Learning Bayesian Nets
Parameters from Partial Data

KF, Chapter 18-18.2

Some material taken from C Guesterin (CMU)
Space of Topics

Learning...
- Parameter, Structure
- Data: Complete, Missing
- Framework: Frequentist, Bayesian
Learning Belief Net Parameters from Partial Data

- Framework
  - Why is the data missing? ... MCAR, MAR, ...
  - Why more challenging?

- Approaches
  - Gradient Ascent
  - EM
  - Gibbs
Learning from Missing data

- To find good $\Theta$, need to compute $P(\Theta, S \mid \mathcal{G})$
- Easy if ..

$$S = \begin{cases} \ c_1 : \begin{pmatrix} \end{pmatrix} \cdots \begin{pmatrix} \end{pmatrix} & \text{incomplete} \\ \ c_2 : \begin{pmatrix} \end{pmatrix} \cdots \begin{pmatrix} \end{pmatrix} & \text{complete} \\ \vdots & \vdots \\ \ c_m : \begin{pmatrix} \end{pmatrix} \cdots \begin{pmatrix} \end{pmatrix} & \end{cases}$$

- What if $S$ is incomplete
  - Some $c_{ij} = *$
  - “Hidden variables” ($X_k$ never seen: $c_{ik} = * \ \forall \ i$)

- Here:
  - Given fixed structure
  - Missing (Completely) At Random:
    Omission not correlated with value, etc.

- Approaches:
  - Gradient Ascent, EM, Gibbs sampling, ..
Why is the data missing?

- Estimating $P(\text{Heads}) = \theta$
  - Earlier: data = [H, T, H, H, ..., T]
  - Now: data = [H, T, ?, ?, H, ..., T]
- Thumbtack falls off table, ...not recorded
  - No information in "?"

VS

- Recorder doesn’t like "Tails", and so omits those values
  - Here, "?" means "Tails" – lots of info!
Formal Model

- \( \mathbf{X} = \{X_1, X_2, \ldots, X_n\} \) : set of r.v.s
- \( \mathbf{O}_\mathbf{X} = \{O_1, O_2, \ldots, O_n\} \) : corresponding set of observability variables
- \( P_{\text{miss}}(\mathbf{X}, \mathbf{O}_\mathbf{X}) = P(\mathbf{X}) \cdot P_{\text{miss}}(\mathbf{O}_\mathbf{X} | \mathbf{X}) \)
- \( P(\mathbf{X}) \) parameterized by \( \theta \)
- \( P_{\text{miss}}(\mathbf{O}_\mathbf{X} | \mathbf{X}) \) parameterized by \( \psi \)
- \( \mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\} \quad \text{Val}(Y_i) = \text{Val}(X_i) \cup \{?\} \)
  \[
  Y_i = \begin{cases} 
  X_i & \text{if } +o_i \\
  ? & \text{if } -o_i 
  \end{cases}
  \]
Uncorrelated Missingness

Here, \( \mathbf{X} \perp \mathbf{O}_X \); \( \theta \perp \psi \mid D \)

- \( P( Y= H ) = \theta \psi \)
- \( P( Y= T ) = (1 - \theta) \psi \)
- \( P( Y= ? ) = 1 - \psi \)

Assuming \( D \) contains \( \#[H], \#[T], \#[?] \)

- \( L[\theta, \psi : D] = \theta^{\#[H]} (1 - \theta)^{\#[T]} \psi^{\#[H]+\#[T]} (1 - \psi)^{\#[?]} \)
- \( \theta^{(\text{MLE})} = \frac{\#[H]}{\#[H] + \#[T]} \)
- \( \psi^{(\text{MLE})} = \frac{\#[H] + \#[T]}{\#[H] + \#[T] + \#[?]}. \)

Simple frequencies!!
Correlated Missingness

Here, \( \neg \left[ \theta \perp \psi \mid D \right] \)
- \( \psi_{Ox|H} = \text{prob of seeing output, if heads} \)
  \( = P(Y=H \mid X=H) \)
- \( \psi_{Ox|T} = P(Y=T \mid X=T) \)

