A Note on the Expectation-Maximization (EM) Algorithm

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

November 2, 2004

1 Introduction

The Expectation-Maximization (EM) algorithm is a general algorithm for maximum-likelihood estimation where the data are "incomplete" or the likelihood function involves latent variables. Note that the notion of "incomplete data" and "latent variables" are related: when we have a latent variable, we may regard our data as being incomplete since we do not observe values of the latent variables; similarly, when our data are incomplete, we often can also associate some latent variable with the missing data. For language modeling, the EM algorithm is often used to estimate parameters of a mixture model, in which the exact component model from which a data point is generated is hidden from us.

Informally, the EM algorithm starts with randomly assigning values to all the parameters to be estimated. It then iterately alternates between two steps, called the expectation step (i.e., the "E-step") and the maximization step (i.e., the "M-step"), respectively. In the E-step, it computes the expected likelihood for the complete data (the so-called Q-function) where the expectation is taken w.r.t. the computed conditional distribution of the latent variables (i.e., the "hidden variables") given the current settings of parameters and our observed (incomplete) data. In the M-step, it re-estimates all the parameters by maximizing the Q-function. Once we have a new generation of parameter values, we can repeat the E-step and another M-step. This process continues until the likelihood converges, i.e., reaching a local maxima. Intuitively, what EM does is to iteratively "augment" the data by "guessing" the values of the hidden variables and to re-estimate the parameters by assuming that the guessed values are the true values.

The EM algorithm is a hill-climbing approach, thus it can only be guanranteed to reach a local maxima. When there are multiple maximas, whether we will actually reach the global maxima clearly depends on where we start; if we start at the "right hill", we will be able to find a global maxima. When there are multiple local maximas, it is often hard to identify the "right hill". There are two commonly used strategies to solving this problem. The first is that we try many different initial values and choose the solution that has the highest converged likelihood value. The second uses a much simpler model (ideally one with a unique global maxima) to determine an initial value for more complex models. The idea is that a simpler model can hopefully help locate a rough region where the global optima exists, and we start from a value in that region to search for a more accurate optima using a more complex model.

There are many good tutorials on the EM algorithm (e.g., [2, 5, 1, 4, 3]). In this note, we introduce the EM algorithm through a specific problem – estimating a simple mixture model.

2 A simple mixture unigram language model

In the mixture model feedback approach [6], we assume that the feedback documents $\mathcal{F} = \{d_1, ..., d_k\}$ are "generated" from a mixture model with two multinomial component models. One component model is the background model $p(w|C)$ and the other is an unknown topic language model $p(w|\theta_F)$ to be estimated. (w is a word.) The idea is to model the common (non-discriminative) words in $\mathcal{F}$ with $p(w|C)$ so that the topic model θ_F would attract more discriminative content-carrying words.

The log-likelihood of the feedback document data for this mixture model is

$$\log L(\theta_F) = \log p(\mathcal{F} \,|\, \theta_F) \quad = \quad \sum_{i=1}^{k} \sum_{j=1}^{|d_i|} \log((1-\lambda)p(d_{ij} \,|\, \theta_F) + \lambda p(d_{ij} \,|\, \mathcal{C}))$$

where d_{ij} is the j-th word in document d_i, $|d_i|$ is the length of d_i, and λ is a parameter that indicates the amount of "background noise" in the feedback documents, which will be set empirically. We thus assume λ to be known, and want to estimate $p(w|\theta_F)$.

3 Maximum Likelihood Estimation

A common method for estimating θ_F is the maximum likelihood (ML) estimator, in which we choose a θ_F that maximizes the likelihood of $\mathcal{F}$. That is, the estimated topic model (denoted by $\hat{\theta}_F$) is given by

$$\begin{aligned} \hat{\theta}_F &= \arg\max_{\theta_F} L(\theta_F) & (1) \\ &= \arg\max_{\theta_F} \sum_{i=1}^{k} \sum_{j=1}^{|d_i|} \log((1-\lambda)p(d_{ij} \,|\, \theta_F) + \lambda p(d_{ij} \,|\, \mathcal{C})) & (2) \end{aligned}$$

The right-side of this equation is easily seen to be a function with $p(w|\theta_F)$ as variables. To find $\hat{\theta}_F$, we can, in principle, use any optimization methods. Since the function involves a logarithm of a sum of two terms, it is difficult to obtain a simple analytical solution via the Lagrange Multiplier approach, so in general, we must rely on numerical algorithms. There are many possibilities; EM happens to be just one of them which is quite natural and guaranteed to converge to a local maxima, which, in our case, is also a global maxima, since the likelihood function can be shown to have one unique maxima.

