

The Expectation-Maximization Algorithm

Charles Elkan
elkan@cs.ucsd.edu

November 16, 2007

This chapter explains the EM algorithm at multiple levels of generality. Section 1 gives the standard high-level version of the algorithm. Section 2 then extends this explanation to make EM applicable to problems with many training examples. Next, Section 3 explains how EM can be used for fitting a mixture of arbitrary component distributions. Finally, Section 4 explains how the EM algorithm can be viewed as a double maximization, and Section 5 explains Jensen's inequality, the basic mathematical fact that underlies all versions of the EM algorithm.

At the lowest, most concrete level, there are different EM algorithms for fitting many particular probabilistic models; a mixture of Gaussians is just one example. The previous chapter describes EM at this lowest level. At an intermediate level, EM for any mixture model involves an E-step that computes degrees of membership, and an M-step that does weighted maximum likelihood; this level is the topic of Section 3 below. More abstractly, EM is an iterative method for maximizing likelihood; this level is explained in Section 1. At an even higher level, EM involves two maximizations; this point of view is explained in Section 4.

Many other tutorial explanations of expectation-maximization exist, including [Bil98, Del02, Bor04]. Three are especially recommended: [Min98, Rus98, Roc07].

1 The general EM algorithm

To simplify notation, assume initially that the entire training data constitute one outcome x of a random variable X . Also let θ be all the parameters of the model

$p(x; \theta)$. The goal, according to the principle of maximum likelihood, is to choose θ to maximize the likelihood function, which is $L(\theta; x) = p(x; \theta)$.

Let Z be any discrete auxiliary random variable whose distribution, like that of X , is a function of θ . Let z range over the possible outcomes of Z and note that by definition $p(x; \theta) = \sum_z p(x, z; \theta)$.

Suppose we have a current estimate θ_t for the parameters. Multiplying inside this sum by $p(z|x; \theta_t)/p(z|x; \theta_t)$ gives that the log likelihood is

$$D = \log p(x; \theta) = \log \sum_z p(x, z; \theta) \frac{p(z|x; \theta_t)}{p(z|x; \theta_t)}.$$

Note that $\sum_z p(z|x; \theta_t) = 1$ and $p(z|x; \theta_t) \geq 0$ for all z . Therefore D is the logarithm of a weighted sum, so we can apply Jensen's inequality, which says $\log \sum_j w_j v_j \geq \sum_j w_j \log v_j$, given $\sum_j w_j = 1$ and each $w_j \geq 0$. Here, we let the sum range over the values z of Z , with the weight w_j being $p(z|x; \theta_t)$. We get

$$D \geq E = \sum_z p(z|x; \theta_t) \log \frac{p(x, z; \theta)}{p(z|x; \theta_t)}.$$

Separating the fraction inside the logarithm to obtain two sums gives

$$E = \left(\sum_z p(z|x; \theta_t) \log p(x, z; \theta) \right) - \left(\sum_z p(z|x; \theta_t) \log p(z|x; \theta_t) \right).$$

Since $E \leq D$ and we want to maximize D , consider maximizing E . The weights $p(z|x; \theta_t)$ do not depend on θ , so we only need to maximize the first sum, which is

$$\sum_z p(z|x; \theta_t) \log p(x, z; \theta).$$

In general, the E-step of an EM algorithm is to compute $p(z|x; \theta_t)$ for all z . The M-step is then to find θ to maximize $\sum_z p(z|x; \theta_t) \log p(x, z; \theta)$.

How do we know that maximizing E actually leads to an improvement in the likelihood? With $\theta = \theta_t$,

$$E = \sum_z p(z|x; \theta_t) \log \frac{p(x, z; \theta_t)}{p(z|x; \theta_t)} = \sum_z p(z|x; \theta_t) \log p(x; \theta_t) = \log p(x; \theta_t)$$

which is the log likelihood at θ_t . So any θ that maximizes E must lead to a likelihood that is better than the likelihood at θ_t .

2 EM with independent training examples

The EM algorithm derived above can be extended to the case where we have a training set $\{x_1, \dots, x_n\}$ such that each x_i is independent. In this case the log likelihood is

$$D = \sum_i \log p(x_i; \theta).$$

Let the auxiliary random variables be a set $\{Z_1, \dots, Z_n\}$ such that the distribution of each Z_i is a function only of x_i and θ . By an argument similar to above,

$$D = \sum_i \log \sum_{z_i} p(x_i, z_i; \theta) \frac{p(z_i|x_i; \theta_t)}{p(z_i|x_i; \theta_t)}.$$

Using Jensen's inequality separately for each i gives

$$D \geq E = \sum_i \sum_{z_i} p(z_i|x_i; \theta_t) \log \frac{p(x_i, z_i; \theta)}{p(z_i|x_i; \theta_t)}.$$

As before, to maximize E we want to maximize the sum

$$\sum_i \sum_{z_i} p(z_i|x_i; \theta_t) \log p(x_i, z_i; \theta).$$

The E-step is to compute $p(z_i|x_i; \theta_t)$ for all z_i for each i . The M-step is then to find

$$\theta_{t+1} = \operatorname{argmax}_{\theta} \sum_i \sum_{z_i} p(z_i|x_i; \theta_t) \log p(x_i, z_i; \theta).$$

3 EM for mixture models

For a mixture model with K components each z_i is between 1 and K . The sum to maximize is

$$\sum_i \sum_k p(k|x_i; \theta_t) [\log p(k; \theta) + \log p(x_i|k; \theta)].$$

