

## COS 424: Interacting with Data

Lecturer: David Blei  
Scribe: Yuhui Luo

Lecture # 11  
March 11, 2008

### 1 Basics of the EM Algorithm

The EM algorithm is a general purpose algorithm for finding the maximum likelihood estimate in latent variable models. In the E-Step, we "fill in" the latent variables using the posterior, and in the M-Step, we maximize the expected complete log likelihood with respect to the complete posterior distribution.

Let  $D \triangleq (x_1, \dots, x_N)$  be the observed data, and let  $Z \triangleq$  hidden random variables. (Note: We are not committing to any particular model.)

Now, let  $\theta \triangleq$  the model parameters. Then:

$$\begin{aligned}\hat{\theta} &= \operatorname{argmax}_{\theta} \log p(x, z|\theta) \\ &= \operatorname{argmax}_{\theta} \log p(z|\theta) + \log p(x|z, \theta).\end{aligned}$$

The expression being maximized on the last line is known as the complete log likelihood. In the latent setting:

$$\hat{\theta} = \operatorname{argmax}_{\theta} \sum_z p(x|\theta)p(x|z, \theta)$$

### 2 Jensen's Inequality

Jensen's inequality is a general result in convexity. It states that for a convex function  $f$ , if  $\lambda \in [0, 1]$ , then:

$$\lambda f(x) + (1 - \lambda)f(y) \geq f(\lambda x + (1 - \lambda)y)$$

This is illustrated by Figure 1 in  $\mathbb{R}^2$ . The points between  $x$  and  $y$  can be represented as  $\lambda x + (1 - \lambda)y$ . Clearly, the red line representing  $\lambda f(x) + (1 - \lambda)f(y)$  will always be larger than the function evaluated at any of the points between  $x$  and  $y$ .

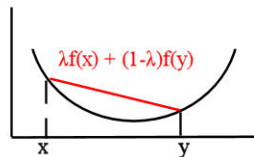


Figure 1: A Convex Function

We can also generalize the result to expectation:  $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ .

### 3 The EM Objective Function

Now, let's re-write the complete log likelihood function by multiplying it by  $\frac{q(z)}{q(z)}$ , where  $q(z)$  represents an arbitrary distribution for the random variable  $Z$ .

$$\begin{aligned}\log p(x|\theta) &= \log \sum_Z p(z|\theta)p(x|z, \theta) \frac{q(z)}{q(z)} \\ &= \log \mathbb{E}_q \left[ \frac{p(z|\theta)p(x|z, \theta)}{q(z)} \right] \\ &\geq \mathbb{E}_q \left[ \log \frac{p(z|\theta)p(x|z, \theta)}{q(z)} \right] \\ \mathcal{L}(\theta; q) &= \mathbb{E}_q[\log p(z|\theta)] + \mathbb{E}_q[\log p(x|z, \theta)] - \mathbb{E}_q[\log q(z)].\end{aligned}$$

We derived the third line by using Jensen's Inequality. The final result is the EM objective function. Note that the final quantity  $\mathbb{E}[\log q(z)]$  is known as entropy.

### 4 The EM Algorithm

The EM algorithm proceeds by coordinate ascent. At each iteration  $t$ , we have the following two values:  $q^{(t)}$  and  $\theta^{(t)}$ .

At the E-Step, we update the posterior value  $q$  of the random variable given the observations while holding  $\theta^{(t)}$  fixed.

$$\begin{aligned}q^{(t+1)} &= \operatorname{argmax}_q \mathcal{L}(q, \theta^{(t)}) \\ &= p(z|x, \theta^{(t)}).\end{aligned}$$

At the M-Step, we update the model parameters to maximize the expected complete log likelihood function.

$$\theta^{(t+1)} = \operatorname{argmax}_\theta \mathcal{L}(q^{(t+1)}, \theta)$$

To see how it works:

$$\begin{aligned}\mathcal{L}(p(z|x, \theta), \theta) &= \sum_Z p(z|x, \theta) \log \frac{p(x, z|\theta)}{p(z|x, \theta)} \\ &= \sum_Z p(z|x, \theta) \log \frac{p(x, z|\theta)p(x|\theta)}{p(x, z|\theta)} \\ &= \sum_Z p(z|x, \theta) \log p(x|\theta) \\ &= \log p(x|\theta) \sum_Z p(z|x, \theta) \\ &= \log p(x|\theta)\end{aligned}$$

Therefore, as we maximize the objective function, we are also maximizing the log likelihood function.