A Family of Probabilistic Kernels Based on Information Divergence

Antoni B. Chan, Nuno Vasconcelos, and Pedro J. Moreno

SVCL-TR 2004/01
June 2004

A Family of Probabilistic Kernels Based on Information Divergence

Antoni B. Chan[1], Nuno Vasconcelos[1], and Pedro J. Moreno[2]

[1] Statistical Visual Computing Lab
Department of Electrical and Computer Engineering
University of California, San Diego
abchan@ucsd.edu nuno@ece.ucsd.edu
[2] Google Inc, 1440 Broadway, New York, NY 10018
pmoreno@gmail.com

June 2004

Abstract

Probabilistic kernels offer a way to combine generative models with discriminative classifiers. We establish connections between probabilistic kernels and feature space kernels through a geometric interpretation of the previously proposed probability product kernel. A family of probabilistic kernels, based on information divergence measures, is then introduced and its connections to various existing probabilistic kernels are analyzed. The new family is shown to provide a unifying framework for the study of various important questions in kernel theory and practice. We exploit this property to design a set of experiments that yield interesting results regarding the role of properties such as linearity, positive definiteness, and the triangle inequality in kernel performance.

Author email: abchan@ucsd.edu

©University of California San Diego, 2004

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of the Statistical Visual Computing Laboratory of the University of California, San Diego; an acknowledgment of the authors and individual contributors to the work; and all applicable portions of the copyright notice. Copying, reproducing, or republishing for any other purpose shall require a license with payment of fee to the University of California, San Diego. All rights reserved.

SVCL Technical reports are available on the SVCL's web page at
http://www.svcl.ucsd.edu

University of California, San Diego
Statistical Visual Computing Laboratory
9500 Gilman Drive, Mail code 0407
EBU 1, Room 5512
La Jolla, CA 92093-0407

1 Introduction

There are many problems, such as speech processing or computer vision, where it is natural to adopt a localized data representation [1]. This consists of representing each observation (e.g. image) as a collection (or bag) of low-dimensional feature vectors (e.g. small image patches or their projection into some linear feature space). This collection is in turn most naturally represented by its probability density function, producing a compact representation that depends on a small number of parameters. Each observation is thus mapped into a probability density function defined on the space of localized feature vectors. Combining a localized representation (or generative models, in general) with discriminative classifiers has shown promise in several classification problems involving proteins [2], audio and speech [3], and images [1]. While this framework has the appeal of combining the generalization guarantees of discriminative learning with the invariance and robustness to outliers induced by the localized representation, its implementation is not trivial. In the domain of kernel-based classifiers, such as support vector machines (SVMs), one source of difficulty is that most commonly used kernels (e.g. polynomial and Gaussian) are not defined on the space of probability distributions. This limitation has motivated the introduction of various probabilistic kernels, including the *Fisher* [2], *TOP* [4], *probability product* [5], and *Kullback-Leibler* [3] kernels, in the recent years.

Different probabilistic kernels can have very different properties and behaviors. One distinctive feature is the existence, or not, of a closed-form expression for the kernel function in terms of the parameters of the generative models on which it acts. The existence of such closed-form solutions is computationally appealing since, in their absence, the kernel function must be evaluated by expensive Monte Carlo methods. While some probabilistic kernels have closed-form solution for most conceivable generative models [5], others can be solved in closed form for certain probability families, e.g. the exponential family, but not for others, e.g. mixture models [1]. At this point, the existence of widespread closed-form solutions appears to be linked to the linearity of the kernel, i.e. the possibility of interpreting the kernel as a linear kernel in the space of probability distributions. It is not clear, however, whether this restriction to linearity has itself a limiting impact on the performance of the resulting classifiers.

