EM 算法与 GMM 参数的极大似然估计

PB19151785 魏嵘

摘要

本文讨论了 EM 算法在解决高斯混合模型中的应用。并针对 EM 算法对初值敏感这一特点，首先采用 K-means 得到参数估计的初值，进而再用 EM 算法进行更加精细优化估计。

关键词： EM 算法；高斯混合模型；K-均值聚类

1 引言

数理统计是统计学入门的基础课之一，它的核心任务即是从样本出发推断总体。在已知样本的分布形式时，统计推断的目标便是确定未知的参数值。极大似然估计法便是通过求解似然函数的极大值来估计参数取定某些值的最大“似然性”。然而，现实中所得到的数据往往是缺失部分信息或者不完整的，即存在“潜在变量”或“隐变量”，如高斯混合模型

$$f(x|\theta) = \sum_{k=1}^{N} \alpha_k p(x|\theta_k)$$

的参数则包括 $\theta = (\mu_k, \sigma_k, \alpha_k)$ 其中由于仅仅通过样本的数据信息，样本数据点来自各个子模型发生概率 α_k 未知, 因此无法通过对似然函数简单的求导求解极大值的方法获得似然性最大的参数估计。此时采用 EM 算法则能很好地解决含有隐变量的概率模型参数的极大似然估计问题，EM 算法又称为期望最大算法，即采用迭代的方法不断收敛逼近局部最优解。

然而，EM 算法对于初值的依赖性较大，聚类结果的不同随初始值不同而波动较大，有时只会收敛到局部的最小值而无法达到全局最优解。在只给出了样本点数据的情况下，如果采用随机选取样本点作为初值的方式最后估计得到的参数值也常常不尽如人意。为解决初值的选取问题, 笔者想到可以先通过“k-均值聚类”的方式将样本点集合划分成 K 个聚簇，使得每个

将每个聚簇的中心作为 EM 算法求解 GMM 模型的初值。即采用 K-means 方法给出 0-1 抉择，将每个待分类的目标对象划给一个聚类，而在此基础上 EM 算法可以看作是一种更加精细的延拓，它假定了每个对象分别属于各个类的概率，考虑进隐变量的分布，进而实现对于给定每个数据一个“概率标签”再生成此标签下的最优模型。

本文将从一维高斯混合模型入手，利用 K-means 聚类的方式给出 GMM 初值的预估，再利用 EM 算法求解对应的参数估计问题。

2 实验环境

本次实验主要目的为采用 python 语言实现 EM 算法求解 GMM 的平均值、方差估计问题，实现的平台为 anaconda 的 jupyter notebook，主要采用的库有：pandas、numpy、matplotlib、seaborn、math、random、scipy

3 EM 算法思路简介

假定完整数据集的密度函数可以表示为 $f(x|\theta)$，观测到的不完全数据集 $g(y|\theta)$，它可以表示为从样本空间 $\mathcal{X}$ 到样本空间 $\mathcal{Y}$ 的一组映射 $x->y(x)$，记 $\mathcal{F}(X)=\mathcal{F}(Y,Z)$。

为了得到 $g(y|\theta)$ 中参数 θ 的极大似然估计，我们记 $p(\theta|Y)$ 为观测数据的似然函数，$p(\theta|Y,Z)$ 为添加隐变量 Z 后的似然函数，$p(Z|\theta,Y)$ 为给定参数 θ 和观测数据 Y 下隐变量 Z 的条件分布函数。

EM 的算法本质即为迭代算法，我们记第 (i+1) 次迭代的两步分别为：

E 步：对隐变量 Z 的条件分布求期望消除隐变量 Z，即

$$Q(\theta|\theta^i,Y)=E_z[p(\theta|Y,Z)|\theta^i,Y]=\int p(\theta|Y,Z)p(Z|\theta^i,Y)dZ$$

M 步：最大化似然函数 $Q(\theta|\theta^i,Y)$，即寻找对应的 θ^{i+1} 满足

$$Q(\theta^{i+1}|\theta^i,Y)=\arg\max Q(\theta|\theta^i,Y)$$

并利用新的到的 θ^{i+1} 进行下一步迭代；

此迭代过程直至 $||Q(\theta^{i+1}|\theta^i,Y)-Q(\theta^i|\theta^i,Y)||$ 充分小为止。

对上述过程，我们有如下注记：

（1）迭代过程中初值 θ 的选取对于最后 EM 算法得到的收敛值有较大的影响, 通常可以通过随机初始化模型运行多次 EM 算法，比较多次收敛结果，从而从收敛结果中选取最优估计值。而本文中笔者通过 K-means 方法优化初值选择从而得到更好的估计，通过查阅相关资料得知还可以通过 binning 法、随机中心法、层次聚类法等方式得到更好的初值估计。

