Learning Stochastic Concepts Efficiently*

Snowbird '94 submission

J. B. Hampshire II
Jet Propulsion Laboratory, 238-420
California Institute of Technology, 4800 Oak Grove Drive
Pasadena, CA 91109-8099
hamps@bvd.jpl.nasa.gov

December 23, 1993

Abstract

I define the efficiency of a supervised learning strategy in classical estimation theoretic terms. Thus, the efficient strategy is much like the efficient parameter estimator defined by Gauss, R. A. Fisher, Rao [5], and Cramer [1]. Proceeding from this basis, there is a simple, unifying relationship between classical "parametric" and connectionist "non-parametric" approaches to statistical pattern recognition[1]; the learning strategy emerges as a key factor determining whether or not the classifier will generalize well for small and/or large training sample sizes.

I described two fundamental strategies for supervised learning: the *probabilistic* strategy seeks to learn class (or concept) probabilities by optimizing a likelihood function or an error measure objective function (i.e., empirical risk measure); the *differential* strategy is discriminative and seeks only to identify the most likely class by optimizing a classification figure-of-merit (CFM) objective function [3]. CFM objective functions are best described as differentiable approximations to a counting function: they count the number of correct classifications (or, equivalently, the number of incorrect classifications) the classifier makes on the training sample.

If the model chosen for the training data is a "proper parametric model" the most efficient learning strategy is likely to be one that is both probabilistic and maximum-likelihood in nature. If, however, the parametric model is an "improper" one, all probabilistic learning strategies prove to be inefficient for both small *and* large training sample sizes. Regardless of whether or not the model is proper, differential learning proves to be asymptotically efficient, guaranteeing the best generalization allowed by the model as long as the training sample size is sufficiently large. Moreover, differential learning requires the least complex model of the data (under a variety of complexity measures) necessary for Bayes-optimal classification, implying best generalization under distribution free VC analysis [6].

These arguments are supported by rigorous proofs from both estimation-theoretic and information-theoretic perspectives [2, part I]; they are also illustrated in a series of real-world pattern recognition experiments [2, part II] (attendees will be invited to explore these experiments interactively on the computer). They lead me to conclude that probabilistic learning is the strategy of choice when a proper model of the training data can be determined (an obvious conclusion), whereas differential learning is the strategy of choice when a proper model of the training data cannot be determined (perhaps not so obvious a conclusion).

I close by discussing the implications of this research for practical autonomous learning machines in light of Kolmogorov's theorem [4], which (arguably) can be interpreted to mean that finding the proper parametric model for a set of stochastic concepts is either pretty easy or really hard — there is no middle ground.

TOPIC: Learning Rules and Generalization **PREFERENCE:** Poster

References

[1] H. Cramer. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1946

[2] J. B. Hampshire II. *A Differential Theory of Learning for Efficient Statistical Pattern Recognition*. PhD thesis, Carnegie Mellon University, Department of Electrical & Computer Engineering, Hammerschlag Hall, Pittsburgh, PA 15213-3890, September 1993.

[3] J. B. Hampshire II and A. H. Waibel. A Novel Objective Function for Improved Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Neural Networks*, 1(2):216-228, June 1990

[4] A. N. Kolmogorov. Three Approaches to the Quantitative Definition of Information *Problems of Information Transmission*, 1(1):1--7, Jan. - Mar. 1965. Faraday Press translation of Problemy Peredachi Informatsii

[5] C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81--91, 1945.

[6] V. N. Vapnik and A. YA. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to their Probabilities. *Theory of Probability and its Applications*, XVI(2):264--280, 1971.

*Most of this research was done in collaboration with Prof. B. V. K. Vijaya Kumar, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213 3890. The collaboration was funded by the Air Force Office of Scientific Research under grant (AFOSR-89-0551). The author's research is currently funded by the National Aeronautics and Space Administration via the Jet Propulsion Laboratory's Office of Telecommunications and Data Acquisition, RTOP-310 3(172.

[1] Indeed, I view both classical and connectionist classifier paradigms as parametric ones, which are either "proper" or "improper" probabilistic models of the data