A Kernel Based Approach to Maximum Entropy Mappings

John W. Fisher III[1]
EECS Dept.
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
Email fisher@ai.mit.edu

Jose C. Principe
ECE Dept.
University of Florida
Gainesville, FL 32611 USA
Email principe@cnel.ufl.edu

Hsiao-Chun Wu
ECE Dept.
University of Florida
Gainesville, FL 32611 USA
Email wu@cnel.ufl.edu

Abstract — **We discuss a kernel based method for learning maximum entropy mappings from exemplars. Information theoretic signal processing has been examined by many authors. The method presented here is related to the approaches of Linsker [1, 2], Bell and Sejnowski [3], and Viola et al [4]. In this paper we discuss the use of this method for deriving maximum entropy mappings in an unsupervised fashion. Extensions to optimizing mutual information are possible. The result of our approach is that maximizing and minimizing entropy for differentiable nonlinear mappings such as a multi-layer perceptron can be accomplished through simple local interactions of the data in the output space. We present empirical results from application of the method to the problem of blind separation of linearly mixed speech sources. We compare our empirical results to the method of Bell and Sejnowski.**

I. Introduction

We discuss a machine learning method for deriving statistically independent features [5, 6, 7]. Our approach is motivated by Linsker's Principle of Information Maximization, which seeks to transfer maximum information about a signal from the input to the output of a mapping, as the criterion for feature extraction [1, 2]. As a result, we seek parameters of a general (differentiable) nonlinear mapping such that the mutual information between the observed output and the signal of interest is maximized.

The method is novel in that entropy is maximized by manipulation of the samples observed at the *output* of differentiable mapping with finite range (i.e. a multi-layer perceptron). As a consequence of the approach, implicit error signals are computed which fit directly into the backpropagation formalism. Furthermore, the technique is extensible to a multilayer perceptrons with an arbitrary number of hidden layers in contrast to previous techniques.

II. Maximum Entropy via an Indirect Measure

Due to the fact that the differentiable mapping used has finite range at the output, the optimal solution for maximum entropy is known (i.e. a uniform distribution). As an indirect measure of entropy the integrated squared error between an estimate of the density at the output and the uniform distribution is used as our learning criterion.

$$J = \int_{\Omega_Y} (f_U(u) - \hat{f}_Y(u, \{y\}))^2 du$$

where Ω_Y is the region of support at the output of the mapping, $f_U(u)$ is the uniform density function, and $\hat{f}_Y(u, \{y\})$ is the estimated density function over the set of data points $\{y\}$ observed at the output of the mapping. It can be shown that integrated squared error criterion is equivalent to a second order Taylor series of Shannon's differentiable entropy expanded about the uniform density.

III. Maximum (Minimum) Entropy as a Local Interaction of the Data

We have shown that using the Parzen window density estimator (with Gaussian kernels) leads to a simple error direction term which is computed via the local interaction between data points and the boundary at the output [6, 7]. The local interaction is either attraction (minimizing entropy) or repulsion (maximizing entropy).

IV. Empirical Results: Application to Blind Source Separation

We have applied the technique above to the problem of blind source separation of instantaneous and linearly mixed speech sources. Our results were compared to the technique of Bell and Sejnowski [3]. In the experiment the kernel-based approach achieved signal-to-noise ratios of approximately 34 dB as compared to 25 dB for that of Bell and Sejnowski.

[1]This work was supported by AFOSR MURI through Boston University GC123919NGD.

References

[1] R. Linsker, "Self-organization in a perceptual system", *Computater*, vol. 21, pp. 105–117, 1988.

[2] R. Linsker, "How to generate ordered maps by maximizing the mutual information between input and output signals", *Neural Computation*, vol. 1, pp. 402–411, 1990.

[3] A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution", *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[4] P. Viola, N. Schraudolph, and T. Sejnowski, "Empirical entropy manipulation for real-world problems", in *Neural Information Processing Systems*, M. Hasselmo D. Touretzky, M. Mozer, Ed., 1995, vol. 8, pp. 851–857.

[5] J.W. Fisher and J.C. Principe, "Unsupervised learning for nonlinear synthetic discriminant functions", in *Proc. SPIE, Optical Pattern Recognition VII*, D. Casasent and T. Chao, Eds., 1996, vol. 2752, pp. 2–13.

[6] J.W. Fisher and J.C. Principe, "Entropy manipulation of arbitrary nonlinear mappings", in *Proc. IEEE Workshop, Neural Networks for Signal Processing VII*, J.C. Principe, Ed., 1997, pp. 14–23.

[7] J.W. Fisher and J.C. Principe, "A methodology for information theoretic feature extraction", in *Proceedings of the IEEE International Joint Conference on Neural Networks*, A. Stuberud, Ed., 1998, pp. ?–?