The EM Algorithm

Ajit Singh

November 20, 2005

1 Introduction

Expectation-Maximization (EM) is a technique used in point estimation. Given a set of observable variables X and unknown (latent) variables Z we want to estimate parameters θ in a model.

Example 1.1 (Binomial Mixture Model). You have two coins with unknown probabilities of heads, denoted p and q respectively. The first coin is chosen with probability π and the second coin is chosen with probability $1 - \pi$. The chosen coin is flipped once and the result is recorded. $x = \{1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1\}$ (Heads = 1, Tails = 0). Let $Z_i \in \{0, 1\}$ denote which coin was used on each toss.

In example 1.1 we added latent variables Z_i for reasons that will become apparent. The parameters we want to estimate are $\theta = (p, q, \pi)$. Two criteria for point estimation are maximum likelihood and maximum a posteriori:

$$\begin{aligned}
\hat{\theta}_{ML} &= \arg\max_{\theta} \log p(x|\theta) \\
\hat{\theta}_{MAP} &= \arg\max_{\theta} \log p(x, \theta) \\
&= \arg\max_{\theta} \left[\log p(x|\theta) + \log p(\theta)\right]
\end{aligned}$$

Our presentation will focus on the maximum likelihood case (ML-EM); the maximum a posteriori case (MAP-EM) is very similar[1].

2 Notation

X	Observed variables
Z	Latent (unobserved) variables
$\theta^{(t)}$	The estimate of the parameters at iteration t.
$\ell(\theta)$	The marginal log-likelihood $\log p(x|\theta)$
$\log p(x, z|\theta)$	The complete log-likelihood, *i.e.*, when we know the value of Z.
$q(z|x, \theta)$	Averaging distribution, a free distribution that EM gets to vary.
$Q(\theta|\theta^{(t)})$	The expected complete log-likelihood $\sum_z q(z|x, \theta) \log p(x, z|\theta)$
$H(q)$	Entropy of the distribution $q(z|x, \theta)$.

[1]In MAP-EM the M-step is a MAP estimate, instead of an ML estimate.

3 Derivation

We could directly maximize $\ell(\theta) = \sum_z \log p(x, z|\theta)$ using a gradient method (*e.g.*, gradient ascent, conjugate gradient, quasi-Newton) but sometimes the gradient is hard to compute, hard to implement, or we do not want to bother adding in a black-box optimization routine.

Consider the following inequality

$$\ell(\theta) = \log p(x|\theta) = \log \sum_z p(x, z|\theta) \tag{1}$$

$$= \log \sum_z q(z|x, \theta) \frac{p(x, z|\theta)}{q(z|x.\theta)} \tag{2}$$

$$\geq \sum_z q(z|x, \theta) \log \frac{p(x, z|\theta)}{q(z|x, \theta)} \equiv F(q, \theta) \tag{3}$$

where $q(z|x, \theta)$ is an arbitrary density over Z. This inequality is foundational to what are called "variational methods" in the machine learning literature[2]. Instead of maximizing $\ell(\theta)$ directly, EM maximizes the lower-bound $F(q, \theta)$ via coordinate ascent:

$$\textbf{E-step}: q^{(t+1)} = \arg\max_q F(q, \theta^{(t)}) \tag{4}$$

$$\textbf{M-step}: \theta^{(t+1)} = \arg\max_\theta F(q^{(t+1)}, \theta) \tag{5}$$

Starting with some initial value of the parameters $\theta^{(0)}$, one cycles between the E and M-steps until $\theta^{(t)}$ converges to a local maxima. Computing equation 4 directly involves fixing $\theta = \theta^{(t)}$ and optimizating over the space of distributions, which looks painful. However, it is possible to show that $q^{(t+1)} = p(z|x, \theta^{(t)})$. We can stop worrying about q as a variable over the space of distributions, since we know the optimal q is a distribution that depends on $\theta^{(t)}$. To compute equation 5 we fix q and note that

$$\ell(\theta) \geq F(q, \theta) \tag{6}$$

$$= \sum_z q(z|x, \theta) \log \frac{p(x, z|\theta)}{q(z|x, \theta)} \tag{7}$$

$$= \sum_z q(z|x, \theta) \log p(x, z|\theta) - \sum_z q(z|x, \theta) \log q(z|x, \theta) \tag{8}$$

$$= Q(\theta|\theta^{(t)}) + H(q) \tag{9}$$

so maximizing $F(q, \theta)$ is equivalent to maximizing the expected complete log-likelihood. Obscuring these details, which explain what EM is doing, we can re-express equations 4 and 5 as

$$\textbf{E-step}: \text{Compute } Q(\theta|\theta^{(t)}) = E_{p(z|x,\theta^{(t)})}[\log p(x, z|\theta)] \tag{10}$$

$$\textbf{M-step}: \theta^{(t+1)} = \arg\max_\theta E_{p(z|x,\theta^{(t)})}[\log p(x, z|\theta)] \tag{11}$$

[2] If you feel compelled to tart it up, you can call equation 3 Gibbs inequality and $F(q, \theta)$ the negative variational free energy.

