A Short Tutorial on Using Expectation-Maximization with Mixture Models

Jason D. M. Rennie
jrennie@csail.mit.edu

March 3, 2004

Abstract

We show how to derive the Expectation-Mazimization (EM) algorithm for mixture models. In a general setting, we show how to obtain a lower bound on the observed data likelihood that is easier to optimize. For a simple mixture example, we solve the update equations and give a "canned" algorithm.

1 EM for Mixture Models

Consider a probability model with unobserved data, $p(x, y|\theta)$, where x represents observed variables and y represents unobserved varaibles. Expectation-Maximization (EM) is an algorithm to find a local maximum of the likelihood of the observed data. It proceeds in rounds. Each round, parameters are chosen to maximize a lower-bound on the likelihood. The lower-bound is then updated so as to be tight for the the new parameter setting.

Let $\theta^{(t)}$ be the current parameter setting. The log-likelihood of the observed data is

$$l(\theta^{(t)}) = \sum_i \log p(x_i|\theta^{(t)}) = \sum_i \log \sum_y p(x_i, y|\theta^{(t)}). \tag{1}$$

We want to find a new parameter setting, $\theta^{(t+1)}$, that increases the log-likelihood of the observed data. In other words, we want to maximize the difference between the original log-likelihood and the new log-likelihood:

$$\theta^{(t+1)} = \arg\max_\theta l(\theta) - l(\theta^{(t)}). \tag{2}$$

Let $Q(\theta, \theta^{(t)}) = l(\theta) - l(\theta^{(t)})$. Note that $p(y|x_i, \theta^{(t)}) = \frac{p(x_i,y|\theta^{(t)})}{\sum_{y'} p(x_i,y'|\theta^{(t)})}$. Consider the following manipulations which result in a lower bound on Q:

$$Q(\theta, \theta^{(t)}) = \sum_i \log \frac{\sum_y p(x_i, y|\theta)}{\sum_{y'} p(x_i, y'|\theta^{(t)})} \tag{3}$$

$$= \sum_i \log \sum_y \frac{p(x_i, y|\theta^{(t)})}{\sum_{y'} p(x_i, y'|\theta^{(t)})} \frac{p(x_i, y|\theta)}{p(x_i, y|\theta^{(t)})} \tag{4}$$

$$= \sum_i \log \sum_y p(y|x_i, \theta^{(t)}) \frac{p(x_i, y|\theta)}{p(x_i, y|\theta^{(t)})} \tag{5}$$

$$= \sum_i \log E_{p(y|x_i,\theta^{(t)})} \left[\frac{p(x_i, y|\theta)}{p(x_i, y|\theta^{(t)})} \right] \tag{6}$$

$$\geq \sum_i E_{p(y|x_i,\theta^{(t)})} \left[\log \frac{p(x_i, y|\theta)}{p(x_i, y|\theta^{(t)})} \right] \tag{7}$$

$$= \sum_i \sum_y p(y|x_i, \theta^{(t)}) \log \frac{p(x_i, y|\theta)}{p(x_i, y|\theta^{(t)})} = L(\theta, \theta^{(t)}). \tag{8}$$

The inequality is a direct result of the concavity of the log function (Jensen's inequality). Call the lower bound $L(\theta, \theta^{(t)})$.

Consider the following (trivial) fact for two arbitrary functions, f and g. Let $x^* = \arg\max_x f(x)$. If $f(x)$ is a lower bound on $g(x)$ (i.e. $f(x) \leq g(x)\ \forall x$), and for some $\overline{x}$, $f(\overline{x}) = g(\overline{x})$, then if $f(x^*) > f(\overline{x})$, then $g(x^*) > g(\overline{x})$. In other words, if moving from $\overline{x}$ to x^* provides an improvement in f, then it also provides an improvement in g. We have constructed L as a lower bound on Q such that $L(\theta^{(t)}, \theta^{(t)}) = Q(\theta^{(t)}, \theta^{(t)})$. Thus, if $L(\theta, \theta^{(t)}) > L(\theta^{(t)}, \theta^{(t)})$, then $Q(\theta, \theta^{(t)}) > Q(\theta^{(t)}, \theta^{(t)})$.

Note that maximizing $L(\theta, \theta^{(t)})$ with respect to θ does not involve the denominator of the log term. In other words, the parameter setting that maximizes L is

$$\theta^{(t+1)} = \arg\max_\theta \sum_i \sum_y p(y|x_i, \theta^{(t)}) \log p(x_i, y|\theta). \tag{9}$$

It is often easier to maximize $L(\theta, \theta^{(t)})$ (with respect to θ) than it is to maximize $Q(\theta, \theta^{(t)})$ (with respect to θ). For example, if $p(x_i, y|\theta)$ is an exponential distribution, $L(\theta, \theta^{(t)})$ is a convex function of θ. For some models, we can solve for the parameters directly, such as in the example discussed in the next section.