A second distinctive feature is the relationship to traditional kernels, such as the popular Gaussian kernel, that are monotonic functions of a metric (e.g. the Euclidean distance for the Gaussian kernel). Such kernels, which we refer to as *metric kernels*, are of particular interest because various probabilistic kernels are defined likewise, but rely on similarity functions that are not metrics, e.g. the symmetric Kullback-Leibler (KL) divergence. At the fundamental level, the only apparent difference between the metric kernels and these probabilistic kernels is that the similarity functions of the latter do not obey the triangle inequality. It is currently not well understood what the role of the triangle inequality plays in even the most elementary kernel properties, e.g. the positive definiteness of the kernel.

This leads us to the third distinctive feature, which is the existence of a proof of positive definiteness for each kernel. While some kernels can be trivially shown to be positive definite (PD) [2, 5], such a proof appears very non-trivial for most. Although positive definiteness is a central concept in kernel theory (in the sense that it enables the

interpretation of the kernel as a dot product in a transform space) a PD kernel does not always seem to be a necessary condition for a SVM to achieve good performance. In some cases, experimental evidence reveals that a small amount of negative definiteness (i.e. a small percentage of negative eigenvalues of small amplitude) does not hurt performance, and may in fact lead to better classifiers than those achievable with provably PD probabilistic kernels. This observation appears to reach beyond the realm of probabilistic kernels [6]. Theoretically, it has been shown that a geometric interpretation of the SVM optimization problem is possible even in the absence of positive definiteness [7]: SVM optimization using a non-PD kernel is equivalent to optimal separation of convex hulls in a pseudo-Euclidean space. In practice, SVM-training is frequently implemented with procedures, such as the SMO algorithm, that only consider 2x2 kernel matrices. For these matrices, most probabilistic kernels can be shown to be PD (and certainly all that we discuss in this work), therefore guaranteeing training convergence, albeit possibly not to the globally optimal solution [6].

The ability to make progress with respect to these open questions seems constrained by the inexistence of a common framework in which they can be addressed in a unified form. To address this problem, we introduce a new family of probabilistic kernels. This family is a direct generalization of the KL kernel, consisting of the set of kernels that are negative exponents of a generic information divergence. It is an interesting family in the sense that 1) it has direct connections to many existing probabilistic kernels, and 2) it contains, as special cases, kernels that exhibit most types of behavior discussed above. For example, while not guaranteed to be PD in general, all kernels in this family can be made PD by appropriate choice of parameters. We consider in detail two cases of special interest, the Rényi and Jensen-Shannon kernels, which are shown to have close connections with the probability product kernel (PPK) and the metric kernels, respectively. These connections enable new insights regarding some of the fundamental questions discussed above (e.g. a new interpretation of kernel nonlinearity as a transformation to a new space where densities are subject to different amounts of smoothing) and establish a unified framework for addressing the others. We explore this framework by designing a set of experiments that produce new, and sometimes surprising, evidence regarding the role of the kernel positive definiteness and linearity, as well as the role of triangle inequality.

2 Probabilistic kernels and SVMs

Given a feature space $\mathcal{X}$, a set of training examples $\{x_1, ..., x_N\} \in \mathcal{X}$, and a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a SVM is a discriminative classifier that constructs a maximum-margin hyperplane in a transformed feature space, defined by the function K, known as the kernel. One interpretation of the kernel function $K(x_i, x_j)$ is that it measures the similarity between two points x_i and x_j in the feature space $\mathcal{X}$. A popular example is the Gaussian kernel, defined as $K_g(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, which is an example of a metric kernel based on the negative exponent of the square of a metric.

