CS 4824/ECE 4424: Graphical Models II

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.
Bayesian network recap

- A Bayes network represents the joint probability distribution over a collection of random variables.
- A Bayes network is a directed acyclic graph and a set of CPD’s:
  - Each node denotes a random variable.
  - Edges denote dependencies.
  - CPD for each node $X_i$ defines $P(X_i | Pa(X_i))$.
  - The joint distribution over all variables is defined as:
    $$P(X_1 \ldots X_n) = \prod_{i} P(X_i | Pa(X_i))$$
Bayesian network recap

- Bayes nets are convenient representation for encoding dependencies / conditional independence
- BN = Graph plus parameters of CPD’s
  - Defines joint distribution over variables
  - Can calculate everything else from that
  - Though inference may be intractable
Inference in Bayes Nets

- In general, intractable (NP-complete)
- For certain cases, tractable
  - Assigning probability to fully observed set of variables
  - Or if just one variable unobserved
  - Or for singly connected graphs (i.e., no undirected loops)
    - Belief propagation
- Sometimes use Monte Carlo methods
  - Generate many samples according to the Bayes Net distribution, then count up the results
Example

- Bird flu and Allegies both cause Sinus problems
- Sinus problems cause Headaches and runny Nose
Prob. of joint assignment: easy

- Suppose we are interested in joint assignment \( <F=f, A=a, S=s, H=h, N=n> \)

- What is \( P(f,a,s,h,n) \)?

\[
P(F=f, A=a, S=s, H=h, N=n) = P(F=f) P(A=a) P(S=s | F=f, A=a) P(H=h | S=s) P(N=n | S=s)
\]

If we have \( k \) RV's, there will be \( k \) terms.

Cost is linear in the number of RV's.

let's use \( P(a,b) \) as shorthand for \( P(A=a, B=b) \)
Prob. of marginals: not so easy

- How do we calculate $P(N=n)$?

$$P(N=n) = \sum_s P(N=n | S=s) \cdot P(S=s)$$

chase up the Bayes Net

$$P(N=n) = \sum_{f,a,h,s} P(F=f, A=a, H=h, S=s, N=n)$$

let's say we have $K$ boolean RV's

How many terms do we have in the sum? $2^{K-1}$

for each of them, we do $K$ multiplications

let's use $P(a,b)$ as shorthand for $P(A=a, B=b)$

$\text{cost} = 2^{K-1} \cdot K$ computations!
Generating a sample from joint distribution: easy

- How can we generate random samples drawn according to $P(F,A,S,H,N)$?

\[
P(N = n) \quad P(S = s | F = f, A = a)
\]

- Randomly draw a value for $F = f$
  - Draw $\tau \in [0, 1]$ uniformly randomly
  - If $\tau < \theta_{F=1}$ then output $f = 1$
  - Else $f = \emptyset$

Monte Carlo Sampling
Random Algorithm

N.B., with a fixed seed, your calculations are reproducible!
Generating a sample from joint distribution: easy

- How can we generate random samples drawn according to $P(F,A,S,H,N)$?

- Random sample of $F$ according to $P(F=1) = \theta_{F=1}$:
  - Draw a value of $r$ uniformly from $[0,1]$
  - If $r < \theta$ then output $F=1$, else $F=0$

- Solution:
  - Draw a random value $f$ for $F$, using its CPD
  - Then draw values for $A$, for $S|A,F$, for $H|S$, for $N|S$

Caution: if prob. is very low then you need to generate LOT of samples!!
Generating a sample from joint distribution: easy

- Note we can estimate marginals like \( P(N=n) \) by generating many samples from joint distribution, then count the fraction of samples for which \( N=n \)

- Similarly, for anything else we care about \( P(F=1 | H=1, N=0) \)

  \[
  
  \frac{P(F=1, H=1, N=0)}{P(H=1, N=0)}
  \]

- weak but general method
Learning of Bayes Nets

- Several types of learning problems
  - Variable values may be fully observed / partly unobserved
  - Easy case: learn parameters when data is fully observed
  - Interesting case: graph known, data partly known

Gruesome case: graph structure is unknown, data partly known

"Structure learning"
Learning CPTs from Fully Observed Data

- Example: Consider learning the parameter
  \[ \theta_{sij} \equiv P(S = 1 | F = i, A = j) \]

- Max Likelihood Estimate is
  \[ \theta_{sij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)} \]

\[
P(\text{data} | \theta) = \prod_{k=1}^{K} P(f_k, a_k, s_k, h_k, n_k)
\]

\[
= \prod_{k=1}^{K} P(f_k) P(a_k) P(S_k | f_k, a_k) P(C_h_k | S_k) P(n_k | S_k)
\]
Learning CPTs from Fully Observed Data

Example: Consider learning the parameter

\[ \theta_{s|i,j} \equiv P(S = 1|F = i, A = j) \]

Max Likelihood Estimate is

\[
\theta_{s|i,j} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}
\]

\[ \delta(x) = 1 \text{ if } x=\text{true}, \]
\[ = 0 \text{ if } x=\text{false} \]
MLE estimate of $\theta_{s|ij}$ from fully observed data

- Maximum likelihood estimate

$$\theta = \arg \max \theta \quad p(\text{data} | \theta)$$

$$\theta = \arg \max \theta \quad \log p(\text{data} | \theta)$$

- Our case

$$\log p(\text{data} | \theta) = \sum_{k=1}^{K} \left[ \log p(f_k) + \log p(a_k) + \log p(s_{k|f_k,a_k}) + \log p(n_k | s_k) \right]$$

$$\frac{\partial \log p(\text{data} | \theta)}{\partial \theta} = \sum_{k=1}^{K} \frac{\partial \log p(s_{k|f_k,a_k})}{\partial \theta_{s|ij}}$$

N.B.
MLE estimate of $\theta_{s|i,j}$ from fully observed data

- Maximum likelihood estimate

$$\theta \leftarrow \arg \max_{\theta} \log P(data|\theta)$$

- Our case

$$P(data|\theta) = \prod_{k=1}^{K} P(f_k, a_k, s_k, h_k, n_k)$$

$$P(data|\theta) = \prod_{k=1}^{K} P(f_k) P(a_k) P(s_k|f_k a_k) P(h_k|s_k) P(n_k|s_k)$$

$$\log P(data|\theta) = \sum_{k=1}^{K} \log P(f_k) + \log P(a_k) + \log P(s_k|f_k a_k) + \log P(h_k|s_k) + \log P(n_k|s_k)$$

$$\frac{\partial \log P(data|\theta)}{\partial \theta_{s|i,j}} = \sum_{k=1}^{K} \frac{\partial \log P(s_k|f_k a_k)}{\partial \theta_{s|i,j}}$$

$$\theta_{s|i,j} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$
Estimate $\theta$ from partly observed data

- What if FAHN observed, but not S?
- Can’t calculate MLE

$$\hat{\theta} \leftarrow \arg \max_{\theta} \log \prod_{k} P(f_k, a_k, s_k, h_k, n_k|\theta)$$

- Let X be all observed variable values (over all examples)
- Let Z be all unobserved variable values
- Can’t calculate MLE

$$\hat{\theta} \leftarrow \arg \max_{\theta} \log P(X, Z|\theta)$$

- What to do?