EM算法估计高斯混合模型参数

刘峻池

2252298@qq.com

摘要

本实验建立了EM算法模型，通过迭代算法中的E-Step与M-Step来估计高斯混合模型的参数，从而分离并逼近各实际高斯模型，对于估计结果进行了图表化表示，并对EM算法的性能进行了测试与分析。

一、介绍

高斯混合模型（Gaussian Mixture Model，GMM）是一种概率模型，用于表示由多个高斯分布组成的复杂概率分布。GMM可以用于聚类、密度估计、异常检测等领域。

GMM的基本假设是，数据是由多个高斯分布组成的混合体，每个高斯分布代表一个聚类中心。对于每个数据点，其属于每个聚类的概率由每个高斯分布在该点处的概率加权得到。GMM的目标是通过最大化似然函数，即最大化观测数据在模型中的概率来确定模型参数，包括每个高斯分布的均值、方差和混合系数。

在本实验中使用EM算法（Expectation-Maximization Algorithm）对于高斯混合模型的参数进行估计。EM算法是一种常用于求解含有隐变量（latent variable）的概率模型参数的迭代算法。其主要思想是通过迭代的方式，交替进行E步和M步操作，最终求得概率模型的参数。

在数据聚类、概率密度估计等领域中，往往存在一些隐变量，也就是不能直接观测到的变量,而隐变量的存在使得直接求解模型参数变得困难。EM算法就是为了解决这种含有隐变量的概率模型参数估计问题而设计的一种迭代算法。其基本思想是，在E步中，利用当前参数的估计值，计算隐变量的后验概率分布，即求解给定样本数据条件下，隐变量的概率分布。在M步中，利用E步计算出的隐变量后验概率，重新估计模型的参数。通过交替执行E步和M步，迭代更新模型的参数，直到收敛为止。

二、方法原理

1、高斯混合模型

高斯分布是拟合随机数据最常用的模型。单变量 x 的高斯分布概率密函数如下:

$$N(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

其中 μ 分布的数学期望，σ 标准差，σ^2 是方差。

但是现实采集的数据较为复杂，通常无法只用一个高斯分布拟合，而是可以看作多个随机过程的混合。可定义高斯混合模型是 K 个高斯分布的组合，用以拟合复杂数据。

假设有一个数据集，包含了 N 个相互独立的数据：$x=x_1,x_2\cdots x_i\cdots x_N$，这些数据看起

来有 K 个峰，这样的数据集可用以下定义的高斯混合模型拟合：

$$p(x|\Theta)=\sum_k \alpha_k N(x;\mu_k,\sigma_k)=\sum_k \alpha_k \frac{1}{\sqrt{2\pi\sigma_k^{\,2}}}\exp\left[-\frac{(x-\mu_k)^2}{2\sigma_k^{\,2}}\right]$$

其中 Θ 代表全体高斯模型参数，α_k 是第 k 个高斯模型的先验概率，各个高斯模型的先验概率加起来等于1。

$$\sum_k \alpha_k = 1$$

2、EM算法

EM 算法是一种迭代的算法，算法解决的问题可如下表述：

1．采集到一组包含 N 个独立数据的数据集 x。

2、预先知道、或者根据数据特点估计可以用 K 个高斯分布混合进行数据拟合。

3、目标任务是估计出高斯混合模型的参数：K 组 $(\alpha_k,\mu_k,\sigma_k)$

对于相互独立的一组数据，最大似然估计(MLE)是最直接的估计方法。对高斯混合模型使用最大似然估计，求得的似然函数是比较的复杂的，单变量和多变量GMM似然函数结果如下，可以看到多变量GMM似然函数涉及多个矩阵的求逆和乘积等运算。所以要估计出 K 组高斯模型的参数准确值是较为困难的。

$$p(x|\Theta)=\prod_i p(x_i|\Theta)=\prod_i\left[\sum_k \alpha_k N(x_i|\mu_k,\sigma_k)\right]$$

GMM 似然函数首先可以通过求对数进行简化，把乘积变成和。和的形式更方便求导和求极值。

$$L(x|\Theta)=\sum_i \ln\left[p(x_i|\mu_k,\sigma_k)\right]=\sum_i \ln\left[\sum_k \alpha_k N(x_i|\mu_k,\sigma_k)\right]$$

然而在上述似然函数种又两重求和，其中一种还是在对数函数种，直接求极值并不可行。而EM算法提出了利用隐参数 z，通过迭代来对最优的告诉混合模型进行逼近。对于GMM模型使用EM算法时，完整的目标函数为：

$$Q(\Theta,\Theta^t)=\sum_i\sum_k \omega_{i,k}^t \ln\frac{\alpha_k}{\omega_{i,k}^t\sqrt{2\pi\sigma_k^{\,2}}}\exp\left[-\frac{(x_i-\mu_k)^2}{2\sigma_k^{\,2}}\right]$$

