ECE 6504: Advanced Topics in Machine Learning
Probabilistic Graphical Models and Large-Scale Learning

Topics
- BN / MRFs
  - Learning from hidden data
  - EM

Readings: KF 19.1-3, Barber 11.1-2

Dhruv Batra
Virginia Tech
Administrativia

• (Mini-)HW4
  – Out now
  – Due: May 7, 11:55pm
  – Implementation:
    • Parameter Learning with Structured SVMs and Cutting-Plane

• Final Project Webpage
  – Due: May 7, 11:55pm
  – Can use late days
  – 1-3 paragraphs
    • Goal
    • Illustrative figure
    • Approach
    • Results (with figures or tables)

• Take Home Final
  – Out: May 8
  – Due: May 13, 11:55pm
  – No late days
  – Open book, open notes, open internet. Cite your sources.
  – No discussions!
Recap of Last Time
Main Issues in PGMs

• Representation
  – How do we store $P(X_1, X_2, \ldots, X_n)$
  – What does my model mean/imply/assume? (Semantics)

• Inference
  – How do I answer questions/queries with my model? such as
    – Marginal Estimation: $P(X_5 \mid X_1, X_4)$
    – Most Probable Explanation: $\text{argmax } P(X_1, X_2, \ldots, X_n)$

• Learning
  – How do we learn parameters and structure of $P(X_1, X_2, \ldots, X_n)$ from data?
  – What model is the right for my data?
# Learning Bayes Nets

## Known structure vs. Unknown structure

<table>
<thead>
<tr>
<th>Data Structure</th>
<th>Known structure</th>
<th>Unknown structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully observable data</td>
<td>Very easy</td>
<td>Hard</td>
</tr>
<tr>
<td>Missing data</td>
<td>Somewhat easy (EM)</td>
<td>Very very hard</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{Data} & : x^{(1)}, \ldots, x^{(m)} \\
\text{structure} & : \text{Bayes Net} \\
\text{CPTs} & : P(X_i | Pa_{Xi})
\end{align*}
\]
Learning the CPTs

For each discrete variable $X_i$

\[
\hat{P}_{MLE}(X_i = a \mid Pa_{X_i} = b) = \frac{\text{Count}(X_i = a, Pa_{X_i} = b)}{\text{Count}(Pa_{X_i} = b)}
\]
Learning Markov Nets

<table>
<thead>
<tr>
<th>Known structure</th>
<th>Unknown structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully observable data</td>
<td>NP-Hard (but doable)</td>
</tr>
<tr>
<td>Missing data</td>
<td>Harder (EM)</td>
</tr>
<tr>
<td></td>
<td>Don’t try this at home</td>
</tr>
</tbody>
</table>

Data: $x^{(1)}, \ldots, x^{(m)}$

Factors: $\Psi_c(x_c)$

Structure + Parameters
Learning Parameters of a BN

• Log likelihood decomposes:

\[ \log P(D | \theta) = m \sum_i \sum_{x_i, Pa_{x_i}} \hat{P}(x_i, Pa_{x_i}) \log P(x_i | Pa_{x_i}) \]

• Learn each CPT independently

• Use counts

\[ \hat{P}(u) = \frac{\text{Count}(U = u)}{m} \]

(C) Dhruv Batra

Slide Credit: Carlos Guestrin
Log Likelihood for MN

• Log likelihood decomposes:

\[ \log P(D | \theta, G) = m \sum_i \sum_{c_i} \hat{P}(c_i) \log \psi_i(c_i) - m \log Z \]

• Doesn’t decompose!
  – \( \log Z \) couples all parameters together
Plan for today

• BN Parameter Learning with Missing Data
  – Why model latent variables?
  – Expectation Maximization (EM)
## Learning Bayes Nets

<table>
<thead>
<tr>
<th></th>
<th>Known structure</th>
<th>Unknown structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully observable data</td>
<td>Very easy</td>
<td>Hard</td>
</tr>
<tr>
<td>Missing data</td>
<td>Somewhat easy (EM)</td>
<td>Very very hard</td>
</tr>
</tbody>
</table>

Data $\{x^{(1)}, \ldots, x^{(m)}\}$

Structure

CPTs – $P(X_i | Pa_{Xi})$

Parameters

---

(C) Dhruv Batra

Slide Credit: Carlos Guestrin
When is data missing?

- Fully Observed Data
- Some hidden variables
  - Never observed
- General hidden pattern
  - Arbitrary entries missing in the data matrix
Why missing data?

• Sometimes no choice
  – sensor error, some data dropped
  – Data collection error, we forgot to ask this question
Why *introduce* hidden variables?

- **Model Sparsity!**
  - Modeling hidden/latent variables can simplify interactions
  - Reduction in #parameters to be learnt

- **Example**
  - On board
Why *introduce* hidden variables?

- Discovering Clusters in data!
  - Modeling different $P(y|x,h)$ for each $h$
Treating Missing Data

• Thought Experiment:
  – Coin Toss: H,T,?,?,H,H,?

