Technical Report
LRDC-ONR-94-1
(1994)

Discrete factor analysis: Learning hidden variables in Bayesian networks.

Joel D. Martin & Kurt VanLehn
3939 O'Hara St.
University of Pittsburgh
Pittsburgh, PA, 15260
(412) 624-0843
martin@cs.pitt.edu

Keywords: Learning; Bayesian nets; Hidden variables; Fast, Heuristic method

Abstract

We present a space of Bayesian network topologies with hidden variables and present a method for rapidly constructing an appropriate such topology given sample data. In the topology, hidden variables can interact and as a result, can explain dependencies and independencies between observable variables. As well, the topology can permit polynomial time probabilistic inference. The learning method can be viewed as both a) factor analysis for discrete variables and b) cluster analysis with overlapping clusters (clumping).

1 Introduction

Even when faced with a flood of interacting variables, an intelligent agent can extract unseen structure that concisely explains its world. This paper presents a space of Bayesian network topologies with hidden variables (or factors) and a method for rapidly learning an appropriate topology from data. The learning method combines techniques from classical statistics and Bayesian motivated approaches to partitioning. It produces simple probabilistic models and can be viewed as both a) factor analysis for discrete variables and b) cluster analysis with overlapping clusters (clumping). Potential applications include creation of predictive domain models, simplification of intractable probabilistic models, and automated scientific discovery.

This approach has many advantages. The class of network topologies is simple and useful.

- It can support polynomial time probabilistic inference.
- Hidden variables can interact to influence observable variables.
- Observable variables are conditionally independent given the hidden variables.
- The resulting network captures conditional independencies among the observable variables.

In addition, the learning method presented here,

- Can approximate a distribution when a simple topology of factors is non-optimal for the data.
- Allows the hidden variables to have an arbitrary number of values.