IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB8920

Chapter I

A Survey of Bayesian Data Mining

Stefan Arnborg
Royal Institute of Technology and
Swedish Institute of Computer Science, Sweden

ABSTRACT

This chapter reviews the fundamentals of inference, and gives a motivation for Bayesian analysis. The method is illustrated with dependency tests in data sets with categorical data variables, and the Dirichlet prior distributions. Principles and problems for deriving causality conclusions are reviewed, and illustrated with Simpson's paradox. The selection of decomposable and directed graphical models illustrates the Bayesian approach. Bayesian and EM classification is shortly described. The material is illustrated on two cases, one in personalization of media distribution, one in schizophrenia research. These cases are illustrations of how to approach problem types that exist in many other application areas.

INTRODUCTION

Data acquired for analysis can have many different forms. We will describe the analysis of data that can be thought of as samples drawn from a population, and the conclusions will be phrased as properties of this larger population. We will focus on very simple models. As the investigator's understanding of a problem area improves, the statistical models tend to become complex. Some examples of such areas are genetic linkage studies, ecosystem studies, and functional MRI investigations, where the signals extracted from measurements are very weak but potentially extremely useful for the application area. Experiments are typically analyzed using a combination of visualization, Bayesian analysis, and conventional test- and confidence-based statistics. In engineering and commercial applications of data

mining, the goal is not normally to arrive at eternal truths, but to support decisions in design and business. Nevertheless, because of the competitive nature of these activities, one can expect well-founded analytical methods and understandable models to provide more useful answers than ad hoc ones.

This text emphasizes characterization of data and the population from which it is drawn with its statistical properties. Nonetheless, the application owners have typically very different concerns: they want to understand; they want to be able to predict and ultimately to control their objects of study. This means that the statistical investigation is a first phase that must be accompanied by activities extracting meaning from the data. There is relatively little theory on these later activities, and it is probably fair to say that their outcome depends mostly on the intellectual climate of the team—of which the analyst is only one member.

Summary

Our goal is to explain some advantages of the Bayesian approach and to show how probability models can display the information or knowledge we are after in an application. We will see that, although many computations of Bayesian data-mining are straightforward, one soon reaches problems where difficult integrals have to be evaluated, and presently only Markov Chain Monte Carlo (MCMC) and expectation maximization (EM) methods are available. There are several recent books describing the Bayesian method from both a theoretical (Bernardo & Smith, 1994) and an application-oriented (Carlin & Louis, 1997) perspective. Particularly, Ed Jaynes' unfinished lecture notes, now available in (Jaynes, 2003) have provided inspiration for me and numerous students using them all over the world. A current survey of MCMC methods, which can solve many complex evaluations required in advanced Bayesian modeling, can be found in the book *Markov Chain Monte Carlo in Practice* (Gilks, Richardson, & Spiegelhalter 1996). Theory and use of graphical models have been explained by Lauritzen (1996) and Cox and Wermuth (1996). A tutorial on Bayesian network approaches to data mining is found in Heckerman (1997). We omit, for reasons of space availability, a discussion of linear and generalized linear models, which are described, e.g., by Hand, Mannila, and Smyth (2001). Another recent technique we omit is optimal recursive Bayesian estimation with particle filters, which is an important new application of MCMC (Doucet, de Freitas & Gordon 2001).

SCHOOLS OF STATISTICS

Statistical inference has a long history, and one should not assume that all scientists and engineers analyzing data have the same expertise and would reach the same type of conclusion using the objectively "right" method in the analysis of a given data set. Probability theory is the basis of statistics, and it links a probability model to an outcome. But this linking can be achieved by a number of different principles. A pure mathematician interested in *mathematical probability* would only consider abstract spaces equipped with a probability measure. Whatever is obtained by analyzing such mathematical structures has no immediate bearing on how we should interpret a data set collected to give us knowledge about the world. When it comes to inference about real-world phenomena, there are two different and complementary views on probability that have competed for the position of "the" statistical