- \( P(Y=H) = \theta \psi_{Ox|H} \)
- \( P(Y=T) = (1 - \theta) \psi_{Ox|T} \)
- \( P(Y=?) = \theta (1 - \psi_{Ox|H}) + (1 - \theta) (1 - \psi_{Ox|T}) \)

Assuming \( D \) contains \#[H], \#[T], \#[?]
- \( L[\theta, \psi : D] = \theta \#[H] (1 - \theta) \#[T] \psi_{Ox|H} \#[H] \psi_{Ox|T} \#[T] \)
  \( \theta (1 - \psi_{Ox|H}) + (1 - \theta) (1 - \psi_{Ox|T}) \#[?] \)

What a mess! Does not factor, so no easy MLE values...
A missing data model $P_{\text{missing}}$ is \textit{missing completely at random (MCAR)} if

$$P \models X \perp O_X$$

- Plausible ...
  - Coffee spills on paper
  - Flecks of dusts in images
- Here, can solve separately for
  - $\theta$ (for $P(X)$)
  - $\psi$ (for $P_{\text{miss}}(O_X | X)$)
MCAR is ... too strong!

- Not MCAR:
  - test results are missing if not ordered... perhaps as patient too sick or too healthy
  - \( \Rightarrow \) Missingness-of-test is correlated with test-outcome

- MCAR is sufficient for decomposition of likelihood...
  - but NOT necessary

- Just need

Observation mechanism is CONDITIONALLY INDEPENDENT of variables, GIVEN OTHER OBSERVATIONS
Weaker Condition

- Flip coin \( X_1, X_2 \)
- If \( X_1 = \text{Heads} \), reveal outcome of \( X_2 \)

- Here, \( P \models O_{X_2} \perp X_2 \mid X_1 \)
  - Outcomes of both coins INDEPENDENT of whether hidden, given observations

- Use \( \theta_{X_1} \theta_{X_2} \psi_{O_{X_2}|H} \psi_{O_{X_2}|T} \) (where \( \theta_{X_1} \perp \theta_{X_2} \))

\[
L(\theta : \mathcal{D}) = \theta_{X_1}^{M[Y_1=\text{Heads}]}(1 - \theta_{X_1})^{M[Y_1=\text{Tails}]} \\
\theta_{X_2}^{M[Y_2=\text{Heads}]}(1 - \theta_{X_2})^{M[Y_2=\text{Tails}]} \\
\psi_{O_{X_2}|H}^{M[Y_1=\text{Heads}, Y_2=\text{Heads}]+M[Y_1=\text{Heads}, Y_2=\text{Tails}]}(1 - \psi_{O_{X_2}|H})^{M[Y_1=\text{Heads}, Y_2=\text{?}]} \\
\psi_{O_{X_2}|Tails}^{M[Y_1=\text{Tails}, Y_2=\text{Heads}]+M[Y_1=\text{Tails}, Y_2=\text{Tails}]}(1 - \psi_{O_{X_2}|\text{Tails}})^{M[Y_1=\text{Tails}, Y_2=\text{?}]} \\
\]

- Four factors, each w/ just 1 parameter
  \( \Rightarrow \) can solve independently!
Missing At Random

- Given tuple of observations $y$, partition variables $X$ into
  - observed $X^y_{\text{obs}} = \{ X_i \mid y_i \neq ? \}$
  - hidden $X^y_{\text{hid}} = \{ X_i \mid y_i = ? \}$
- Missing data model $P_{\text{miss}}$ is \textit{missing at random (MAR)} if
  \[ \forall y \text{ w/ } P_{\text{miss}}(y) > 0 \text{ and } \forall x^y_{\text{hid}} \in \text{Val}(x^y_{\text{hid}}) \]
  \[ P_{\text{miss}} \models O_X \perp x^y_{\text{hid}} \mid x^y_{\text{obs}} \]

\[ P_{\text{miss}}(x^y_{\text{hid}} \mid x^y_{\text{obs}}, O_X) = P_{\text{miss}}(x^y_{\text{hid}} \mid x^y_{\text{obs}}) \]
Meaning of MAR...