EM算法在计算给过程中总共分为两个步骤，E-Step和M-Step。

E-step目标就是计算隐参数的值，也就是对每一个数据点，分别计算其属于每一种高斯模型的概率。所以隐参量 ω 是一个 $N\times K$ 矩阵。每一次迭代后 $\omega_{i,k}$ 都可以用最新的高斯参数 $(\alpha_k,\mu_k,\sigma_k)$ 进行更新，得到目标函数的最新表达。$\omega_{i,k}$ 的计算公式如下：

$$\omega_{i,k}^t=\frac{\alpha_k^t N(x_i|\mu_k^t,\sigma_k^t)}{\sum_k \alpha_k^t N(x_i|\mu_k^t,\sigma_k^t)}$$

M-step的任务就是最大化目标函数，从而求出高斯参数的估计。通过对目标函数的各个

参数进行偏导计算并求得相应的极值，从而得出EM算法中参数的迭代公式，最终 $(\alpha_k, \mu_k, \sigma_k)$ 参数的迭代公式分别如下。

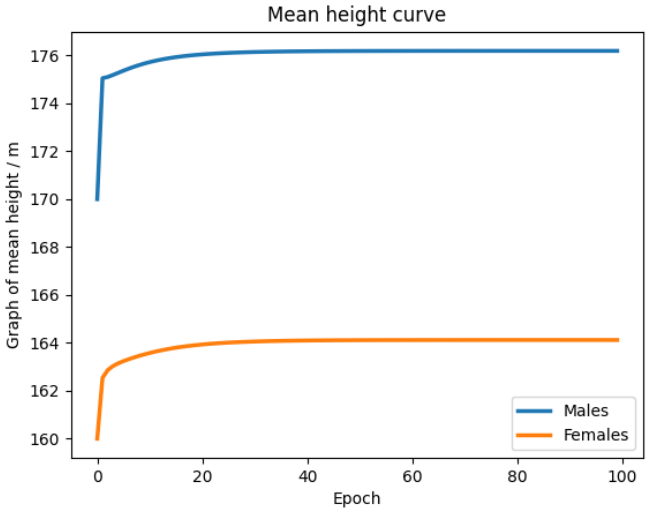
$$\alpha_k^{t+1} = \frac{\sum_i \omega_{i,k}^t}{N} \qquad \mu_k^{t+1} = \frac{\sum_i \omega_{i,k}^t x_i}{\sum_i \omega_{i,k}^t} \qquad (\sigma_k^2)^{t+1} = \frac{\sum_i \omega_{i,k}^t (x_i - \mu_k^{t+1})^2}{\sum_i \omega_{i,k}^t}$$

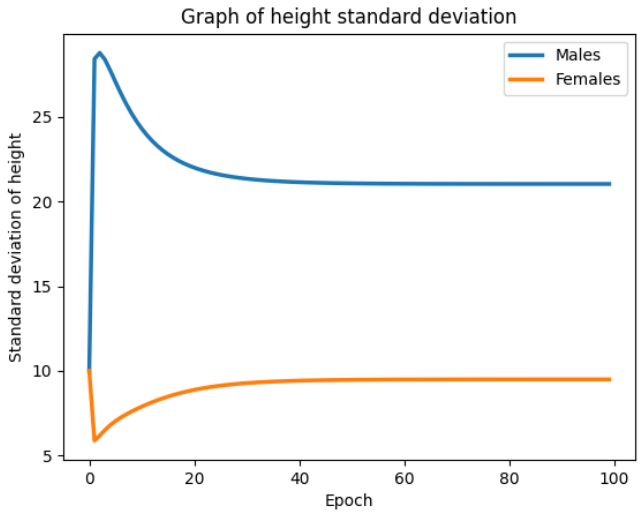
重复上述E-step和M-step即可逼近最优的高斯混合模型参数，从而获得最优的高斯混合模型。

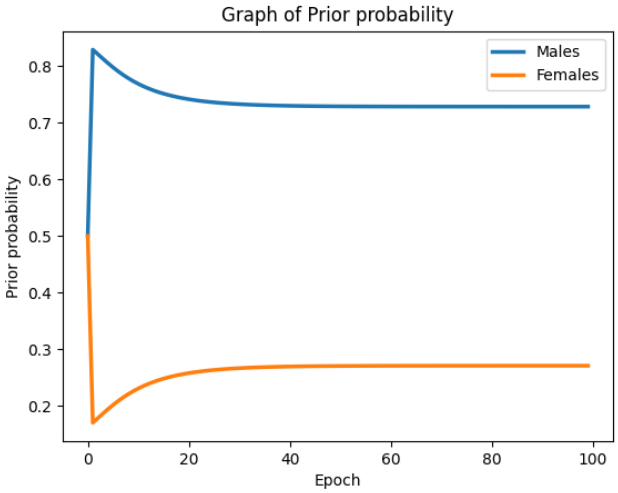
三、实验结果与分析

本实验使用Python进行编程实现，数据集由500个符合正态分布 $N(164, 9)$ 的女生身高数据和1500个符合正态分布 $N(176, 25)$ 的男生身高数据组成，数据使用Numpy随机生成。

