EM Algorithm

May 1, 2010

1 The EM algorithm in general

The expectation maximization algorithm (EM algorithm) is a general technique for finding maximum likelihood solutions for probabilistic models which have latent variables.

Consider a probabilistic model with observed value $X$ and hidden variables $Z$. The joint distribution is $p(X, Z|\theta)$ which has parameter $\theta$. Our goal is to maximize the likelihood function that is given by

$$p(X|\theta) = \sum_{z} p(X, Z|\theta)$$

Here we assume $Z$ is discrete, although it would be similar if $Z$ is continuous variable. Suppose that direct optimization of $p(X|\theta)$ is difficult, but the optimization of the complete-data likelihood function $p(X, Z|\theta)$ is significantly easier. Then for any distribution $q(Z)$ over variable $Z$, we have

$$\ln p(X|\theta) = \sum_{Z} q(Z) \ln \left\{ \frac{p(X, Z|\theta)}{q(Z)} \right\} - \sum_{Z} q(Z) \ln \left\{ \frac{p(Z|X, \theta)}{q(Z)} \right\}$$

$$\triangleq \mathcal{L}(q, \theta) + KL(q||p) \quad (1)$$

where $\mathcal{L}(q, \theta)$ is a functional of the distribution $q(Z)$ and a function of the parameter $\theta$. To verify the decomposition, we should make use the fact

$$\ln p(X, Z|\theta) = \ln p(Z|X, \theta) + \ln p(X|\theta)$$

and

$$\sum_{Z} q(Z) = 1$$

$KL(q||p)$ is known as the Kullback-Leibler divergence. It’s easy to show that $KL(q||p) \geq 0$ using Jensen’s inequality, with equality if and only if $q(Z) = p(Z|X, \theta)$. That means $\mathcal{L}(q, \theta)$ is a lower bound of $\ln p(X|\theta)$.

The relationships between those three variables can be shown in Figure 1. The EM algorithm is a two-stage

\[1\text{Part of Christopher M. Bishop’s book ‘Pattern Recognition and Machine Learning’}\]
iterative optimization technique for finding maximum likelihood. We can use the decomposition defined at Equation (1).

In the E step, the lower bound $\mathcal{L}(q, \theta_{old})$ is maximized with respect to $q(Z)$ while holding $\theta_{old}$ fixed. Since the value $\ln p(X|\theta_{old})$ does not depend on $q(Z)$. So the largest value of $\mathcal{L}(q, \theta_{old})$ will occur when $KL(q||p) = 0$, or $q(Z) = p(Z|X, \theta_{old})$, as illustrated in Figure 2.

In the subsequent M step, the distribution $q(Z)$ is held fixed and the lower bound $\mathcal{L}(q, \theta)$ is maximized with respect to $\theta$ to give some new value $\theta_{new}$. This will cause the lower bound $\mathcal{L}$ to increase (unless it is already at a maximum). In this case, $KL(q||p) > 0$ since in this case,

$$q(Z) = p(Z|X, \theta_{old}) \neq p(Z|X, \theta_{new})$$

So the log-likelihood function will increase after E and M steps. As illustrated in Figure 3.
(Notice: in the M step, the $\mathcal{L}(q, \theta)$ can be expressed as

$$\mathcal{L}(q, \theta) = \sum_Z q(Z) \ln \left\{ \frac{p(X, Z | \theta)}{q(Z)} \right\}$$

$$= \sum_Z p(Z | X, \theta^{old}) \ln \left\{ \frac{p(X, Z | \theta)}{p(Z | X, \theta^{old})} \right\}$$

$$= \sum_Z p(Z | X, \theta^{old}) \ln \{p(X, Z | \theta)\} - \sum_Z p(Z | X, \theta^{old}) \ln \{p(Z | X, \theta^{old})\}$$

The second part is dose not depend on $\theta$. So maximize $\mathcal{L}(q, \theta)$ with respect to $\theta$ is equivalent to maximize $\sum_Z p(Z | X, \theta^{old}) \ln \{p(X, Z | \theta)\}$ with respect to $\theta$. Compare this result to the one the professor proposed in his lecture notes.)

![Figure 3: Illustration of the M step](image)

Also, the EM steps can be demonstrated in Figure 4. Here the red curve is the (incomplete data) log-likelihood function whose value we wish to maximize. We start with some initial parameter value $\theta^{old}$, and in the E step we evaluate the posterior distribution over latent variables, which gives rise to a lower bound $\mathcal{L}(q, \theta^{old})$ whose value equals the log-likelihood at $\theta^{old}$, as shown by the blue curve. In the M step, the bound is maximized giving the value $\theta^{new}$.

## 2 EM for Gaussian Mixtures

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters.

1. **E step**: Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_n | \mu_j, \Sigma_j)}$$
2. **M step**: Re-estimate the parameters using the current responsibilities

\[
\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n
\]

\[
\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k^{\text{new}})(x_n - \mu_k^{\text{new}})^t
\]

\[
\pi_k^{\text{new}} = \frac{N_k}{N}
\]

where

\[
N_k = \sum_{n=1}^{N} \gamma(z_{nk})
\]

3. Evaluate the log likelihood

\[
\ln p(X|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) \right\}
\]

and check for convergence of either the parameters or the log-likelihood. If the convergence criterion is not satisfied, return to step 2.