Notes on Expectation Maximization Algorithm

Liangjie Hong

February 6, 2012

1 Standard EM Algorithm

In general, we are usually interested in the following setting. Let $X = (x_1, x_2, ..., x_n)$ as our observation data points with the parameter θ. We want to find the parameter by maximizing the likelihood function:

$$\hat{\theta} = \arg\max_{\theta} \; \log P(X|\theta)$$

However, sometimes, the data points are missing or we need some latent variables to model the data and thus we really want to model the following complete likelihood fucntion:

$$\hat{\theta} = \arg\max_{\theta} \; \log \int_h P(X, h|\theta) \, dh \tag{1}$$

We can further manipulate Equation 1 as follows:

$$\log \int_h P(X, h|\theta) \, dh = \log \int_h \frac{P(X, h|\theta)}{q(h)} q(h) \, dh = \log E_q \left[\frac{P(X, h|\theta)}{q(h)} \right] \tag{2}$$

where $\int_h q(h) = 1$. By using following Jensen's inequality:

$$E[f(x)] \geq f(E[x])$$

(which is applied to convex functions while $\log$ is a concave function, so we need to flip the inequality) Equation 2 can be re-written as:

$$\log E_q \left[\frac{P(X, h|\theta)}{q(h)} \right] \geq E_q \left[\log \frac{P(X, h|\theta)}{q(h)} \right] = E_q \left[\log P(X, h|\theta) - \log q(h) \right] = G(q, \theta) \tag{3}$$

Note, here we formulate a lower-bound for **each** data point. We want to maximize the bound on current θ. Therefore, the problem is converted to maximize the lower-bound:

$$\hat{q} = \arg\max_{q} \; G(q, \theta)$$

By applying Lagrange Multiplier, we can obtain the objective function as follows

$$\widetilde{G} = \int_h q(h) \log P(X, h|\theta) \, dh - \int_h q(h) \log q(h) \, dh + \lambda \left(1 - \int_h q(h) \, dh \right)$$

Taking the derivative respect to $q(h)$, we can obtain:

$$\frac{\partial \widetilde{G}}{\partial q(h)} = \log P(X, h|\theta) - \log q(h) - 1 - \lambda = 0$$

Therefore, we have:

$$\begin{aligned} e^{\lambda+1} q(h) = P(X,h|\theta) &\Rightarrow \int_h [e^{\lambda+1} q(h)]\, dh = \int_h P(X,h|\theta)\, dh \\ e^{\lambda+1} &= \int_h P(X,h|\theta)\, dh \end{aligned}$$

Hence:

$$q(h) = \frac{P(X,h|\theta)}{\int_h P(X,h|\theta)} = p(h|X,\theta) \tag{4}$$

This equation suggests that how E-step is performed: Compute the posterior distribution of the hidden variables, given the data and the current guess of parameter θ.

Another point of view, which also validates that the posterior indeed maximize the likelihood in this case. We start from Equation 3:

$$\begin{aligned} E_q\left[\log \frac{P(X,h|\theta)}{q(h)}\right] &= E_q\left[\log \frac{P(X|\theta)P(h|X,\theta)}{q(h)}\right] \\ &= E_q\left[\log \frac{P(h|X,\theta)}{q(h)}\right] + E_q[\log P(X|\theta)] \\ &= -E_q\left[\log \frac{q(h)}{P(h|X,\theta)}\right] + E_q[\log P(X|\theta)] \\ &= -D(q(h)||P(h|X,\theta)) + \log P(X|\theta) \end{aligned}$$

where D is the KL divergence between two distributions. If $q(h) = P(h|X,\theta)$, the distance is minimized and therefore the likelihood is maximized.

The M-step is usually problem-dependent. From Equation 3, we fix $q(h)$ and maximize θ.

2 Generalizations

2.1 Generalized EM

Sometimes, it might be difficult to maximize θ during the M-step. As long as the new θ still increases the lower bound, the algorithm will still converge to a local optimum.

2.2 Variational EM

Variational EM relaxes the requirement that $q(h) = P(h|X,\theta)$ during the E-step. It may be that the true posterior is intractable, so we use a simplified family of $q(h)$ distributions (e.g., fully factorized $q(h) = \sum_i q(h_i)$) to approximate the true posterior distribution.

Acknowledgement

I thank Dr. Yan Fu for detailed discussions. Other similar notes are here: [2, 1, 3].

References

[1] T. Minka. Expectation-maximization as lower bound maximization. Technical report, Microsoft Research Cambridge, 1999.

[2] R. M. Neal and G. E. Hinton. *A view of the EM algorithm that justifies incremental, sparse, and other variants*, pages 355–368. MIT Press, Cambridge, MA, USA, 1999.

[3] C. Zhai. A note on the expectation-maximization (em) algorithm. Technical report, UIUC, 2007.