Information Theoretic Learning

*Deniz Erdogmus, *derdogmus@ieee.org*

Department of Computer Science and Electrical Engineering,
Oregon Health and Science University, USA

Jose C. Principe, *principe@cnel.ufl.edu*

Department of Electrical and Computer Engineering,
University of Florida, USA

INTRODUCTION

Learning systems depend on three interrelated components: topologies, cost/performance functions, and learning algorithms. Topologies provide the constraints for the mapping, and the learning algorithms offer the means to find an optimal solution; but the solution is optimal with respect to what? Optimality is characterized by the criterion and in neural network literature, this is the least addressed component, yet it has a decisive influence in generalization performance. Certainly, the assumptions behind the selection of a criterion should be better understood and investigated.

Traditionally, least squares has been the benchmark criterion for regression problems; considering classification as a regression problem towards estimating class posterior probabilities, least squares has been employed to train neural network and other classifier topologies to approximate correct labels. The main motivation to utilize least squares in regression simply comes from the intellectual comfort this criterion provides due to its success in traditional linear least squares regression applications – which can be reduced to solving a system of linear equations. For nonlinear regression, the assumption of Gaussianity for the measurement error combined with the maximum likelihood principle could be emphasized to promote this criterion. In nonparametric regression, least squares principle leads to the conditional expectation solution, which is intuitively appealing. Although these are good reasons to use the mean squared error as the cost, it is inherently linked to the assumptions and habits stated above. Consequently, there is information in the error signal that is not captured during the training of nonlinear adaptive systems under non-Gaussian distribution conditions when one insists on second-order statistical criteria. This argument extends to other linear-second-order techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), and canonical correlation analysis (CCA). Recent work tries to generalize these techniques to nonlinear scenarios by utilizing kernel techniques or other heuristics. This begs the question: *what other alternative cost functions could be used to train adaptive systems and how could we establish rigorous techniques for extending useful concepts from linear and second-order statistical techniques to nonlinear and higher-order statistical learning methodologies*?

BACKGROUND

This seemingly simple question is at the core of recent research on information theoretic learning (ITL) conducted by the authors, as well as research by others on

alternative optimality criteria for robustness to outliers and faster convergence, such as different L_p-norm induced error measures (Sayed, 2005), the epsilon-insensitive error measure (Scholkopf & Smola, 2001), Huber's robust m-estimation theory (Huber, 1981), or Bregman's divergence based modifications (Bregman, 1967). Entropy is an uncertainty measure that generalizes the role of variance in Gaussian distributions by including information about the higher-order statistics of the probability density function (pdf) (Shannon & Weaver, 1964; Fano, 1961; Renyi, 1970; Csiszár & Körner, 1981). For on-line learning, information theoretic quantities must be estimated nonparametrically from data. A nonparametric expression that is differentiable and easy to approximate stochastically will enable importing useful concepts such as stochastic gradient learning and backpropagation of errors. The natural choice is kernel density estimation (KDE) (Parzen, 1967), due its smoothness and asymptotic properties. The plug-in estimation methodology (Gyorfi & van der Meulen, 1990) combined with definitions of Renyi (Renyi, 1970), provides a set of tools that are well-tuned for learning applications – tools suitable for supervised and unsupervised, off-line and on-line learning. Renyi's definition of entropy for a random variable X is

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\int p^\alpha(x)dx \qquad (1)$$

This generalizes Shannon's linear additivity postulate to exponential additivity resulting in a parametric family. Dropping the logarithm for optimization simplifies algorithms. Specifically of interest is the quadratic entropy (α=2), because its sample estimator requires only one approximation (the density estimator itself) and an analytical expression for the integral can be obtained for kernel density estimates. Consequently, a sample estimator for quadratic entropy can be derived for Gaussian kernels of standard deviation σ on an iid sample set $\{x_1,\ldots,x_N\}$ as the sum of pairwise sample (particle) interactions (Principe et al, 2000):

$$\hat{H}_2(X) = -\log\left(\frac{1}{N^2}\sum_{i=i}^{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_i - x_j)\right). \qquad (2)$$

