1 EM Algorithm

The goal of the EM algorithm is to find maximum likelihood solutions for models having latent variables. Consider a probabilistic model in which we collectively denote all of the observed variabels by $\mathbf{X}$ and all of the hidden variables by $\mathbf{Z}$. The joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ is governed by a set of parameters denoted $\boldsymbol{\theta}$. Our goal is to maximize the likelihood function that is given by

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

This method based on the assumption that the computation of $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ is easier than $p(\mathbf{X}|\boldsymbol{\theta})$. Let $q(\mathbf{Z})$ be the distribution of $\mathbf{Z}$. Then is obvious that

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}|\boldsymbol{\theta})$$

We've known that

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) + \ln p(\mathbf{X}|\boldsymbol{\theta})$$

Then

$$\begin{aligned}
\ln p(\mathbf{X}|\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} \ln q(\mathbf{Z}) p(\mathbf{X}|\boldsymbol{\theta}) \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\
&\equiv \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)
\end{aligned}$$

Since $\mathrm{KL}(q\|p) \geq 0$, $\mathcal{L}(q, \boldsymbol{\theta})$ is the lower bound of $\ln p(\mathbf{X}|\boldsymbol{\theta})$, i.e.

$$\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbf{X}|\boldsymbol{\theta})$$

Note that $\mathrm{KL}(q\|p) = 0$ if and only if $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.

The EM algorithm is a two-stage iterative optimization technique for finding maximum likelihood solutions.

1.1 E Step

Let $\boldsymbol{\theta}^{\mathrm{old}}$ be the current parameter vector. Maximize $\mathcal{L}(q, \boldsymbol{\theta})$ with respect to $q(\mathbf{Z})$ with holding $\boldsymbol{\theta}^{\mathrm{old}}$ fixed. $\mathcal{L}(q, \boldsymbol{\theta})$ will reach its maximum $\ln p(\mathbf{X}|\boldsymbol{\theta})$ when $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})$. That is,

$$\max \mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{\mathrm{old}}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})$$

1.2 M Step

Maximize $\mathcal{L}(q, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ with holding $q(\mathbf{Z})$ fixed. Here we get new parameter $\boldsymbol{\theta}^{\mathrm{new}}$. This will cause $\mathcal{L}(q, \boldsymbol{\theta})$ increase and then cause $\ln p(\mathbf{X}|\boldsymbol{\theta})$ increase subsequently. Furthermore, the $q(\mathbf{Z})$ in this step is determined in the previous step, $\mathrm{KL}(q, \boldsymbol{\theta}^{\mathrm{new}})$ will be nonzero. Hence $\ln p(\mathbf{X}|\boldsymbol{\theta})$ increases more than its lowe bound $\mathcal{L}(q, \boldsymbol{\theta})$. From above we have,

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})$$

$$= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \text{constant}$$

Since $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) = q(\mathbf{Z})$, the last term is a constant. This tells us that we actually maximize the expectation of the complete-data log likelihood.