Under the localized data representation, each example is reduced to a probability density function (defined on a space of localized feature vectors) and the kernel becomes a measure of similarity between probability distributions. A probabilistic kernel

is thus defined as a mapping $K : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$, where $\mathcal{P}$ is the space of probability distributions. One example is the *probability product kernel* (PPK) [5]

$$K_\rho(p,q) = \int_\Omega p(x)^\rho q(x)^\rho dx \tag{1}$$

where ρ is a parameter and $p(x)$ and $q(x)$ are probability distributions defined on the space Ω. In [5], the PPK is proven to be trivially positive definite. The PPK has two interesting special cases. The first, obtained with $\rho = 1/2$, is the *Bhattacharyya kernel*

$$K_{\frac{1}{2}}(p,q) = \int_\Omega \sqrt{p(x)}\sqrt{q(x)}dx \tag{2}$$

The second, obtained with $\rho = 1$, is the *expected likelihood kernel* (or *correlation kernel*), which measures the correlation between the two distributions,

$$K_1(p,q) = \int_\Omega p(x)q(x)dx. \tag{3}$$

While the PPK has a closed-form solution for distributions in the exponential family [5], a closed-form solution for mixture distributions only exists when $\rho = 1$ (the linear case). Hence, the correlation kernel has the appeal that it is computationally efficient on mixture distributions. This is an important attribute since many models of interest can be seen as mixtures or mixture extensions [5]. However, the correlation kernel suffers from two serious problems: 1) when the dimensionality of Ω is large, the kernel values are badly scaled; and 2) it is unnormalized in the sense that the auto-correlation of a probability distribution can take a wide range of values[1]. These two problems frequently lead to kernel matrices that are ill-conditioned.

The limitations of the correlation kernel can be avoided by normalizing the correlation function, leading to the *normalized correlation kernel*,

$$K_N(p,q) = \frac{\int_\Omega p(x)q(x)dx}{\sqrt{\int_\Omega p(x)^2dx}\sqrt{\int_\Omega q(x)^2dx}} \tag{4}$$

which can be shown to be positive definite with a proof similar to that of [5]. This kernel has a closed-form solution for mixture distributions whenever the same holds for the corresponding mixture components. Noting that $K_1(p,q)$ is itself a valid inner product between functions, i.e. $K_1 = \langle p(x), q(x)\rangle_\Omega$, where

$$\langle p(x), q(x)\rangle_\Omega = \int_\Omega p(x)q(x)dx, \tag{5}$$

the correlation kernel can be seen as the extension (to the space of probability distributions) of the linear kernel, $K_L(x_i, x_j) = \langle x_i, x_j\rangle = x_i^T x_j$, where $x_i, x_j \in \mathbb{R}^n$. This enables an interesting geometric interpretation of the PPK and the normalized correlation kernel. In particular, by rewriting

$$K_\rho(p,q) = \langle \Phi[p(x)], \Phi[q(x)]\rangle_\Omega \tag{6}$$

[1]For example, the maximum and minimum values of auto-correlation for the data set used in this study was $3.7346\text{x}10^{-04}$ and $3.0314\text{x}10^{-96}$ for a feature space of 64 dimensions.

where $\Phi[p(x)] = p(x)^\rho$, the former can be interpreted as a linear kernel in a transformed feature space. The feature transformation, $\Phi[p(x)]$, is a smoothing operation on $p(x)$ controlled by the parameter ρ. The function $p(x)$ is unsmoothed when $\rho = 1$, and as ρ decreases, $p(x)$ becomes more smoothed, eventually becoming constant when $\rho = 0$. For the normalized correlation kernel we have

$$K_N(p,q) = \frac{\langle p(x), q(x)\rangle_\Omega}{\|p(x)\|_\Omega \|q(x)\|_\Omega} = \cos(p,q), \tag{7}$$

i.e. the kernel computes the cosine of the angle between $p(x)$ and $q(x)$ in the inner product space of (5). Due to their connection to the standard linear kernel we refer to the correlation and normalized correlation kernels as *linear probabilistic kernels*.