MAR $\Rightarrow$

$P_{\text{miss}}(x_{\text{hid}}^y \mid x_{\text{obs}}^y, o_X) = P_{\text{miss}}(x_{\text{hid}}^y \mid x_{\text{obs}}^y)$

$\Rightarrow$

$P_{\text{miss}}(y) = P_{\text{miss}}(o_X \mid x_{\text{obs}}^y) P(x_{\text{obs}}^y)$

- Depends on $\psi$
- Depends on $\theta$

If $P_{\text{miss}}$ is MAR, then

$L(\theta, \psi; D) = L(\theta; D) L(\psi; D)$

MAR $\Rightarrow$

Can ignore observation model when learning model parameters!
Comments on MAR...

- There are many MAR situations but ...
- **BP_Sensor** measures blood pressure
  - BP_Sensor can fail if patient is overweight
  - Obesity is relevant to blood pressure
  - So... “non-observation” is informative – not MAR
  - (But if we know Weight & Height, then $O_B \perp B \mid \{W,H\}$ )
- Probably no X-ray $X$ if no broken bones,
  - So $\neg (O_x \perp X)$, not MAR
  - But if “primary complaint” $C$ known, $O_x \perp X \mid C$ ... MAR!

- We will assume MAR from now on...
Bayesian Learning for 2-node BN

- Every path between $\theta_X - \theta_{Y|X}$ is:
  - $\theta_X \rightarrow X[m] \rightarrow Y[m] \leftarrow \theta_{Y|X}$
    - Partial
- Complete data
  - $\Rightarrow$ values for $D = \{X[1], \ldots, X[M], Y[1], \ldots, Y[M]\}$
  - $\Rightarrow$ path is NOT active

$\Rightarrow \theta_X \perp \theta_{Y|X} \perp D$
Example ...

- Complete data:
- Likelihood:
  - $\theta_x^{29}(1 - \theta_x)^{14} \theta_{y|x=+}^{10} \theta_{y|x=0}^{4} \theta_{y|x=-1}^{13}(1 - \theta_{y|x=-1})^{16}$
  - Easy to solve
- What if don’t know $X[1]$
  - (Assume $Y[1]=+$ )
  - Likelihood:
    - $\theta_x^{29}(1 - \theta_x)^{13} \theta_{y|x=+}^{10} \theta_{y|x=0}^{4} \theta_{y|x=-1}^{12} [\theta_x \theta_{y|x=+} + (1- \theta_x) \theta_{y|x=-1}]$
    - Not as nice...
- If $k$ missing values, $L(...; D)$ could have many terms...
Geometric Visualization

- Complete data: *unimodal*
- Incomplete data:
  ... sum of unimodals...
  which is *multimodal*!
Problems with Hidden Variables

- Observe $X, Y$... but not $H$
  - $P(+x, -y) = \sum_h P(h) P(+x|h) P(-y|h)$
- Likelihood
  
  \[
  L(\theta : D) = \prod_{x,y} \left[ \sum_h P(h) P(x|h) P(y|h) \right]^{\#(x,y)}
  \]

- Cannot decouple estimate of $P(x|h)$ from $P(y|h)$
Problems with Partial Data

- In general, likelihood over iid data:
  \[ L(\theta : D) = \prod_m (\sum_{h[m]} P(o[m], h[m] | \theta) \]

- Involves *evaluating likelihood function* ... can be arbitrary BN inference \( \Rightarrow \) INTRACTABLE!

- More bad news: Likelihood function is...
  - *not* unimodal
  - does *not* have closed form representation
  - is *not* decomposable as product of likelihoods for diff parameters
Learning Belief Net Parameters from Partial Data

- Framework
  - Why is the data missing? ... MCAR, MAR, ...
  - Why more challenging?