• Case 1: Missing at Random

• Case 2: Missing with bias

• BN illustration of the two cases
  – On board
  – Takeaway message: Need to model missing data
Likelihood with Complete/Missing Data

- Example on board $X \rightarrow Y$
  - One variable $X$; parameter $\theta_X$
  - Two variables $X,Y$; parameters $\theta_X$, $\theta_{Y|X}$

- Takeaway Messages:
  - Parameters get coupled ($LL = \text{sum-log-sum} \text{ doesn’t factorize}$)
  - Computing $LL$ requires marginal inference!
Data likelihood for BNs

- Given structure, log likelihood of fully observed data:

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G})$$
Marginal likelihood

- What if $S$ is hidden?

$$\log P(\mathcal{D} \mid \theta_G, G)$$
Log likelihood for BNs with hidden data

- Marginal likelihood – $O$ is observed, $H$ is hidden

$$\ell(\theta : D) = \sum_{j=1}^{m} \log P(o^{(j)} | \theta)$$

$$= \sum_{j=1}^{m} \log \sum_{h} P(h, o^{(j)} | \theta)$$
EM Intuition

- **Chicken & Egg problem**
  - If we knew $h$, then learning $\theta$ would be easy
  - If we knew theta, then finding $P(h \mid o, \theta)$ would be “easy”
    - Sum-product inference

- **EM solution**
  - Initialize
  - Fix $\theta$, find $P(h \mid o, \theta)$
  - Use these to learn $\theta$
E-step for BNs

- E-step computes probability of hidden vars $h$ given $o$

$$Q^{(t+1)}(h \mid o) \leftarrow P(h \mid o, \theta^{(t)})$$

- Corresponds to inference in BN
The M-step for BNs

• Maximization step:

$$\theta^{(t+1)} \leftarrow \arg \max_{\theta} \sum_{j=1}^{m} \sum_{\mathbf{h}} Q^{(t+1)}(\mathbf{h} \mid \mathbf{o}^{(j)}) \log P(\mathbf{h}, \mathbf{o}^{(j)} \mid \theta)$$

• Use expected counts instead of counts:
  – If learning requires $\text{Count}(\mathbf{h}, \mathbf{o})$
  – Use $E_{Q^{(t+1)}}[\text{Count}(\mathbf{h}, \mathbf{o})]$
M-step for each CPT

- M-step decomposes per CPT
  - Standard MLE:

\[
P(X_i = x_i \mid Pa_{X_i} = z) = \frac{\text{Count}(X_i = x_i, Pa_{X_i} = z)}{\text{Count}(Pa_{X_i} = z)}
\]

\[
P(X_i = x_i \mid Pa_{X_i} = z) = \frac{\text{ExCount}(X_i = x_i, Pa_{X_i} = z)}{\text{ExCount}(Pa_{X_i} = z)}
\]
The general learning problem with missing data

- Marginal likelihood – \( o \) is observed, \( h \) is missing:

\[
ll(\theta : \mathcal{D}) = \log \prod_{j=1}^{M} P(o^j | \theta) \\
= \sum_{j=1}^{M} \log P(o^j | \theta) \\
= \sum_{j=1}^{M} \log \sum_{h} P(o^j, h | \theta)
\]
Applying Jensen’s inequality

• Use: \( \log \sum_h P(h) f(h) \geq \sum_h P(h) \log f(h) \)

\[
ll(\theta : D) = \sum_{j=1}^{M} \log \sum_h Q_j(h) \frac{P(o^j, h | \theta)}{Q_j(h)}
\]
Convergence of EM

• Define potential function $F(\theta, Q)$:

$$ll(\theta : D) \geq F(\theta, Q_j) = \sum_{j=1}^{M} \sum_{h} Q_j(h) \log \frac{P(o^j, h | \theta)}{Q_j(h)}$$

• EM corresponds to coordinate ascent on $F$
  – Fix $\theta$, maximize $Q$
  – Fix $Q$, maximize $\theta$
  – Thus, maximizes lower bound on marginal log likelihood
EM is coordinate ascent

\[ ll(\theta : \mathcal{D}) \geq F(\theta, Q_j) = \sum_{j=1}^{M} \sum_h Q_j(h) \log \frac{P(o^j, h | \theta)}{Q_j(h)} \]

- **E-step**: Fix \( \theta^{(t)} \), maximize F over Q:
  - On board

  - “Realigns” F with likelihood: \( Q_j(h) = P(h | o^j, \theta^{(t)}) \)

  \[ F(\theta^{(t)}, Q^{(t)}) = ll(\theta^{(t)} : \mathcal{D}) \]
EM is coordinate ascent

$$ll(\theta : D) \geq F(\theta, Q_j) = \sum_{j=1}^{M} \sum_{h} Q_j(h) \log \frac{P(o^j, h | \theta)}{Q_j(h)}$$

- **M-step**: Fix $Q^{(t)}$, maximize $F$ over $\theta$

- Corresponds to weighted dataset:
  - $<o^1, h=1>$ with weight $Q^{(t+1)}(h=1|o^1)$
  - $<o^1, h=2>$ with weight $Q^{(t+1)}(h=2|o^1)$
  - $<o^1, h=3>$ with weight $Q^{(t+1)}(h=3|o^1)$
  - $<o^2, h=1>$ with weight $Q^{(t+1)}(h=1|o^2)$
  - $<o^2, h=2>$ with weight $Q^{(t+1)}(h=2|o^2)$
  - $<o^2, h=3>$ with weight $Q^{(t+1)}(h=3|o^2)$
EM Intuition

\[
Q(\theta, \theta_t)
\]
\[
Q(\theta, \theta_{t+1})
\]
\[
l(\theta)
\]
What you need to know about learning BNs with missing data

- EM for Bayes Nets

- E-step: inference computes expected counts
  - Only need expected counts over $X_i$ and $\text{Pa}_{xi}$

- M-step: expected counts used to estimate parameters

- Which variables are hidden can change per datapoint
  - Also, use labeled and unlabeled data → some data points are complete, some include hidden variables