The pairwise interaction of samples through the kernel intriguingly provides a connection to entropy of particles in physics. Particles interacting trough *information forces* (as in the *N*-body problem in physics) can employ computational techniques developed for simulating such large scale systems. The use of entropy in training multilayer structures can be studied in the backpropagation of information forces framework (Erdogmus et al, 2002). The quadratic entropy estimator was employed in measuring divergences between probability densities and blind source separation (Hild et al, 2006), blind deconvolution (Lazaro et al, 2005), and clustering (Jenssen et al, 2006). Quadratic expressions with mutual-information-like properties were introduced based on the Euclidean and Cauchy-Schwartz distances (ED/CSD). These are advantageous with computational simplicity and statistical stability in optimization (Principe et al, 2000).

Following the conception of information potential and force and principles, the pairwise-interaction estimator is generalized to use arbitrary kernels and any order α of entropy. The stochastic information gradient (SIG) is developed (Erdogmus et al, 2003) to train adaptive systems with a complexity comparable to the LMS (least-mean-square) algorithm - essential for training complex systems with large data sets. Supervised and

unsupervised learning is unified under information-based criteria. Minimizing error entropy in supervised regression or maximizing output entropy for unsupervised learning (factor analysis), minimization of mutual information between the outputs of a system to achieve independent components or maximizing mutual information between the outputs and the desired responses to achieve optimal subspace projections in classification is possible. Systematic comparisons of ITL with conventional MSE in system identification verified the advantage of the technique for nonlinear system identification and blind equalization of communication channels. Relationships with instrumental variables techniques were discovered and led to the error-whitening criterion for unbiased linear system identification in noisy-input-output data conditions (Rao et al, 2005).

SOME IDEAS IN AND APPLICATIONS OF ITL

Kernel Machines and Spectral Clustering: KDE has been motivated by the smoothness properties inherent to reproducing kernel Hilbert spaces (RKHS). Therefore, a practical connection between KDE-based ITL, kernel machines, and spectral machine learning techniques was imminent. This connection was realized and exploited in recent work that demonstrates an information theoretic framework for pairwise similarity (spectral) clustering, especially normalized cut techniques (Shi & Malik, 2000). Normalized cut clustering is shown to determine an *optimal* solution that maximizes the CSD between clusters (Jenssen, 2004). This connection immediately allows one to approach kernel machines from a density estimation perspective, thus providing a robust method to select the *kernel size*, a problem still investigated by some researchers in the kernel and spectral techniques literature. In our experience, kernel size selection based on suitable criteria aimed at obtaining the *best* fit to the training data - using Silverman's regularized squared error fit (Silverman, 1986) or leave-one-out cross-validation maximum likelihood (Duin, 1976), for instance - has proved to be convenient, robust, and accurate techniques that avoid many of the computational complexity and load issues. Local data spread based modifications resulting in variable-width KDE are also observed to be more robust to noise and outliers.

An illustration of ITL clustering by maximizing the CSD between the two estimated clusters is provided in Figure 1. The samples are labeled to maximize

$$D_{CS}(p,q) = -\log \frac{< p,q >_f}{\| p \|_f \| q \|_f} \tag{3}$$

where p and q are KDE for two candidate clusters, f is the overall data KDE and the weighted inner product to measure angular distance between clusters is

$$< p,q >_f = \int p(x)q(x) f^{-1}(x)dx\,. \tag{4}$$

When estimated using a weighted KDE variant, this criterion becomes equivalently

$$D_{CS}(p,q) \approx \frac{\sum_{x_i \in p, y_j \in q} K_{1/f}(x_i, y_j)}{\sqrt{\sum_{x_i \in p, x_i \in p} K_{1/f}(x_i, x_j) \sum_{y_i \in q, y_j \in q} K_{1/f}(y_i, y_j)}} \tag{5}$$

