Graphical Models

Lecture 19:
Partially Observed Data – Parameter Estimation

Andrew McCallum
mccallum@cs.umass.edu

Thanks to Noah Smith and Carlos Guestrin for some slide materials.
Partially Observed, Incomplete Data

• Until now, we have always assumed during learning that the data are completely observed: \((x^{(1)}, x^{(2)}, ..., x^{(T)})\).

• Today we consider learning when the data are incomplete.
  – Missing values
  – Truly hidden variables
Example

• Two binary variables, $X$ and $Y$.
• Three binomial distributions: $\theta_X$, $\theta_{Y|X=1}$, $\theta_{Y|X=0}$.
• Let $\#\{\ldots\}$ be a sufficient statistic function that counts values in the data.

\[
L(\theta_X, \theta_{Y|X=1}, \theta_{Y|X=0}) = (\theta_X)^{\#\{1,*\}} \times (1 - \theta_X)^{\#\{0,*\}} \times \\
(\theta_{Y|X=1})^{\#\{1,1\}} \times (1 - \theta_{Y|X=1})^{\#\{1,0\}} \times \\
(\theta_{Y|X=0})^{\#\{0,1\}} \times (1 - \theta_{Y|X=0})^{\#\{0,0\}}
\]
Example

- log L is concave, with a unique global optimum, and we know we can solve for it in closed form.

\[
L(\theta_X, \theta_Y|X=1, \theta_Y|X=0) = (\theta_X)^{\#\{1,*\}} \times (1 - \theta_X)^{\#\{0,*\}} \times \\
(\theta_Y|X=1)^{\#\{1,1\}} \times (1 - \theta_Y|X=1)^{\#\{1,0\}} \times \\
(\theta_Y|X=0)^{\#\{0,1\}} \times (1 - \theta_Y|X=0)^{\#\{0,0\}}
\]
Example

- Consider observation of one additional example that is *incomplete*: \((X = ?, Y = 1)\).

- Likelihood now has to sum over both assignments of the unknown variable.

\[
L(\theta_X, \theta_Y|X=1, \theta_Y|X=0) = \left( \theta_X \right)^{\{1,*\}+1} \times (1 - \theta_X)^{\{0,*\}} \times \\
(\theta_Y|X=1)^{\{1,1\}+1} \times (1 - \theta_Y|X=1)^{\{1,0\}} \times \\
(\theta_Y|X=0)^{\{0,1\}} \times (1 - \theta_Y|X=0)^{\{0,0\}} + \\
(\theta_X)^{\{1,*\}} \times (1 - \theta_X)^{\{0,*\}+1} \times \\
(\theta_Y|X=1)^{\{1,1\}} \times (1 - \theta_Y|X=1)^{\{1,0\}} \times \\
(\theta_Y|X=0)^{\{0,1\}+1} \times (1 - \theta_Y|X=0)^{\{0,0\}}
\]
Missing Data

• In general, the likelihood function will now be a summation over all possible assignments to all missing (latent, hidden) variables.

• There could be exponentially many!
  – You shouldn’t be too worried, though: this is really just marginalization, given some evidence.

• Note: every example could have a different set of variables that are observed or hidden.
Effects of Missing Data

\[ L(\theta) = \prod_t P(x^{(t)}_{\text{observed}} | \theta) \]

\[ = \prod_t \sum_{x_{\text{missing}} \in \text{Val}(X^{(t)}_{\text{missing}})} P(x^{(t)}_{\text{observed}}, x_{\text{missing}} | \theta) \]

• Each term in the summation is log-concave (unimodal; there is a single optimal value of \( \theta \)).
• The sum of these terms may be multimodal!
Sum of Concave Terms
Effects of Missing Data

• Likelihood decomposability was really helpful in both MLE and Bayesian estimation when our data were fully observed.
  – Also in structure learning.
  – Recall that this went away when learning Markov networks.
Simple Example

• Consider two binary random variables.

• Step 1: Global parameter independence.

\[ P(\theta) = \prod_i P(\theta_{X_i} | \text{Parents}(X_i)) \]
Simple Example

• Step 2: *Local* parameter independence.

\[
P(\theta_{X_i}|\text{Parents}(X_i)) = \prod_{u \in \text{Val}(\text{Parents}(X_i))} P(\theta_{X_i}|\text{Parents}(X_i)=u)
\]
Simple Example

• Does local parameter independence cause problems for global parameter independence?
Simple Example

- Does local parameter independence cause problems for global parameter independence?
Simple Example

- Good news: given $X_m$, one of the edges becomes inactive.
  - Context-sensitive independence!

