CS229 Lecture notes

原作者：[Andrew Ng](http://cs229.stanford.edu/)（[吴恩达](http://open.163.com/movie/2008/1/M/C/M6SGF6VB4_M6SGHFBMC.html)）

翻译：[CycleUser](https://zhuanlan.zhihu.com/python-kivy)

Part IX

# 期望最大化算法(EM algorithm)

在前面的若干讲义中，我们已经讲过了期望最大化算法（EM algorithm），使用场景是对一个高斯混合模型进行拟合（fitting a mixture of Gaussians）。在本章里面，我们要给出期望最大化算法（EM algorithm）的更广泛应用，并且演示如何应用于一个大系列的具有潜在变量（latent variables）的估计问题（estimation problems）。我们的讨论从 Jensen 不等式（Jensen’s inequality）开始，这是一个非常有用的结论。

# 1 Jensen 不等式（Jensen’s inequality）

设 f 为一个函数，其定义域（domain）为整个实数域（set of real numbers）。这里要回忆一下，如果函数 f 的二阶导数 f′′(x) ≥ 0 (其中的 x ∈ R)，则函数 f 为一个凸函数（convex function）。如果输入的为向量变量，那么这个函数就泛化了，这时候该函数的海森矩阵（hessian） H 就是一个半正定矩阵（positive semi-definite H ≥ 0）。如果对于所有的 x ，都有二阶导数 f′′(x) > 0，那么我们称这个函数 f 是严格凸函数（对应向量值作为变量的情况，对应的条件就是海森矩阵必须为正定，写作 H > 0）。这样就可以用如下方式来表述 Jensen 不等式：

**定理（Theorem）**：设 f 是一个凸函数，且设 X 是一个随机变量（random variable）。然后则有：

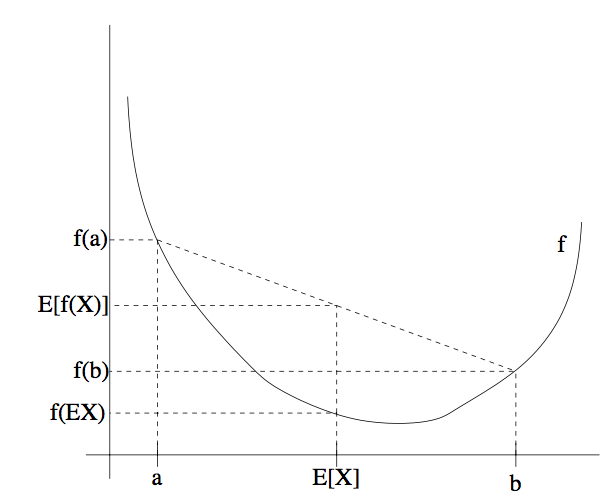
E[f(X)] ≥ f(EX).

（译者注：函数的期望等于期望的函数值）

此外，如果函数 f 是严格凸函数，那么 E[f(X)] = f(EX) 当且仅当 X = E[X] 的概率（probability）为 1的时候成立（例如 X 是一个常数。）

还记得我们之前的约定（convention）吧，写期望（expectations）的时候可以偶尔去掉括号（parentheses），所以在上面的定理中， f(EX) = f(E[X])。

为了容易理解这个定理，可以参考下面的图：



上图中，f 是一个凸函数，在图中用实线表示。另外 X 是一个随机变量，有 0.5 的概率（chance）取值为 a，另外有 0.5 的概率取值为 b（在图中 x 轴上标出了）。这样， X 的期望值就在图中所示的 a 和 b 的中点位置。

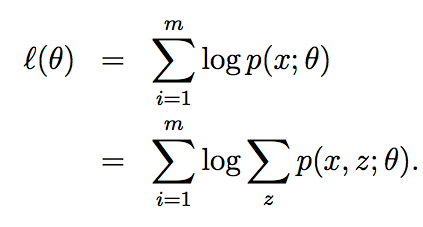
图中在 y 轴上也标出了 f(a), f(b) 和 f(E[X])。接下来函数的期望值 E[f(X)] 在 y 轴上就处于 f(a) 和 f(b) 之间的中点的位置。如图中所示，在这个例子中由于 f 是凸函数，很明显 E[f(X)] ≥ f(EX)。

顺便说一下，很多人都记不住不等式的方向，所以就不妨用画图来记住，这是很好的方法，还可以通过图像很快来找到答案。

**备注。**回想一下，当且仅当 –f 是严格凸函数（[strictly] convex）的时候，f 是严格凹函数（[strictly] concave）（例如，二阶导数 f′′(x) ≤ 0 或者其海森矩阵 H ≤ 0）。Jensen 不等式也适用于凹函数（concave）f，但不等式的方向要反过来，也就是对于凹函数，E[f(X)] ≤ f(EX)。

