浙江大学数学科学学院

前沿数学专题讨论 PR03

26 Oct 2022

Submitted To:

张南松老师

22-23 秋冬学期

Submitted By :

吴凡

农业工程 + 统计学

Contents

1 概述

1.1 简介

EM，英文全称为 Expectation-Maximization Algorithm，中文名为最大期望算法，是在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐性变量。

最大期望算法经过两个步骤交替进行计算：

1. 第一步是计算期望（E），利用对隐藏变量的现有估计值，计算其最大似然估计值；
2. 第二步是最大化（M），最大化在 E 步上求得的最大似然值来计算参数的值。M 步上找到的参数估计值被用于下一个 E 步计算中，这个过程不断交替进行。

1.2 算法思想

假定你是一五星级酒店的厨师，现在需要把锅里的菜平均分配到两个碟子里。如果只有一个碟子乘菜那就什么都不用说了，但问题是有 2 个碟子，正因为根本无法估计一个碟子里应该乘多少菜，所以无法一次性把菜完全平均分配。

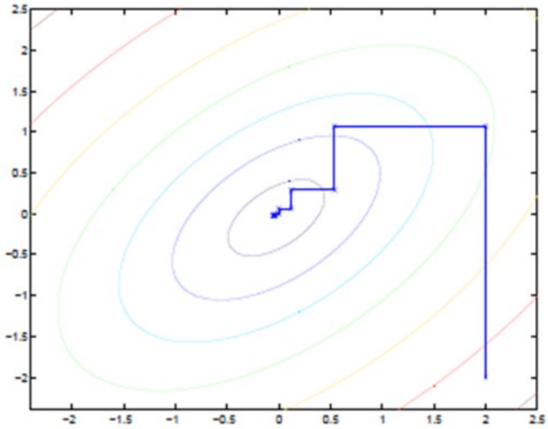
解法：大厨先把锅里的菜一股脑倒进两个碟子里，然后看看哪个碟子里的菜多，就把这个碟子中的菜往另一个碟子中匀匀，之后重复多次匀匀的过程，直到两个碟子中菜的量大致一样。上面的例子中，平均分配这个结果是“观测数据 z”，为实现平均分配而给每个盘子分配多少菜是“待求参数 ”，分配菜的手感就是“概率分布”。

EM 算法的思想：

1. 给 自主规定个初值 (既然我不知道想实现“两个碟子平均分配锅里的菜”的话每个碟子需要有多少菜，那我就先估计个值)；
2. 根据给定观测数据和当前的参数 ，求未观测数据 z 的条件概率分布的期望 (在上一步中，已经根据手感将菜倒进了两个碟子，然后这一步根据“两个碟子里都有菜”和“当前两个碟子都有多少菜”来判断自己倒菜的手感)；

3. 上一步中 z 已经求出来了，于是根据极大似然估计求最优的 ’（手感已经有了，那就根据手感判断下盘子里应该有多少菜，然后把菜匀匀）；
4. 因为第二步和第三步的结果可能不是最优的，所以重复第二步和第三步，直到收敛（重复多次匀匀的过程，直到两个碟子中菜的量大致一样）。

