Lecture 17: Parameter Learning with Missing Data

Machine Learning
Queens College
Today

• Parameter learning with missing data
Incomplete Data

- Hidden variables
- Missing values

- Challenges
  - Foundational – is the learning task well defined?
  - Computational – how can we learn with missing data?
Treating Missing Data

• How Should we treat missing data?

• **Case I:** A coin is tossed on a table, occasionally it drops and measurements are not taken
  – Sample sequence: H,T,?,?,?,T,?,H
  – Treat missing data by ignoring it

• **Case II:** A coin is tossed, but only heads are reported
  – Sample sequence: H,?,?,?,?,H,?,?,H
  – Treat missing data by filling it with Tails
Treating Missing Data

• When can we ignore the missing data mechanism and focus only on the likelihood?
  – For every $X_i$, $\text{Ind}(X_i; O_{X_i})$
  – Missing at Random (MAR) is sufficient
    • The probability that the value of $X_i$ is missing is independent of its actual value given other observed values
Hidden (Latent) Variables

- Attempt to learn a model with hidden variables
  - In this case, MAR always holds (variable is always missing)

- Why should we care about unobserved variables?

17 parameters

59 parameters
Hidden (Latent) Variables

- Hidden variables also appear in clustering.

- **Naïve Bayes** model:
  - Class variable is hidden.
  - Observed attributes are independent given the class.

![Diagram of Hidden (Latent) Variables]

**Diagram Notes:**
- Hidden
- Observed
- Possible missing values
Likelihood for Complete Data

Input Data:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>x₀</td>
<td>y₀</td>
</tr>
<tr>
<td>x₀</td>
<td>y₁</td>
</tr>
<tr>
<td>x₁</td>
<td>y₀</td>
</tr>
</tbody>
</table>

Likelihood:

\[
L(D : \theta) = P(x[1], y[1]) \cdot P(x[2], y[2]) \cdot P(x[3], y[3]) \\
= P(x^0, y^0) \cdot P(x^0, y^1) \cdot P(x^1, y^0) \\
= \theta_{x^0} \cdot \theta_{y^0|x^0} \cdot \theta_{x^0} \cdot \theta_{y^1|x^0} \cdot \theta_{x^1} \cdot \theta_{y^0|x^1} \\
= \left( \theta_{x^0} \cdot \theta_{x^0} \cdot \theta_{x^1} \right) \cdot \left( \theta_{y^0|x^0} \cdot \theta_{y^1|x^0} \right) \cdot \left( \theta_{y^0|x^1} \right)
\]

- Likelihood decomposes by variables
- Likelihood decomposes within CPDs
Likelihood for Incomplete Data

Input Data:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>$y^0$</td>
</tr>
<tr>
<td>$x^0$</td>
<td>$y^1$</td>
</tr>
<tr>
<td>?</td>
<td>$y^0$</td>
</tr>
</tbody>
</table>

Likelihood:

\[
P(Y|X)
\]

\[
\begin{array}{c|c|c}
X & P(Y|X) & \theta_{y0|x0} & \theta_{y1|x0} \\
\hline
x^0 & y^0 & \theta_{y0|x0} & \theta_{y1|x0} \\
\hline
x^1 & y^1 & \theta_{y0|x1} & \theta_{y1|x1} \\
\end{array}
\]
Likelihood for Incomplete Data

Input Data:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>y^0</td>
</tr>
<tr>
<td>x^0</td>
<td>y^1</td>
</tr>
<tr>
<td>?</td>
<td>y^0</td>
</tr>
</tbody>
</table>

Likelihood:

\[ L(D; \theta) = P(y^0) \cdot P(x^0, y^1) \cdot P(y^0) \]
\[ = \left( \sum_{x \in X} P(x, y^0) \right) \cdot P(x^0, y^1) \cdot \left( \sum_{x \in X} P(x, y^0) \right) \]
\[ = \left( \theta_{x^0} \cdot \theta_{y^0|x^0} + \theta_{x^1} \cdot \theta_{y^0|x^1} \right) \cdot \theta_{x^0} \cdot \theta_{y^1|x^0} \cdot \left( \theta_{x^0} \cdot \theta_{y^0|x^0} + \theta_{x^1} \cdot \theta_{y^0|x^1} \right) \]
\[ = \left( \theta_{x^0} \cdot \theta_{y^0|x^0} + \theta_{x^1} \cdot \theta_{y^0|x^1} \right)^2 \cdot \theta_{x^0} \cdot \theta_{y^1|x^0} \]

- Likelihood does not decompose by variables
- Likelihood does not decompose within CPDs
- Computing likelihood per instance requires inference!
Identifiability

- Likelihood can have multiple global maxima

- Example:
  - We can rename the values of the hidden variable H
  - If H has two values, likelihood has two global maxima

- With many hidden variables, there can be an exponential number of global maxima

- Multiple local and global maxima can also occur with missing data (not only hidden variables)
MLE from Incomplete Data

- Nonlinear optimization problem

---

**Expectation Maximization (EM):**
- Use “current point” to construct alternative function (which is “nice”)
- Guaranty: maximum of new function has better score than current point
Expectation Maximization (EM)

• Tailored algorithm for optimizing likelihood functions

• **Intuition**
  – Parameter estimation is easy given complete data
  – Computing probability of missing data is “easy” (=inference) given parameters

• **Strategy**
  – Pick a starting point for parameters
  – “Complete” the data using current parameters
  – Estimate parameters relative to data completion
  – Iterate
  – Procedure guaranteed to improve at each iteration
Expectation Maximization (EM)

- Initialize parameters to $\theta^0$

- **Expectation (E-step):**
  - For each data case $o[m]$ and each family $X,U$ compute $P(X,U \mid o[m], \theta^i)$
  - Compute the expected counts for each $x,u$
    \[
    \overline{M}_{\theta^i}[x,u] = \sum_m P(x,u \mid o[m], \theta^i)
    \]

- **Maximization (M-step):**
  - Treat the expected sufficient statistics as observed and set the parameters to the MLE with respect to the expected counts
    \[
    \theta_{x|u}^{i+1} = \frac{\overline{M}_{\theta^i}[x,u]}{\overline{M}_{\theta^i}[u]}
    \]
Expectation Maximization (EM)

Training data

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>y^0</td>
</tr>
<tr>
<td>x^0</td>
<td>y^1</td>
</tr>
<tr>
<td>?</td>
<td>y^0</td>
</tr>
</tbody>
</table>

Initial network

E-Step (inference)

Expected counts

<table>
<thead>
<tr>
<th>N(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N(X,Y)</td>
</tr>
</tbody>
</table>

M-Step (reparameterize)

Iterate
EM Example

Input Data:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>

| X | P(Y|X) |
|---|-------|
| x^0 | y^0   |
| x^1 | y^1   |
Expectation Maximization (EM)

- **Formal Guarantees:**
  - $L(D; \Theta_{i+1}) \geq L(D; \Theta_i)$
    - Each iteration improves the likelihood
  - If $\Theta_{i+1} = \Theta_i$, then $\Theta_i$ is a stationary point of $L(D; \Theta)$
    - Usually, this means a local maximum

- **Main cost:**
  - Computations of expected counts in E-Step
  - Requires inference for each instance in training set
EM – Practical Considerations

• **Initial parameters**
  – Highly sensitive to starting parameters
  – Choose randomly
  – Choose by guessing from another source

• **Stopping criteria**
  – Small change in data likelihood
  – Small change in parameters

• **Avoiding bad local maxima**
  – Multiple restarts
  – Early pruning of unpromising starting points
Next Time

• Reinforcement learning