EM, Sufficient Statistics

Ji Ma

April 22, 2014

1 The Complete log likelihood and expected log likelihood

When the hidden variables are observable, then we can directly maximize the complete log likelihood (CLL) which is:

$$\log p(x, y) \tag{1}$$

The sufficient statistics can be read directly from the complete log likelihood.

Example 1, Guassian Mixture:

$$\log p(X = x, Y = y_k) = -\frac{1}{2}(x - \mu_k)^t(x - \mu_k) - 0.5 * \log 2 * \pi + \log p(y_k) \tag{2}$$

$$= -\frac{1}{2}(x^t x - 2x^t \mu_k + \mu_k^t \mu_k) - 0.5 * \log 2 * \pi + q_k \log \pi_k \tag{3}$$

Example 2, HMM, suppose there are totally M hidden states.

$$\log p(X = x, Y = y) = \log\{\pi_{q0} \prod_{t=0}^{T-1} p(y_{t+1}|y_t) \prod_{t=0}^{T} p(x_t|y_t)\} \tag{4}$$

$$= \sum_{i=1}^{M} q_0^i \log \pi_i + \sum_{t=0}^{T-1} \sum_{i,j}^{M} q_t^i q_{t+1}^j \log a_{ij} + \sum_{t=0}^{T} \sum_{i,j}^{M} q_t^i y_t^j \log e_{ij} \tag{5}$$

Here we see that q_0^i is the sufficient statistic for π_i, $\sum_{t=0}^{T-1} q_t^i q_{t+1}^j$ is the sufficient statistic for a_{ij}, $\sum_{t=0}^{T} q_t^i y_t^j$ is the sufficient statistic for e_{ij}.

2 Expected Likelihood and Sufficient Statistics

The Expected likelihood (ELL) is the expectation of the complete log likelihood w.r.t. the distribution of the hidden variables given the observed data (See Jordan and Morphy for the tight lower bound derivation).

$$\sum_y p(y|x) \log p(x, y) \tag{6}$$

The difference in computing the sufficient statistics is that, for CLL, values of the sufficient statistics can be counted from the training data, for ELL, we need to compute the expectation of sufficient statistics.

Back to our HMM example, the ELL is

$$\sum_y p(y|x) \log p(x,y) = \sum_y p(y|x) \log\{\pi_{q0} \prod_{t=0}^{T-1} p(y_{t+1}|y_t) \prod_{t=0}^{T} p(x_t|y_t)\} \tag{7}$$

$$= \sum_y p(y|x) \sum_{i=1}^{M} q_0^i \log \pi_i + \sum_y p(y|x) \sum_{t=0}^{T-1} \sum_{i,j}^{M} q_t^i q_{t+1}^j \log a_{ij} + \dots \tag{8}$$

Alternatively, we can now treat $\sum_y p(y|x) * q_0^i = E(q_0^i)$ as the sufficient statistic for π_i, $\sum_y p(y|x) \sum_{t=0}^{T-1} q_t^i q_{t+1}^j = E(q_t^i q_{t+1}^j)$ as the sufficient statistic for a_{ij}.

In summary, to apply EM, 1. Find the sufficient statistics of CLL for the parameters to be estimated 2. Compute the expectation of the sufficient statistics w.r.t the distribution P(Y—X) 3. Use these expectations as partial count to compute the MLE.

OK, sufficient statistic, MLE, EM, we done! Oh yeah!