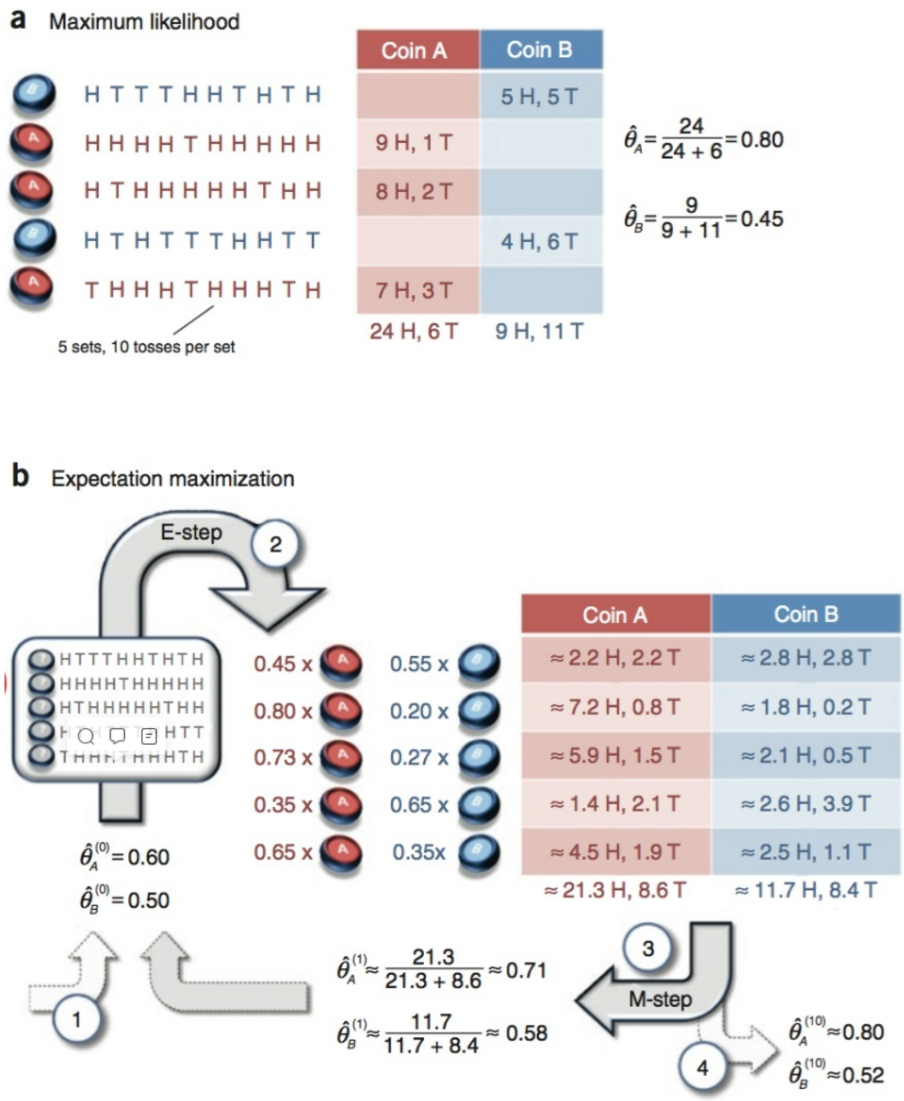
而上面的第二步被称作 E 步（求期望），第三步被称作 M 步（求极大化），从而不断的 E、M。



EM 迭代优化的路径是直线式的，可以看到每一步都会向最优值前进一步，而且前进路线是平行于坐标轴的，因为每一步只优化一个变量。

这犹如在 x-y 坐标系中找一个曲线的极值，然而曲线函数不能直接求导，因此什么梯度下降方法就不适用了。但固定一个变量后，另外一个可以通过求导得到，因此可以使用坐标上升法，一次固定一个变量，对另外的求极值，最后逐步逼近极值。对应到 EM 上，E 步：固定 ，优化 Q；M 步：固定 Q，优化 ；交替将极值推向最大。

1.3 EM 和 MLE



- MLE 是在“模型已定，参数未知”的情况下根据给定观察序列（所有序列服从同一分布）估计模型参数的估计方法。模型参数的准确性，跟观察序列直接相关。
- EM 算法就是含有隐变量的 MLE。

2 数学模型

假设现在有 m 个独立样本 $x=(x_1,x_2,\cdots,x_m)$, 模型参数为 θ

$$\theta_{MLE}=argmax_\theta logp(x|\theta)$$

对于 EM 算法:

$$\theta^{(t+1)}=argmax_\theta E_{z|x,\theta^{(t)}}[logP(x,z|\theta)]$$

EM 算法的迭代过程为，$\theta_0,\theta^{(1)},\theta^{(1)},\cdots,\theta^{(n)}$ 至收敛

下证:

$$P(x|\theta^{(t)})\leq P(x|\theta^{(t+1)})$$

证明:

