EM算法

请允许笔者在本文最后一次用EM算法的全称，Expectation-Maximization Algorithm, 期望-最大化算法。

# EM算法用来解决什么问题

## 题外话：密度函数估计和统计推断

我们涉及到的很多基础统计问题都可以归纳为密度函数估计（参数估计）和统计推断问题。比如给你100个学生的身高，你会马上算均值方差，然后做假设检验，看p值。其背后的道理是你假设这些人的身高服从正态分布，或者其他分布也可以，基于这个假设，你用均值和方差两个矩估计作为分布参数估计值，拟合出参数曲线，基于这条曲线才有p值这个连续函数的统计推断。

矩估计这种不太精致的算法之所以广泛运用，是因为它在简单分布之中有着和最大似然估计往往有相同的结果，而在计算上却远比最大似然估计简单得多，是条殊途同归的捷径。如果面对混合分布，矩估计将彻底失效。试想一下给你下面的这个双峰柱状图描述男女生的身高，但你没有任何性别信息，于是矩估计对你来说不再那么迷人。喜欢玩数据的人会告诉你，假设这是两个正态分布的混合分布，去寻求最大似然来估计这条分布曲线，这就是混合高斯模型（Mixture Gaussian Model）。然后你会基于这条混合曲线做新的统计推断，男女分开，回到最简单统计推断问题。而EM算法就是解决这个混合模型的密度函数估计问题的一种方法。

## EM算法：一把解决混合模型参数估计的快刀

EM算法到底针对的是哪一类问题，很值得玩味。有的用于隐含变量估计，有的用于存在缺失值情况下的统计推断，但这些都可以归为混合模型的参数估计问题，然后用最大似然的思路，用期望-最大化期望的迭代方式解出参数。以下给出一个灯泡寿命的问题来阐述EM算法所要解决的问题以及思路。

假设我们通过随机抽样取1000个节能灯泡来测试这批灯泡的寿命。记灯泡的寿命为，服从参数为的指数分布, 待估计，且。现让1000个灯泡连续工作300个小时观测灯泡的使用寿命。记观测值为，则1000个灯泡的观测值分布如下图。

用最直观的方法理解这个问题，这是个含有缺失值的估计问题，所有灯泡寿命大于300小时的都无法被实验所观测。但这个问题同样也可以理解为混合分布的估计，其密度函数如下：

以现有的数据应用最大似然估计的思路和EM的方法可以把的最大似然解求出来。注意，这里选择了一个非常简单的混合分布，实际上大家直接把似然函数写出来求导就能找到最优的。但大家不妨类比以下函数求极值的问题，如果是简单函数求极值通过求导可以找到解，但多元函数或者复杂函数就要求助类似梯度下降的迭代方法，而EM算法的地位就相当于这些迭代算法，迭代的终点就是最大似然解。

在运用混合高斯模型解决男女身高的统计推断问题时，我们把男女身高分布的均值和方差作为估计的目标，把两个分布的比例当做隐变量。似乎隐变量有不一样的参数地位，但在混合高斯分布中，我们可以一视同仁地看待它们为待估参数。一个身高180cm的人我们只认为他很有可能是男生，但他是男是女并不影响我们作为总体的推断。

这里插一句，笔者觉得混合分布是统计人的思维方式，而数据挖掘和CS出身的人更愿意将其看为是聚类问题，上文提到男女分开作为未知变量的模拟结果，这不就是个聚类问题吗。EM算法的思维可扩展性极强，后文会谈到EM是怎么把K-means囊括进去的。

所以总结EM算法的思路如下：

依据真实数据构建理论分布->明确待估参数及迭代代数式->EM迭代->最大似然解

EM算法的优势在于可以实现混合分布的多参数估计，参数知道的越多，对实际问题了解也就越深，故事也就讲的越好。其算法运用的难点在于理论分布的构建和迭代公式的推导，理论分布的构建是EM算法的艺术性所在。试想一下作为一个数据科学的咨询师，理论分布的构建是你打动受众的关键。而迭代公式的推导则会因数学逻辑的严谨和几何的优美性质感动自己。这里陶醉得有点恶心。