![Diagram showing the relationship between $X_m$, $Y_m$, and $M$ with variables $\theta_{y|x=1}$, $\theta_{y|x=0}$, $\theta_x$, $\theta_{y|x=1}$, and $\theta_{y|x=0}$]
Simple Example

- Global *and* local parameter independence hold.

\[
P(\theta) = \prod_i \prod_{u \in \text{Val}(\text{Parents}(X_i))} P(\theta_{X_i|\text{Parents}(X_i)=u})
\]
Local Decomposability

- But now, suppose $X_m$ is hidden, and (for simplicity) that $\theta_X$ is known.
  - V-structure! $X_m$ depends on parameters and vice versa.
  - Context-sensitive independence is lost; the two $\theta_{Y|X}$ distributions now depend on each other because of $X$. 
Global Decomposability

• Also lost, since estimates of all parameters depend on how we “reconstruct” H for each example.

\[
L(\theta) = \prod_{x,y} \left( \sum_h P(h) P(x \mid h) P(y \mid h) \right)^{\#\{x,y\}}
\]
In General

• More missing information implies more active trails.
  – Conditional independence assumptions weaken.
• Once data go missing, we lose the closed-form solution, the global concavity of log L, and decomposition.
• Learning just got harder.
Some Other Issues

- Sometimes data are missing at random, and the probability of a random variable’s value being missing is independent of the value itself.

- If not, then things get harder, because the observation *pattern* may tell us something about the missing data.
  - See K&F 19.1.

- Often the data are of one kind (all missing the same parts) or two kinds (some complete data, some incomplete data all missing the same parts).
Naïve Bayes Model
Clustering

$X_1 \rightarrow C \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow \ldots \rightarrow X_n$
Identifiability

- Is there a single parameter setting that maximizes likelihood?

- **Identifiability**: changing the parameters changes the likelihood.
  - Single global maximum.

- Local identifiability: within a small neighborhood, changing the parameters changes the likelihood.
  - But there could be different models in different parts of the parameter space that achieve equal likelihood.
Dealing with Missing Data is Hard

- All kinds of challenges.
- This doesn’t mean we shouldn’t attempt to do it!
  - Consider the payoff if we get it to work.

- We’ll consider two approaches to optimizing log L with respect to the parameters:
  - gradient ascent (and related)
  - expectation-maximization (EM)
Log-Likelihood Objective

\[
\theta_{\text{MLE}} = \arg \max_\theta \sum_t \log P(x^{(t)}_{\text{observed}} | \theta)
\]

\[
= \arg \max_\theta \sum_t \log \sum_{x^{(t)}_{\text{missing}} \in \text{Val}(X^{(t)}_{\text{missing}})} P(x^{(t)}_{\text{observed}}, x^{(t)}_{\text{missing}} | \theta)
\]

• Taking the derivative with respect to one parameter, \( P(x | u) = \theta_{x|u} \) (assume nonzero) ...
First Derivative of Marginal w.r.t. A Parameter

\[
\frac{\partial P(x_{\text{observed}})}{\partial \theta_{x_i|u}} = \frac{\partial}{\partial \theta_{x_i|u}} \sum_{x_{\text{missing}}} P(x_{\text{observed}}, x_{\text{missing}})
\]

\[
= \sum_{x_{\text{missing}}} \frac{\partial}{\partial \theta_{x_i|u}} P(x_{\text{observed}}, x_{\text{missing}})
\]

\[
= \sum_{x_{\text{missing}}} \left\{ \begin{array}{ll} 
\frac{P(x_{\text{observed}}, x_{\text{missing}})}{\theta_{x_i|u}} & \text{if } x_{\text{observed}}, x_{\text{missing}} \text{ are compatible with } x \text{ and } u \\
0 & \text{otherwise} \end{array} \right.
\]

\[
= \frac{1}{\theta_{x_i|u}} \sum_{x_{\text{missing}} \text{ compatible}(x_{\text{missing}}; x, u)} P(x_{\text{observed}}, x_{\text{missing}})
\]

The division is really just a shorthand for dividing out the parameter; if \(\theta_{x|u} = 0\), the first derivative just involves multiplying the other probabilities together.
First Derivative of log L w.r.t. $\theta_{x|u}$

\[
\begin{align*}
\theta_{\text{MLE}} &= \arg \max_{\theta} \sum_t \log P(x^{(t)}_{\text{observed}} | \theta) \\
&= \arg \max_{\theta} \sum_t \log \sum_{x_{\text{missing}} \in \text{Val}(X_{\text{missing}}^{(t)})} P(x^{(t)}_{\text{observed}}, x_{\text{missing}} | \theta)
\end{align*}
\]

\[
\begin{align*}
\frac{\partial \log L}{\partial \theta_{x|u}} &= \sum_t \frac{\partial \log P(x^{(t)}_{\text{observed}} | \theta)}{\partial \theta_{x|u}} \\
&= \sum_t \frac{1}{P(x^{(t)}_{\text{observed}} | \theta)} \frac{\partial P(x^{(t)}_{\text{observed}} | \theta)}{\partial \theta_{x|u}} \\
&= \sum_t P(x, u | x^{(t)}_{\text{observed}}, \theta) \theta_{x|u}
\end{align*}
\]
Gradient and Inference