3.1 Limitations of EM

EM is useful for several reasons: conceptual simplicity, ease of implementation, and the fact that each iteration improves $\ell(\theta)$. The rate of convergence on the first few steps is typically quite good, but can become excruciatingly slow as you approach a local optima. Generally, EM works best when the fraction of missing information is small[3] and the dimensionality of the data is not too large. EM can require many iterations, and higher dimensionality can dramatically slow down the E-step.

4 Using the EM algorithm

Applying EM to example 1.1 we start by writing down the expected complete log-likelihood

$$\begin{aligned}
Q(\theta|\theta^{(t)}) &= E\left[\log \prod_{i=1}^{n} [\pi p^{x_i}(1-p)^{1-x_i}]^{z_i}[(1-\pi)q^{x_i}(1-q)^{1-x_i}]^{1-z_i}\right] \\
&= \sum_{i=1}^{n} E[z_i|x_i,\theta^{(t)}][\log \pi + x_i \log p + (1-x_i)\log(1-p)] \\
&\quad + (1-E[z_i|x_i,\theta^{(t)}])[\log(1-\pi) + x_i \log q + (1-x_i)\log(1-q)]
\end{aligned}$$

Next we compute $E[z_i|x_i,\theta^{(t)}]$

$$\begin{aligned}
\mu_i^{(t)} = E[z_i|x_i,\theta^{(t)}] &= p(z_i = 1|x_i,\theta^{(t)}) \\
&= \frac{p(x_i|z_i,\theta^{(t)})p(z_i=1|\theta^{(t)})}{p(x_i|\theta^{(t)})} \\
&= \frac{\pi[p^{(t)}]^{x_i}[(1-p^{(t)})]^{1-x_i}}{\pi^{(t)}[p^{(t)}]^{x_i}[(1-p^{(t)}]^{1-x_i} + (1-\pi^{(t)})[q^{(t)}]^{x_i}[(1-q^{(t)})]^{1-x_i}}
\end{aligned}$$

Maximizing $Q(\theta|\theta^{(t)})$ w.r.t. θ yields the update equations

$$\begin{aligned}
\frac{\partial Q(\theta|\theta^{(t)})}{\partial \pi} = 0 &\implies \pi^{(t+1)} = \frac{1}{n}\sum_i \mu_i^{(t)} \\
\frac{\partial Q(\theta|\theta^{(t)})}{\partial p} = 0 &\implies p^{(t+1)} = \frac{\sum_i \mu_i^{(t)} x_i}{\sum_i \mu_i^{(t)}} \\
\frac{\partial Q(\theta|\theta^{(t)})}{\partial q} = 0 &\implies q^{(t+1)} = \frac{\sum_i (1-\mu_i^{(t)})x_i}{\sum_i (1-\mu_i^{(t)})}
\end{aligned}$$

4.1 Constrained Optimization

Sometimes the M-step is a constrained maximization, which means that there are constraints on valid solutions not encoded in the function itself. An example of a constrained optimization is to

[3] The statement "fraction of missing information is small" can be quantified using Fisher information.

maximize

$$H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{12}$$

$$\text{such that} \quad \sum_{i=1}^{n} p_i = 1 \tag{13}$$

Such problems can be solved using the method of Lagrange multipliers. To maximize a function $f(p_1, \ldots, p_n)$ on the open set $\mathbf{p} = (p_1, \ldots, p_n) \subset \mathbb{R}^n$ subject to the constraint $g(\mathbf{p}) = 0$ it suffices to maximize the unconstrained function

$$\Lambda(\mathbf{p}, \lambda) = f(\mathbf{p}) - \lambda g(\mathbf{p})$$

To solve equation 12 we encode the constraint as $g(\mathbf{p}) = \sum_i p_i - 1$ and maximize

$$\Lambda(\mathbf{p}, \lambda) = -\sum_{i=1}^{n} p_i \log_2 p_i - \lambda \left(\sum_{i=1}^{n} p_i - 1 \right)$$

in the unusual unconstrained manner, by solving the system of equations

$$\frac{\partial \Lambda(\mathbf{p}, \lambda)}{\partial p_i} = 0, \quad \frac{\partial \Lambda(\mathbf{p}, \lambda)}{\partial \lambda} = 0$$

which leads to the solution $p_i = \frac{1}{n}$.

Acknowledgements: The idea of EM as coordinate ascent was first presented in "A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants", by R.M. Neal and G.E. Hinton. This presentation is also indebited to an unpublished manuscript by M.I. Jordan and C. Bishop.