The EM Algorithm

Seminar "Theoretical Topics in Data Science"

Aniket Phutane

January 12, 2022

RWTH Aachen University

If we define a joint distribution over observed and latent variables, the corresponding distribution of the observed variables alone is obtained by marginalization. This allows relatively complex marginal distributions over observed variables to be expressed in terms of more tractable joint distributions over the expanded space of observed and latent variables. The introduction of latent variables thereby allows complicated distributions to be formed from simpler components. A general technique for finding maximum likelihood estimators in latent variable models is the expectation-maximization (EM) algorithm.

The EM algorithm is used for obtaining maximum likelihood estimates of parameters when some of the data is missing. More generally, however, the EM algorithm can also be applied when there is latent, i.e. unobserved, data that was never intended to be observed in the first place. In that case, we simply assume that the latent data is missing and proceed to apply the EM algorithm.

The EM algorithm is an iterative algorithm that has two main steps. In the E-step, it tries to "guess" the values of the latent random variables. In the M-step, it updates the parameters of our model based on our guesses. Since in the M-step we are pretending that the guesses in the first part were correct, the maximization becomes easy.

The EM algorithm has many applications throughout statistics. It is often used for example, in machine learning and data mining applications, and in Bayesian statistics where it is often used to obtain the mode of the posterior marginal distributions of parameters

1 Introduction

The EM algorithm is used to find (local) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either missing values exist among the data, or the model can be formulated more simply by assuming the existence of further unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component to which each data point belongs.

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values, the parameters, and the latent variables, and simultaneously solving the resulting equations. In statistical models with latent variables, this is usually impossible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation.

The EM algorithm proceeds from the observation that there is a way to solve these two sets of equations numerically. One can simply pick arbitrary values for one of the two sets of unknowns, use them to estimate the second set, then use these new values to find a better estimate of the first set, and then keep alternating between the two until the resulting values both converge to fixed points. It's not obvious that this will work, but it can be proven in this context. Additionally, it can be proven that the derivative of the likelihood is (arbitrarily close to) zero at that point, which in turn means that the point is either a local maximum or a saddle point.[13] In general, multiple maxima may occur, with no guarantee that the global maximum will be found. Some likelihoods also have singularities in them, i.e., nonsensical maxima. For example, one of the solutions that may be found by EM in a mixture model involves setting one of the components to have zero variance and the mean parameter for the same component to be equal to one of the data points.

2 The Classical EM Algorithm

We begin by assuming that the complete data-set consists of Z = (X , Y) but that only X is observed. The complete-data log-likelihood is then denoted by $l(\theta; X, Y)$ where θ is the unknown parameter vector for which we wish to find the MLE.

E-Step: The E-step of the EM algorithm computes the expected value of $l(\theta; X, Y)$ given the observed data, X, and the current parameter estimate, θold say. In particular, we define

$$Q(\theta; \theta_{old}) := E[l(\theta; X, Y)|X, \theta_{old}]$$

$$Q(\theta; \theta_{old}) = \int l(\theta; X, y)p(y|X, \theta_{old})dy \tag{1}$$

$p(.|X, \theta_{old})$ is the conditional density of Y given the observed data, X, and assuming $\theta = \theta_{old}$.
M-Step: The M-step consists of maximizing over θ the expectation computed in Equation (1). That is, we set

$$\theta_{new} := max_{\theta} Q(\theta; \theta_{old}).$$

We then set $\theta_{old} = \theta_{new}$.

The two steps are repeated as necessary until the sequence of θ_{new}'s converges. Indeed under very general circumstances convergence to a local maximum can be guaranteed and we explain why this is the case below. If it is suspected that the log-likelihood function has multiple local maximums then the EM algorithm should be run many times, using a different starting value of θ_{old} on each occasion. The ML estimate of θ is then taken to be the best of the set of local maximums obtained from the various runs of the EM algorithm.

3 Required Terminologies

3.1 Kullback-Leibler Divergence

Let P and Q be two probability distributions such that if P(x) = 0 then Q(x) = 0. The Kullback-Leibler (KL) divergence or relative entropy of Q from P is defined to be

$$KL(P||Q) = \int_x P(x) ln \frac{P(x)}{Q(x)} \tag{2}$$

With the understanding that 0 log 0 = 0. The KL divergence is a fundamental concept in information theory and machine learning. One can imagine P representing some true but unknown distribution that we approximate with Q and that KL(P || Q) measures the "distance" between P and Q. This interpretation is valid because we will see below that KL(P || Q) $\geq$ 0 with equality if and only if P = Q. Note, however that the KL divergence is not a true measure of distance since it is asymmetric in that KL(P || Q) $\neq$ KL(Q || P) and does not satisfy the triangle inequality.

