Mach Learn (2012) 86:209–231
DOI 10.1007/s10994-011-5262-7

Linear classifiers are nearly optimal when hidden variables have diverse effects

Nader H. Bshouty · Philip M. Long

Received: 8 November 2010 / Accepted: 12 August 2011 / Published online: 5 September 2011

Abstract We analyze classification problems in which data is generated by a two-tiered random process. The class is generated first, then a layer of conditionally independent hidden variables, and finally the observed variables. For sources like this, the Bayes-optimal rule for predicting the class given the values of the observed variables is a two-layer neural network. We show that, if the hidden variables have non-negligible effects on many observed variables, a linear classifier approximates the error rate of the Bayes optimal classifier up to lower order terms. We also show that the hinge loss of a linear classifier is not much more than the Bayes error rate, which implies that an accurate linear classifier can be found efficiently.

Keywords Learning theory · Bayes-optimal · Linear classification · Hidden variables

1 Introduction

In many classification problems, groups of features are positively associated, even among examples of a given class. For example, when classifying news articles as to whether they are about sports or not, words about soccer tend to appear in the same articles. Similarly, diseases often coordinately affect the production rates of members of biomolecular pathways.

One way to model this phenomenon is to use a probability distribution with hidden variables (Hofmann 2001; Blei et al. 2003; Zhang 2004a; Papadimitriou et al. 2000; Langseth and Nielsen 2006). In one model of this type, the class designation directly and conditionally independently affects the hidden variables, each of which in turn drives a set

Editor: Nicolo Cesa-Bianchi.

N.H. Bshouty
Department of Computer Science, Technion, 32000 Haifa, Israel
e-mail: bshouty@cs.technion.ac.il

P.M. Long (✉)
Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
e-mail: plong@google.com