Journal of Machine Learning Research X (2002) X-X Submitted 4/2003; Published X/XX

CCIT Report #421 April 2003

Generalization Error Bounds for Bayesian Mixture Algorithms

Ron Meir RMEIR@EE.TECHNION.AC.IL
Department of Electrical Engineering
Technion, Haifa 32000, Israel

Tong Zhang TZHANG@WATSON.IBM.COM
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA

Editor: Unknown

Abstract

Bayesian approaches to learning and estimation have played a significant role in the Statistics literature over many years. While they are often provably optimal in a frequentist setting, and lead to excellent performance in practical applications, there have not been many precise characterizations of their performance for finite sample sizes under general conditions. In this paper we consider the class of Bayesian mixture algorithms, where an estimator is formed by constructing a data-dependent mixture over some hypothesis space. Similarly to what is observed in practice, our results demonstrate that mixture approaches are particularly robust, and allow for the construction of highly complex estimators, while avoiding undesirable overfitting effects. Our results, while being data-dependent in nature, are insensitive to the underlying model assumptions, and apply whether or not these hold. At a technical level, the approach applies to unbounded functions, constrained only by certain moment conditions. Finally, the bounds derived can be directly applied to non-Bayesian mixture approaches such as Boosting and Bagging.

1. Introduction and Motivation

The standard approach to Computational Learning Theory is usually formulated within the so-called frequentist approach to Statistics. Within this paradigm one is interested in constructing an estimator, based on a finite sample, which possesses a small loss (generalization error). While many algorithms have been constructed and analyzed within this context, it is not clear how these approaches relate to standard optimality criteria within the frequentist framework. Two classic optimality criteria within the latter approach are *minimaxity* and *admissibility*, which characterize optimality of estimators in a rigorous and precise fashion (Robert, 2001). Except for some special cases (e.g. Yang (1999)), it is not known whether any of the approaches used within the Machine Learning community lead to optimality in either of the above senses of the word. On the other hand, it is known that under certain regularity conditions, Bayesian estimators lead to either minimax or admissible estimators, and thus to well-defined optimality in the classical (frequentist) sense. In fact, it can be shown that Bayes estimators, or limits thereof, are essentially the only estimators which