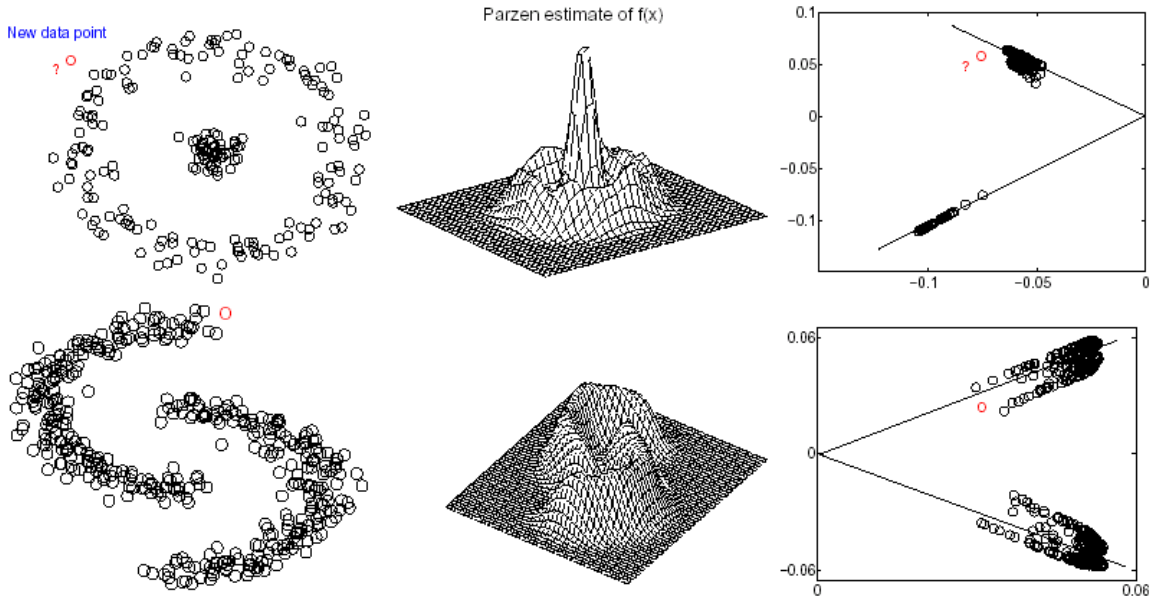

Figure 1. Maximum CSD clustering of two synthetic benchmarks: training and novel test data (left), KDE using Gaussian kernels with Silverman-kernel-size (center), and spectral projections of data on two dominant eigenfunctions of the kernel. The eigenfunctions are approximated using the Nystrom formula.

where $K_{1/f}$ is an equivalent kernel generated from the original kernel K (Gaussian here). One difficulty with kernel machines is their nonparametric nature, the requirement to solve for the eigendecomposition of a large positive-definite matrix that has size $N \times N$, for N training samples. The solution is a weighted sum of kernels evaluated over each training sample, thus the test procedure for each novel sample involves evaluating the sum of N kernels: $y_{test} = \sum_{k=1}^{N} w_k K(x_{test} - x_k)$. The Fast Gauss Transform (FGT) (Greengard, 1991), which uses the polynomial expansions for a Gaussian (or other) kernel has been employed to overcome this difficulty. FGT carefully selects few center points around which truncated Hermite polynomial expansions approximate the kernel machine. FGT still requires heavy computational load in off-line training (minimum $O(N^2)$, typically $O(N^3)$). The selection of expansion centers is typically done via clustering (e.g., Ozertem & Erdogmus, 2006).

Correntopy as a Generalized Similarity Metric: The main feature of ITL is that it preserves the universe of concepts we have in neural computing, but allows the adaptive system to *extract more information* from the data. For instance, the general Hebbian principle is reduced into a second order metric in traditional artificial neural network literature (input-output product), thus becoming a synonym for second order statistics. The learning rule that maximizes output entropy (instead of output variance), using SIG with Gaussian kernels is $\Delta w(n) = \eta(x(n) - x(n-1))(y(n) - y(n-1))$ (Erdogmus et al, 2002), which still obeys the Hebbian principle, yet extracts more information from the data (leading to the error-whitening criterion for input-noise robust learning).