# 2 期望最大化算法（EM algorithm）

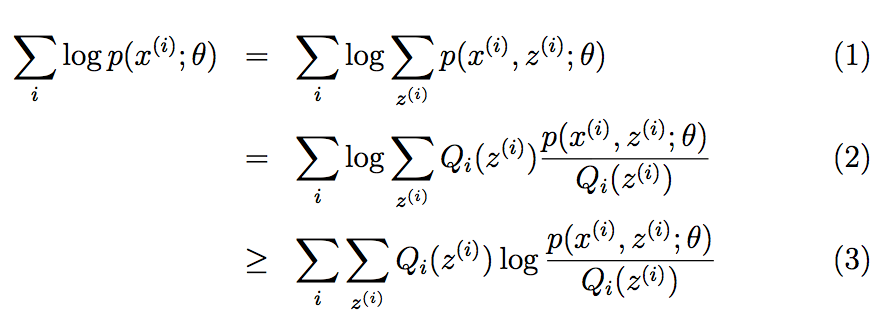
Suppose we have an estimation problem in which we have a training set {x(1), . . ., x(m)} consisting of m independent examples. We wish to fit the parameters of a model p(x, z) to the data, where the likelihood is given by



But, explicitly finding the maximum likelihood estimates of the parameters θ may be hard. Here, the z(i)’s are the latent random variables; and it is often the case that if the z(i)’s were observed, then maximum likelihood estimation would be easy.

In such a setting, the EM algorithm gives an efficient method for maximum likelihood estimation. Maximizing l(θ) explicitly might be difficult, and our strategy will be to instead repeatedly construct a lower-bound on l (E-step), and then optimize that lower-bound (M-step).

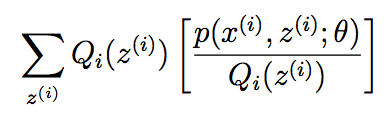
For each i, let Qi be some distribution over the z’s ( Qi(z) = 1, Qi(z) ≥ 0). Consider the following:1



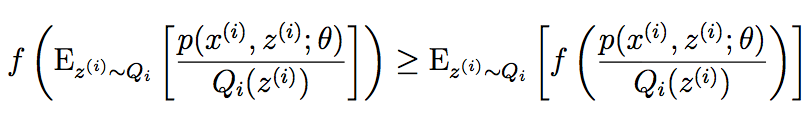


The last step of this derivation used Jensen’s inequality. Specifically, f(x) = log x is a concave function, since f′′(x) = −1/x2 < 0 over its domain x ∈ R+.

Also, the term



in the summation is just an expectation of the quantity p(x(i), z(i); θ)/Qi(z(i)) with respect to z(i) drawn according to the distribution given by Qi. By Jensen’s inequality, we have



where the “z(i) ∼ Qi” subscripts above indicate that the expectations are with respect to z(i) drawn from Qi. This allowed us to go from Equation (2) to Equation (3).

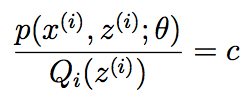
Now, for any set of distributions Qi, the formula (3) gives a lower-bound on l(θ). There’re many possible choices for the Qi’s. Which should we choose? Well, if we have some current guess θ of the parameters, it seems

1If z were continuous, then Qi would be a density, and the summations over z in our discussion are replaced with integrals over z.

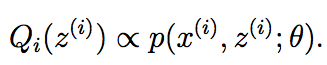
  

natural to try to make the lower-bound tight at that value of θ. I.e., we’ll make the inequality above hold with equality at our particular value of θ. (We’ll see later how this enables us to prove that l(θ) increases monotonically with successive iterations of EM.)

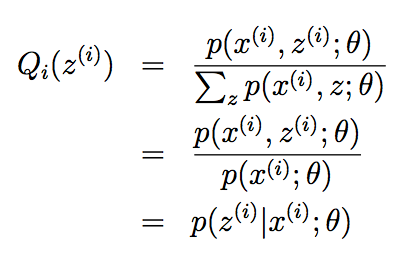
To make the bound tight for a particular value of θ, we need for the step involving Jensen’s inequality in our derivation above to hold with equality. For this to be true, we know it is sufficient that that the expectation be taken over a “constant”-valued random variable. I.e., we require that



for some constant c that does not depend on z(i). This is easily accomplished by choosing



Actually, since we know Qi(z(i)) = 1 (because it is a distribution), this further tells us that

****

Thus, we simply set the Qi’s to be the posterior distribution of the z(i)’s given x(i) and the setting of the parameters θ.