（2）通过相关数学推导可知，EM 算法在每次迭代均能提高所估计似然函数的值 $g(\theta|y)$，若 $g(\theta|y)$ 有上界，则 EM 算法可以确保收敛到一个点，但是无法保证是全局的最大值点，只是局部极大似然的估计。

4 EM 算法的 python 实现

4.1 高斯混合模型的生成

通过测试输入的 n 的值产生具体的 n 个高斯分布，其中混合比例同样通过随机数生成，均值取值范围为-10-10，标准差取值范围为 1-5，并利用 python 中的 numpy 库生成 n*1000 个样本点，并最终输出对应的图像。后续将均以 3 个高斯分布的混合为例，并通过 random.seed 确保每次生成的混合分布相同以便于不断检验修改模型。

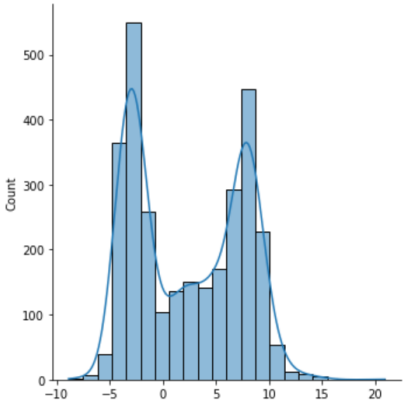

图 1: 样本数据点的分布

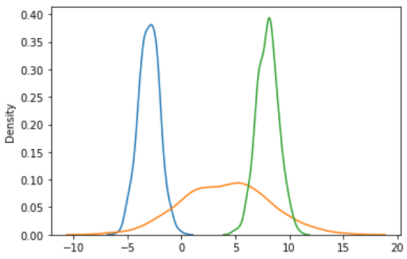

图 2: 高斯混合模型的离散子模型

4.2 GMM 模型初值的确定

通过《python 编程导论》中对 k-means 方法的介绍，第一步先随机选取了 k 个样本点作为初始的簇质心，再将每个样本点分配给距离最近的质心从而建立起 k 个簇；第二步则是对每个簇中所有的样本点取均值，作为新的 k 个质心；并不断迭代上述两个步骤直到两次质心相差很小时，返回当前的簇集合。并将质心作为 EM 算法迭代的均值的初值，以此作为参数得到拟合图像：

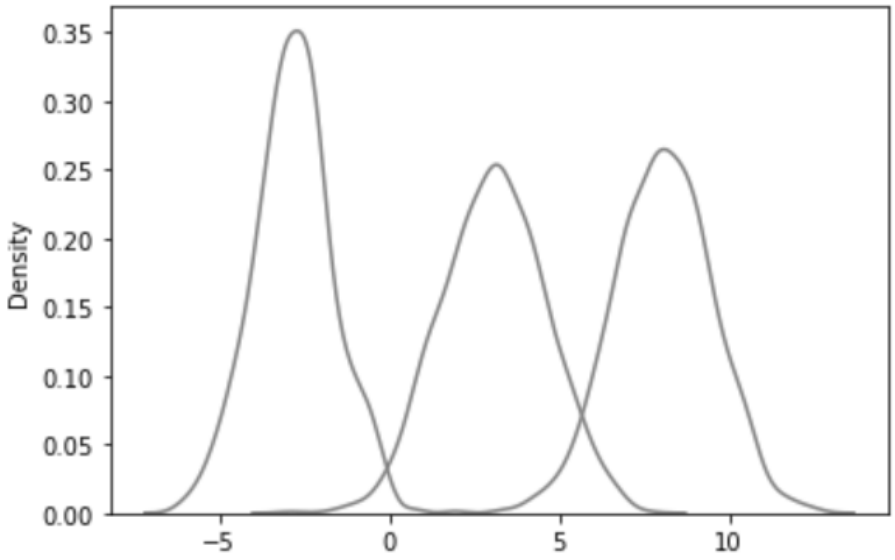

图 3: K-means 得到的聚类结果

但从图中明显可以发现，由于分簇时已将欧氏距离较近的数据分为一簇，因此直接各簇的标准差作为每个子模型的标准差是不合适的，结合高斯分布顶点的表达式 $\frac{1}{\sqrt{2\pi}\sigma}$，即标准差 σ 越大，数据的离散程度越大，反映在绘制出的图像上则对称中心的高度越低。由此得到修改的初值数据如下：