ITL quantifies global properties of the data, but will it be possible to apply it to functions, specifically those in RKHS? A concrete example is on similarity between

random variables, which is typically expressed as second order correlation. Correntropy generalizes similarity to include higher order moment information. The name indicates the strong relation to correlation, but also stresses the difference – the average over the lags (for random processes) or over dimensions (for multidimensional random variables) is the information potential, i.e. the argument of second order Renyi's entropy. For random variables X and Y with joint density $p(x,y)$, correntropy is defined as

$$V(X,Y) = \int\int \delta(x-y)\, p(x,y)\, dxdy \tag{6}$$

and measures how dense the two random variables are along the line $x=y$ in the joint space. Notice that it is similar to correlation, which also asks the same question in a second moment framework. However, correntropy is local to the line $x=y$, while correlation is quadratically dependent upon distances of samples in the joint space. Using a KDE with Gaussian kernels

$$V(X,Y) = \frac{1}{N}\sum_{i=1}^{N} G(x_i - y_i)\,. \tag{7}$$

Correntropy is a positive-definite function, thus defines a RKHS. Unlike correlation, RKHS is nonlinearly related to the input, because all moments of the random variable are included in the transformation. It is possible to analytically solve for least squares regression and principal components in this space, yielding nonlinear fits in input space. Correntopy induced metric (CIM) behaves as the L_2-norm for small distances and progressively approaches the L_1-norm and then converges to L_0 at infinity. Thus robustness to outliers is automatically achieved and equivalence to Huber's robust estimation can be proven (Santamaria, 2006). Unlike conventional kernel methods, correntropy solutions remain in the same dimensionality as the input vector. This might indicate built-in regularization properties, yet to be explored.

Nonparametric Learning in the RKHS: It is possible to obtain robust solutions to a variety of problems in learning using the nonparametric and local nature of KDE and its relationship with RKHS theory. Recently, we explored the possibility of designing nonparametric solutions to the problem of identifying nonlinear dimensionality reduction schemes that maintain maximal discriminative information in a pattern recognition problem (quite appropriately measured by the mutual information between the data and the class labels as agreed upon by many researchers). Using the RKHS formalism and based on the KDE, results were obtained that consistently outperformed the alternative rather heuristic kernel approaches such as kernel PCA and kernel LDA (Scholkopf & Smola, 2001). The conceptual oversight in the latter two is that, both PCA and LDA procedures are most appropriate for Gaussian distributed data (although acceptable for other symmetric unimodal distributions and are commonly but possibly inappropriately used for arbitrary data distributions).

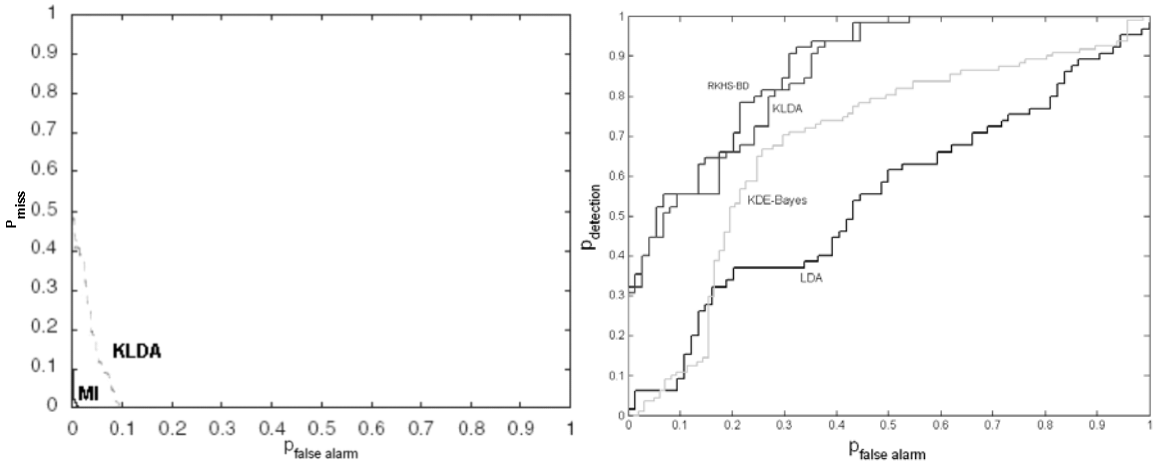

Figure 2. Maximum mutual information projection versus kernel LDA test ROC results on hand-written digit recognition shown in terms of type-1 and type-2 errors (left); ROC results (P_{detect} vs P_{false}) compared for various techniques on sonar data. Both data are from the UCI Machine Learning Repository (2007).