# EM算法如何实现

## 混合分布参数估计的似然思维

只要涉及混合分布一般都有未知变量或者隐参数，比如灯泡例子中未被观测到的灯泡寿命和身高问题中的性别信息，但这些未知变量只是求解混合分布参数的桥梁，未知变量的真实值并不重要。参数才是EM算法的目标。

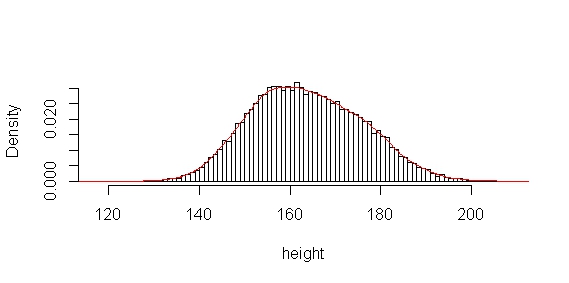
为了求解这个混合分布我们要构造似然函数。假设已知的样本信息为，未知的样本信息为，混合分布的参数为，则似然函数可以表示为, 亦即 因为已知。所以求最大似然解是一个纯粹的函数求极值问题。如果大家有看到关于EM算法的一些文章会说明EM收敛的过程是一个坐标轴收敛(Coordinate descent)，从上述的角度就不难理解，基于这个多元函数，E步骤其实是基于优化，亦即

似然函数延收敛。而M步骤则是基于优化，亦即

似然函数延收敛。至于为什么会收敛到最大似然解，后面将给出证明。

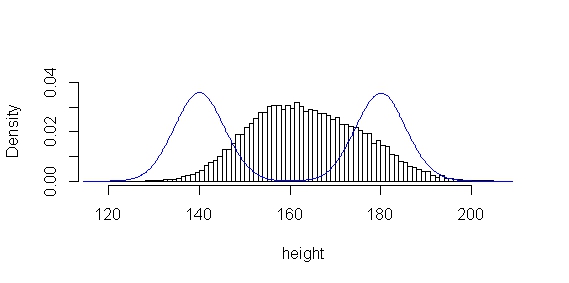
抽象的语言该告一段落了，我们回到前文所述的男女生身高一例来看EM算法怎么求解高斯混合模型。

为了方便叙述和可视化我们取一个比较大的样本（当然，数据是用模拟出来的），现已知22000人的身高，这22000人的性别未知。22000个样本点就是我们所知的全部信息，我们的目标就是那条红色的分布曲线，换言之就是想知道男女身高的具体分布参数，包括均值方差以及男女比例。身高分布如下。

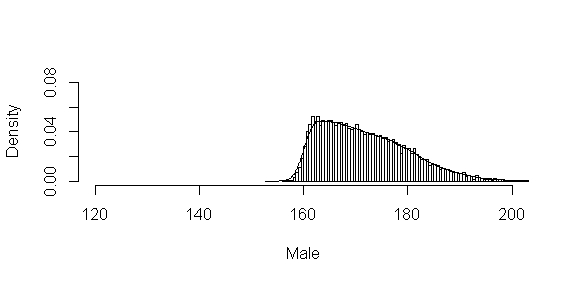
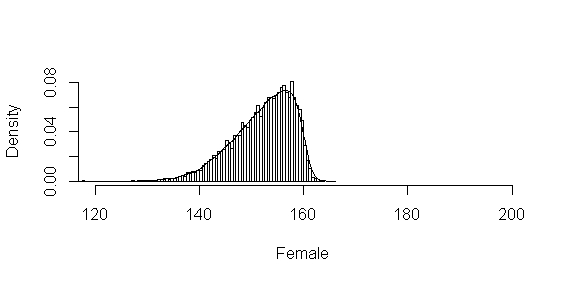