4 Incomplete vs. Complete Data

The main idea of the EM algorithm is to "augment" our data with some latent/hidden variables so that the "complete" data has a much simpler likelihood function – simpler for the purpose of finding a maxima. The original data are thus treated as "incomplete". As we will see, we will maximize the incomplete data likelihood (our original goal) through maximizing the expected complete data likelihood (since it is much easier to maximize) where expectation is taken over all possible values of the hidden variables (since the complete data likelihood, unlike our original incomplete data likelihood, would contain hidden variables).

In our example, we introduce a binary hidden variable z for each *occurrence* of a word w to indicate whether the word has been "generated" from the background model $p(w|C)$ or the topic model $p(w|\theta_F)$. Let d_{ij} be the j-th word in document d_i. We have a corresponding variable z_{ij} defined as follows:

$$z_{ij} = \begin{cases} 1 & \text{if word } d_{ij} \text{ is from background} \\ 0 & \text{otherwise} \end{cases}$$

We thus assume that our complete data would have contained not only all the words in $\mathcal{F}$, but also their corresponding values of z. The log-likelihood of the complete data is thus

$$\begin{aligned} L_c(\theta_F) &= \log p(\mathcal{F}, \mathbf{z} \,|\, \theta_F) \\ &= \sum_{i=1}^{k} \sum_{j=1}^{|d_i|} [(1 - z_{ij}) \log((1-\lambda) p(d_{ij} \,|\, \theta_F)) + z_{ij} \log(\lambda p(d_{ij} \,|\, \mathcal{C}))] \end{aligned}$$

Note the difference between $L_c(\theta_F)$ and $L(\theta_F)$: the sum is outside of the logarithm in $L_c(\theta_F)$, and this is possible because we assume that we *know* which component model has been used to generated each word d_{ij}.

What is the relationship between $L_c(\theta_F)$ and $L(\theta_F)$? In general, if our parameter is θ, our original data is X, and we augment it with a hidden variable H, then $p(X, H|\theta) = p(H|X, \theta) p(X|\theta)$. Thus,

$$L_c(\theta) = \log p(X, H|\theta) = \log p(X|\theta) + \log p(H|X, \theta) = L(\theta) + \log p(H|X, \theta)$$

5 A Lower Bound of Likelihood

Algorithmically, the basic idea of EM is to start with some initial guess of the parameter values $\theta^{(0)}$ and then iteratively search for better values for the parameters. Assuming that the current estimate of the parameters is $\theta^{(n)}$, our goal is to find another $\theta^{(n+1)}$ that can improve the likelihood $L(\theta)$.

Let us consider the difference between the likelihood at a potentially better parameter value θ and the likelihood at the current estimate $\theta^{(n)}$, and relate it with the corresponding difference in the complete likelihood:

$$L(\theta) - L(\theta^{(n)}) = L_c(\theta) - L_c(\theta^{(n)}) + \log \frac{p(H|X, \theta^{(n)})}{p(H|X, \theta)} \tag{3}$$

Our goal is to maximize $L(\theta) - L(\theta^{(n)})$, which is equivalent to maximizing $L(\theta)$. Now take the expectation of this equation w.r.t. the conditional distribution of the hidden variable given the data X and the current estimate of parameters $\theta^{(n)}$, i.e., $p(H|X, \theta^{(n)}$. We have

$$L(\theta) - L(\theta^{(n)}) = \sum_H L_c(\theta) p(H|X, \theta^{(n)}) - \sum_H L_c(\theta^{(n)}) p(H|X, \theta^{(n)}) + \sum_H p(H|X, \theta^{(n)}) \log \frac{p(H|X, \theta^{(n)})}{p(H|X, \theta)}$$

