EM in two pages

Brendan O'Connor

January 31, 2017

We would like to maximize *incomplete data loglikelihood* for seen data x, latent variables z, and model parameters θ. (Latent variables are local to an instance, but parameters cut across the dataset.) Unfortunately the sum inside the log is hard to deal with.

$$\ell(\theta) = \log p(x|\theta) = \log \sum_z p(x, z|\theta) = \sum_i \log \sum_{z_i} p(x_i, z_i|\theta)$$

Assume we have a model where, if only we knew the z's, it would be easy. That is, assume we have a good algorithm to maximize the *complete* likelihood $\max_\theta \log p(x, z|\theta)$, when z is known and fixed. This motivates why we want to derive EM in the first place.

EM derivation: add in a $q_i(z)/q_i(z)$ term (weird special local probabilities for the latent variables, for every instance) and apply Jensen's inequality to get the EM bound.

$$\ell(\theta) = \sum_i \log \sum_{z_i} \frac{q_i(z_i)}{q_i(z_i)} p(x_i, z_i|\theta) \tag{1}$$

$$\geq \sum_i \sum_{z_i} q_i(z_i)[\log p(x_i, z_i|\theta) - \log q_i(z_i)] \tag{2}$$

$$\equiv J(q, \theta) \;\; (\equiv \sum_i J_i) \tag{3}$$

J consists of weighted complete-loglikelihood, plus the entropy of q. EM is coordinate ascent on J. Maximizing for θ (the M-step) is simply maximizing weighted log-likelihood, since the q entropy term drops out. (For counting-based estimation, this is simply weighted counting. For gradient-based estimation, this is simply weighted gradient calculation.) Maximizing for q (the E-step) leads to setting q to local posteriors. This is because, since $p(x, z) = p(z|x)p(x)$, rewrite as

$$J_i = \sum_z q(z) \log[p(z|x)p(x)/q(z)] = -D(q(z) \;||\; p(z|x)) + \log p(x)$$

where the first term is the negative KL divergence between $q(z)$ and $p(z|x)$; to maximize J_i, the $p(x)$ term is irrelevant to q, so set $q(z) := p(z|x)$ to minimize the KL divergence to zero. Thus the EM steps are

- E-step: set all $q_i(z_i) := p(z_i|x_i, \theta)$
- M-step: learn new $\theta := \arg\max_\theta \sum_i \sum_{z_i} q_i(z_i) \log p(x_i, z_i|\theta)$

Or as one big equation,

$$\theta^{(new)} := \arg\max_\theta \sum_i \sum_{z_i} p(z_i|x_i, \theta^{(old)}) \log p(x_i, z_i|\theta)$$

In mixture models, $p(x, z|\theta) = p(z|\theta)p(x|z, \theta)$. The $q_i(z)$ local posteriors therefore are different than the mixture priors $p(z|\theta)$. When you update the prior parameters (mixture proportions), that's the average across all instances—at any local instance your belief is much different. Also in the E-step you'll want to use Bayes rule, $p(z|x) \propto p(z)p(x|z)$.

There are two ways to break down the J objective:

$$
\begin{aligned}
\ell(\theta) \equiv \log p_\theta(X) &= \sum_i \log p_\theta(x_i) = \sum_i \log \sum_{z_i} p_\theta(x_i, z_i) \\
&= \sum_i \log \sum_{z_i} \frac{q_i(z_i)}{q_i(z_i)} p_\theta(x_i, z_i) \\
\ell(\theta) &\overset{\text{Jensen's Ineq}}{\geq} \sum_i \sum_{z_i} q_i(z_i) \log \left[\frac{1}{q_i(z_i)}\ p_\theta(z_i|x_i)\ p_\theta(x_i) \right] \\
J(Q,\theta) &\equiv -E_Q \log Q(Z) + E_Q \log p_\theta(Z|X) + \log p_\theta(X)
\end{aligned}
$$

$$
\underbrace{-E_Q \log Q(Z)}_{\substack{H(Q) \\ \text{Entropy}}} + \underbrace{E_Q \log p_\theta(Z|X) + \log p_\theta(X)}_{\substack{E_Q \log p_\theta(X,Z) \\ \text{Weighted LL}}}
$$

M-step
$\max_\theta J(Q,\theta)$

$$
\underbrace{-E_Q \log Q(Z) + E_Q \log p_\theta(Z|X)}_{\substack{-KL(Q \,||\, p_\theta(Z|X)) \\ \text{Nonpositive}}} + \underbrace{\log p_\theta(X)}_{\ell(\theta)}
$$

E-step
$\max_Q J(Q,\theta)$
$\Rightarrow Q := P_\theta(Z|X)$

The E-step makes the bound tight for the current θ since it achieves $KL = 0$. Thus after the E-step, a Q has been chosen such that

$$J(Q, \theta^{(cur)}) = \ell(\theta^{(cur)})$$

Thus every iteration of EM results in a θ with a higher log-likelihood.

See also Neal and Hinton 1998, Murphy 2012 chs 11 and 21, MacKay 2003, and `http://cs229.stanford.edu/notes/cs229-notes8.pdf`.