Now, for this choice of the Qi’s, Equation (3) gives a lower-bound on the log likelihood l that we’re trying to maximize. This is the E-step. In the M-step of the algorithm, we then maximize our formula in Equation (3) with respect to the parameters to obtain a new setting of the θ’s. Repeatedly carrying out these two steps gives us the EM algorithm, which is as follows:

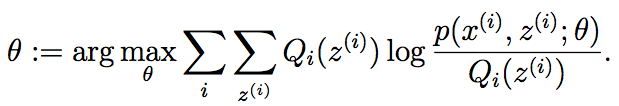
Repeat until convergence {

(E-step) For each i, set





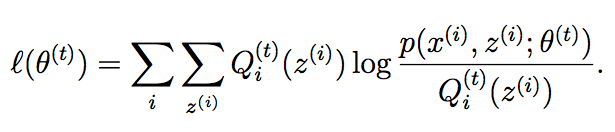
(M-step) Set

}

How we know if this algorithm will converge? Well, suppose θ(t)

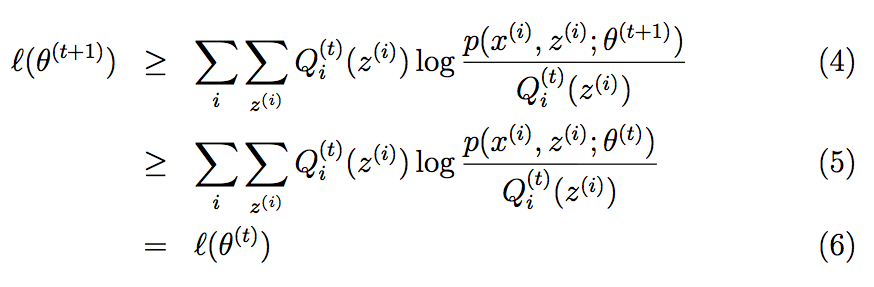
and θ(t+1) are the parameters from two successive iterations of EM. We will now prove that l(θ(t)) ≤ l(θ(t+1)), which shows EM always monotonically improves the log-likelihood. The key to showing this result lies in our choice of the Qi’s. Specifically, on the iteration of EM in which the parameters had started out as θ(t), we would have chosen .

We saw earlier that this choice ensures that Jensen’s inequality, as applied to get Equation (3), holds with equality, and hence

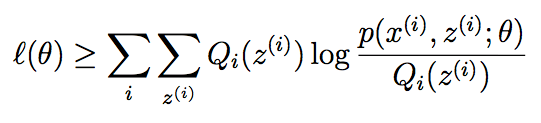


The parameters θ(t+1) are then obtained by maximizing the right-hand side

of the equation above. Thus,

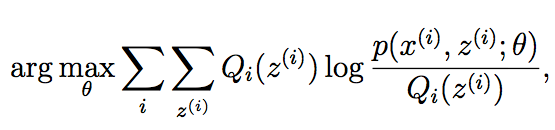


 This first inequality comes from the fact that





holds for any values of Qi and θ, and in particular holds for Qi = Q i (t), θ = θ(t+1). To get Equation (5), we used the fact that θ(t+1) is chosen explicitly to be





and thus, this formula evaluated at θ(t+1) must be equal to or larger than the

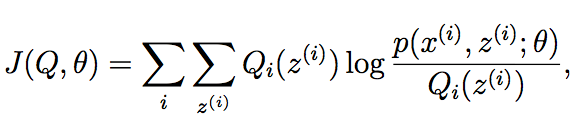
same formula evaluated at θ(t). Finally, the step used to get (6) was shown

earlier, and follows from Q(t) having been chosen to make Jensen’s inequality

i hold with equality at θ(t).

Hence, EM causes the likelihood to converge monotonically. In our description of the EM algorithm, we said we’d run it until convergence. Given the result that we just showed, one reasonable convergence test would be to check if the increase in l(θ) between successive iterations is smaller than some tolerance parameter, and to declare convergence if EM is improving l(θ) too slowly.

Remark. If we define

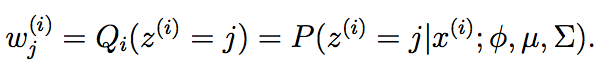


then we know l(θ) ≥ J(Q, θ) from our previous derivation. The EM can also be viewed a coordinate ascent on J, in which the E-step maximizes it with respect to Q (check this yourself), and the M-step maximizes it with respect to θ.

3 Mixture of Gaussians revisited

Armed with our general definition of the EM algorithm, let’s go back to our old example of fitting the parameters φ, μ and Σ in a mixture of Gaussians. For the sake of brevity, we carry out the derivations for the M-step updates only for φ and μj, and leave the updates for Σj as an exercise for the reader.