Note that the left side of the equation remains the same as the variable H does not occur there. The last term can be recognized as the KL-divergence of $p(H|X, \theta^{(n)})$ and $p(H|X, \theta)$, which is always non-negative. We thus have

$$L(\theta) - L(\theta^{(n)}) \geq \sum_H L_c(\theta) p(H|X, \theta^{(n)}) - \sum_H L_c(\theta^{(n)}) p(H|X, \theta^{(n)})$$

or

$$L(\theta) \quad \geq \quad \sum_H L_c(\theta)p(H|X,\theta^{(n)}) + L(\theta^{(n)}) - \sum_H L_c(\theta^{(n)})p(H|X,\theta^{(n)}) \tag{4}$$

We thus obtain a lower bound for the original likelihood function. The main idea of EM is to maximize this lower bound so as to maximize the original (incomplete) likelihood. Note that the last two terms in this lower bound can be treated as constants as they do not contain the variable θ, so the lower bound is essentially the first term, which is the expectation of the complete likelihood, or the so-called "Q-function" denoted by $Q(\theta;\theta^{(n)})$.

$$Q(\theta;\theta^{(n)}) = E_{p(H|X,\theta^{(n)})}[L_c(\theta)] = \sum_H L_c(\theta)p(H|X,\theta^{(n)})$$

The Q-function for our mixture model is the following

$$\begin{aligned} Q(\theta_F;\theta_F^{(n)}) &= \sum_{\mathbf{z}} L_c(\theta_F)p(\mathbf{z}|\mathcal{F},\theta_F^{(n)}) & (5)\\ &= \sum_{i=1}^{k}\sum_{j=1}^{|d_i|}[p(z_{ij}=0|\mathcal{F},\theta_F^{(n)})\log((1-\lambda)p(d_{ij}\,|\,\theta_F)) + p(z_{ij}=1|\mathcal{F},\theta_F^{(n)})\log(\lambda p(d_{ij}\,|\,\mathcal{C}))] & (6)\end{aligned}$$

6 The General Procedure of EM

Clearly, if we find a $\theta^{(n+1)}$ such that $Q(\theta^{(n+1)};\theta^{(n)}) > Q(\theta^{(n)};\theta^{(n)})$, then we will also have $L(\theta^{(n+1)}) > L(\theta^{(n)})$. Thus the general procedure of the EM algorithm is the following

1. Initialize $\theta^{(0)}$ randomly or heuristically according to any prior knowledge about where the optimal parameter value might be.
2. Iteratively improve the estimate of θ by alternating between the following two-steps:
   (a) The E-step (expectation): Compute $Q(\theta;\theta^{(n)})$
   (b) The M-step (maximization): Re-estimate θ by maximizing the Q-function:
   $$\theta^{(n+1)} = argmax_\theta Q(\theta;\theta^{(n)})$$
3. Stop when the likelihood $L(\theta)$ converges.

As mentioned earlier, the complete likelihood $L_c(\theta)$ is much easier to maximize as the values of the hidden variable are assumed to be known. This is why the Q-function, which is an expectation of $L_c(\theta)$, is often much easier to maximize than the original likelihood function. In cases when there does not exist a natural latent variable, we often introduce a hidden variable so that the complete likelihood function is easy to maximize.

The major computation to be carried out in the E-step is to compute $p(H|X,\theta^{(n)})$, which is sometimes very complicated. In our case, this is simple:

$$p(z_{ij}=1|\mathcal{F},\theta_F^{(n)}) \quad = \quad \frac{\lambda p(d_{ij}|C)}{\lambda p(d_{ij}|C) + (1-\lambda)p(d_{ij}|\theta_F^{(n)})} \tag{7}$$

And of course, $p(z_{ij} = 0|\mathcal{F}, \theta_F^{(n)}) = 1 - p(z_{ij} = 1|\mathcal{F}, \theta_F^{(n)})$. Note that, in general, z_{ij} may depend on all the words in $\mathcal{F}$. In our model, however, it only depends on the corresponding word d_{ij}.

