Abstract

Title of Dissertation: Model Selection and Universal Data Compression

Sanjeev Khudanpur, Doctor of Philosophy, 1997

Dissertation directed by: Professor Prakash Narayan
Electrical Engineering Department and
Institute for Systems Research

Statistical models play a vital role in the analysis of virtually every engineering problem and are central, for instance, to the design of communication systems. At the same time they are mathematical abstractions of our understanding of the physical phenomena they represent and, but for rare exceptions, selection of a particular statistical model for an application is a matter of choice. Two conflicting criteria drive the choice of a statistical model: consistency of the model with past observations of the process, and the ease of using the model in design and analysis. The former leans in favor of complex models that "explain" every idiosyncracy of the observed phenomena, while the latter calls for simple models that are computationally tractable.

In this dissertation, we first consider two model selection problems. In the first, a hierarchical family of parametric models, namely hidden Markov sources, is considered. The hierarchy is in terms of an increasing number of states of the underlying Markov chain. Models with an equal number of states can each be described by the same number of parameters and are interpreted to have the same degree of complexity. We define in a standard manner a criterion for selecting the optimal model size from this family. We then study the problem of identifying this model size from observed data. We propose an estimator of the number of states of a hidden Markov source which we show is strongly consistent. Our main result here is a lower bound on the

error in this estimation problem for a large class of estimators. We also show that for several interesting cases, this bound is achieved by our estimator of model size which is based on a maximum likelihood criterion. One such special case is the order estimation of finite Markov sources, while another is the order estimation of a class of renewal processes. In the second model selection problem, we consider a variation of the minimum information divergence criterion for selection of a particular model from a nonparametric class of models.

Sometimes, there are situations when a statistical model based on data has limited utility because some one-time action is to be performed on the given data. Compression of a given sequence is one such example. In the second part of this dissertation, we consider the lossy compression of an individual sequence and characterize the optimal compressibility in a manner similar to the distortion rate characterization of probabilistic sources.