0.1 Introduction

Bayesian Graphical models can be a useful tool for inference and knowledge representation. While not providing causal interpretations directly, they can be useful in developing causal relationships. These models became popular in the 90's when the message passing algorithm was developed that allowed fitting of large graphical models. In a graphical model the conditional probability distribution is represented as a graphical tree structure.

We investigate tools that will allow for non experts to interact with graphical models.

BayesDB provides a SQL like interface to graphical models using a non parametric approach to fitting the CPD.

"This paper introduces an approach for searching structured data based on probabilistic programming and nonparametric Bayes. Users specify queries in a probabilistic language that combines standard SQL database search operators with an information theoretic ranking function called predictive relevance. Predictive relevance can be calculated by a fast sparse matrix algorithm based on posterior samples from CrossCat, a nonparametric Bayesian model for high-dimensional, heterogeneously-typed data tables. The result is a flexible search technique that applies to a broad class of information retrieval problems, which we integrate into BayesDB, a probabilistic programming platform for probabilistic data analysis. This paper demonstrates applications to databases of US colleges, global macroeconomic indicators of public health, and classic cars. We found that human evaluators often prefer the results from probabilistic search to results from a standard baseline." **?**

CrossCat: A Fully Bayesian Nonparametric Method for Analyzing Heterogeneous, High Dimensional Data

"There is a widespread need for statistical methods that can analyze high-dimensional datasets with- out imposing restrictive or opaque modeling assumptions. This paper describes a domain-general data analysis method called CrossCat. CrossCat infers multiple non-overlapping views of the data, each consisting of a subset of the variables, and uses a separate nonparametric mixture to model each view. CrossCat is based on approximately Bayesian inference in a hierarchical, nonparamet- ric model for data tables. This model consists of a Dirichlet process mixture over the columns of a data table in which each mixture component is itself an independent Dirichlet process mixture over the rows; the inner mixture components are simple parametric models whose form depends on the types of data in the table. CrossCat combines strengths of mixture modeling and Bayesian net- work structure learning. Like mixture modeling, CrossCat can model a broad class of distributions by positing latent variables, and produces representations that can be efficiently conditioned and sampled from for prediction. Like Bayesian networks, CrossCat represents the dependencies and independencies between variables, and thus remains accurate when there are multiple statistical signals. Inference is done via a scalable Gibbs sampling scheme; this paper shows that it works well in practice. This paper also includes empirical results on heterogeneous tabular data of up to 10 million

cells, such as hospital cost and quality measures, voting records, unemployment rates, gene expression measurements, and images of handwritten digits. CrossCat infers structure that is consistent with accepted findings and common-sense knowledge in multiple domains and yields predictive accuracy competitive with generative, discriminative, and model-free alternatives."

?

"Most natural domains can be represented in multiple ways: we can categorize foods in terms of their nutritional content or social role, animals in terms of their taxonomic groupings or their ecological niches, and musical instruments in terms of their taxonomic categories or social uses. Previous approaches to modeling human categorization have largely ignored the problem of cross-categorization, focusing on learning just a single system of categories that explains all of the features. Cross-categorization presents a difficult problem: how can we infer categories without first knowing which features the categories are meant to explain? We present a novel model that suggests that human cross-categorization is a result of joint inference about multiple systems of categories and the features that they explain. We also formalize two commonly proposed alternative explanations for cross-categorization behavior: a features-first and an objects-first approach. The features-first approach suggests that cross-categorization is a consequence of attentional processes, where features are selected by an attentional mechanism first and categories are derived second. The objects-first approach suggests that cross-categorization is a consequence of repeated, sequential attempts to explain features, where categories are derived first, then features that are poorly explained are recategorized. We present two sets of simulations and experiments testing the models' predictions about human categorization. We find that an approach based on joint inference provides the best fit to human categorization behavior, and we suggest that a full account of human category learning will need to incorporate something akin to these capabilities."

?

Bibliography

Vikash K. Mansinghka, Patrick Shafto, Eric Jonas, Cap Petschulat, Max Gasner, and Joshua B. Tenenbaum. Crosscat: A fully bayesian nonparametric method for analyzing heterogeneous, high dimensional data. *CoRR*, abs/1512.01272, 2015. URL `http://arxiv.org/abs/1512.01272`.

Feras Saad, Leonardo Casarsa, and Vikash K. Mansinghka. Probabilistic search for structured data via probabilistic programming and nonparametric bayes. *CoRR*, abs/1704.01087, 2017. URL `http://arxiv.org/abs/1704.01087`.