Clearly, the distribution of the data in the kernel induced feature space could not be Gaussian for all typically exploited kernel selections (such as the Gaussian kernel), since these are usually translation invariant, therefore the data is, in principle, mapped to an infinite dimensional hypersphere on which the data could not have been Gaussian distributed (nor symmetrically distributed in general for the ideal kernel for a given problem since these are positive definite functions). Consequently, the hasty use of kernel extensions of second-order techniques is not necessarily optimal in a meaningful statistical sense. Nevertheless, these techniques have found *successful* applications in various problems; however, their suboptimality is clear from comparisons with more carefully designed solutions. In order to illustrate how drastic the performance difference could be, we present a comparison of a mutual information based nonlinear nonparametric projection approach (Ozertem et al, 2006) and kernel LDA in a simplified two-class handwritten digit classification case study and sonar mine detection case study. The ROC curves of both algorithms on the test set after being trained with the same data is shown in Figure 2. The kernel is assumed to be a circular Gaussian with size set to Silverman's rule-of-thumb. For the sonar data, we also include KDE-based approximate Bayes classifier and linear LDA for reference. In this example, KLDA performs close to mutual information projections, as observed occasionally.

FUTURE TRENDS

Nonparametric Snakes, Principal Curves and Surfaces: More recently, we have been investigating the application of KDE and RKHS to nonparametric clustering, principal curves and surfaces. Interesting mean-shift-like fixed-point algorithms have been obtained; specifically interesting is the concepts of *nonparametric snakes* (Ozertem & Erdogmus, 2007) and *local principal manifolds* (Erdogmus & Ozertem, 2007) that we developed recently. The nonparametric snake approach overcomes the principal difficulties experienced by snakes (active contours) for image segmentation, such as low capture range, data curvature inhomogeneity, and noisy and missing edge information.

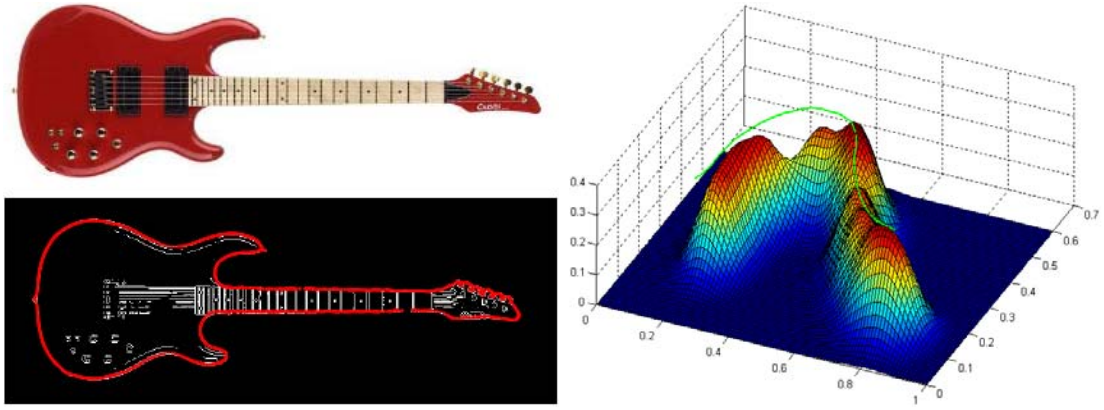

Figure 3. Nonparametric snake after convergence from an initial state that was located at the boundary of the guitar image rectangle (left). The global principal curve of a mixture of ten Gaussians obtained according to the local subspace maximum definition for principal manifolds (right).

Similarly, the local conditions for determining whether a point is in a principal manifold or not provide guidelines for designing fixed point and other iterative learning algorithms for identifying such important structures.

Specifically in nonparametric snakes, we treat the edgemap of an image as samples and the values of the edginess as weights to construct a weighted KDE, from which, a fixed point iterative algorithm can be devised to detect the boundaries of an object in background. The designed algorithm can be easily made robust to outlier edges, converges very fast, and can penetrate into concavities, while not being trapped into the object at missing edge localities. The guitar image in Figure 3 emphasizes these advantages as the image exhibits both missing edges and concavities, while background complexity is trivially low as that was not the main concern in this experiment – the variable width KDE easily avoids textured obstacles. The algorithm could be utilized to detect the ridge-boundary of a structure in any dimensional data set in other applications.

