CS 4824/ECE 4424: Graphical Models III

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

- 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
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
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

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

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

\( k \)th training example

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

- Maximum likelihood estimate
- Our case
MLE estimate of $\theta_{s|ij}$ from fully observed data

- Maximum likelihood estimate

\[
\theta \leftarrow \arg \max_{\theta} \log P(\text{data}|\theta)
\]

- Our case

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

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

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

\[
\frac{\partial \log P(\text{data}|\theta)}{\partial \theta_{s|ij}} = \sum_{k=1}^{K} \frac{\partial \log P(s_k|f_k a_k)}{\partial \theta_{s|ij}}
\]

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

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

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

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

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

- What to do?
Estimate $\theta$ from partly observed data

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

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

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

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

- EM seeks* to estimate:

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

* EM guaranteed to find local optima
Estimate $\theta$ from partly observed data

- EM seeks to estimate:

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

- here, observed $X=$\{F,A,H,N\}, unobserved $Z=$\{S\}

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

$$
E_{P(Z|X,\theta)} \log P(X, Z|\theta) =
$$
Estimate $\theta$ from partly observed data

- EM seeks to estimate:

$$\theta \leftarrow \arg \max_{\theta} \mathbb{E}_{Z|X,\theta}[\log P(X, Z|\theta)]$$

- Here, observed $X=\{F,A,H,N\}$, unobserved $Z=\{S\}$

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

$$E_{P(Z|X,\theta)} \log P(X, Z|\theta) = \sum_{k=1}^{K} \sum_{i=0}^{1} P(s_k = i|f_k, a_k, h_k, n_k)$$

$$[\log P(f_k) + \log P(a_k) + \log P(s_k|f_ka_k) + \log P(h_k|s_k) + \log P(n_k|s_k)]$$
EM algorithm — informally

- EM is a general procedure for learning from partly observed data
- Given observed variables $X$, unobserved $Z$ (\(X=\{F,A,H,N\}, Z=\{S\}\))

Begin with arbitrary choice for parameters $\theta$

Iterate until convergence:

- **E Step**: estimate the values of unobserved $Z$ conditioned on $X$ using $\theta$
- **M Step**: use observed values plus E-step estimates to derive a better $\theta$

- Guaranteed to find local maximum. Each iteration increases

\[
E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]
\]
EM algorithm — precisely

- EM is a general procedure for learning from partly observed data
- Given observed variables $X$, unobserved $Z$ ($X=\{F,A,H,N\}$, $Z=\{S\}$)
- Define $Q(\theta'|\theta) = E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$

Begin with arbitrary choice for parameters $\theta$

Iterate until convergence:

- E Step: Use $X$ and current $\theta$ to calculate $P(Z|X,\theta)$
- M Step: Replace current $\theta$ by $\theta \leftarrow \arg \max_{\theta'} Q(\theta'|\theta)$

- Guaranteed to find local maximum. Each iteration increases

$$E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$$