In order to see that KL(P || Q) $\geq$ 0, we first recall that a function f($\cdot$) is convex on R if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) for all \alpha \in [0,1]. \tag{3}$$

We also recall Jensen's inequality:

3.2 Jensen's Inequality

Let f($\cdot$) be a convex function on R and suppose E[X] and E[f(X)] $< \infty$.
Then

$$f(E[X]) \leq E[f(X)]. \tag{4}$$

Noting that $-ln(x)$ is a convex function we have

$$KL(P||Q) = \int_x P(x) ln \frac{P(x)}{Q(x)}$$

$$KL(P||Q) \geq -ln(\int_x P(x) ln \frac{P(x)}{Q(x)} \longrightarrow by Jensen's inequality$$

Moreover it is clear from Equation (2) that $KL(P||Q) = 0$ if $P = Q$. In fact because $ln(x)$ is strictly convex it is easy to see that $KL(P||Q) = 0$ only if $P = Q$.
For an interpretation of the theorem, consider the figure below.

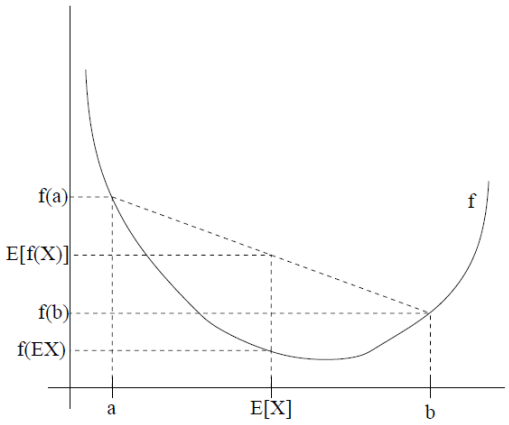

Figure 1: Here, f is a convex function shown by the solid line. Also, X is a random variable that has a 0.5 chance of taking the value a, and a 0.5 chance of taking the value b (indicated on the x-axis). Thus, the expected value of X is given by the midpoint between a and b. We also see the values $f(a), f(b)$ and $f(E[X])$ indicated on the y-axis. Moreover, the value $E[f(X)]$ is now the midpoint on the y-axis between $f(a)$ and $f(b)$. From our example, we see that because f is convex, it must be the case that $E[f(X)] \geq f(EX)$.

4 The EM Algorithm in General

The EM algorithm is often stated more generally using the language of information theory. In this section1 we will describe this more general formulation and relate it back to EM algorithm as described in Section 1. As before the goal is to maximize the likelihood function, L(θ; X), which is given by

$$L(\theta; X) = p(X|\theta) = \int_y p(X, y|\theta)dy \tag{5}$$

The implicit assumption underlying the EM algorithm is that it is difficult to optimize p(X | θ) with respect to θ but that it is much easier to optimize p(X , Y | θ). We first introduce an arbitrary distribution, q(Y), over the latent variables, Y, and note that we can decompose the log-likelihood, $l(\theta; X)$, into two terms according to

$$l(\theta; X) := lnp(X|\theta) = L(q, \theta) + KL(q||p_{Y|X}) \tag{6}$$

where $L(q, \theta)$ and $KL(q||p_{Y|X})$ are the likelihood and Kullback-Leibler (KL) divergence which are given by

$$L(q, \theta) = \int_Y q(Y) ln \left(\frac{p(X, Y|\theta)}{q(Y)} \right) \tag{7}$$

$$KL(q||p_{Y|X}) = -\int_Y q(Y) ln \left(\frac{p(Y|X, \theta)}{q(Y)} \right) \tag{8}$$

Proof for the above:

$$L(q, \theta) + KL(q||p) = \sum_y q(Y) \left\{ ln \frac{p(X, Y|\theta)}{q(Y)} \right\} - ln \left(\frac{p(Y|X, \theta)}{q(Y)} \right)$$

$$L(q,\theta) + KL(q||p) = \sum_y q(Y) \left\{ ln \frac{p(X,Y|\theta)}{p(Y|X,\theta)} \right\}$$

$$L(q,\theta) + KL(q||p) = \sum_y q(Y) lnp\,(X|\theta)$$

$$L(q,\theta) + KL(q||p) = lnp\,(X|\theta)$$

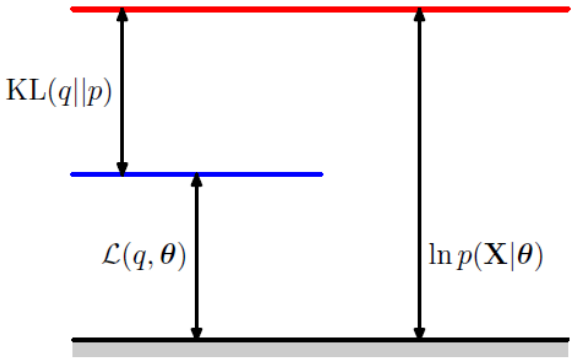

Figure 2: Illustration of the decomposition, which holds for any choice of distribution q(Y). Because the Kullback-Leibler divergence satisfies $KL(q||p) \geq 0$, we see that the quantity $L(q,\theta)$ is a lower bound on the log likelihood function $lnp(X|\theta)$.

