

## Lecture 22 The EM Algorithm

The Expectation-Maximization (EM) algorithm is a device to find MLEs when some unobserved latent variables exist. Let  $X$  and  $Y$  denote latent variables and observed data, and  $\theta$  be a parameter. In general,  $X$ ,  $Y$  and  $\theta$  are all multidimensional. The goal is to find a MLE  $\hat{\theta}$  based on  $Y$ . Suppose  $f(x, y|\theta)$  is the joint density for  $(X, Y)$ , and  $f_Y(y|\theta)$  is the marginal density for  $Y$ .  $L(\theta|x, y) = \log f(x, y|\theta)$  and  $L(\theta|y) = \log f_Y(y|\theta)$  are usually referred to as the likelihood functions based on the *complete* data  $(x, y)$  and the *observed* data  $y$  respectively. The goal is to find a MLE  $\hat{\theta}$  based on the observed data  $y$ .

Here is a challenge: Finding  $\hat{\theta}$  requires maximizing the integral  $L(\theta|y) = \log \int_{\mathcal{X}} f(x, y|\theta) \nu_X(dx)$  as a function of  $\theta$ .

The EM algorithm starts with an initial guess  $\theta^{(0)}$  and alternates the following two steps at  $t = 0, 1, 2, \dots$

**E-step:** Compute the target function

$$Q(\theta, \theta^{(t)}) = E[L(\theta|X, y)|y; \theta^{(t)}] = \int_{\mathcal{X}} \log f(x, y|\theta) f_X(x|y; \theta^{(t)}) \nu_{\mathcal{X}}(dx);$$

**M-step:** Find  $\theta^{(t+1)}$  to maximize  $Q(\theta, \theta^{(t)})$ .

The key idea in the E-step is “imputation”. Since the likelihood function based on the complete data is usually easy to handle, it is appealing to simply “fill in” a set of missing data  $x$  and maximize  $L(\theta|x, y)$ . However, considering the variability of  $X$ , a correct “fill in” approach is to average over all possible values of  $X$ . This is implemented by integrating  $\log f(x, y|\theta)$  according to the current “predictive density”  $f_X(x|y; \theta^{(t)})$  for the missing data  $x$ .

Now we show the monotonicity of the EM algorithm, i.e. it always increases the likelihood. Denote by  $E_t(\cdot)$  the conditional expectation under the density  $f_X(x|y; \theta^{(t)})$ . Note that

$$L(\theta|y) = L(\theta|X, y) - \log f_X(X|y; \theta)$$

implies

$$L(\theta|y) = E_t L(\theta|X, y) - E_t[\log f_X(X|y; \theta)] = Q(\theta, \theta^{(t)}) - E_t[\log f_X(X|y; \theta)].$$

Let  $h(X) = \frac{f_X(X|y; \theta)}{f_X(X|y; \theta^{(t)})}$ . Then Jensen’s inequality implies  $E_t \log h(X) \leq E_t h(X) - 1 = 0$ . Hence  $E_t[\log f_X(X|y; \theta)] \leq E_t[\log f_X(X|y; \theta^{(t)})]$ . Suppose  $Q(\theta, \theta^{(t)}) > Q(\theta^{(t)}, \theta^{(t)})$ . Then

$$L(\theta|y) > Q(\theta^{(t)}, \theta^{(t)}) - E_t[\log f_X(X|y; \theta^{(t)})] = L(\theta^{(t)}|y).$$

The increase in likelihood at the iteration  $t$  will be positive provided  $Q(\theta, \theta^{(t)}) > Q(\theta^{(t)}, \theta^{(t)})$ , and this is indeed the case unless  $\theta^{(t)}$  is already a maximizer.

**Example 22.1** (a mixture of Gaussian distributions) Let  $Y = Y^n = (Y_1, \dots, Y_n)$  be iid samples from the density  $f_{Y_1}(y_1|\theta) = \sum_{k=1}^K w_k g(y_1|\mu_k, \sigma_k^2)$  where the weights  $w_k \geq 0$ ,  $k = 1, \dots, K$ ,  $w_1 + \dots + w_K = 1$ , and  $g(y_1|\mu_k, \sigma_k^2)$  is the normal density of  $N(\mu_k, \sigma_k^2)$ ,  $k = 1, \dots, K$ . The interpretation: for each  $i = 1, \dots, n$ , a label  $X_i$  is drawn with  $P(X_i = k) = w_k$ ; given  $X_i = k$ , sample  $Y_i$  from  $N(\mu_k, \sigma_k^2)$ . Here the component labels  $X = X^n = (X_1, \dots, X_n)$  are latent variables, and  $\theta = (w_1, \dots, w_K; \mu_1, \dots, \mu_K; \sigma_1^2, \dots, \sigma_K^2)$  is an unknown parameter. We want to find a MLE  $\hat{\theta}$  based on  $Y$ .

