Alex Smola
Barnabas Poczos

Ina Fiterau
CMU - MLD

Recitation 10: Bayes Nets

Part 2: Structure Learning

Main Source: F2010 Probabilistic Graphical Models, instructor: Noah Smith

Suggested Reading: Shay Cohen's recitation notes
http://select.cs.cmu.edu/class/10701-F09/recitations/recitation10.pdf

Learning Bayesian Networks

	Known Structure	Unknown Structure
Fully Observed Data	EASY (estimate CPT)	HARD (structure + CPT)
Missing Data	HARD (Variational Methods)	VERY HARD

BN Learning for Known Structure

- MLE for a BN whose CPDs (Conditional Probability Distributions) have disjoint parameters =

  MLEs for each of its CPDs

- => Estimate MLEs for the parameters of the conditionals

Decomposability

$$\begin{aligned}
\theta_{\mathrm{MLE}} &= \arg\max_{\theta} \prod_t P(X = x^{(t)} \mid \theta) \\
&= \arg\max_{\boldsymbol{\theta}} \prod_t \prod_i P(X_i = x_i^{(t)} \mid \mathrm{Parents}(X_i) = \mathrm{Parents}(x_i), \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_t \sum_i \log P(X_i = x_i^{(t)} \mid \mathrm{Parents}(X_i) = \mathrm{Parents}(x_i), \boldsymbol{\theta})
\end{aligned}$$

If the parameters **θ** are partitioned by CPT …

$$= \arg\max_{\boldsymbol{\theta}} \sum_i \sum_t \log P(X_i = x_i^{(t)} \mid \mathrm{Parents}(X_i) = \mathrm{Parents}(x_i), \boldsymbol{\theta}_i)$$

Deriving the MLE

- Most common distributions have closed forms for the MLE – you've used them
- Useful to know what distributions you obtain when you condition
- Solve analytically, for every parameter:

$$\frac{\partial}{\partial \theta_j} \sum_t \log P(X_i = x_i^{(t)} \mid \text{Parents}(X_i) = \text{Parents}(x_i^{(t)})) \quad = \quad 0$$

- Convex optimization

Learning Structure

- Same principle: maximizing the likelihood of the data
- Alternative:
  - use stat. tests to det. cond. independencies
  - construct the corresponding PDAG
- Idea: use likelihood to score structures

Likelihood and BN structures

$$\max_{\mathcal{G},\boldsymbol{\theta}} \log P_{\mathcal{G},\boldsymbol{\theta}}(X = x) = \max_{\mathcal{G}} \max_{\boldsymbol{\theta}} \log P_{\mathcal{G},\boldsymbol{\theta}}(X = x)$$
$$= \max_{\mathcal{G}} \log P_{\mathcal{G},\boldsymbol{\theta}_{\mathrm{MLE}}(\mathcal{G})}(X = x)$$

- For every possible structure, consider it with its best possible parameters (MLE)
- Optimistic, but correct if the overall goal is maximizing likelihood

Deriving the structure score for G

$$\begin{aligned}
\log P_{\mathcal{G},\boldsymbol{\theta}}(X = \boldsymbol{x}) &= \sum_i \sum_{x_i \in \mathrm{Val}(X_i)} \sum_{\boldsymbol{u} \in \mathrm{Val}(\mathrm{Parents}(X_i))} \mathrm{count}(x_i, \boldsymbol{u}; \boldsymbol{x}) \log \theta_{x_i \mid \boldsymbol{u}} \\
&= \sum_i \sum_{x_i \in \mathrm{Val}(X_i)} \sum_{\boldsymbol{u} \in \mathrm{Val}(\mathrm{Parents}(X_i))} m\hat{P}(x_i, \boldsymbol{u}) \log \hat{P}(x_i \mid \boldsymbol{u}) \\
&= \sum_i \sum_{x_i \in \mathrm{Val}(X_i)} \sum_{\boldsymbol{u} \in \mathrm{Val}(\mathrm{Parents}(X_i))} m\hat{P}(x_i, \boldsymbol{u}) \log \left(\frac{\hat{P}(x_i, \boldsymbol{u})}{\hat{P}(\boldsymbol{u})} \frac{\hat{P}(x_i)}{\hat{P}(x_i)} \right)
\end{aligned}$$