$$P(x|\theta)=\frac{P(x,z|\theta)}{P(z|x,\theta)}$$

$$\Rightarrow logP(x|\theta)=logP(x,z|\theta)-logP(z|x,\theta) \tag{1}$$

$$\int_Z P(z|x,\theta^{(t)})logP(x|\theta)\,dz=logP(x|\theta)\int_Z P(z|x,\theta^{(t)})\,dz=logP(x|\theta) \tag{2}$$

$$\int_Z P(z|x,\theta^{(t)})[logP(x,z|\theta)-logP(z|x,\theta)]\,dz \tag{3}$$

$$=\int_Z P(z|x,\theta^{(t)})logP(x,z|\theta)\,dz-\int_Z P(z|x,\theta^{(t)})logP(z|x,\theta)\,dz \tag{4}$$

令 $Q(\theta,\theta^{(t)})=\int_Z P(z|x,\theta^{(t)})logP(x,z|\theta)\,dz$

令 $H(\theta,\theta^{(t)})=\int_Z P(z|x,\theta^{(t)})logP(z|x,\theta)\,dz$

则

$$logP(x|\theta)=Q(\theta,\theta^{(t)})-H(\theta,\theta^{(t)})$$

$$logP(x|\theta^{(t)})=Q(\theta^{(t)},\theta^{(t)})-H(\theta^{(t)},\theta^{(t)}) \tag{5}$$

$$logP(x|\theta^{(t+1)})=Q(\theta^{(t+1)},\theta^{(t)})-H(\theta^{(t+1)},\theta^{(t)}) \tag{6}$$

要证：

$$P(x|\theta^{(t)}) \leq P(x|\theta^{(t+1)})$$

等价于证：

$$logP(x|\theta^{(t)}) \leq logP(x|\theta^{(t+1)})$$

等价于证：

$$\begin{aligned} Q(\theta^{(t)}, \theta^{(t)}) &\leq Q(\theta^{(t+1)}, \theta^{(t)}) \\ H(\theta^{(t)}, \theta^{(t)}) &\geq H(\theta^{(t+1)}, \theta^{(t)}) \end{aligned} \tag{7}$$

$$\begin{aligned} \theta^{(t+1)} &= argmax_\theta Q(\theta, \theta^{(t)}) \\ Q(\theta^{(t+1)}, \theta^{(t)}) &\geq Q(\theta, \theta^{(t)}) \end{aligned} \tag{8}$$

由 θ 的任意性，知 $Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta^{(t)}, \theta^{(t)})$

由 Jenson 不等式：$E[logX] \leq logE[X]$

$$H(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)})$$

$$= \int_Z P(z|x, \theta^{(t)}) log \frac{P(z|x, \theta^{(t+1)})}{P(z|x, \theta^{(t)})} \, dz \tag{9}$$

$$= E[log \frac{P(z|x, \theta^{(t+1)})}{P(z|x, \theta^{(t)})}] \tag{10}$$

$$\leq logE(\frac{P(z|x, \theta^{(t+1)})}{P(z|x, \theta^{(t)})}) \tag{11}$$

$$= log \int_Z P(z|x, \theta^{(t)}) \frac{P(z|x, \theta^{(t+1)})}{P(z|x, \theta^{(t)})} \, dz \tag{12}$$

$$= log \int_Z P(z|x, \theta^{(t+1)}) \, dz \tag{13}$$

$$= log1 = 0 \tag{14}$$

3 EM 的应用

3.1 EM 的优缺点

1. 优点：算法简单，稳定上升的步骤能非常可靠地找到“最优的收敛值”；
2. 缺点
   (a) EM 算法的收敛速度，非常依赖初始值的设置，设置不当，计算时的代价是相当大的
   (b) 对于大规模数据和多维高斯分布，其总的迭代过程，计算量大，迭代速度易受影响
   (c) EM 算法中的 M-Step 依然是采用求导函数的方法, 所以它找到的是极值点, 即局部最优解, 而不一定是全局最优解。

3.2 EM 的应用

- K-means 聚类
- GMM(Gaussian Mixture Model, 高斯混合模型)
- HMM(Hidden Markov Model, 隐马尔可夫模型)