[1] is the original Expectation-Maximization paper. [2] discuss the convergence properties and suggest a hybrid algorithm that switches between EM and Conjugate Gradients based on an estimate of the "missing information."

2 A Simple Mixture Example

Consider a two-component mixture model where the observations are sequences of heads and tails. The unobserved variable takes on one of two values, $y \in$

$\{1, 2\}$. Three parameters define the joint distribution, $\theta = \{\lambda_1, \phi_1, \phi_2\}$. λ_1 is the probability of using component #1 to generate the observations. ϕ_1 is the probability of heads for component #1; ϕ_2 is the probability of heads for component #2. We define $\lambda_2 = 1 - \lambda_1$ for convenience. Let n_i be the length of observed sequence i; let h_i be the number of heads. The joint likelihood is

$$p(x_i, y|\theta) = \lambda_y \phi_y^{h_i} (1 - \phi_y)^{(n_i - h_i)}. \tag{10}$$

To maximize the observed data likelihood, we start from an initial setting of the parameters, $\theta^{(0)}$, and iteratively maximize the lower bound. Let

$$\begin{aligned} J(\theta, \theta^{(t)}) &= \sum_i \sum_y p(y|x_i, \theta^{(t)}) \log p(x_i, y|\theta) & (11) \\ &= \sum_i \sum_y p(y|x_i, \theta^{(t)}) \log \lambda_y \phi_y^{h_i} (1 - \phi_y)^{(n_i - h_i)} & (12) \end{aligned}$$

Due to the structure of the function, we can solve for the optimal parameter settings by simply setting the partial derivatives to zero. Let $p_{1i} = p(y = 1|x_i, \theta^{(t)})$, $p_{2i} = p(y = 2|x_i, \theta^{(t)})$. The partial derivative of J with respect to λ_1 is

$$\frac{\partial J}{\partial \lambda_1} = \frac{\sum_i (p_{1i} - \lambda_1)}{\lambda_1 (1 - \lambda_1)} \tag{13}$$

Thus, the maximizing setting of λ_1 is $\lambda_1^* = \frac{1}{m} \sum_{i=1}^m p_{1i}$. The partial of J wrt ϕ_1 is

$$\frac{\partial J}{\partial \phi_1} = \frac{\sum_i p_{1i} h_i - \phi_1 \sum_i p_{1i} n_i}{\phi_1 (1 - \phi_1)} \tag{14}$$

Thus, the maximizing setting of ϕ_1 is $\phi_1^* = \frac{\sum_i p_{1i} h_i}{\sum_i p_{1i} n_i}$. Similarly, the maximizing setting of ϕ_2 is $\phi_2^* = \frac{\sum_i p_{2i} h_i}{\sum_i p_{2i} n_i}$. We set $\theta^{(t+1)} = (\lambda_1^*, \phi_1^*, \phi_2^*)$ and repeat. Figure 1 gives a concise summary of the implementation of EM for this example.

The "canned" algorithms given in [3] (Appendix B) provide useful criteria for determining convergence.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society series B*, 39:1–38, 1977.

[2] Ruslan Salakhutdinov, Sam Rowies, and Zoubin Ghahramani. Optimization with EM and expectation-conjugate-gradient. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, 2003.

[3] Jonathan Richard Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. http://www.cs.cmu.edu/~jrs/jrspapers.html, 1994.

- Randomly choose an inital parameter setting, $\theta^{(0)}$.
- Let $t = 0$. Repeat until convergence.
  - Let $(\lambda_1, \phi_1, \phi_2) := \theta^{(t)}$, $\lambda_2 := 1 - \lambda_1$.
  - Let $p_{yi} := \frac{\lambda_y \phi_y^{h_i}(1-\phi_y)^{(n_i - h_i)}}{\sum_{y'} \lambda_{y'} \phi_{y'}^{h_i}(1-\phi_{y'})^{(n_i - h_i)}}$ for $y \in \{1, 2\}$, $i \in \{1, \ldots, m\}$.
  - Let $\lambda_1^* := \frac{1}{m} \sum_{i=1}^{m} p_{1i}$
  - Let $\phi_1^* := \frac{\sum_i p_{1i} h_i}{\sum_i p_{1i} n_i}$.
  - Let $\phi_2^* := \frac{\sum_i p_{2i} h_i}{\sum_i p_{2i} n_i}$.
  - Let $\theta^{(t+1)} := (\lambda_1^*, \phi_1^*, \phi_2^*)$.
  - Let $t := t + 1$.

Figure 1: A summary of using the EM algorithm for the simple mixture example.