It is known that the KL divergence satisfies $KL(q||p_{Y|X}) \geq 0$ and equals 0 if and only if $q(Y) = p_{Y|X}$. It therefore follows that $L(q,\theta) \leq l(\theta; X)$ for all distributions, q(·).

We can now use the decomposition of Equation (6) to define the EM algorithm. We begin with an initial parameter estimate, θ_{old}.

4.1 Algorithm

E-Step: The E-step maximizes the lower bound, L(q, θ_{old}), with respect to q(·) while keeping θ_{old} fixed. In principle this is a variational problem since we are optimizing a functional, but the solution is easily found.

First note that l(θ_{old}; X) does not depend on q(·). It then follows from Equation (5) (with $\theta = \theta_{old}$) that maximizing L(q, θ_{old}) is equivalent to minimizing $KL(q||p_{Y|X})$. Since this latter term is always non-negative we see that L(q, θ_{old}) is optimized when $KL(q||p_{Y|X}) = 0$ which, by our earlier observation, is the case when we take q(Y) = p(Y | X , θ_{old}).

At this point we see that the lower bound, L(q, θ_{old}), will now equal the current value of the log-likelihood, l(θ_{old}; X).

M-Step: In the M-step we keep q(Y) fixed and maximize L(q, θ) over θ to obtain θ_{new}. This will therefore cause the lower bound to increase (if it is not already at a maximum) which in turn means that the log-likelihood must also increase.

Moreover, at this new value θ_{new} it will no longer be the case that $KL(q||p_{Y|X}) = 0$ and so by Equation (6) the increase in the log-likelihood will be greater than the increase in the lower bound.

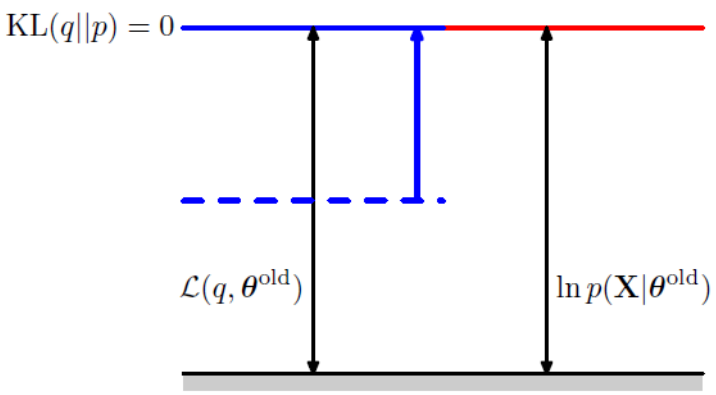

Figure 3: Illustration of the E step of the EM algorithm. The q distribution is set equal to the posterior distribution for the current parameter values θ_{old}, causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.

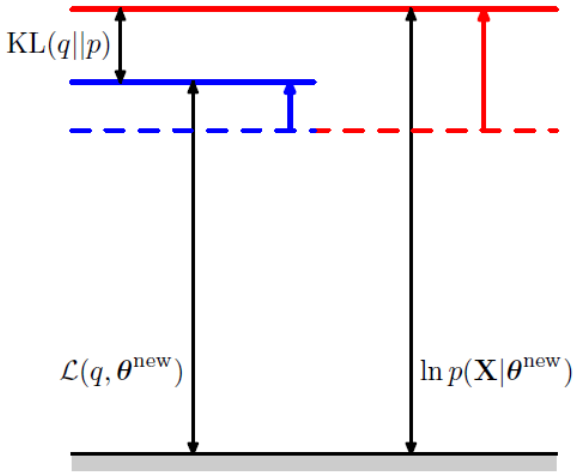

Figure 4: Illustration of the M step of the EM algorithm. The distribution $q(Z)$ is held fixed and the lower bound $L(q, \theta)$ is maximized with respect to the parameter vector θ to give a revised value θ_new. Because the KL divergence is non-negative, this causes the log likelihood $lnp(X|\theta)$ to increase by at least as much as the lower bound does.

4.2 Geometrical Interpretation

The operation of the EM algorithm can also be viewed in the space of parameters, as illustrated schematically. Here the red curve depicts the (in-complete data) log likelihood function whose value we wish to maximize.
We start with some initial parameter value θ_{old}, and in the first E step we evaluate the posterior distribution over latent variables, which gives rise to a lower bound $L(\theta, \theta_{old})$ whose value equals the log likelihood at θ_{old}, as shown by the blue curve.
Note that the bound makes a tangential contact with the log likelihood at θ_{old}, so that both curves have the same gradient. This bound is a convex function having a unique maximum (for mixture components from the exponential family).
In the M step, the bound is maximized giving the value θ_{new}, which gives a larger value of log likelihood than θ_{old}. The subsequent E step then constructs a bound that is tangential at θ_{new} as shown by the green curve.

