Structure Learning

EECS 474 Fall 2016
Road Map

- Basics of Probability and Statistical Estimation
- Bayesian Networks
- Markov Networks
- Inference
- Learning
  - Parameters, \textbf{Structure}, EM
- HMMs
Learning Structure

- **Hard problem**
  - Finding the BN structure with the highest “score” among those structures with at most $k$ parents is NP hard for $k>1$ (Chickering, 1995)

- **Inputs**
  - Data (potentially incomplete)

- **Outputs**
  - Graphical model structure (we’ll focus on Bayes Nets)

- **Approaches**
  - Constraint-based
  - Score-based approaches
    - Local search
  - Bayesian Model Averaging
Learning Structure

- **Hard problem**
  - Finding the BN structure with the highest “score” among those structures with at most $k$ parents is NP hard for $k>1$ (Chickering, 1995)

- **Inputs**
  - Data (potentially incomplete)

- **Outputs**
  - Graphical model structure (we’ll focus on Bayes Nets)

- **Approaches**
  - Constraint-based
  - Score-based approaches
    - Local search
  - Bayesian Model Averaging
Constraint-based Approaches

- Idea: we know how to construct a Bayes Net if we can perform **independence tests**
  - \((A \perp B \mid C)\)?

- Naïve construction
  - depends on variable ordering
  - Issues potentially large number of independence queries

- A more sophisticated PDAG construction process works better (see book)
Constraint-based approach guarantees

- Can uncover a *perfect* map using a polynomial # of tests if:
  - Bounded in-degree $d$ in $G^*$ (the true graph)
  - Perfect independence queries up to size $2d + 2$
    (Strong)
  - $P^*$ (true dist.) is *faithful* to $G^*$
    (Also strong)
    - i.e., any independencies in $P^*$ reflected as $d$-separation in $G^*$
Learning Structure

- Approaches
  - Constraint-based
  - Score-based approaches
    - Local search
  - Bayesian Model Averaging
Scoring Structures

- **Maximum likelihood $G$**
  - Choose $G = \arg \max_G \max_{\theta} P(\text{Data} \mid \theta)$

- **Or MAP:**
  - Choose $G = \arg \max_G \max_{\theta} P(\text{Data} \mid \theta) \ P(\theta)$

- …what’s wrong with these?
Bayesian Score

Bayesian Score for $G =$

prior for $G$

+ 

likelihood integrated over all parameters for $G$

BayesianScore($G : \text{Data}$) = $\log P(\text{Data} \mid G) + \log P(G)$

$P(\text{Data} \mid G) = \int_{\Theta_G} P(\text{Data} \mid \theta_G, G) P(\theta_G \mid G) \, d\theta_G$
Integrating over parameters
Training (x-axis) vs. Test (y-axis) Perf.

\[ \frac{1}{M} \log P(D \mid G) \]

500 instances

\[ \frac{1}{M} \log P(D \mid G) \]

10,000 instances
Bayesian Information Criterion

- **Bayes Score includes:**
  
  \[ P(\text{Data} \mid G) = \int_{\Theta_G} P(\text{Data} \mid \theta_G, G) \, P(\theta_G \mid G) \, d\theta_G \]

- Integral sometimes difficult

- **Approximation:**
  
  \[ \text{score}_{BIC}(G) = - (\text{Dim}[G]/2) \log M + \log \max_{\theta_G} P(\text{Data} \mid \theta_G) \]
Finding the BN structure with the highest score among those structures with at most \( k \) parents is NP hard for \( k > 1 \) (Chickering, 1995)

**Heuristic methods**
- Greedy
- Greedy with restarts
Structure priors

- Lots of options
  - All possible structures equally likely
  - Partial ordering, required / prohibited arcs
  - $\text{Prior}(G) \propto \text{Similarity}(G, G_{\text{prior}})$
Learning Structure

- Approaches
  - Constraint-based
  - Score-based approaches
    - Local search
  - Bayesian Model Averaging
Bayesian Model Averaging

- Previous methods all find a single graph $G$

- *Bayesian model averaging* instead makes predictions by averaging over structures:

$$P(\text{test example} \mid \text{Data}) = \sum_G P(\text{test example} \mid \text{Data}, G) \ P(G \mid \text{Data})$$