Stat 471: Lecture 26
Expectation-Maximization Algorithm I.

Moo K. Chung
mchung@stat.wisc.edu

November 9, 2003

1. *Expectation-maximization* (EM) method is an iterative method for maximizing difficult likelihood problems. It was first introduced by Dempster *et al.* (J. Roy. Statist. Soc. 1997). Suppose we have a random sample $X = (X_1, \cdots, X_n)$ iid from $f(x|\theta)$. We wish to find the maximum likelihood estimator

   $$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i|\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \ln f(x_i|\theta).$$

   This optimization problem is difficult and it gives a motivation for EM algorithm. We augment the data with with *latent* (unobserved or missing) data X^m such that the complete data $X^c = (X, X^m)$. The density of X^c is

   $$X^c = (X, X^m) \sim f(x^c) = f(x, x^m).$$

   The conditional density for the missing data X^m is

   $$f(x^m|x,\theta) = \frac{f(x, x^m|\theta)}{f(x|\theta)}.$$

   Rearranging terms,

   $$f(x|\theta) = \frac{f(x, x^m|\theta)}{f(x^m|x,\theta)}.$$

   Taking logarithm, we get log-likelihood

   $$\ln f(X|\theta) = \ln f(X^c|\theta) - \ln f(X^m|X,\theta).$$

   Take expectation with respect to $f(x^m|x,\theta_0)$ so that X are considered as constants.

   $$\begin{align} \ln f(X|\theta) &= \mathbb{E}[\ln f(X^c|\theta)|X,\theta_0] \tag{1} \\ &\quad -\mathbb{E}[\ln f(X^m|X,\theta)|X,\theta_0]. \tag{2} \end{align}$$

2. . Let us denote the log-likelihood by

   $$Q(\theta|\theta_0, x) = \mathbb{E}[\ln f(X^c|\theta)|X,\theta_0].$$

   EM algorithm
   1. the E-step: compute $Q(\theta|\hat{\theta}_{j-1}, x)$
   2. the M-step: maximize $Q(\theta|\hat{\theta}_{j-1}, x)$ and take

   $$\hat{\theta}_j = \arg\max_{\theta} Q(\theta|\hat{\theta}_{j-1}, x).$$

   Proof. If the above procedure is iterated, we get the sequence of estimators $\hat{\theta}_0, \hat{\theta}_1, \cdots$ and it can be shown that it converges to the maximum likelihood estimator $\hat{\theta}$.

   Note that $Q(\hat{\theta}_{j+1}|\hat{\theta}_j, x) \geq Q(\hat{\theta}_j|\hat{\theta}_j, x)$. Now let

   $$R(\theta|\theta_0, x) = \mathbb{E}[\ln f(X^m|X,\theta)|X,\theta_0].$$

   It can be shown that $\theta_0 = \arg\max_\theta R(\theta|\theta_0, x)$ For the proof, use Jensen's inequality (see Carsella & Berger's Statistical Inference). Then

   $$R(\hat{\theta}_{j+1}|\hat{\theta}_j, x) \leq R(\hat{\theta}_j|\hat{\theta}_j, x).$$

   Consequently

   $$\ln f(X|\hat{\theta}_j) \leq \ln f(X|\hat{\theta}_{j+1}).$$

   This inequality guarantee the the sequence of estimators $\hat{\theta}_j$ monotonically increase the likelihood. To guarantee that the limit converges to the maximum likelihood estimator, we need the condition of continuity of $Q(\theta|\theta_0, x)$ in θ and θ_0.

3. The difficulty of EM algorithm is at the E-step where we need to compute the conditional expectation $Q(\theta|\hat{\theta}_{j-1}, x)$. The *Monte-Carlo EM* algorithm overcome this by simulating missing data $X^m \sim f(x^m|x,\theta)$. so that

   $$\widehat{Q}(\theta|\theta_0, x) = \frac{1}{l} \sum_{j=1}^{l} \ln f(X, X^m|\theta).$$