- Approaches
  - Gradient Descent
  - EM
  - Gibbs
Gradient Ascent

- Want to maximize likelihood
  - $\theta^{(\text{MLE})} = \arg\max_{\theta} L(\theta : D)$

- Unfortunately...
  - $L(\theta : D)$ is nasty, non-linear, multimodal fn

- So...
  - Gradient-Ascent
  - ... 1st-order Taylor series

\[ f_{\text{obj}}(\theta^{-}) \approx f_{\text{obj}}(\theta^{0}) + (\theta - \theta^{0})^{T} \nabla f_{\text{obj}}(\theta^{0}) \]

Need derivative!
Gradient Ascent [APN]

View: \( P_\Theta(S) = P(S | \Theta, G) \) as fn of \( \Theta \)

\[
\frac{\partial \ln P_\Theta(S)}{\partial \theta_{ijk}} = \sum_{\ell=1}^{m} \frac{\partial \ln P_\Theta(c_\ell)}{\partial \theta_{ijk}} = \sum_{\ell=1}^{m} \frac{\partial P_\Theta(c_\ell)/\partial \theta_{ijk}}{P_\Theta(c_\ell)}
\]

\[
\frac{\partial P_\Theta(c_\ell)/\partial \theta_{ijk}}{P_\Theta(c_\ell)} = \frac{P_\Theta(c_\ell | v_{ik}, p_{aij}) P_\Theta(p_{aij})}{P_\Theta(c_\ell)} = \frac{P_\Theta(v_{ik}, p_{aij} | c_\ell)}{\theta_{ijk}}
\]

Alg: fn Basic-APN( BN = \( \langle G, \Theta \rangle, S \) ): (modified) CPTables
- inputs: BN, a Belief net with CPT entries
  D, a set of data cases
- repeat until \( \Delta \Theta \approx 0 \)
  \( \Delta \Theta \leftarrow 0 \)
  for each \( c_r \in S \)
    - Set evidence in BN to \( c_r \)
    - For each \( X_i \) w/ value \( v_{ik} \), parents w/ \( j^{th} \) value \( p_{aij} \)
      \( \Delta \Theta_{ijk} += P(v_{ik}, p_{aij} | c_r) / \theta_{ijk} \)
    - \( \Theta += \alpha \Delta \Theta \)
    - \( \Theta \leftarrow \text{project } \Theta \text{ onto constraint region} \)
- return(\( \Theta \))

Note: Computed \( P(v_{ik}, p_{aij} | c_r) \) to deal with \( c_r \)
\( \Rightarrow \) can “piggyback” computation
Issues with Gradient Ascent

- Lots of Tricks for efficient ascent
  - Line Search
  - Conjugate Gradient
  - ...
  - Take Cmput551, or optimization

- Constraints
  - $\Theta_{ijk} \in [0, 1]$
  - $\sum_r \Theta_{ijr} = 1$
  - But ... $\Theta_{ijk} + = \alpha \Delta \Theta_{ijk}$ could violate
  - Use $\Theta_{ijk} = \exp(\lambda_{ijk}) / \sum_r \exp(\lambda_{ijr})$
  - Find best $\lambda_{ijk}$ ... unconstrained ...
Expectation Maximization (EM)

- EM is designed to find most likely $\theta$, given incomplete data!
- Recall simple Maximization needs counts:
  $\#(+x, +y)$, ...
- But is instance $[?, +y]$ in
  ... $\#(+x, +y)$? ... $\#(-x, +y)$?
- Why not put it in BOTH... fractionally?
  - What is weight of $\#(+x, +y)$?
  - $P_\theta(+x \mid +y)$, based on current value of $\theta$
EM Approach – E Step

Sample $S = \begin{pmatrix}
0 & 0 & 1 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{pmatrix}$

Set $S^{(0)} = \begin{pmatrix}
0 & 0 & 1 & 1.0 \\
0 & 1 & 0 & 0.7 \\
1 & 0 & 0 & 0.3 \\
0 & 0 & 1 & 0.1 \\
0 & 1 & 1 & 0.9 \\
0 & 0 & 1 & 0.7 \times 0.1 \\
0 & 1 & 1 & 0.7 \times 0.9 \\
1 & 0 & 1 & 0.3 \times 0.1 \\
1 & 1 & 1 & 0.3 \times 0.9
\end{pmatrix}$
EM Approach – M Step

- Use fractional data:

\[ S^{(0)} = \]