3 Kernels based on divergence measures

Under the interpretation of the kernel as a measure of similarity, it is possible to define kernels based on information divergences, which are measures of dissimilarity between probability distributions. One common measure is the the Kullback-Leibler (KL) divergence

$$KL(p||q) = \int_\Omega p(x) \log \frac{p(x)}{q(x)} dx. \tag{8}$$

This is a non-negative function, equal to zero when $p(x) = q(x)$. The *Kullback-Leibler kernel* [3] is obtained by exponentiating the symmetric KL divergence,

$$K_{KL}(p,q) = e^{-a(KL(p||q)+KL(q||p))} \tag{9}$$

where $a > 0$ is a kernel parameter akin to the variance of the standard Gaussian kernel. More generally, it is possible to create a family of kernels by exponentiating any divergence between two probability densities,

$$K(p,q) = e^{-aF(p||q)} \tag{10}$$

where $F(p||q)$ is a non-negative function, equal to zero if and only if $p(x) = q(x)$, and symmetric in the arguments $p(x)$ and $q(x)$. Note that, when $G(p||q) = \sqrt{F(p||q)}$ obeys the triangle inequality (i.e. $G(p||q) \leq G(p||r) + G(r||q)$), F is the square of a metric, and the kernel is a metric kernel. In the remainder of this section we consider two information divergence kernels that establish connections with the PPK and the family of metric kernels, and address the issue of positive definiteness.

3.1 Rényi kernel

The Rényi divergence [8] is an alternative divergence measure based on a relaxed set of the information postulates that define the Shannon entropy. The Rényi divergence of order-α is defined as

$$D_\alpha(p||q) = \frac{1}{\alpha - 1} \log \int_\Omega p(x)^\alpha q(x)^{1-\alpha} dx \tag{11}$$

where $\alpha > 0$ and $\alpha \neq 1$. The Rényi divergence is a generalization of the KL divergence, and it can be shown that they are equal as $\alpha \to 1$ [8]. The *Rényi kernel* is obtained by exponentiating the symmetric Rényi divergence,

$$K_D(p,q) = e^{-a(D_\alpha(p||q)+D_\alpha(q||p))} \tag{12}$$

$$= \left[\int_\Omega p(x)^\alpha q(x)^{1-\alpha}dx \int_\Omega p(x)^{1-\alpha}q(x)^\alpha dx\right]^{\frac{a}{1-\alpha}}. \tag{13}$$

Setting $a = \frac{1-\alpha}{2}$ leads to a form similar to the PPK

$$K_D(p,q) = \sqrt{\int_\Omega p(x)^\alpha q(x)^{1-\alpha}dx \int_\Omega p(x)^{1-\alpha}q(x)^\alpha dx}. \tag{14}$$

It is clear from this form that, for $\alpha = 1/2$, the Rényi kernel is the Bhattacharyya kernel, which is a special case of the PPK.

The role of α in the Rényi kernel is similar to that of ρ in the PPK. Both parameters control the amount of smoothing applied to the probability densities. In the case of the PPK, both $p(x)$ and $q(x)$ receive the same amount of smoothing. This constraint is relaxed by the Rényi kernel which supports a different amount of smoothing for each density. Since, in the first integral of (14), $p(x)$ is smoothed by α and $q(x)$ by $1-\alpha$, for small α the integral is the correlation of a smoothed $p(x)$ and an unsmoothed (or slightly smoothed) $q(x)$. The second integral of (14) reverses the roles of the two densities, therefore ensuring the symmetry of the kernel. The Rényi kernel is the geometric mean of the two correlations.

3.2 Jensen-Shannon kernel

The Jensen-Shannon (JS) divergence [9] is a measurement of whether two samples, defined by their empirical distributions, are drawn from the same source distribution. The JS divergence is defined as

$$JS(p||q) = H[\beta p(x) + (1-\beta)q(x)] - \beta H[p(x)] - (1-\beta)H[q(x)] \tag{15}$$

where β is a parameter and $H[p(x)] = -\int_\Omega p(x)\log p(x)dx$ is the Shannon entropy of $p(x)$. Substituting for H and setting $\beta = 1/2$, the JS divergence becomes

