the 2003 Pattern Recognition Journal Best Paper Award," and the "2005 IEEE ICASSP Outstanding Student Paper Award."

Torbjørn Eltoft received the Cand. Real. (M.S.) and Dr. Scient. (Ph.D.) degrees from the University of Tromsø, Norway, in 1981 and 1984, respectively. His early research was on the application of modern signal processing techniques in experimental ionospheric physics. Since 1984, he has been working with remote sensing, with a special interest in the nonlinear SAR imaging of ocean waves and the scattering of microwaves from the ocean surface. He joined the Faculty of Physics, University of Tromsø, in 1988, where he is currently a Professor with the group of Electrical Engineering. His current research interests include remote sensing, image and signal processing, and artificial neural networks. Dr. Eltoft was awarded the "Year 2000 Outstanding Paper Award in Neural Networks" by the IEEE Neural Networks Council and "Honorable Mention for the 2003 Pattern Recognition Journal Best Paper Award."

Deniz Erdogmus received the B.S. in Electrical & Electronics Engineering (EEE) and the B.S. in Mathematics, both in 1997, and the M.S. in EEE in 1999 from the Middle East Technical University, Turkey. He received his Ph.D. in Electrical & Computer Engineering from the University of Florida (UF) in 2002. He worked as a research engineer at TUBITAK-SAGE, Turkey from 1997 to 1999, focusing on the design of navigation, guidance, and flight control systems. He was also a research assistant and a postdoctoral research associate at UF from 1999 to 2004, concentrating on signal processing, adaptive systems, machine learning, and information theory, specifically with applications in biomedical engineering including brain machine interfaces. Currently, he is holding an Assistant Professor position jointly at the Computer Science and Electrical Engineering Department and the Biomedical Engineering Department of the Oregon Health and Science University. His reseach focuses on information theoretic adaptive signal processing and its applications to biomedical signal processing problems. Dr. Erdogmus has over 35 articles in international scientific journals and numerous conference papers and book chapters. He has also served as associate editor and guest editor for various journals, participated in various conference organization and scientific committees, and he is a member of Tau Beta Pi, Eta Kappa Nu, IEEE, and IEE.



Jose C. Principe is Distinguished Professor of Electrical and Biomedical Engineering at the University of Florida since 2002. He joined the University of Florida in 1987, after an eight-year appointment as Professor at the University of Aveiro, in

Portugal. Dr. Principe holds degrees in electrical engineering from the University of Porto (Bachelors), Portugal, University of Florida (Master and Ph.D.), USA, and a Laurea Honoris Causa degree from the Universita Mediterranea in Reggio Calabria, Italy. Dr. Principe interests lie in nonlinear non-Gaussian optimal signal processing and modeling and in biomedical engineering. He created, in 1991, the Computational NeuroEngineering Laboratory to synergistically focus the research in biological information processing models. Dr. Principe is a Fellow of the IEEE, past President of the International Neural Network Society, editor in chief of the Transactions of Biomedical Engineering since 2001, and a former member of the Advisory Science Board of the FDA. He holds five patents and has submitted seven more. Dr. Principe was supervisory committee chair of 47 Ph.D. and 61 Master students, and he is the author of more than 400 refereed publications (3 books, 4 edited books, 14 book chapters, 116 journal papers, and 276 conference proceedings).