Information Theoretic Learning

Deniz Erdogmus
Northeastern University, USA

Jose C. Principe
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

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

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):

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

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