- New estimates:

\[ \hat{\theta}_{+c}^{(1)} = \frac{\#(a_+c)}{\#(c)} = \frac{1.0 + (1.0) + (1.0)}{4} = 0.75 \]

\[ \hat{\theta}_{+a|+c}^{(1)} = \frac{(0.3 \times 0.1) + (0.3 \times 0.9)}{1 + 0.1 + 0.9 + (0.7 \times 0.1) + (0.7 \times 0.9) + (0.3 \times 0.1) + (0.3 \times 0.9)} = 0.1 \]

\[ \hat{\theta}_{+b|+c}^{(1)} = \frac{(0.3 \times 0.1) + (0.3 \times 0.9)}{0.1 + (0.7 \times 0.9) + (0.3 \times 0.9)} = 0.33 \]
EM Approach – M Step

• Use fractional data:

\[ S^{(0)} = \]

\[ \begin{array}{ccc}
A & B & C \\
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 1 & 0 \\
0 & 0 & 1 \\
0 & 1 & 1 \\
0 & 0 & 1 \\
0 & 1 & 1 \\
1 & 0 & 1 \\
1 & 1 & 1 \\
\end{array} \]

• New estimates:

\[ \hat{\theta}_{+c}^{(1)} = \frac{\#(+c)}{\#(\{\})} = \frac{(0.3\times0.1)+(0.3\times0.9)}{1+0.1+0.9+(0.7\times0.1)+(0.7\times0.9)+(0.3\times0.1)+(0.3\times0.9)} = 0.1 \]

\[ \hat{\theta}_{+a|+c}^{(1)} = \frac{\#(+a,+c)}{\#(+c)} = \frac{(a\times c)}{1+0.1+0.9+(0.7\times0.1)+(0.7\times0.9)+(0.3\times0.1)+(0.3\times0.9)} = \]

Then

- E-step: re-estimate distributions over the missing values based on these new \( \theta^{(1)} \) values
- M-step: compute new \( \theta^{(2)} \) values, using statistics based on these new distribution
EM Steps

**E step:**
- Given parameters $\theta$,
- find probability of each missing value
  - ... so get $E[ N_{ijk} ]$

**M step:**
- Given completed (fractional) data
  - based on $E[ N_{ijk} ]$
- find max-likely parameters $\theta$
EM Approach

- Assign $\Theta^{(0)} = \{\theta_{ijk}^{(0)}\}$ randomly.

- Iteratively, $k = 0, \ldots$
  
  **E step:** Compute EXPECTED value of $N_{ijk}$, given $\langle G, \Theta^k \rangle$
  
  $$\tilde{N}_{ijk} = E_{P(x|S, G)}(N_{ijk}) = \sum_{\alpha_i \in S} P(x_i^k, p_i^j | c_\ell, \Theta^k, S)$$

  **M step:** Update values of $\Theta^{k+1}$, based on $\tilde{N}_{ijk}$
  
  $$\theta_{ijk}^{k+1} = \frac{\tilde{N}_{ijk} + 0}{\sum_{k=1}^n (\tilde{N}_{ijk} + 0)}$$

  ... until $\|\Theta^{k+1} - \Theta^k\| \approx 0$.

- Return $\Theta^k$

1. This is ML computation; MAP is similar
   
   "Q" $\rightarrow$ $\alpha_{ijk}$

2. Finds local optimum

3. Used for HMM

4. Views each tuple with $k$ "s"s as $O(2^k)$ partial-tuples
Facts about EM ...

- Always converges
- Always improve likelihood
  - \[ L( \theta^{(t+1)} : D ) > L( \theta^{(t)} : D ) \]
  - ... except at stationary points...