Deriving the structure score for G

$$
\begin{aligned}
\log P_{\mathcal{G},\boldsymbol{\theta}}(X=\boldsymbol{x}) &= \sum_i \sum_{x_i \in \mathrm{Val}(X_i)} \sum_{\boldsymbol{u} \in \mathrm{Val}(\mathrm{Parents}(X_i))} \mathrm{count}(x_i, \boldsymbol{u}; \boldsymbol{x}) \log \theta_{x_i|\boldsymbol{u}} \\
&= \sum_i \sum_{x_i \in \mathrm{Val}(X_i)} \sum_{\boldsymbol{u} \in \mathrm{Val}(\mathrm{Parents}(X_i))} m\hat{P}(x_i, \boldsymbol{u}) \log \hat{P}(x_i \mid \boldsymbol{u}) \\
&= \sum_i \sum_{x_i \in \mathrm{Val}(X_i)} \sum_{\boldsymbol{u} \in \mathrm{Val}(\mathrm{Parents}(X_i))} m\hat{P}(x_i, \boldsymbol{u}) \log \left(\frac{\hat{P}(x_i, \boldsymbol{u})}{\hat{P}(\boldsymbol{u})} \frac{\hat{P}(x_i)}{\hat{P}(x_i)} \right) \\
&= m \sum_i \sum_{x_i \in \mathrm{Val}(X_i)} \sum_{\boldsymbol{u} \in \mathrm{Val}(\mathrm{Parents}(X_i))} \hat{P}(x_i, \boldsymbol{u}) \left(\log \left(\frac{\hat{P}(x_i, \boldsymbol{u})}{\hat{P}(\boldsymbol{u})\hat{P}(x_i)} + \log \hat{P}(x_i) \right) \right) \\
&= m \sum_i I_{P_{\mathcal{G},\boldsymbol{\theta}}}(X_i; \mathrm{Parents}(X_i)) + m \sum_i \sum_{x_i \in \mathrm{Val}(X_i)} \sum_{\boldsymbol{u} \in \mathrm{Val}(\mathrm{Parents}(X_i))} \hat{P}(x_i, \boldsymbol{u}) \log \hat{P}(x_i) \\
&= m \sum_i I_{P_{\mathcal{G},\boldsymbol{\theta}}}(X_i; \mathrm{Parents}(X_i)) + m \sum_i \sum_{x_i \in \mathrm{Val}(X_i)} \hat{P}(x_i) \log \hat{P}(x_i) \\
&= m \sum_i I_{P_{\mathcal{G},\boldsymbol{\theta}}}(X_i; \mathrm{Parents}(X_i)) - m \sum_i H_{P_{\mathcal{G},\boldsymbol{\theta}}}(X_i)
\end{aligned}
$$

Decomposition

- Structure's likelihood decomposes by family => increased efficiency

$$\log P_{\mathcal{G},\boldsymbol{\theta}}(X = x) \quad = \quad m \underbrace{\sum_i I_{P_{\mathcal{G},\boldsymbol{\theta}}}(X_i; \text{Parents}(X_i))}_{\text{directly related to structure}} - m \underbrace{\sum_i H_{P_{\mathcal{G},\boldsymbol{\theta}}}(X_i)}_{\text{doesn't depend on structure}}$$

Problem with Mutual Information

$$I(X;Y) \leq I(X;Y \cup Z)$$

- Unless conditional independence holds exactly in the data, more connections are always better!
- For structure, MLE is guaranteed to

OVERFIT

Possible Solution: ***Chow-Liu***

- Each node can have at most one parent
- Structure will have at most n-1 edges
- Decision is where to place the edges
- Algorithm:
  - Consider $I(X_i, X_j)$ to be the score of putting an edge between X_i and X_j
  - Find the maximum spanning tree
  - Number of trees? $O(2^{nlog(n)})$

Possible Solution: ***Chow-Liu***

- Each node can have at most one parent
- Structure will have at most n-1 edges
- Decision is where to place the edges
- Algorithm:
  - Consider $I(X_i, X_j)$ to be the score of putting an edge between X_i and X_j
  - Find the maximum spanning tree
  - Pick root, traverse to get structure

Possible Solution: ***Chow-Liu***

- Maximum-scoring spanning trees gives the skeleton
- Trees with the same skeleton have
  - The same conditional independence assertions
  - The same mutual information score
- The resulting model has no V-structures

Not covered (yet)

- Being Bayesian: priors on structure
- Consistent* Scores:
  - Bayesian Score and modularity
  - Bayesian Information Criterion (BIC)
    - Penalizes model dimension by log(m)/2

* as the number of samples goes to infinity, the recovered structure is 'I-equivalent' to the map of the true distribution

- Structure search