The M-step involves maximizing the Q-function. This may sometimes be quite complex as well. But, again, in our case, we can find an analytical solution. In order to achieve this, we use the Lagrange multiplier method since we have the following contraint on the parameter variables $\{p(w|\theta_F)\}_{w\in V}$, where V is our vocabulary.

$$\sum_{w\in V} p(w|\theta_F) = 1$$

We thus consider the following auxiliary function

$$g(\theta_F) = Q(\theta_F; \theta_F^{(n)}) + \mu(1 - \sum_{w\in V} p(w|\theta_F))$$

. and take its derivative w.r.t. each parameter variable $p(w|\theta_F)$.

$$\frac{\partial g(\theta_F)}{\partial p(w|\theta_F)} = [\sum_{i=1}^{k} \sum_{j=1,d_{ij}=w}^{|d_i|} \frac{p(z_{ij} = 0|\mathcal{F}, \theta_F^{(n)})}{p(w \mid \theta_F)}] - \mu \tag{8}$$

Setting this derivative to zero and solving the equation for $p(w|\theta_F)$, we obtain

$$\begin{aligned} p(w|\theta_F) &= \frac{\sum_{i=1}^{k}\sum_{j=1,d_{ij}=w}^{|d_i|} p(z_{ij} = 0|\mathcal{F}, \theta_F^{(n)})}{\sum_{i=1}^{k}\sum_{j=1}^{|d_i|} p(z_{ij} = 0|\mathcal{F}, \theta_F^{(n)})} && (9) \\ &= \frac{\sum_{i=1}^{k} p(z_w = 0|\mathcal{F}, \theta_F^{(n)})c(w, d_i)}{\sum_{i=1}^{k}\sum_{w\in V} p(z_w = 0|\mathcal{F}, \theta_F^{(n)})c(w, d_i)} && (10) \end{aligned}$$

Note that we changed the notaion so that the sum over each word position in document d_i is now a sum over all the distinct words in the vocabulary. This is possible, because $p(z_{ij}|\mathcal{F}, \theta_F^{(n)})$ depends only on the corresponding word d_{ij}. Using word w, rather then the word *occurrence* d_{ij}, to index z, we have

$$p(z_w = 1|\mathcal{F}, \theta_F^{(n)}) = \frac{\lambda p(w|C)}{\lambda p(w|C) + (1-\lambda)p(w|\theta_F^{(n)})} \tag{11}$$

We therefore have the following EM updating formulas for our simple mixture model:

$$\begin{aligned} p(z_w = 1|\mathcal{F}, \theta_F^{(n)}) &= \frac{\lambda p(w|C)}{\lambda p(w|C) + (1-\lambda)p(w|\theta_F^{(n)})} \quad \text{E-step} && (12) \\ p(w|\theta_F^{(n+1)}) &= \frac{\sum_{i=1}^{k}(1 - p(z_w = 1|\mathcal{F}, \theta_F^{(n)}))c(w, d_i)}{\sum_{i=1}^{k}\sum_{w\in V}(1 - p(z_w = 1|\mathcal{F}, \theta_F^{(n)})c(w, d_i))} \quad \text{M-step} && (13) \end{aligned}$$

Note that we never need to *explicitely* compute the Q-function; instead, we compute the distribution of the hidden variable z and then directly obtain the new parameter values that will maximize the Q-function.

References

[1] J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. 1997. Technical Report, University of Berkeley, ICSI-TR-97-021, 1997.

[2] J. Lafferty. Notes on the em algorithm. *Online article*. http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/11761-s97/WWW/tex/em.ps.

[3] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley and Sons, Inc., 1997.

[4] T. P. Minka. Expectation-maximization as lower bound maximization. *Online article*. http://citeseer.nj.nec.com/minka98expectationmaximization.html.

[5] R. Rosenfeld. The em algorithm. *Online article*. http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/11761-s97/WWW/tex/EM.ps.

[6] C. Zhai and J. Lafferty. Model-based feedback in the KL-divergence retrieval model. In *Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, pages 403–410, 2001.