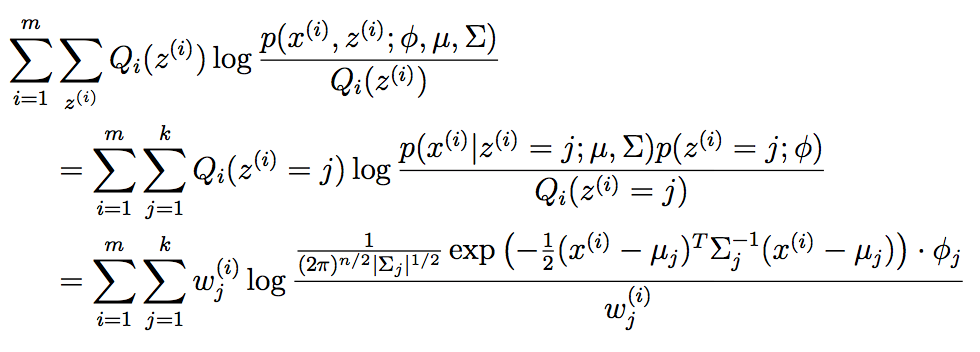
The E-step is easy. Following our algorithm derivation above, we simply calculate



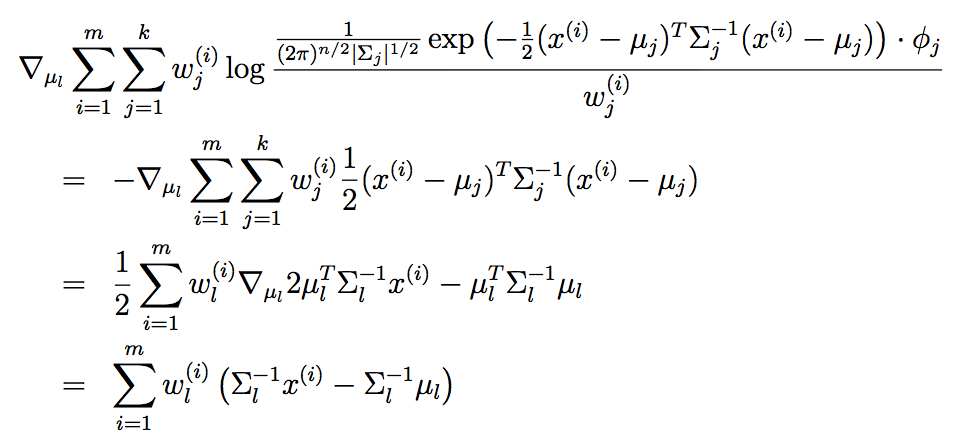


Here, “Qi(z(i) = j)” denotes the probability of z(i) taking the value j under the distribution Qi.

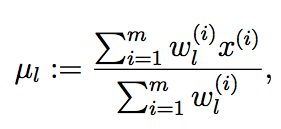
Next, in the M-step, we need to maximize, with respect to our parameters φ, μ, Σ, the quantity



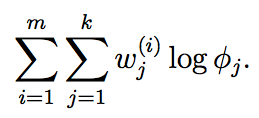
Let’s maximize this with respect to μl. If we take the derivative with respect to μl, we find



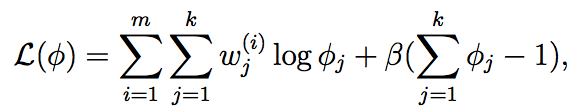
Setting this to zero and solving for μl therefore yields the update rule



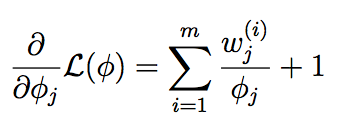
which was what we had in the previous set of notes. Let’s do one more example, and derive the M-step update for the parameters φj. Grouping together only the terms that depend on φj, we find that we need to maximize



However, there is an additional constraint that the φj’s sum to 1, since they represent the probabilities φj = p(z(i) = j;φ). To deal with the constraint that kj=1 φj = 1, we construct the Lagrangian



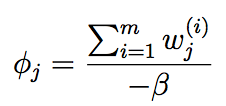
where β is the Lagrange multiplier.2 Taking derivatives, we find



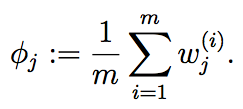


2We don’t need to worry about the constraint that φj ≥ 0, because as we’ll shortly see, the solution we’ll find from this derivation will automatically satisfy that anyway.

Setting this to zero and solving, we get



I.e.,  Using the constraint that Σj φj = 1, we easily find −β= Σi=1m Σj=1k w j (i) = Σi=1m 1=m. (This used the fact that w j (i) =Qi(z(i) = j).) We therefore have our M-step updates for the parameters φj ：



The derivation for the M-step updates to Σj are also entirely straightforward.

and since probabilities sum to 1, have our M-step updates for the parameters φj: