Probabilistic Graphical Models

David Sontag

New York University

Lecture 11, April 12, 2012

Today: learning with partially observed data

- Overview of EM (expectation maximization) algorithm
- Application to mixture models
- Derivation of EM algorithm
- Variational EM
- Application to learning parameters of LDA

Maximum likelihood

- Recall from last week, that the *density estimation* approach to learning leads to *maximizing* the **empirical log-likelihood**

$$\max_{\theta} \quad \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \log p(\mathbf{x}; \theta)$$

- Suppose that our joint distribution is

$$p(\mathbf{X}, \mathbf{Z}; \theta)$$

  where our samples $\mathbf{X}$ are observed and the variables $\mathbf{Z}$ are never observed in $\mathcal{D}$
- That is, $\mathcal{D} = \{(0,1,0,?,?,?),(1,1,1,?,?,?),(1,1,0,?,?,?),\ldots\}$
- Assume that the hidden variables are *missing completely at random* (otherwise, we should explicitly model *why* the values are missing)

Maximum likelihood

- We can still use the same maximum likelihood approach. The objective that we are maximizing is

$$\ell(\theta) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta)$$

- Because of the sum over $\mathbf{z}$, there is no longer a closed-form solution for θ^* in the case of Bayesian networks
- Furthermore, the objective is no longer convex, and potentially can have a different mode for every possible assignment $\mathbf{z}$
- One option is to apply (projected) gradient ascent to reach a local maxima of $\ell(\theta)$

Expectation maximization

- The expectation maximization (EM) algorithm is an alternative approach to reach a local maximum of $\ell(\theta)$
- Particularly useful in settings where there exists a closed form solution for θ^{ML} if we had fully observed data
- For example, in Bayesian networks, we have the closed form ML solution

$$\theta^{ML}_{x_i|\mathbf{x}_{pa(i)}} = \frac{N_{x_i,\mathbf{x}_{pa(i)}}}{\sum_{\hat{x}_i} N_{\hat{x}_i,\mathbf{x}_{pa(i)}}}$$

where $N_{x_i,\mathbf{x}_{pa(i)}}$ is the number of times that the (partial) assignment $x_i, \mathbf{x}_{pa(i)}$ is observed in the training data

Expectation maximization

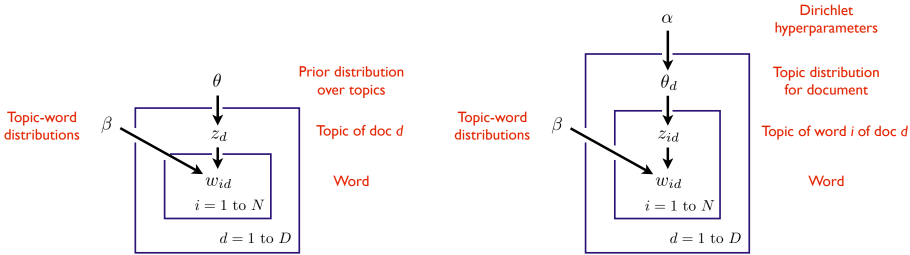
Algorithm is as follows:

1. Write down the **complete log-likelihood** $\log p(\mathbf{x}, \mathbf{z}; \theta)$ in such a way that it is linear in $\mathbf{z}$
2. Initialize θ_0, e.g. at random or using a good first guess
3. Repeat until convergence:

$$\theta_{t+1} = \arg\max_{\theta} \sum_{m=1}^{M} E_{p(\mathbf{z}_m|\mathbf{x}_m;\theta_t)}[\log p(\mathbf{x}_m, \mathbf{Z}; \theta)]$$

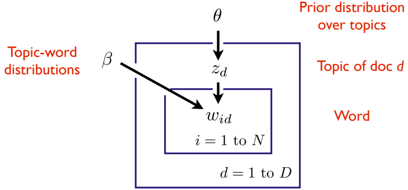
- Notice that $\log p(\mathbf{x}_m, \mathbf{Z}; \theta)$ is a random function because $\mathbf{Z}$ is unknown
- By linearity of expectation, objective decomposes into expectation terms and data terms
- "E" step corresponds to computing the objective (i.e., the **expectations**)
- "M" step corresponds to **maximizing** the objective

Application to mixture models



- Model on left is a **mixture model**
  - Document is generated from a single topic
- Model on right (latent Dirichlet Allocation) is an **admixture model**
  - Document is generated from a distribution over topics

EM for mixture models



- The complete likelihood is $p(\mathbf{w}, \mathbf{Z}; \theta, \beta) = \prod_{d=1}^{D} p(\mathbf{w}_d, Z_d; \theta, \beta)$, where

$$p(\mathbf{w}_d, Z_d; \theta, \beta) = \theta_{Z_d} \prod_{i=1}^{N} \beta_{Z_d, w_{id}}$$

- Trick #1: re-write this as

$$p(\mathbf{w}_d, Z_d; \theta, \beta) = \prod_{k=1}^{K} \theta_k^{1[Z_d=k]} \prod_{i=1}^{N} \prod_{k=1}^{K} \beta_{k, w_{id}}^{1[Z_d=k]}$$

EM for mixture models

- Thus, the complete log-likelihood is:

$$\log p(\mathbf{w}, \mathbf{Z}; \theta, \beta) = \sum_{d=1}^{D} \left(\sum_{k=1}^{K} 1[Z_d = k] \log \theta_k + \sum_{i=1}^{N} \sum_{k=1}^{K} 1[Z_d = k] \log \beta_{k, w_{id}} \right)$$

- In the "E" step, we take the expectation of the complete log-likelihood with respect to $p(\mathbf{z} \mid \mathbf{w}; \theta^t, \beta^t)$, applying linearity of expectation, i.e.

$$E_{p(\mathbf{z} \mid \mathbf{w}; \theta^t, \beta^t)}[\log p(\mathbf{w}, \mathbf{z}; \theta, \beta)] =$$

$$\sum_{d=1}^{D} \left(\sum_{k=1}^{K} p(Z_d = k \mid \mathbf{w}; \theta^t, \beta^t) \log \theta_k + \sum_{i=1}^{N} \sum_{k=1}^{K} p(Z_d = k \mid \mathbf{w}; \theta^t, \beta^t) \log \beta_{k, w_{id}} \right)$$

- In the "M" step, we maximize this with respect to θ and β

EM for mixture models

- Just as with complete data, this maximization can be done in closed form
- First, re-write expected complete log-likelihood from

$$\sum_{d=1}^{D}\left(\sum_{k=1}^{K} p(Z_d = k \mid \mathbf{w}; \theta^t, \beta^t) \log \theta_k + \sum_{i=1}^{N}\sum_{k=1}^{K} p(Z_d = k \mid \mathbf{w}; \theta^t, \beta^t) \log \beta_{k,w_{id}}\right)$$

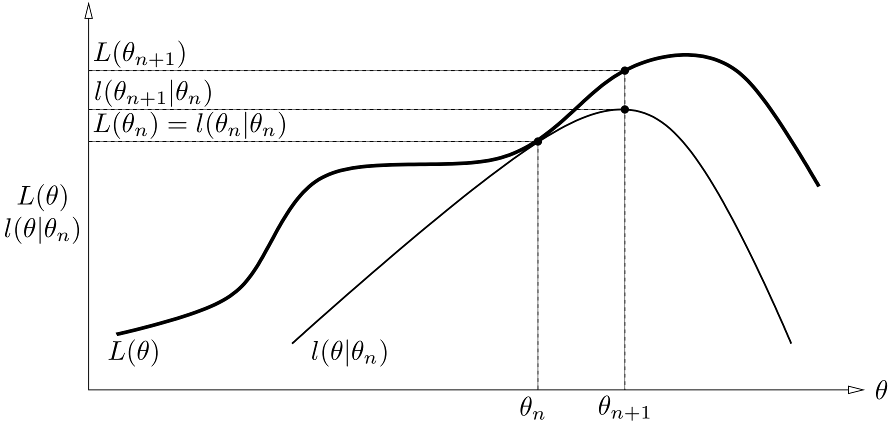
to

$$\sum_{k=1}^{K} \log \theta_k \sum_{d=1}^{D} p(Z_d = k \mid \mathbf{w}_d; \theta^t, \beta^t) + \sum_{k=1}^{K}\sum_{w=1}^{W} \log \beta_{k,w} \sum_{d=1}^{D} N_{dw} p(Z_d = k \mid \mathbf{w}_d; \theta^t, \beta^t)$$

- We then have that

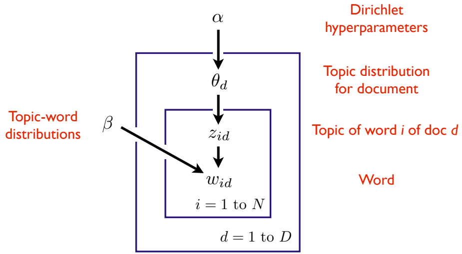
$$\theta_k^{t+1} = \frac{\sum_{d=1}^{D} p(Z_d = k \mid \mathbf{w}_d; \theta^t, \beta^t)}{\sum_{\hat{k}=1}^{K}\sum_{d=1}^{D} p(Z_d = \hat{k} \mid \mathbf{w}_d; \theta^t, \beta^t)}$$

Derivation of EM algorithm



(Figure from tutorial by Sean Borman)

Application to latent Dirichlet Allocation



- Parameters are α and β
- Both θ_d and $\mathbf{z}_d$ are unobserved
- The difficulty here is that **inference** is intractable
- Could use Monte carlo methods to approximate the expectations

Variational EM

- Mean-field is ideally suited for this type of approximate inference together with learning
- Use the variational distribution

$$q(\theta_d, \mathbf{z}_d | \gamma_d, \phi_d) = q(\theta_d \mid \gamma_d) \prod_{n=1}^{N} q(z_n \mid \phi_{dn})$$

- We then lower bound the log-likelihood using Jensen's inequality:

$$\begin{aligned}
\log p(\mathbf{w} \mid \alpha, \beta) &= \sum_d \log \int \sum_{\mathbf{z}_d} p(\theta_d, \mathbf{z}_d, \mathbf{w}_d \mid \alpha, \beta) d\theta_d \\
&= \sum_d \log \int \sum_{\mathbf{z}_d} \frac{p(\theta_d, \mathbf{z}_d, \mathbf{w}_d \mid \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta_d \\
&\geq \sum_d E_q[\log p(\theta_d, \mathbf{z}_d, \mathbf{w}_d \mid \alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})].
\end{aligned}$$

- Finally, we maximize the lower bound with respect to α, β, and q.