- For CPtable for Belief net:
  - Need to perform general BN inference
  - Use Click-tree or ClusterGraph
    - ... just needs one pass
    - (as \( N_{ijk} \) depends on node+parents)
Gibbs Sampling

- Let $S^{(0)}$ be COMPLETED version of $S$, randomly filling-in each missing $c_{ij}$
  
  Let $d_{ij}^{(0)} = c_{ij}$
  
  If $c_{ij} = \ast$, then $d_{ij}^{(0)} = \text{Random}[\text{Domain}(X_i)]$

- For $k = 0..$
  
  - Compute $\Theta^{(k)}$ from $S^{(k)}$ [frequencies]
  
  - Form $S^{(k+1)}$ by...
    * $d_{ij}^{k+1} = c_{ij}$
    * If $c_{ij} = \ast$ then
      
      Let $d_{ij}^{k+1}$ be random value for $X_i$,
      based on current distr $\Theta^k$ over $Z - X_i$

- Return average of these $\Theta^{(k)}$s

Note: As $\Theta^{(k)}$ based on COMPLETE DATA $S^{(k)}$

$\Rightarrow \Theta^{(k)}$ can be computed efficiently!

“Multiple Imputation”
Gibbs Sampling – Example

Guess initial values $\theta^0$

<table>
<thead>
<tr>
<th></th>
<th>0.8</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.9</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

A | B | C
---|---|---
\(\theta_{+a|+c}\) | \(\theta_{-a|+c}\) | \(\theta_{+b|+c}\) | \(\theta_{-b|+c}\)
\(\theta_{+a|-c}\) | \(\theta_{-a|-c}\) | \(\theta_{+b|-c}\) | \(\theta_{-b|-c}\)

\[S^{(1)} = \]

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

New

- Flip 0.3-coin:
- Flip 0.9-coin:
- Flip 0.8-coin:
- Flip 0.9-coin:

Then
- Use $S^{(1)}$ to get new $\theta^{(2)}$ parameters
- Form new $S^{(2)}$ by drawing new values from $\theta^{(2)}$
Gibbs Sampling (con't)

- Algorithm: Repeat
  - Given COMPLETE data $S^{(i)}$, compute new ML values for $\{\theta^{(i+1)}_{ijk}\}$
  - Using NEW parameters, impute (new) missing values $S^{(i+1)}$

- Q: What to return?
  AVERAGE over separated $\Theta^{(i)}$'s
  - eg, $\Theta^{(500)}$, $\Theta^{(600)}$, $\Theta^{(700)}$, ...

- Q: When to stop?
  When distribution over $\Theta^{(i)}$s have converged

- Comparison: Gibbs vs EM
  - + EM "splits" each instance
    ...into $2^k$ parts if $k$ "s
  - – EM knows when it is done, and what to return
General Issues

- All alg’s are heuristic...
- Starting values $\theta$
- Stopping criteria
- Escaping local maxima

- So far, trying to optimize likelihood. Could try to optimize APPROXIMATION to likelihood...
Gaussian Approximation

(Assumes large amounts of data)

- Let $g(\Theta) = \log [P(S|\Theta, G) P(\Theta|G)]$
  Let $\hat{\Theta}_{BN} = \arg\max_{\Theta} g(\Theta)$
  ...also maximizes $P(\Theta|G, S)$.

  With many samples,
  $\hat{\Theta}_{BN} \approx \arg\max_{\Theta}\{P(S|\Theta, G)\}$

- $g(\Theta) \approx g(\hat{\Theta}_{BN}) - \frac{1}{2}(\Theta - \hat{\Theta}_{BN}) A(\Theta - \hat{\Theta}_{BN})^t$
  (2nd-order Taylor; $A$ is neg. Hessian of $g(\hat{\Theta}_{BN})$)

  So...
  $P(\Theta|G, S) \propto P(S|\Theta, G) P(\Theta|G)$
  $\approx P(S|\hat{\Theta}_{BN}, G) P(\hat{\Theta}_{BN}|G) e^{((\Theta - \hat{\Theta}_{BN}) A(\Theta - \hat{\Theta}_{BN})^t)}$

  ...which looks (approximately) Gaussian!

- Now use
  \underline{gradient descent or EM}

Note: Can often use values computed during inference!
Summary

- Missingness: MCAR vs MAR
- Approaches
  - Gradient Ascent
  - EM
  - Gibbs sampling
    - Multiple imputation

Note covered: Bayesian methods
  - MCMC, Variational, Particles, ...