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 180 | 5 | 140 | 5 | 0.5 |

给定如上表的初值进行EM迭代，这条假设的混合分布曲线如下图中的蓝线，经过不断迭代会收敛到如上图的红线。



首先是E步，暂不提恼人的数学公式，纯粹从理念的角度理解E步就是基于现有的假设（现有的混合分布参数值），样本最有可能的展现是什么。基于现有的分布，这堆身高就可以分成来自男性和女性的身高，我们可以用模拟的方式给这些人一个临时的性别标记。在给定的参数下，未知变量的以最有可能的方式分布，这是似然函数曲面沿着未知参数方向的一次前进。这里遵从的还是最大似然的逻辑。



然后我们进行M步，基于这个男女的分布就可以重新估计混合分布的参数。前文说道简单分布的矩估计和最大似然估计之间的关系，这里就可以偷懒，直接用矩估计的方式对两个独立的分布参数进行估计。M步就是基于所有的已知和未知变量进行最大似然估计，其实质是似然函数曲面沿着分布参数的方向前进。一样，还是最大似然估计思想。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 171.72 | 8.24 | 152.13 | 5.75 | 0.59 |

经过不断的迭代，给一个收敛的标准，就可以获得混合分布参数的近似值。值得注意的是，在E步之中我们用了随机模拟这一投机取巧的方法绕过了数学推导，而且随机模拟的构造十分粗糙，但从理念上理解，随机模拟能从样本个性当中反映求期望的本质，这也就是为什么蒲丰投针这样的随机试验看起来是多么迷人。类似的随机模拟思维在复杂分布中反而是所向披靡的利剑。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| Sim | 171 | 10 | 155 | 8 | 0.55 |
| Result | 170.68 | 10.11 | 154.85 | 7.90 | 0.56 |

## E步

开始上数学公式了，不喜欢的可以跳过，反正笔者认为这篇文章最有趣的部分已经写完了。

关于EM的数学表达式有太多种，笔者这里就使用最初的学习EM算法时用的公式，顺便感谢一下Prof. Ajay Jasra老师。这里先给出构造函数公式，然后再谈我的理解。

假设为，在步时的值，则

注意：

1. 实际上是， 即关于变量的函数；
2. 红色部分是迭代到步的似然函数，步的都是变量；
3. 蓝色部分分母为常数，整个部分正比于，亦可看做的分布；
4. 通过求期望，“最大似然”的被确定，似然函数关于的不确定性被消除，从到步，到的直接关系被建立。

整个EM算法的技术难点在于构造函数，函数如果能解得出来，EM算法只剩下求极值的问题。理想的EM算法使用分布函数的推导方法来构建函数，这无疑提高了EM算法的使用门槛。其实现实情况没那么悲观，细心的朋友一看到就已经想好用随机模拟的方法去解上述的函数，这也是为什么笔者要用这个表达式的原因：对于E步而言，似然函数很容易写出来，而基于已有的和去可以构造出函数。关于这方面内容后文将会给出说明。

## M步

有了函数，M步就很简单，只要对函数求极值便可，即

只要迭代以上的E步和M步，直到函数收敛，便可以求出来。

# EM算法收敛的数学证明

## 坐标轴收敛

Prof. Andrew Ng在视频中对于坐标轴收敛的说明特别形象。用如下的一张图来解释如何用建立最低边界的方式逐步前进最终找到似然凹函数的极值（关于函数凹凸的说法让人觉得脑子不够用）。每一条红线都是一次迭代的函数，每一次函数的跳跃都是的更新。

其实换一种空间思维的方式来理解这个问题你就知道上面这张图其实是多维空间的曲面在二维图像中的简单投影。高维空间的想象是一件容易走火入魔的事情，因为这个似然函数曲线（或者叫超平面），可以一次延几个维度前进收敛于极值。

## Jensen's不等式

EM算法的妙处在于其目标是混合分布的参数估计，而其过程是参数在每一步迭代中的更新，其过程和结果是统一的。EM算法迭代的两个步骤在数理推导上并不很对称，因为直接建立了到的直接关系，M步的递进收敛是不言而喻的。只要每一个E步都是在向最大似然解收敛整个理论就完整了，这里不难，也没什么营养，直接用Jensen's不等式就能证明。

# EM算法的一些解决办法