STAT 598L
Learning Bayesian Network Structure

Sergey Kirshner

Department of Statistics
Purdue University
skirshne@purdue.edu

November 2, 2009

Acknowledgements: some of the slides were based on Luo Si's viewgraphs (http://www.cs.purdue.edu/homes/lsi/CS590M_Spring_08/CS590M.html)

Outline

Because often it is not known.

Density estimation: The joint distribution from the data can be used for prediction/inference.

Knowledge discovery: Dependence structure may shed light on relation between the variables in the domain.

Limitations:

- Identifiability: At best, we can recover the structure up to the I-equivalence class.
- Noise: Distribution often cannot be reconstructed perfectly from a relatively small and noisy data.
  - Some edges are “borderline”. Adding them may introduce spurious edges adding more free parameters.
  - Not adding them may lead to missing some dependence relations.
- Data is limited.
  - Factors with lots of parents are difficult to establish because of data fragmentation.
  - As a result, we often prefer sparser structures as they correspond to models with fewer parameters (simpler).

Because often it is not known.

Density estimation: The joint distribution from the data can be used for prediction/inference.

Knowledge discovery: Dependence structure may shed light on relation between the variables in the domain.

Limitations:

- Identifiability: At best, we can recover the structure up to the I-equivalence class.
- Noise: Distribution often cannot be reconstructed perfectly from a relatively small and noisy data.
  - Some edges are “borderline”. Adding them may introduce spurious edges adding more free parameters.
  - Not adding them may lead to missing some dependence relations.
- Data is limited.
  - Factors with lots of parents are difficult to establish because of data fragmentation.
  - As a result, we often prefer sparser structures as they correspond to models with fewer parameters (simpler).

Score-based structure learning

- Performing model selection.
- Pose an optimization problem: define the scoring (loss) function of the model's fit to the data.
- Search: find the structure $\mathcal{G}$ maximizing the scoring function.
- Limitation: the search space of graphs is super-exponential. For the general case, the problem is NP-hard.

Constraint-based structure learning

- Bayesian network represents dependencies.
- Want to find the network that represents the conditional independence relations represented in the data.
- Limitation: sensitive to failures in independence tests. True dependence model may not fall within the search space.

Approaches to Learning BN Structure (Bayesian Model Averaging)

Bayesian model averaging (BMA): treat the structure as a random variable and integrate it out. Alternatively, consider a distribution over structures (and/or parameters) and compute the posterior over structures.

- True dependence may not be represented by a BN (or MN) structure anyway.
- Because of the noise and finite sample size, lots of models have about the same score.
- In many tasks, we are after a distribution over variables and may not be interested in a structure. Integrating the structure out removes one source of uncertainty.
- In a Bayesian setting, we may be able to estimate the probabilities or confidence measures for having particular edge/structures.
- Limitation: The number of possible structures is huge ($2^{\mathcal{O}(n^2)}$). Averaging over them can be extremely difficult.

Constraint-Based Structure Learning

Goal: Want to find a minimal I-map satisfying the conditional independence relations in the data set.

- Fixing the order.
- For each node, looking for the minimal set of parents.
  - Dependence/independence is established using independence tests.

Limitations:

- Independence tests involve a large number of variables.
- Construction involves a large number of queries.
- Constructed network is sensitive to the chosen ordering, and the true ordering is unknown.

Looking for a P-map

Assumptions:

- Network has bounded in-degree d for each node.
- The independence test can answer the query perfectly for up to $2d + 2$ variables.
- Underlying distribution has a P-map (yeah, right...)

Algorithm:

1. Find skeleton
2. Find immoral sets of v-structures
3. Direct constrained edges

How To Test for Independence

Basic Setup: given data, determine whether two variables are independent. This is a classical statistical problem of hypothesis testing.

- H_0: null hypothesis that the variables are independent, $P(X,Y) = P(X)\,P(Y)$.
- Want a procedure to either accept or reject the hypothesis based on the evidence (data).
  - χ^2 statistic: $d_{\chi^2}(\mathcal{D}) = \sum_{X,Y,Z} \frac{\left(M[x,y,z] - M \times \hat{P}(z)\hat{P}(x|z)\hat{P}(y|z)\right)^2}{M \times \hat{P}(z)\hat{P}(x|z)\hat{P}(y|z)}$
  - Conditional mutual information between X and Y given Z.
- Accept H_0 if the value is low; reject if the value is high.

Score-based Structure Learning

Goal: Want to find a structure $\mathcal{G}$ (and, possibly, parameters $\boldsymbol{\theta}|\mathcal{G}$) that maximize the fit of the model to the data.

- First, we need to define scoring function.
- Need to search over the space of all possible graphs (DAGs) to find the graph that maximizes the scoring function.

Key: Scoring function plays a very significant role.

- Likelihood-based scores: $ScoreL(\mathcal{G};\mathcal{D}) = \ln P(\mathcal{D}|\mathcal{G},\boldsymbol{\theta}_{MLE})$
- Bayesian scores: integrate out the multinomial parameters given structure

Likelihood-based Scores

$$\mathcal{G}^\star = \operatorname*{argmax}_{\mathcal{G}} \ln P\left(\mathcal{D}|\mathcal{G}, \boldsymbol{\theta}_{MLE}\right) \text{ where } \boldsymbol{\theta}_{MLE} = \operatorname*{argmax}_{\boldsymbol{\theta}} \ln P\left(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G}\right)$$

- Not very useful: if the structure is not restricted, adding edges will always increase the objective function.
  - Solution: Add regularizer (prior over structures).
  - Another solution: Restrict models within a certain class (no more than d parents).
- Bad news: If $d > 1$, the problem is NP-hard.
- Good news: Tractable for maximum spanning trees.

Chow-Liu Trees (Incomplete)

Problem:

Given a complete data set $\mathcal{D} = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m\}$ of n-variate vectors, find a tree-structured Bayesian network $(\mathcal{G}, \boldsymbol{\theta})$ maximizing the likelihood of the data.

Solution:

It is easier go back and forth between Bayesian and Markov networks, estimating the parameters $\boldsymbol{\theta}_{MLE}$ using Bayesian networks, but to search through the space of all equivalent spanning trees using the equivalent Markov network.

- Given a tree structure $\mathcal{G}$, the sufficient statistics are defined over the pairs of variables corresponding to the edges.