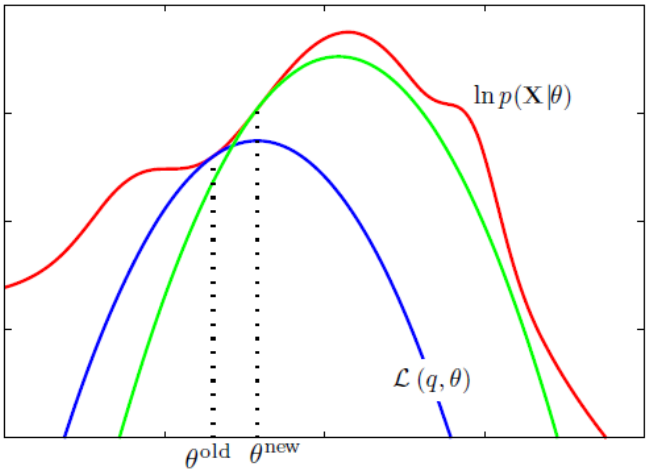

Figure 5: The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values.

Comparing the General EM Algorithm with the Classical EM Algorithm

To do this, first substitute $q(Y) = p(Y|X, \theta_{old})$ into Equation (7) and Equation (8) to obtain

$$L(q, \theta) = Q(\theta; \theta_{old}) + constant \tag{9}$$

where Q(θ; θ_{old}) is the expected complete-date log-likelihood as defined in Equation (1) where the expectation is taken assuming $\theta = \theta_{old}$.
The M-step of the general EM algorithm is therefore identical to the M-step of Section 1 since the constant term in Equation (9) does not depend on θ.
The E-step in the general EM algorithm takes

$$q(Y) = p(Y|X, \theta_{old})$$

which, at first glance, appears to be different to the E-step in Section 1.

But there is no practical difference: the E-step in Section 1 simply uses p(Y | X , θ_{old}) to compute Q(θ; θ_{old}) and, while not explicitly stated, the general E-step must also do this since it is required for the M-step.

5 Application : Computation of Mode in a Bayesian Setting

The EM algorithm can also be used to compute the mode of the posterior distribution, $p(\theta|X)$, in a Bayesian setting where we are given a prior, $p(\theta)$, on the unknown parameter (vector), θ .

To see this, first note that we can write $p(\theta|X) = p(X|\theta)p(\theta)/p(X)$ which upon taking logs yields

$$lnp(\theta|X) = lnp(X|\theta) + lnp(\theta) - lnp(X) \tag{10}$$

If we now use Equation (6) to substitute for $lnp(X|\theta)$ on the right-hand-side of Equation (10) we obtain

$$lnp(\theta|X) = L(q, \theta) + KL(q||p_{Y|X}) + lnp(\theta) - lnp(X). \tag{11}$$

We can now find the posterior mode of $lnp(\theta|X)$ using a version of the EM algorithm.

The E-step is exactly the same as before since the final two terms on the right-hand-side of Equation (11) do not depend on q(·). The M-step, where we keep q(·) fixed and optimize over θ , must be modified however to include the $lnp(\theta)$ term. There are also related methods that can be used to estimate the variance-covariance matrix, $\sum$, of θ. In this case it is quite common to approximate the posterior distribution of θ with a Gaussian distribution centered at the mode and with variance-covariance matrix, $\sum$. This is called a Laplacian approximation and it is a simple but commonly used framework for deterministic inference.

It only works well, however, when the posterior is unimodal with contours that are approximately elliptical. It is also worth pointing out, however, that it is often straightforward to compute the mode of the posterior and determine a suitable $\sum$ for the Gaussian approximation so that the Laplacian approximation need not rely on the EM algorithm. We also note in passing that the decomposition in Equation (6) also forms the basis of another commonly used method of deterministic inference called variational Bayes.

The goal with variational Bayes is to select q(·) from some parametric family of distributions, Q, to approximate $p(Y|X)$. The dependence on θ is omitted since we are now in a Bayesian setting and θ can be subsumed into the latent or hidden variables, Y. In choosing q(·) we seek to maximize the lower bound, $L(q)$, or equivalently by Equation (6), to minimize $KL(q||p_{Y|X})$. A common choice of Q is the set of distributions under which the latent variables are independent.

References

[1] C. Bishop, "Mixture models and em," in *Pattern Recognition and Machine Learning*, 2006, pp. 423–455.