First, MLEs based on the complete data  $(X, Y)$  can be easily obtained: for each  $k$ ,

$$\widehat{w}_k = \frac{S_n^{(k)}}{n}, \quad \widehat{\mu}_k = \frac{\sum_{i=1}^n I_{\{X_i=k\}} Y_i}{S_n^{(k)}}, \quad \widehat{\sigma}_k^2 = \frac{\sum_{i=1}^n I_{\{X_i=k\}} (Y_i - \widehat{\mu}_k)^2}{S_n^{(k)}}, \quad (22.1)$$

where  $S_n^{(k)} = \sum_{i=1}^n I_{\{X_i=k\}}$ .

Second, since  $X$  is unobservable, it is reasonable to replace  $I_{\{X_i=k\}}$  by

$$z_{ki}^{(t)} \triangleq P(X_i = k | y_i; \theta^{(t)}) = \frac{w_k^{(t)} g(y_i | \mu_k^{(t)}, (\sigma_k^2)^{(t)})}{\sum_{j=1}^K w_j^{(t)} g(y_i | \mu_j^{(t)}, (\sigma_j^2)^{(t)})} \quad (22.2)$$

at the iteration  $t$  with  $\theta^{(t)} = (w_1^{(t)}, \dots, w_K^{(t)}; \mu_1^{(t)}, \dots, \mu_K^{(t)}; (\sigma_1^2)^{(t)}, \dots, (\sigma_K^2)^{(t)})$ .

Third, following (22.1) and (22.2), we would expect the EM estimates at the iteration  $t + 1$  to be

$$w_k^{(t+1)} = \frac{\sum_{i=1}^n z_{ki}^{(t)}}{n} \quad (22.3)$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^n z_{ki}^{(t)} y_i}{\sum_{i=1}^n z_{ki}^{(t)}} \quad (22.4)$$

$$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^n z_{ki}^{(t)} (y_i - \mu_k^{(t+1)})^2}{\sum_{i=1}^n z_{ki}^{(t)}} \quad (22.5)$$

To verify (22.3) – (22.5), note the complete data likelihood

$$L(\theta | x^n, y^n) = \log \prod_{i=1}^n w_{x_i} g(y_i | \mu_{x_i}, \sigma_{x_i}^2).$$

Hence  $Q(\theta, \theta^{(t)}) = Q_1(\theta) + Q_2(\theta)$  where

$$Q_1(\theta) = \sum_{i=1}^n \sum_{k=1}^K z_{ki}^{(t)} \log w_k$$

and

$$Q_2(\theta) = \sum_{i=1}^n \sum_{k=1}^K z_{ki}^{(t)} \log g(y_i | \mu_k, \sigma_k^2)$$

Since  $\{w_k\}$  only appears in  $Q_1(\theta)$ , it follows from Jensen's inequality that

$$\sum_{k=1}^K w_k^{(t+1)} \log \frac{w_k}{w_k^{(t+1)}} \leq \sum_{k=1}^K w_k^{(t+1)} \frac{w_k}{w_k^{(t+1)}} - 1 = 0,$$

which implies that

$$Q_1(\theta) = n \sum_{k=1}^K w_k^{(t+1)} \log w_k \leq n \sum_{k=1}^K w_k^{(t+1)} \log w_k^{(t+1)},$$

hence  $w_k^{(t+1)}$  is indeed the EM estimate of  $w_k$  at the iteration  $t + 1$ . Standard calculus applied to  $Q_2(\theta)$  verifies (22.4) and (22.5).

#### References:

- [1] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *J. Royal Stat. Society, Series B* **39**, 1-38.
- [2] Meng, X.L. and van Dyk, D. (1997). The EM algorithm: an old folk-song sung to a fast new tune (with discussion). *J. Royal Stat. Society, Series B* **59**, 511-568.