$$JS(p||q) = \frac{1}{2}KL(p||r) + \frac{1}{2}KL(q||r) \tag{16}$$

where $r(x) = \frac{1}{2}p(x) + \frac{1}{2}q(x)$. This divergence can be interpreted as the average distance (in the KL sense) between each probability distribution and the average distribution, or equivalently as the *diversity* [10] of two distributions with equal priors. Using the fact that H is a concave function, it can be shown that the JS divergence is non-negative and equal to zero when $p(x) = q(x)$. Exponentiating this divergence measure leads to the *Jensen-Shannon kernel*

$$K_{JS}(p,q) = e^{-aJS(p||q)}. \tag{17}$$

It is proven in [11] that (16) is the square of a metric from which it follows that the JS kernel is a metric kernel. Furthermore, (16) is a negative definite kernel [12], and by applying simple properties of positive definite kernels [13], the JS kernel is positive definite.

3.3 Positive definiteness

There is currently little understanding regarding the positive definiteness (PD) of the KL and Rényi kernels. Nevertheless, one aspect that makes the family of information divergence kernels interesting is that they can always be made PD by appropriate choice of parameters. In particular, as long as no two densities in the training set are exactly alike (zero divergence), it is always possible to make the kernel matrix diagonally dominant [14], and therefore PD, by making a sufficiently large. In the extreme, as $a \to \infty$, the kernel matrix approaches the identity matrix. Interestingly, the kernel matrix is also guaranteed to be PD as $a \to 0$, since all the entries converge to 1. Clearly, none of these PD extremes is desirable since they imply making all the examples alike or making each example similar only to itself. Nevertheless, they illustrate how forcing a kernel to be PD can reduce its expressiveness, and hurt classification accuracy. In practice, the positive definiteness of a kernel can usually be checked by evaluating the positive definiteness of the kernel matrices obtained with particular data sets. For probabilistic kernels this is not always true since, in the absence of closed-form solutions, the non positive definiteness of a kernel matrix can result from inaccurate approximations[2]. With regards to the kernels discussed above, we have found that the KL and Rényi kernels can produce matrices with negative eigenvalues, albeit usually only a few and of much smaller amplitude than their positive counterparts.

4 Experiments and results

We used the COREL database to experiment with the different probabilistic kernels in an image classification task. The COREL database contains a variety of image classes, including landscapes, animals, underwater scenes, and structures. We selected 15 image classes from the database, each class containing 100 images, for a total of 1500 images. Each image was represented using a localized representation [1] by scanning each color channel with an 8x8 window shifted every 4 pixels. A feature vector was created from each 192-pixel window by computing the 64 lowest frequency coefficients of its discrete cosine transform (DCT). Finally, a mixture of 8 Gaussians of diagonal covariance was fit to the collection of DCT vectors extracted from the image.

4.1 Experiment setup

The image database was split with 80% of the images for training and the remaining 20% for testing. A multi-class SVM was constructed using the 1-v-1 MaxVotes

[2]Even though the Bhattacharyya kernel is provably PD, we have found it to sometimes produce kernel matrices with negative eigenvalues, when evaluated via Monte Carlo approximations.