$$initial\_dic = 0 : (7.88, 1.11), 1 : (-2.82, 1.22), 2 : (2.73, 4)$$

并通过样本点的分布得到三个子模型的大致初始混合比例

$$[0.38, 0.48, 0.14]$$

4.3 E 步迭代函数

E 步的作用是在给定估计的初始参数的条件下，估计出样本点来自于第 k 个高斯分布的概率，也就是相应的隐变量。假定初值比例为 $ratio[k] = \alpha_k, k = 1, 2 \cdot n$, 并带入假定的初始均值、方差可以算出对应样本点的概率密度为 $g(y_i|\theta_k)$，从而得到在给定参数 $\theta = (\alpha, \mu, \sigma)$ 下每个样本点 y_i 来自于各个高斯混合模型模型的概率，又称为模型 k 对于观测数据 y_i 的响应度。其计算公式为

$$r[k][i] = \frac{\alpha_k g(y_i|\theta_k)}{\sum_{k=1}^{n} \alpha_k g(y_i|\theta_k)}$$

并返回二维数组 $r_{3\times n}$

4.4 M 步迭代函数

利用 E 步返回的数组 r 计算得到新的估计值, 即

$$\mu_k = \frac{\Sigma_{i=1}^{n} r_{ik} y_i}{\Sigma_{i=1}^{n} r_{ik}}$$

$$\sigma_k^2 = \frac{\Sigma_{i=1}^{n} r_{ik} (y_i - \nu_k)^2}{\Sigma_{i=1}^{n} r_{ik}}$$

$$\alpha_k = \frac{\Sigma_{i=1}^{n} r_{ik}}{n}$$

并返回更新过后的比例与参数估计

4.5 运行结果

得到如下输出值:

```
真实的混合参数为： {0: (-3, 1), 1: (4, 4), 2: (8, 1)}
采用EM算法得到的混合参数为： {0: (7.82, 1.28), 1: (-2.83, 1.21), 2: (3.43, 3.83)}
真实的混合比例为： [0.37, 0.38, 0.24]
采用EM算法得到的混合比例为： [0.34, 0.41, 0.24]
```

图 4: EM 算法结果与真实值的比较

并利用 EM 算法估计的参数值得到相应的拟合曲线为:

