April 29, 2018

Let's restate the general EM algorithm. First, we had the E-step:

$$Q(\theta, \hat{\theta}^{(t)}) = \mathbb{E}_{z|x;\hat{\theta}^{(t)}}\big[\ln L(\theta; x, z)\big]$$

Where z are latent variables, and x are observed variables. Then, the M-step is

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta, \hat{\theta}^{(t)})$$

We want to apply this algorithm to our Gaussian Mixture Model, where

$$p(x_i \mid z_i) = \prod_{j=1}^{J} \mathcal{N}(x_i \mid \mu_j, \Sigma_j)^{z_{ij}}$$

where J is the number of labels (or Gaussians) we have. We have

$$p(z_{ij} = 1) = \tau_j$$

where τ_j is a probability vector. This gives us

$$p(z_i) = \prod_{j=1}^{J} = \tau_j^{z_{ij}}$$

We can draw the variable diagram as

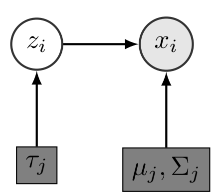

Figure 1: $\theta_j = (\tau_j, \mu_j, \Sigma_j)$

Next, we can compute the likelihood as

$$
\begin{aligned}
L(\theta; x, z) &= \prod_{i=1}^{N} p(x_i \mid z_i; \theta) p(z_i; \theta) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{J} \tau_j^{z_{ij}} \mathcal{N}(x_i; \mu_j, \Sigma_j)^{z_{ij}}
\end{aligned}
$$

Then, for our E-step we need to compute the log likelihood,

$$
\ln(L(\theta; x, z)) = \sum_{i=1}^{N} \sum_{j=1}^{J} z_{ij} \Big(\ln \tau_j + \ln \mathcal{N}(x_i; \mu_j, \Sigma_j) \Big)
$$

And we need to compute

$$
\begin{aligned}
p(z_{ij} = 1 \mid x_i; \hat{\theta}^{(t)}) &= \frac{p(x_i \mid z_{ij}) P(z_{ij} = 1)}{p(x_i)} \\
&= \frac{\hat{\tau}_j \mathcal{N}(x_i; \hat{\mu_j}, \hat{\Sigma_j})}{\sum_{j=1}^{J} \hat{\tau}_j \mathcal{N}(x_i; \hat{\mu_j}, \hat{\Sigma_j})} \\
&\hat{=} p_{ij}
\end{aligned}
$$

Then

$$
\begin{aligned}
Q(\theta, \hat{\theta}^{(t)}) &= \mathbb{E}_{z|x;\hat{\theta}^{(t)}} \left[\sum_{i=1}^{N} \sum_{j=1}^{J} z_{ij} \Big(\ln \tau_j + \ln \mathcal{N}(x_i; \mu_j, \Sigma_j) \Big) \right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{J} \Big(\ln \tau_j + \ln \mathcal{N}(x_i; \mu_j, \Sigma_j) \Big) \mathbb{E}_{z|x;\hat{\theta}^{(t)}}[z_{ij}]
\end{aligned}
$$

and $\mathbb{E}_{z|x;\hat{\theta}^{(t)}}[z_{ij}] = p_{ij}$.

Now for the M-step, we want to calculate $\frac{d}{\mu_j}$, $\frac{d}{\Sigma_j}$, and $\frac{d}{\tau_j}$.

$$
\begin{aligned}
0 &= \frac{d}{\mu_j} \sum_{i=1}^{N} \sum_{j=1}^{J} \Big(\ln \tau_j + \ln \mathcal{N}(x_i; \mu_j, \Sigma_j) \Big) p_{ij} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij} (\mu_j - x_i) \\
&= \sum_{i=1}^{N} p_{ij} \mu_j - \sum_{i=1}^{N} p_{ij} x_i \\
\mu_j &= \frac{\sum_{i=1}^{N} p_{ij} x_i}{\sum_{i=1}^{N} p_{ij}}
\end{aligned}
$$

And similarly

$$0 = \frac{d}{\Sigma_j} \sum_{i=1}^{N} \sum_{j=1}^{J} \Big(\ln \tau_j + \ln \mathcal{N}(x_i; \mu_j, \Sigma_j) \Big) p_{ij}$$

$$\Sigma_j = \frac{\sum_{i=1}^{N} p_{ij} (x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{N} p_{ij}}$$

Finally to find $\frac{d}{\tau_j}$, we need $\frac{j=1}{J} \tau_j =$, therefore we have to use Lagrange multipliers to enforce this constraint.

$$0 = \frac{d}{d\tau_j} \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij} \ln \tau_j + \lambda (\sum_{j=1}^{J} \tau_j - 1)$$

$$= \sum_{i=1}^{N} \frac{p_{ij}}{\tau_j} + \lambda \tau_j \qquad = \frac{\sum_{i=1}^{N} p_{ij}}{-\lambda}$$

And we set $\lambda = -N$, giving us

$$\tau_j = \frac{\sum_{i=1}^{N} p_{ij}}{N}$$

Using these derivation for the parameters gives us the M-step.

Now, we can use both E and M steps to get an algorithm to classify our data.