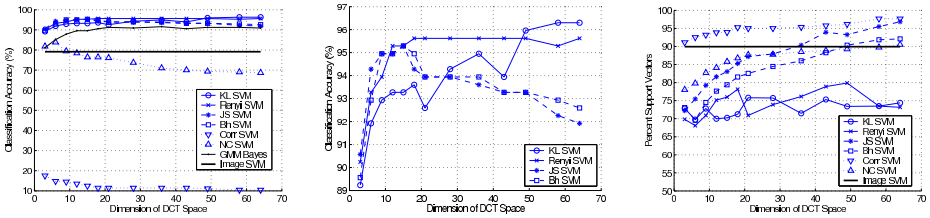
method [15]. Preliminary results showed it to outperform other multiclass SVM methods, such as the 1-v-1 DAG-SVM or the 1-v-all SVM [15]. The SVM was trained with six probabilistic kernels: the Kullback-Leibler (KL), Rényi, Jensen-Shannon (JS), Bhattacharyya (Bh), correlation (Corr), and normalized correlation (NC). Kernel parameters were selected by cross-validation, using 70% of the training set for training and the remaining 30% validation. Once the best parameters were found, the SVM was trained using all the training data. For the divergence-based kernels (KL, Rényi, and JS), the parameters were selected from $a \in \{2^{-10}, 2^{-9}, ..., 2^{4}\}$, and additionally for the Rényi kernel, $\alpha \in \{0.1, 0.2, ..., 0.8\}$. The C-SVM formulation was used, with $C \in \{2^{-2}, 2^{-1}, ..., 2^{12}\}$, and the SVMs were trained and tested using `libsvm` [16]. All probabilistic kernels without closed-form solution were evaluated using a Monte Carlo approximation with 10,000 points.

Two additional classifiers were trained as baseline comparisons. The first, a standard Bayesian classifier (GMM Bayes) was trained using the DCT features generated by the localized representation, modeling each class conditional density as a mixture of 32 Gaussians. The second was an SVM using the image pixels as feature vector (Image SVM). Each image was cropped and downsampled into an 88x54 thumbnail, and converted into a 4752-dimensional vector by concatenating the rows of the thumbnail. The SVM was trained with the Gaussian kernel with γ selected by cross-validation over $\gamma \in \{2^{-10}, 2^{-9}, ..., 2^{4}\}$. Finally, all experiments were repeated for a range of dimensions of the DCT space (i.e. the number of DCT coefficients used in the localized representation).

4.2 Experiment results

The results of classification accuracy versus the dimension of the DCT space are presented in Figures 1a and 1b. Clearly, normalizing the correlation kernel improves the classification performance significantly (improvement of about 60% in the worst case). Nonetheless, the non-linear probabilistic kernels (KL, Rényi, JS, and Bh) outperform the two linear probabilistic kernels for all dimensions. These results suggest that the linear feature space of probability distributions is not expressive enough for image classification. On the other hand, the non-linear probabilistic kernels induce a transformed feature space (e.g. the Bhattacharyya and Rényi kernels apply density smoothing) that appears to improve classification significantly. In the Bhattacharyya case, a direct geometric interpretation is possible: the smoothing transformation expands the support of each density, increasing the correlation with its neighbors. This improves the generalization of the kernel, by making it capable of capturing similarities between densities that are close but do not have extensive overlap.

The non-linear probabilistic kernels performed similarly, as seen in Figure 1b, (note that the vertical scale is different) but the Rényi and KL kernels appear to have slightly better performance. The fact that the JS kernel is a metric kernel does not seem to lead to any clear advantage in terms of classification accuracy. Interestingly, the performance of the two provably PD kernels (JS and Bh) drops slightly as the dimension of the DCT space increases. On the other hand, the performance of the two non-PD kernels (KL and Rényi) improves slightly. To further test the importance of positive definiteness we examined the performance of the KL kernel in more detail. The percent



(a) (b) (c)

Figure 1: Classification accuracy versus feature space dimension for (a) all the classifiers and (b) the non-linear probabilistic kernel SVMs, and (c) the percent of the training data used as support vectors.

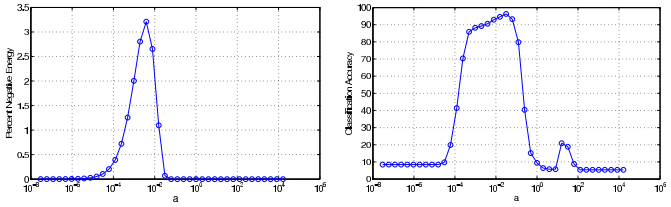


Figure 2: (left) Negative eigenvalue energy and (right) classification accuracy versus a for the KL kernel.

magnitude of negative eigenvalues was calculated for a range of values of a, and averaged over the kernel matrices used in training. For each value of a, the classification accuracy was measured, using $C = 1$. Figure 2 shows the plot of (left) the percent negative energy of the eigenvalues and (right) the classification accuracy versus a. Clearly, while the kernel matrix can be forced to be PD, the classification accuracy actually decreases. In fact, the best classifiers seem to occur when the kernel contains some amount of negative eigenvalue energy. These results support the conjecture that strict enforcement of properties such as positive definiteness or the satisfaction of the triangle inequality may not always lead to the best classifier performance.

