Stochastic Complexity in Learning

J. Rissanen

IBM Almaden Research Center
San Jose, CA 95120-6099

Abstract: This is an expository paper on the latest results in the theory of stochastic complexity and the associated MDL principle with special interest in modeling problems arising in machine learning. As an illustration we discuss the problem of designing MDL decision trees, which are meant to improve the earlier designs in two ways: First, by use of the sharper formula for the stochastic complexity at the nodes the earlier found tendency of getting too small trees appears to be overcome. Secondly, a dynamic programming based pruning algorithm is described for finding the optimal trees, which generalizes an algorithm described in Nohre (1994).

1. Introduction

By learning from data one generally means the process of gaining knowledge about or understanding of the mechanism that generates the data, the 'go of it', as expressed by Maxwell. This can be done by use of 'models', which serve as the language in which the constraints predicated to the data can be described. In a definite sense an ultimate model of data is the shortest program in a universal programming language that generates the data, the length of which defines the algorithmic complexity of the data, Solomonoff (1964), Kolmogorov (1965), Chaitin (1969) (related work Chaitin (1966)), also called the Kolmogorov complexity, although the notion was introduced by Solomonoff in a clear and unambiguous manner; see Li and Vitanyi (1993) for a comprehensive discussion of the fascinating algorithmic theory of information and the pio-
neers. However, any hope of founding inductive inference on the algorithmic notion of information or, synonymously, complexity, is shattered by the noncomputability of the Kolmogorov complexity. Neither do the ingenious further constructs of semicomputable universal measures overcome the problem of inductive inference, which is inherently nonformalizable.

The idea of measuring the strength of constraints in terms of the code length with which data can be encoded by use of models, as suggested by Kolmogorov complexity, appeared to us as too good to be abandoned. In fact, we can avoid the noncomputability of Kolmogorov complexity either by restricting the permitted encoding operations or the model classes and weakening the requirement that the sought-for complexity is shortest for every string. Instead, we require it to be such that for practical purposes it cannot be beaten. How exactly the resulting notion of stochastic complexity, given a class of models, is to be defined has been a problem for which a satisfactory solution, in fact, in the form of a formula, has been found only recently, Rissanen (1994). The sense in which it is ‘practically’ unbeatable amounts to the following: If we imagine a long string being generated by any member in an uncountable class of models, each describing a random process, then both in the mean and almost surely the code length for the string with any code cannot be shorter than the stochastic complexity, except when the string is generated by models in a subset of measure zero. The subset in question depends on the code being used. Although this leaves open the possibility that, by a wild guess of the data generating model one could construct a better code, the chance of success will be nil.

The yardstick provided by the stochastic complexity with which different model classes are judged changes the way statistics can be done. Rather than trying to estimate the nonexistent ‘true’ data generating distribution, as the case is in traditional statistics with the unavoidable difficulties that can be overcome only by ad hoc means, the objective now is the sensible one of searching for better and better model classes. This is the MDL (Minimum Description Length) principle or, equivalently, a global maximum likelihood principle, global in the sense that any two model classes can be compared, whether or not they have the same number of