CS 6347

Lecture 15

Expectation Maximization
Unobserved Variables

- **Latent or hidden variables** in the model are never observed
  - We may or may not be interested in their values, but their existence is crucial to the model
- Some observations in a particular sample may be **missing**
  - Missing information on surveys or medical records (quite common)
  - We may need to model how the variables are missing
Hidden Markov Models

\[ p(x_1, \ldots, x_T, y_1, \ldots, y_T) = p(y_1)p(x_1|y_1) \prod_{t} p(y_t|y_{t-1})p(x_t|y_t) \]

- \( X \)’s are observed variables, \( Y \)’s are latent

- Example: \( X \) variables correspond sizes of tree growth rings for one year, the \( Y \) variables correspond to average temperature
Missing Data

• Data can be missing from the model in many different ways

  – Missing completely at random: the probability that a data item is missing is independent of the observed data and the other missing data

  – Missing at random: the probability that a data item is missing can depend on the observed data

  – Missing not at random: the probability that a data item is missing can depend on the observed data and the other missing data
Handling Missing Data

• Discard all incomplete observations
  – Can introduce bias

• Imputation: actual values are substituted for missing values so that all of the data is fully observed
  – E.g., find the most probable assignments for the missing data and substitute them in (not possible if the model is unknown)
  – Use the sample mean/mode

• Explicitly model the missing data
  – For example, could expand the state space
  – The most sensible solution, but may be non-trivial if we don’t know how/why the data is missing
Modelling Missing Data

• Add additional binary variable $m_i$ to the model for each possible observed variable $x_i$ that indicates whether or not that variable is observed

$$p(x_{obs}, x_{mis}, m) = p(m|x_{obs}, x_{mis})p(x_{obs}, x_{mis})$$
Add additional binary variable $m_i$ to the model for each possible observed variable $x_i$ that indicates whether or not that variable is observed.

\[
p(x_{obs}, x_{mis}, m) = p(m|x_{obs}, x_{mis})p(x_{obs}, x_{mis})
\]

Explicit model of the missing data
(missing not at random)
Add additional binary variable $m_i$ to the model for each possible observed variable $x_i$ that indicates whether or not that variable is observed:

$$p(x_{obs}, x_{mis}, m) = p(m|x_{obs})p(x_{obs}, x_{mis})$$

Missing at random
• Add additional binary variable $m_i$ to the model for each possible observed variable $x_i$ that indicates whether or not that variable is observed

\[ p(x_{obs}, x_{mis}, m) = p(m)p(x_{obs}, x_{mis}) \]

Missing completely at random
Modelling Missing Data

• Add additional binary variable $m_i$ to the model for each possible observed variable $x_i$ that indicates whether or not that variable is observed.

$$p(x_{obs}, x_{mis}, m) = p(m)p(x_{obs}, x_{mis})$$

Missing completely at random

How can you model latent variables in this framework?
Learning with Missing Data

• In order to design learning algorithms for models with missing data, we will make two assumptions
  
  – The data is missing at random
  
  – The model parameters corresponding to the missing data ($\delta$) are separate from the model parameters of the observed data ($\theta$)

• That is

$$p(x_{obs}, m|\theta, \delta) = p(m|x_{obs}, \delta)p(x_{obs}|\theta)$$
Learning with Missing Data

\[ p(x_{obs}, m|\theta, \delta) = p(m|x_{obs}, \delta)p(x_{obs}|\theta) \]

- Under the previous assumptions, the log-likelihood of samples \((x^1, m^1), \ldots, (x^K, m^K)\) is equal to

\[
l(\theta, \delta) = \sum_{k=1}^{K} \log p(m^k|x^k_{obs}, \delta) + \sum_{k=1}^{K} \log \sum_{x_{mis_k}} p(x^k_{obs_k}, x_{mis_k}|\theta)\]
Learning with Missing Data

\[
p(x_{obs}, m|\theta, \delta) = p(m|x_{obs}, \delta)p(x_{obs}|\theta)
\]

- Under the previous assumptions, the log-likelihood of samples \((x^1, m^1), \ldots, (x^K, m^K)\) is equal to

\[
l(\theta, \delta) = \sum_{k=1}^{K} \log p(m^k|x_{obs}^k, \delta) + \sum_{k=1}^{K} \log \sum_{x_{mis}^k} p(x_{obs}^k, x_{mis}^k|\theta)
\]

Separable in \(\theta\) and \(\delta\), so if we don’t care about \(\delta\), then we only have to maximize the second term over \(\theta\)
Learning with Missing Data

\[ l(\theta) = \sum_{k=1}^{K} \log \sum_{x_{mis_k}} p(x_{obs_k}^k, x_{mis_k} | \theta) \]

- This is NOT a concave function of \( \theta \)
  - In the worst case, could have a different local maximum for each possible value of the missing data
  - No longer have a closed form solution, even in the case of Bayesian networks
Expectation Maximization

- The expectation-maximization algorithm (EM) is a method to find a local maximum or a saddle point of the log-likelihood with missing data.

- Basic idea:

\[
l(\theta) = \sum_{k=1}^{K} \log \sum_{x_{mis_k}} p(x_{obs_k}, x_{mis_k} | \theta) = \sum_{k=1}^{K} \log \sum_{x_{mis_k}} q_k(x_{mis_k}) \cdot \frac{p(x_{obs_k}, x_{mis_k} | \theta)}{q_k(x_{mis})} \geq \sum_{k=1}^{K} \sum_{x_{mis_k}} q_k(x_{mis_k}) \log \frac{p(x_{obs_k}, x_{mis_k} | \theta)}{q_k(x_{mis_k})}
\]
Expectation Maximization

\[ F(q, \theta) \equiv \sum_{k=1}^{K} \sum_{x_{mis_k}} q_k(x_{mis_k}) \log \frac{p(x^k_{obs_k}, x_{mis_k} \mid \theta)}{q_k(x_{mis_k})} \]

• Maximizing \( F \) is equivalent to the maximizing the log-likelihood

• Could maximize it using coordinate ascent

\[
q^{t+1} = \arg \max_{q_{1, \ldots, q_K}} F(q, \theta^t)
\]

\[
\theta^{t+1} = \arg \max_{\theta} F(q^{t+1}, \theta)
\]
Expectation Maximization

\[
\sum_{x_{mis_k}} q_k(x_{mis_k}) \log \frac{p(x^k_{obs_k}, x_{mis_k} | \theta)}{q_k(x_{mis_k})}
\]

- This is just \(-d \left( q_k \| p(x^k_{obs_k}, \cdot | \theta) \right)\)
- Maximized when \(q_k(x_{mis_k}) = p(x_{mis_k} | x^k_{obs_k}, \theta)\)
- Can reformulate the EM algorithm as

\[
\theta^{t+1} = \arg\max_{\theta} \sum_{k=1}^{K} \sum_{x_{mis_k}} p(x_{mis_k} | x^k_{obs_k}, \theta^t) \log p(x^k_{obs_k}, x_{mis_k} | \theta)
\]
An Example: Bayesian Networks

• Recall that MLE for Bayesian networks without latent variables yielded

\[ \theta_{x_i|x_{\text{parents}(i)}} = \frac{N_{x_i,x_{\text{parents}(i)}}}{\sum_{x_i'} N_{x_i',x_{\text{parents}(i)}}} \]

• Let’s suppose that we are given observations from a Bayesian network in which one of the variables is hidden

— What do the iterations of the EM algorithm look like?

(on board)