The generalization capability of the SVM is known to be related to the number of support vectors used by the classifier. For example, Vapnik has shown (Theorem 5.2 of [17]) that the probability of error is bounded by the ratio of the number of support vectors to the number of training samples. Figure 1c shows the percentage of the training data used as support vectors by the different kernels. The number of support vectors used by the KL and Rényi kernels was less than the other kernels and stayed approximately constant with the dimension of the DCT space. This suggests that the generalization capability of the KL and Rényi kernels is better than the other probabilistic kernels, which is confirmed by the performance curves in Figures 1a and 1b.

Regarding the baseline methods, the non-linear probabilistic kernels outperformed both the image-based SVM and the GMM Bayes classifier. This was expected since, for the image-based SVM, the feature space is highly variant (in the sense that two spatially transformed versions of the same image may lie very far away in feature

space) and the hence classification problem is more difficult [1]. For GMM Bayes this was also expected, given the well known improved generalization ability of large-margin classifiers.

Overall, the experimental results support the following conclusions: 1) linear probabilistic kernels are not expressive enough for image classification; 2) non-linear probabilistic kernels, whether based on density smoothing or on information divergences, are better for this task; 3) strict kernel positive definiteness is not always required for good classification results; and 4) some negative eigenvalue energy does not hinder, and can even help, classification performance.

References

[1] N. Vasconcelos, P. Ho, and P. Moreno. The kullback-leibler kernel as a framework for discriminant and localized representations for visual recognition. In *European Conference on Computer Vision*, Prague, Czech, 2004.

[2] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, 1998.

[3] P. J. Moreno, P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec 2003.

[4] K. Tsuda, M. Kawanabe, G. Ratsch, S. Sonnenburg, and K. Muller. A new discriminative kernel from probabilistic models. *Neural Computation*, 14(10):2397–2414, 2002.

[5] T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *Proceedings of the Annual Conference on Computational Learning Theory and Kernel Workshop*, 2003.

[6] C. Bahlmann, B. Haasdonk, and H. Burkhardt. On-line handwriting recognition with support vector machines – a kernel approach. In *8th International Workshop on Frontiers in Handwriting Recognition*, pages 49–54, 2002.

[7] B. Haasdonk. Feature space interpretation of svms with non positive definite kernels. Technical report, IIF-LMB, University Freiburg, Germany, Oct 2003.

[8] A. Rényi. On measures of entropy and information. In *Berkeley Symposium on Mathematical Statistics and Probability, Proceedings v.1*, pages 547–561, 1960.

[9] J. Lin. Divergence measures based on shannon entropy. *IEEE Transactions on Information Theory*, 37(14):145–51, Jan 1991.

[10] N. Vasconcelos. Feature selection by maximum marginal diversity. In *Proceedings of Neural Information Processing Systems*, Vancouver, Canada, 2002.

[11] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–60, July 2003.

[12] F. Topsøe. Jensen-shannon divergence and norm-based measures of discrimination and variation. Technical report, Department of Mathematics, University of Copenhagen, 2003.

[13] M. G. Genton. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, Dec 2001.

[14] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[15] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–25, Mar 2002.

[16] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[17] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, 1995.

SVCL-TR 2004/01
June 2004

A Family of Probabilistic Kernels Based on Information Divergence

Antoni B. Chan, Nuno Vasconcelos, and Pedro J. Moreno