In defining principal manifolds, we avoided the traditional least-squares error reconstruction type criteria, such as Hastie's self-consistent principal curves (Hastie & Stuetzle, 1989), and proposed a local subspace maximum definition for principal manifolds inspired by differential geometry. This definition lends itself to a uniquely defined principal manifold hierarchy such that one can use inflation and deflation to obtain a d-dimensional principal manifold from a $(d+1)$-dimensional principal manifold. The rigorous and local definition lends itself to easy algorithm design and multiscale principal structure analysis for probability densities. We believe that in the near future, the community will be able to prove maximal information preserving properties of principal manifolds obtained using this definition in a manner similar to mean-shift clustering solving for minimum information distortion clustering (Rao et al, 2006) and maximum likelihood modelling achieving minimum Kullback-Leibler divergence asymptotically (Carreira-Perpinan & Williams, 2003; Erdogmus & Principe, 2006).

CONCLUSION

The use of information theoretic learning criteria in neural networks and other adaptive system solutions have so far clearly demonstrated a number of advantages that

arise due to the increased information content of these measures relative to second-order statistics (Erdogmus & Principe, 2006). Furthermore, the use of kernel density estimation with smooth kernels allows one to obtain continuous and differentiable criteria suitable for iterative descent/ascent-based learning and the nonparametric nature of KDE and its variants (such as variable-size kernels) allow one to achieve simultaneously robustness, global optimization through *kernel annealing*, and data modeling flexibility in designing neural networks and learning algorithms for a variety of benchmark problems. Due to lack of space, detailed mathematical treatments cannot be provided in this article; the reader is referred to the literature for details.

REFERENCES

Bregman, L.M., (1967). The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming. *USSR Computational Mathematics and Physics*, (7), 200-217.

Carreira-Perpinan, M.A., Williams, C.K.I., (2003). On the Number of Modes of a Gaussian Mixture. *Proceedings of Scale-Space Methods in Computer Vision*. 625-640.

Csiszár, I., Körner, J. (1981). *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press.

Duin, R.P.W., On the Choice of Smoothing Parameter for Parzen Estimators of Probability Density Functions. *IEEE Transactions on Computers*, (25) 1175-1179.

Erdogmus, D., Ozertem, U., (2007). Self-Consistent Locally Defined Principal Surfaces. *Proceedings of ICASSP 2007*. to appear.

Erdogmus, D., Principe, J.C., From Linear Adaptive Filtering to Nonlinear Information Processing. *IEEE Signal Processing Magazine*, (23) 6, 14-33.

Erdogmus, D., Principe, J.C., Hild II, K.E., (2002). Do Hebbian Synapses Estimate Entropy? *Proceedings of NNSP'02*, 199-208.

Erdogmus, D., Principe, J.C., Hild II, K.E., (2003). On-Line Entropy Manipulation: Stochastic Information Gradient. *IEEE Signal Processing Letters*, (10) 8, 242-245.

Erdogmus, D., Principe, J.C., Vielva, L. Luengo, D., (2002). Potential Energy and Particle Interaction Approach for Learning in Adaptive Systems. *Proceedings of ICANN'02*, 456-461.

Fano, R.M. (1961). *Transmission of Information: A Statistical Theory of Communications*, MIT Press.

Greengard, L., Strain, J., (1991). The Fast Gauss Transform. *SIAM Journal of Scientific and Statistical Computation*, (12) 1, 79–94.

Gyorfi, L., van der Meulen, E.C. (1990). On Nonparametric Estimation of Entropy Functionals. *Nonparametric Functional Estimation and Related Topics*, (G. Roussas, ed.), Kluwer Academic Publisher, 81-95.

Hastie, T., Stuetzle, W., (1989). Principal Curves. *Journal of the American Statistical Association*, (84) 406, 502-516.

Hild II, K.E., Erdogmus, D., Principe, J.C., (2006). An Analysis of Entropy Estimators for Blind Source Separation. *Signal Processing*, (86) 1, 182-194.

