1 E-M for latent variable models

The lecture notes and slides over-complicate E-M for mixture models. The derivation is quite simple.

Let $f(x;\theta)$ be the full distribution of a latent variable model. The density for mixture models can be decomposed as follows (assuming independence):

$$f(x;\theta) = f(x|z;\theta_x)f(z;\theta_z)$$

Where Z follows a discrete distribution and θ_x and θ_z are parameters specifically for X and Z respectively. Since Z is discrete, this function can be written as

$$f(x;\theta) = \sum_j p_j f(x|Z = p_j;\theta_{x,j}),\ \sum_j p_j = 1$$

For notational simplicity, I will drop the notation $\theta_{x,j}$ and simply revert to θ_j.

Intuitively, we should maximize $f(x;\theta)$ over the observed data to estimate θ. We proceed down the path, though we should feel some discomfort since we don't observe z and so the problem may be tricky.

2 Maximize the likelihood

Given data $x_i, i \in [1, 2, ..., n]$, we can try to maximize the log-likelihood:

$$\max_{\theta,p} \sum_i \log\left(\sum_j p_j f(x_i|Z = p_j;\theta_j)\right)$$

We typically do this by taking derivatives with respect to the parameters and setting the derivatives equal to zero. However, we have the additional constraint that the probabilities for Z must sum to one, which requires us to create the Lagrangian:

$$\ell(\theta, p, q) = \sum_i \log\left(\sum_j p_j f(x_i|Z = p_j;\theta_j)\right) + q\left(\sum_j p_j - 1\right)$$

Now we can start taking derivatives of the Lagrangian:

$$\begin{aligned}\frac{\partial \ell}{\partial \theta_j} &= \sum_i \left(p_j \frac{\partial f(x_i|Z = p_j;\theta_j)}{\partial \theta_j}\right) \frac{1}{\sum_j p_j f(x_i|Z = p_j;\theta_j)} \\ &= \sum_i \frac{\partial \log\left(f(x_i|Z = p_j;\theta_j)\right)}{\partial \theta_j} \frac{p_j f(x_i|Z = p_j;\theta_j)}{\sum_j p_j f(x_i|Z = p_j;\theta_j)}\end{aligned}$$

$$= \sum_i h_{i,j} \frac{\partial \log\left(f(x_i|Z = p_j; \theta_j)\right)}{\partial \theta_j}$$

Where $h_{i,j} = \frac{p_j f(x_i|Z=p_j;\theta_j)}{\sum_j p_j f(x_i|Z=p_j;\theta_j)}$ Note that for many distributions we already have the partial derivative of the log likelihood with respect to the parameters; these are derived for standard maximum likelihood calculations.

$$\frac{\partial \ell}{\partial q} = \sum_j p_j - 1$$

$$\frac{\partial \ell}{\partial p_j} = \sum_i \frac{f(x_i|Z = p_j; \theta_j)}{\sum_j p_j f(x_i|Z = p_j; \theta_j)} + q$$

$$= \sum_i \frac{h_{i,j}}{p_j} + q$$

Setting each of these to zero (and treating $h_{i,j}$ like a constant!) yields

$$q = -\sum_j \sum_i h_{i,j}$$

$$p_j = \frac{\sum_i h_{i,j}}{\sum_j \sum_i h_{i,j}} = \frac{1}{n} \sum_i h_{i,j}$$

The E-M algorithm now becomes obvious: simply maximize the likelihood (for a given set of initial p), then set $h_{i,j}$ with the updated p, θ, and repeat until convergence.

3 Example

Assume an mixture of two exponential distributions. The maximization step (fixing $h_{i,j}$), is

$$\frac{\partial \ell}{\partial \theta_j} = \sum_i h_{i,j} \left(\frac{1}{\theta_j} - x_i \right)$$

Setting this equal to zero yields

$$\theta_j = \frac{\sum_i h_{i,j}}{\sum_i h_{i,j} x_i}$$

Updating the p_j are trivial.

Then the next iteration of $h_{i.j}$ can be found using the (new) θ_j, p_j.

4 Takeaway

If maximizing the parameters of a mixture model, the E-M algorithm is a simple extension of the process for finding the maximum likelihood parameters. Most of the work has already been performed through finding the MLE for the base distribution(s).