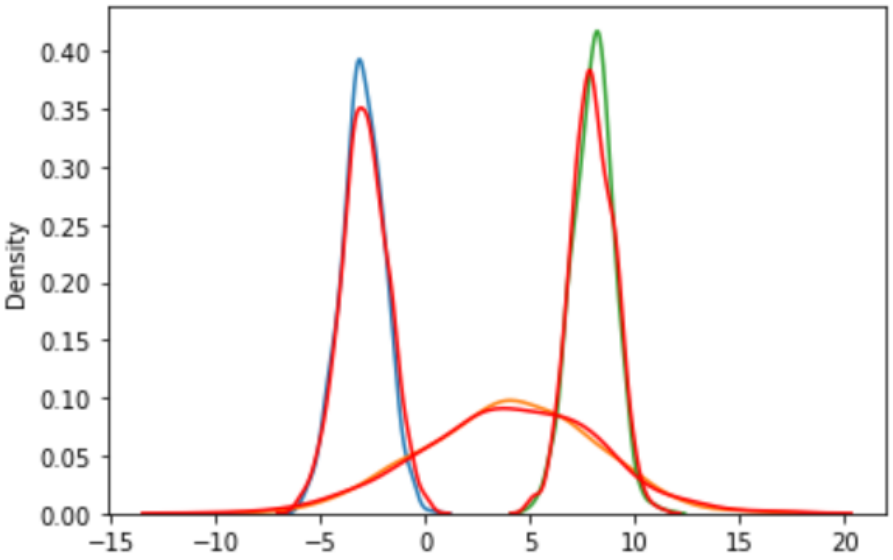

图 5: EM 算法的拟合结果

从最终的输出结果来看，采用 EM 算法得到的参数估计明显优于初始粗糙的参数估计，达到预期结果。

5 总结

此结课作业的灵感来源于 2021 年春季学期数理统计课上老师在讲解“极大似然估计法”时给出的一道例题，不同于以往的简单求导解决极大似然估计的相关计算题，乍一看上去此题并不能计算得到确定的表达式，而需采用迭代的方式不断优化参数的估计。由于仍然不太明白上课时的理论讲解与推导，笔者决定采用编程实践的方式加深对“极大似然估计法”及 EM 算法的理解。在实践过程中所遇到的大部分困难都通过 2021 春季学期面向科学问题求解的编程实践课上得到解决——在老师的建议下采用 anaconda 的 jupyter 笔记本作为实验的运行环境完成本次作业，并在助教的帮助下完成实验运行平台的搭建；采用上课提到的“k-means”聚类方法解决 EM 算法初值随机选取的收敛性问题等等，收获颇丰。

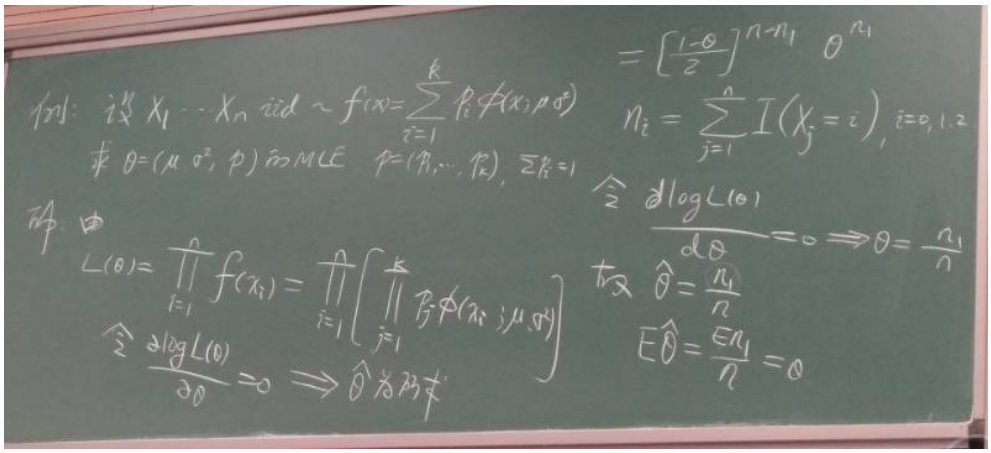

图 6: 数理统计课堂练习

从统计的角度来看，EM 算法巧妙地给出了有缺失数据的参数估计问题的优化解法。而在当下火热的人工智能领域，EM 算法曾入选“数据挖掘十大算法”之一，是众多机器学习领域算法的基础，影响范围极广。

通过 2021 年秋季学期的课程简介，笔者发现下学期的专业课“凸优化”中会进一步拓展 EM 算法的理论部分，将其解释为 F 函数的极大-极大算法，并基于此解释对 EM 算法做出若干变形推广，如广义极大期望算法（GEM）。同时，在下学期即将学习的“统计实用软件”中更是可以直接调用 R 语言的相关库和代码包实现 EM 算法，种种有趣的数据分析与处理方法有待进一步深入的学习与思考。

6 致谢

感谢 2021 春季学期面向科学问题求解的编程实践的孙广中老师在上课过程中不断强调的优化思想，课堂内容更是融入了计算思维、数据结构、数据库、优化方式、机器学习、动态规划等等众多领域的核心算法简介，受益匪浅。

同时感谢 2021 春季学期数理统计的张伟平老师与王学钦老师，他们对课本内容的补充以及对高阶内容的点拨让我感受到了统计学的魅力，激发了我的学习热情与兴趣，确实如张老师第一节课所说：统计是一门艺术。

参考文献

[1] 岳佳, 王士同. 高斯混合模型聚类中 EM 算法及初始化的研究 [J]. 微计算机信息,2006,22(33):244-246,302.DOI:10.3969/j.issn.1008-0570.2006.33.086.

[2] P. DEMPSTER, N. M. LAIRD, D. B. RUBIN. Maximum Likelihood from Incomplete Data Via the EM Algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological),1977,39(1):1-22. DOI:10.1111/j.2517-6161.1977.tb01600.x.

[3] 李航. 统计学习方法 [M]. 北京: 清华大学出版社, 2012.

[4] 茆诗松，王静龙，濮晓龙. 高等数理统计 [M]. 北京: 高等教育出版社, 1998.