Huber, P.J., (1981). *Robust Statistics*. Wiley.

Jenssen, R., Erdogmus, D., Principe, J.C., Eltoft, T., (2004). The Laplacian PDF Distance: A Cost Function for Clustering in a Kernel Feature Space. *Advances in NIPS'04*, 625-632.

Jenssen, R., Erdogmus, D., Principe, J.C., Eltoft, T., (2006). Some Equivalences Between Kernel Methods and Information Theoretic Methods. *JVLSI Signal Processing Systems*, (45) 1-2, 49-65.

Lazaro, M., Santamaria, I., Erdogmus, D., Hild II, K.E., Pantaleon, C., Principe, J.C., (2005). Stochastic Blind Equalization Based on PDF Fitting Using Parzen Estimator. *IEEE Transactions on Signal Processing*, (53) 2, 696-704.

Ozertem, U., Erdogmus, D., (2006). Maximum Entropy Approximation for Kernel Machines. *Proceedings of MLSP 2005*.

Ozertem, U., Erdogmus, D., Jenssen, R., (2006). Spectral Feature Projections that Maximize Shannon Mutual Information with Class Labels. *Pattern Recognition*, (39) 7, 1241-1252.

Ozertem, U., Erdogmus, D., (2007). A Nonparametric Approach for Active Contours. *Proceedings of IJCNN 2007*, to appear.

Parzen, E., (1967). On Estimation of a Probability Density Function and Mode. *Time Series Analysis Papers*, Holden-Day, Inc.

Principe, J.C., Fisher, J.W., Xu, D., (2000). Information Theoretic Learning. *Unsupervised Adaptive Filtering*, (S. Haykin, ed.), Wiley, 265-319.

Rao, Y.N., Erdogmus, D., Principe, J.C., (2005). Error Whitening Criterion for Adaptive Filtering: Theory and Algorithms. *IEEE Transactions on Signal Processing*, (53) 3, 1057-1069.

Rao, S., de Madeiros Martins, A., Liu, W., Principe, J.C., (2006). Information Theoretic Mean Shift Algorithm. *Proceedings of MLSP 2006*.

Renyi, A., (1970). *Probability Theory*, North-Holland Publishing Company.

Sayed, A.H. (2005). *Fundamentals of Adaptive Filtering*. Wiley & IEEE Press.

Scholkopf, B., Smola, A.J. (2001). *Learning with Kernels*. MIT Press.

Shannon, C.E., Weaver, W. (1964). *The Mathematical Theory of Communication*, University of Illinois Press.

Shi, J., Malik, J., (2000). Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (22) 8, 888-905.

Silverman, B.W., (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall.

Santamaria, I., Pokharel, P.P., Principe, J.C., (2006). Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization. *IEEE Transactions on Signal Processing*, (54) 6, 2187-2197.

UCI Machine Learning Repository (2007). *http://mlearn.ics.uci.edu/MLRepository.html*. last accessed in June 2007.

TERMS AND DEFINITIONS

Cauchy-Schwartz Distance: An angular density distance measure in the Euclidean space of probability density functions that approximates information theoretic divergences for nearby densities.

Correntropy: A statistical measure that estimates the similarity between two or more random variables by integrating the joint probability density function along the main diagonal of the vector space (line along ones). It relates to Renyi's entropy when averaged over sample-index lags.

Information Theoretic Learning: A technique that employs information theoretic optimality criteria such as entropy, divergence, and mutual information for learning and adaptation.

Information Potentials and Forces: Physically intuitive pairwise particle interaction rules that emerge from information theoretic learning criteria and govern the learning process, including backpropagation in multilayer system adaptation.

Kernel Density Estimate: A nonparametric technique for probability density function estimation.

Mutual Information Projections: Maximally discriminative nonlinear nonparametric projections for feature dimensionality reduction based on the reproducing kernel Hilbert space theory.

Renyi Entropy: A generalized definition of entropy that stems from modifying the additivity postulate and results in a class of information theoretic measures that contain Shannon's definitions as special cases.

Stochastic Information Gradient: Stochastic gradient of nonparametric entropy estimate based on kernel density estimation.