EM Shortcut for the Exponential Family

Yun Wang
(Based on a 2001 version by Guy Lebanon)

Language and Statistics, Spring 2013

Let x_i be the i-th data point, and $z_i = (z_{i1}, \dots, z_{il})$ be the latent variables associated with x_i. Denote the set of parameters by θ. The complete likelihood $P(x_i, z_i|\theta)$ is said to belong to the **exponential family** if it has the following form:

$$L(x_i, z_i|\theta) = c(x_i, \theta) \exp \left[\sum_{j=1}^{l} z_{ij} g_j(x_i, \theta) \right] \tag{1}$$

In this case, the log-likelihood is linear in the latent variables:

$$\log L(x_i, z_i|\theta) = \log c(x_i, \theta) + \sum_{j=1}^{l} z_{ij} g_j(x_i, \theta) \tag{2}$$

The auxiliary function of the EM algorithm, which is the expectation of the log-likelihood, can be obtained by replacing the latent variables with their expectations:

$$E_{z_i|x_i,\theta^{(k)}}[\log L(x_i, z_i|\theta^{(k+1)})] = \log c(x_i, \theta^{(k+1)}) + \sum_{j=1}^{l} E[z_{ij}|x_i, \theta^{(k)}] g_j(x_i, \theta^{(k+1)}) \tag{3}$$

As a result, the EM algorithm boils down to:

- Write down the maximum likelihood estimator of the parameters as if the latent variables z_{ij} are known;
- Replace the latent variables z_{ij} by their expectations $E[z_{ij}|x_i, \theta^{(k)}]$ to get the recursive formula for $\theta^{(k+1)}$, and iterate until convergence.

The above procedure can always be used if the model is a mixture of sub-models. We can choose the latent variables z_{ij} to be mutually exclusive indicators, i.e. if the i-th data point came from the j-th sub-model, let $z_{ij} = 1$ and all the other $z_{ik} = 0$ ($k \neq j$). Furthermore, let $c(x_i, \theta) = 1$ and $g_j(x_i, \theta)$ be the log-likelihood of the i-th data point if it came from the j-th sub-model. Then we can see that the complete likelihood function does have the form of Eq. (1).

For example, let's consider estimating the parameters of a Gaussian mixture model (GMM) with unit variances but unknown means $\mu_1, \dots, \mu_l$ and priors $\lambda_1, \dots, \lambda_l$. For each data point x_i, we associate it with l latent variables $z_{i1}, \dots, z_{il}$, where $z_{ij} = 1$ if x_i came from the j-th Gaussian and 0 otherwise. If the latent variables were known (i.e. we knew which data points came from which Gaussians), the maximum likelihood estimates of the means and priors would be:

$$\hat{\mu}_{j,\mathrm{ML}} = \frac{\sum_{i=1}^{n} z_{ij} x_i}{\sum_{i=1}^{n} z_{ij}} \tag{4}$$

$$\hat{\lambda}_{j,\mathrm{ML}} = \frac{\sum_{i=1}^{n} z_{ij}}{n} \tag{5}$$

where n is the total number of data points. When the latent variables are not known, we replace the z_{ij} in the formulas above with its expectation $E[z_{ij}|x_i, \theta^{(k)}]$:

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} E[z_{ij}|x_i, \theta^{(k)}] x_i}{\sum_{i=1}^{n} E[z_{ij}|x_i, \theta^{(k)}]} \tag{6}$$

$$\lambda_j^{(k+1)} = \frac{\sum_{i=1}^{n} E[z_{ij}|x_i, \theta^{(k)}]}{n} \tag{7}$$

The expectation can be calculated as follows:

$$E[z_{ij}|x_i, \theta^{(k)}] = \frac{\lambda_j^{(k)} \exp[-\frac{1}{2}(x_i - \mu_j^{(k)})^2]}{\sum_{j'=1}^{l} \lambda_{j'}^{(k)} \exp[-\frac{1}{2}(x_i - \mu_{j'}^{(k)})^2]} \tag{8}$$

The entire EM algorithm runs as follows:

1. **Initialize** the means $\mu_j^{(0)}$ and priors $\lambda_j^{(0)}$ $(j = 1, \ldots, l)$;
2. **E-step**: Calculate the expectations of latent variables with Eq. (8);
3. **M-step**: Update the means and priors with Eqs. (6) and (7);
4. Terminate if converged, otherwise go to step 2.