在本实验中用 $(\mu_{男}, \mu_{女}, \sigma_{男}, \sigma_{女}, \alpha_{男}, \alpha_{女})$ 来代表待估计参数集合，当使用 $(170, 160, 10, 10, 0.5, 0.5)$ 作为初始值时，所得到的参数估计值变化曲线如下。

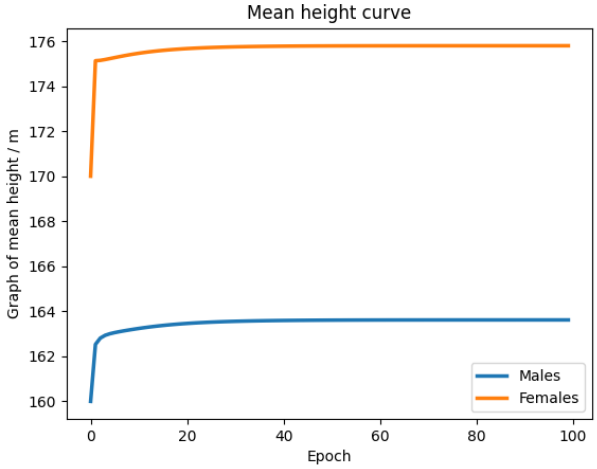






可见在初值 $(170,160,10,10,0.5,0.5)$ 下，EM算法可以较好的估计高斯混合模型的参数。

但是当初始变为 $(160,170,10,10,0.5,0.5)$ 时，即将男女生身高初值交换，换为距离对方的真实值较近的初值时，对于男女身高参数的估计不再准确了。



可见在初值 $(160,170,10,10,0.5,0.5)$ 下，对于男女生身高的估计呈现相反的结果，可见EM算法对于初值有较高的要求。

本实验选取多组不同的初值集合来探究初值对于EM算法结果的影响，相应的初值集及结果如下。

	初始值	$\mu_{男}$	$\mu_{女}$	$\sigma_{男}$	$\sigma_{女}$	$\alpha_{男}$	$\alpha_{女}$
真实值	–	176	164	25	9	-	-
标准初始值	**(170, 160, 10, 10, 0.5, 0.5)**	176.01	163.85	24.72	8.36	0.76	0.24
均值变化	(168, 168, 10, 10, 0.5, 0.5)	173.13	173.13	47.56	47.56	0.50	0.50
	(160, 180, 10, 10, 0.5, 0.5)	163.85	176.01	8.37	24.72	0.24	0.76
方差变化	(170, 160, 5, 20, 0.5, 0.5)	176.48	163.90	24.73	8.36	0.76	0.24
	(170, 160, 20, 5, 0.5, 0.5)	176.01	163.85	24.73	8.36	0.76	0.24
先验概率变化	(170, 160, 10, 10, 0.2, 0.8)	176.01	163.85	24.72	8.36	0.76	0.24
	(170, 160, 10, 10, 0.8, 0.2)	176.01	163.85	24.73	8.36	0.76	0.24

由上述实验结果可以得出：

1、EM算法对于均值的初值较为敏感，在估计高斯混合模型的参数时，若各高斯模型的均值初值相同时，EM算法失效；若各高斯模型的均值初值与真实值的大小关系不对应时，EM算法估计的模型参数虽然有效，但是在各高斯模型之间的对应关系会混乱，从而导致算法失效。同时EM算法对于方差初值与先验概率初值不敏感，初值的选定对于估计的结果影响甚微。

2、在初值选定合理的前提下，EM算法在估计高斯混合模型的参数时，可以较好且较快的收敛逼近真实的参数值，从而可以得到与实际高斯混合模型相近的模型。

四、结论

本实验建立了EM算法模型，通过迭代算法中的E-Step与M-Step来估计高斯混合模型的参数，从而分离并逼近各实际高斯模型，对于估计结果进行了图表化表示，并对EM算法的性能进行了测试与分析。

实验结果得出，EM算法对于均值的初值较为敏感，当初值选定不佳时算法易失效，在初值选定合理的前提下，EM算法在估计高斯混合模型的参数时，可以较好且较快的收敛逼近真实的参数值，从而可以得到与实际高斯混合模型相近的模型。

参考资料

[1] 高斯混合模型（GMM） https://zhuanlan.zhihu.com/p/30483076

[2] EM算法原理及推导 https://zhuanlan.zhihu.com/p/326055752