• The gradient depends on (scaled) marginal probabilities.
• This is a key application of inference: for each example, and for each variable $X_i$, we need to infer

$$P(X_i, \text{Parents}(X_i) \mid \mathbf{x}_{\text{observed}})$$

• We can do this with one clique tree calibration per example! (Exploiting family preservation property.)
Gradient Ascent on Log-Likelihood

• Need to do a little work to deal with the constraints on parameters (e.g., summing to one, nonnegativity).
  – Reparameterize, or use Lagrange multipliers.
• If parameters are not multinomials, use the chain rule:

\[
\frac{\partial \log L}{\partial \theta} = \sum_{x,u} \frac{\partial \log L}{\partial P(x \mid u)} \frac{\partial P(x \mid u)}{\partial \theta}
\]
Expectation-Maximization
Expectation-Maximization

- Gradient ascent and friends are general algorithms.
- EM is specifically for maximizing likelihood in the presence of incomplete data!
  - Not a general technique for non-convex problems.
Intuition Behind EM

• If only we had complete data, parameter estimation would be easy!
  – Sufficient statistics.
  – Idea: randomly fill in missing values! (What’s wrong?)

• We are really solving two problems at the same time:
  – estimating parameters
  – hypothesizing missing values
Chicken and Egg

• If we had the complete data, parameter estimation by MLE would be easy.

• If we had the parameters, inferring an assignment for the missing information would be easy: probabilistic inference.
Expectation Maximization

- Initialize parameters: $\theta^{(0)}$

- Repeat:
  - **E step**: Infer distribution over missing values (inference); gather *expected* sufficient statistics. For discrete distributions, this looks like “fractional” counting.
    \[
    \text{ess}^{(i)}(x, u) = \sum_t P(x, u \mid x_{\text{observed}}^{(t)}, \theta^{(i)})
    \]
  - **M step**: Estimate parameters using the complete data distribution just inferred.
    \[
    \theta_x^{(i+1)} = \frac{\text{ess}^{(i)}(x, u)}{\sum_{x'} \text{ess}^{(i)}(x', u)}
    \]
Behavior of EM

• EM works: the log-likelihood will improve on each iteration.

• Easiest way to understand it: coordinate ascent.
  – E step finds missing data distribution to match current value of P: “best Q” (really expected sufficient statistics) for fixed $\theta$.
  – M step: fix Q, find $\theta$. 
M Step: Maximizing a Lower Bound on $\log L$

$$
\log L(\theta) = \sum_t \log \sum_{x_{\text{missing}}} P(x_{\text{observed}}^{(t)}, x_{\text{missing}} | \theta)
$$

$$
= \sum_t \sum_{x_{\text{missing}}} Q(x_{\text{missing}} | x_{\text{observed}}^{(t)}) \frac{P(x_{\text{observed}}^{(t)}, x_{\text{missing}} | \theta)}{Q(x_{\text{missing}} | x_{\text{observed}}^{(t)})}
$$

$$
= \sum_t \log \mathbb{E}_{Q_t}[f_t]
$$

Jensen’s inequality

$$
\geq \sum_t \mathbb{E}_{Q_t}[\log f_t]
$$

$$
= \sum_t \sum_{x_{\text{missing}}} Q(x_{\text{missing}} | x_{\text{observed}}^{(t)}) \log \frac{P(x_{\text{observed}}^{(t)}, x_{\text{missing}} | \theta)}{Q(x_{\text{missing}} | x_{\text{observed}}^{(t)})}
$$

$$
= \sum_t \sum_{x_{\text{missing}}} Q(x_{\text{missing}} | x_{\text{observed}}^{(t)}) \log P(x_{\text{observed}}^{(t)}, x_{\text{missing}} | \theta) + \text{constant}
$$

“complete data” distribution as stand-in for empirical distribution
Local Optima

- Both gradient ascent and EM will converge only on a local optimum.
  - But that’s often pretty good.
  - Some techniques exist to try to avoid this problem, e.g., multiple runs at random initial points.
  - Clever initialization can go a long way.

- Numerical convergence is always an issue.
  - In practice, pick a threshold for relative change in log-likelihood.
  - Training too long can lead to overfitting.
Variations

• For some kinds of priors, we can alter EM to do Bayesian estimation.

• If we use MAP inference instead of marginal inference on the E step, we get “hard” EM.
  – Example: K-means clustering.
  – Sometimes works well; different objective function.

• EM for Markov networks? Yes.