基于混合高斯模型的 EM 算法估计男女分布

——ZY2203106 李文雯

一. 问题描述

由给定的 python 代码按照高斯分布随机生成男女样本分别为 1500 和 500 个，并将身高数据拼接得到混合身高的数据集，通过 csv 库导入数据集，在初始化参数之后，通过 EM 算法迭代去估计混合数据集中的男女人数占比和两个高斯分布的参数，并与实际数据集进行对比验证。

二. 理论基础

高斯混合模型：

高斯混合模型可以看作是由 K 个单高斯模型组合而成的模型，这个子模型是混合模型的隐变量(Hidden variable)。一般来说，一个混合模型可以使用任何概率分布，这里使用高斯混合模型是因为高斯分布具备很好的数学性质以及良好的计算性能。

高斯混合模型的概率分布为：

$$P(x|\theta) = \sum_{k=1}^{K} \alpha_k \phi(x|\theta_k)$$

对于这个模型而言，参数 $\theta = (\tilde{\mu_k}, \tilde{\sigma_k}, \tilde{\alpha_k})$，也就是每个子模型的期望、方差（或协方差）、在混合模型中发生的概率。

EM 算法求解高斯混合模型：

EM 是一种用于含有隐变量的模型参数的最大似然估计，是一种迭代算法，通过 E 步和 M 步交替迭代直至收敛估计出模型参数。

（1）首先初始化参数；

（2）E-step:依据当前参数，计算每个数据 j 来自子模型 K 的可能性：

$$\gamma_{jk}=\frac{\alpha_k\phi\left(x_j|\ \theta_k\right)}{\sum_{k=1}^{K}\alpha_k\phi\left(x_j|\ \theta_k\right)},j=1,2,\ldots,N;k=1,2,\ldots,K$$

（3）M-step:计算新一轮迭代的模型参数：

$$\mu_k=\frac{\sum_{j}^{N}\left(\gamma_{jk}x_j\right)}{\sum_{j}^{N}\gamma_{jk}},k=1,2,\ldots,K$$

$$\Sigma_k=\frac{\sum_{j}^{N}\gamma_{jk}\left(x_j-\mu_k\right)\left(x_j-\mu_k\right)^T}{\sum_{j}^{N}\gamma_{jk}},k=1,2,\ldots,K$$

$$\alpha_k=\frac{\sum_{j=1}^{N}\gamma_{jk}}{N},k=1,2,\ldots,K$$

（4）重复计算 E-step 和 M-step 直至收敛。

至此，就找到了高斯混合模型的参数。需要注意的是，EM 算法具备收敛性，但并不保证找到全局最大值，有可能找到局部最大值。本文解决方法是初始化几次不同的参数进行迭代，取方差之和更小的那次。

三． 实验结果

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。

下面展示两组不同初始参数时的实验结果：

第一组：（初值与真实值偏差较大时）

类别	男生占比	男生均值	男生方差	女生占比	女生均值	女生方差
初始值	0.5	172.55	1	0.5	173.07	1
EM 估计值	0.22	163.55	2.79	0.78	175.53	5.12
相对偏差	-70.74%	-7.07%	-44.14%	212.23%	7.03%	70.62%

第二组：（初值与真实值偏差较小时）

类别	男生占比	男生均值	男生方差	女生占比	女生均值	女生方差
初始值	0.5	173	1	0.5	162	1
EM 估计值	0.78	175.53	5.12	0.22	163.55	2.79
相对偏差	4.08 %	-0.27%	2.37%	-12.23%	-0.27%	-6.90%

由上表可得，选取不同的初始值，参数收敛值也不同，选取的初值与真实值越接近，EM 算法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。

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。

五． 参考文档

EM 算法（Expectation Maximization Algorithm）详解_zhihua_oba 的博客-CSDN 博客