Using the mixture model notation from above, we have $p(k; \theta) = \alpha_k$ and $p(x_i|k; \theta) = f(x_i; \lambda_k)$. The sum to maximize is then

$$E = \sum_i \sum_k p(k|x_i; \theta_t) [\log \alpha_k + \log f(x_i; \lambda_k)].$$

For the E-step we use Bayes' rule:

$$w_{ik} = p(k|x_i; \theta_t) = \frac{f(x_i; \lambda_k)\alpha_k}{\sum_j f(x_i; \lambda_j)\alpha_j}.$$

For the M-step, the two terms inside the square brackets in E involve disjoint sets of parameters, so we can do two separate maximizations. The first one is to maximize

$$\sum_i \sum_k w_{ik} \log \alpha_k = \sum_k c_k \log \alpha_k$$

where $c_k = \sum_i w_{ik}$, subject to the constraint $\sum_k \alpha_k = 1$. Using a Lagrange multiplier, one can show that the solution is

$$\alpha_k = \frac{c_k}{\sum_j c_j}.$$

The second one is to maximize

$$\sum_i \sum_k w_{ik} \log f(x_i; \lambda_k).$$

This can be divided into K separate maximizations, each of the form

$$\lambda_k = \operatorname{argmax}_\lambda \sum_i w_{ik} \log f(x_i; \lambda).$$

Each of these maximizations is a weighted maximum-likelihood problem, as claimed previously.

4 EM as double maximization

This section gives a simplified explanation of a point of view on the EM algorithm due originally to [NH98].

The standard EM algorithm uses the weights $p(z|x; \theta_t)$, but other weights g_z may also be used. For any such weights, the log likelihood can be written

$$D = \log p(x; \theta) = \log \sum_z p(x, z; \theta) \frac{g_z}{g_z}$$

and Jensen's inequality is applicable if $\sum_z g_z = 1$ and $g_z \geq 0$ for all z . In this case

$$D \geq E = \sum_z g_z \log \frac{p(x, z; \theta)}{g_z}.$$

The overall goal is to maximize D , so consider choosing g_z and choosing θ to maximize E . However, rather than choosing g_z and θ simultaneously, suppose we choose g_z first based on $\theta = \theta_t$, and then choose θ based on the new g_z .

The first maximization is of

$$E = \sum_z g_z (\log p(x, z; \theta_t) - \log g_z).$$

with θ_t fixed. Introducing a Lagrange multiplier λ for the constraint $\sum_z g_z = 1$ gives the unconstrained objective function

$$F = \lambda(1 - \sum_z g_z) + \sum_z g_z (\log p(x, z; \theta_t) - \log g_z).$$

The partial derivatives are

$$\frac{\partial F}{\partial g_z} = -\lambda + (-1) + \log p(x, z; \theta_t) - \log g_z.$$

Solving for when the partial derivatives equal zero yields

$$\log g_z = \text{constant} + \log p(x, z; \theta_t).$$

The constraint $\sum_z g_z = 1$ gives

$$g_z = \frac{p(x, z; \theta_t)}{\sum_z p(x, z; \theta_t)} = \frac{p(x, z; \theta_t)}{p(x; \theta_t)} = p(z|x; \theta_t)$$

which are the weights used in the standard EM algorithm. As shown above, with these weights and with $\theta = \theta_t$, $E = \log p(x; \theta_t)$ which is the log likelihood D at θ_t .

In the double maximization version of EM, both the E and M steps are maximizations. The E-step is to solve

$$w_z = \operatorname{argmax}_{g_z} \sum_z g_z \log \frac{p(x, z; \theta_t)}{g_z}$$

while the M-step solves

$$\theta_{t+1} = \operatorname{argmax}_{\theta} \sum_z w_z \log p(x, z; \theta).$$

5 Jensen's inequality

The mathematical fact on which the EM algorithm is based is known as Jensen's inequality. It is the following lemma.

Lemma: Suppose the weights w_j are nonnegative and sum to one, and let each x_j be any real number for $j = 1$ to $j = n$. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be any concave function. Then

$$f\left(\sum_j w_j x_j\right) \geq \sum_j w_j f(x_j).$$

Proof: The proof is by induction on n . For the base case $n = 2$, the definition of being concave says that

$$f(wa + (1 - w)b) \geq wf(a) + (1 - w)f(b).$$

The logarithm function is concave, so Jensen's inequality applies to it.

References

- [Bil98] Jeff A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report TR-97-021, U. C. Berkeley, April 1998.
- [Bor04] Sean Borman. The expectation maximization algorithm: A short tutorial. Unpublished paper available at <http://www.seanborman.com/publications>, 2004.
- [Del02] Frank Dellaert. The expectation maximization algorithm. Technical Report GIT-GVU-02-20, Georgia Institute of Technology, 2002.
- [Min98] Thomas P. Minka. Expectation-maximization as lower bound maximization. Unpublished paper available at <http://research.microsoft.com/~minka>, 1998.
- [NH98] Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pages 355–368, Norwell, MA, USA, 1998. Kluwer Academic Publishers.

- [Roc07] Alexis Roche. EM algorithms and variants: An informal tutorial. Unpublished paper available at <http://www.madic.org/people/roche/>, 2007.
- [Rus98] Stuart Russell. The EM algorithm. Unpublished note available at <http://www.cs.berkeley.edu/~russell/classes/cs281/s98/em.ps>, 1998.