

作业 2 基于 EM 算法的高斯混合模型

一、实验内容

给定男女生身高均值和方差，通过 EM 算法和高斯混合模型得到全班男女生的平均身高、方差和男女比例。

二、实验原理

EM 算法估计 GMM 参数。

分 2 步 ① 估计参数的期望值 (E-step)
② 使用 ① 的值最大化似然函数 (M-step)

假设有一个二值高斯混合模型: $(\mu_1, \sigma_1^2, \pi_1; \mu_2, \sigma_2^2, \pi_2)$ 为待估计参数, 满足:

$$p(x|z, \mu, \sigma) = \sum_{k=1}^2 \pi_k N(x|\mu_k, \sigma_k^2) \quad (1)$$

下面求解 μ_k, σ_k, π_k 的最大似然函数:

将 (1) 式取对数再对 μ_k 求导, 令导数为 0, 整理得:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N r(z_k) x_n \quad (2)$$

其中 $N_k = \sum_{n=1}^N r(z_k)$, N 表示数据量, $r(z_k)$ 表示 x_n 属于 k 类的后验概率。

同理可得 σ_k, π_k 的最大似然函数:

$$\sigma_k^2 = \frac{1}{N_k} \sum_{n=1}^N r(z_k) (x_n - \mu_k)(x_n - \mu_k)^T \quad (3)$$
$$\pi_k = \frac{N_k}{N} \quad (4)$$

$r(z_k)$ 可由贝叶斯定理得到: $r(z_k) = \frac{\pi_k N(x|\mu_k, \sigma_k^2)}{\sum_{j=1}^2 \pi_j N(x|\mu_j, \sigma_j^2)} \quad (5)$

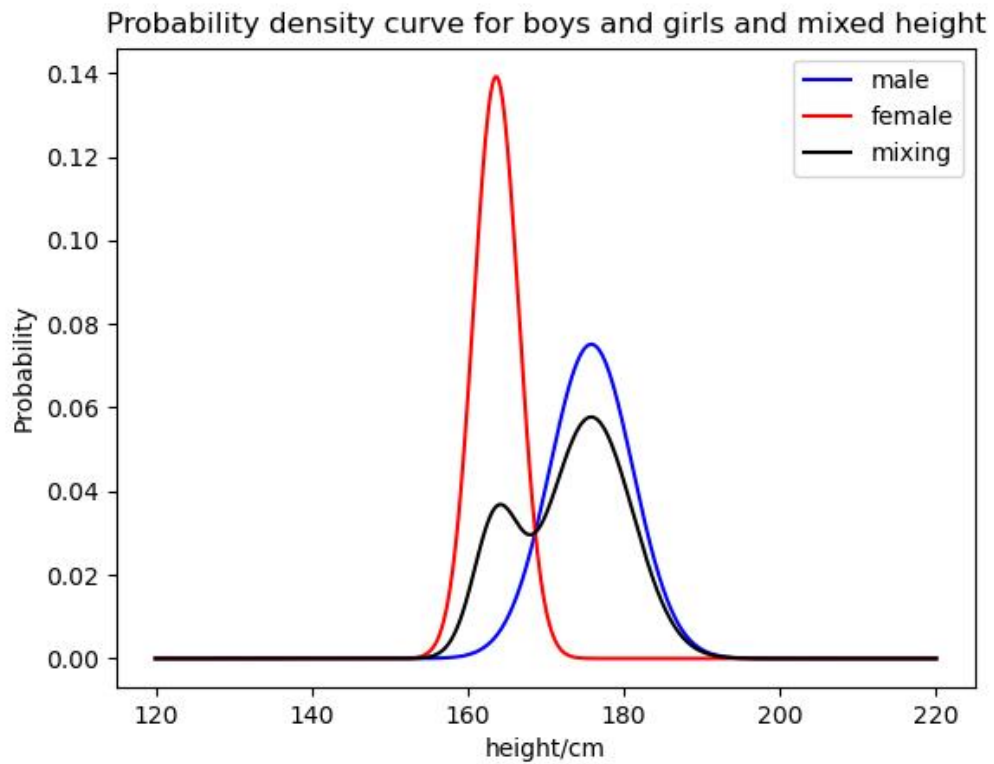
在使用 EM 算法时, 首先根据经验指定 μ, σ 的初值, 代入 (5) 式计算 $r(z_k)$, 再代入 (2) (3) (4) 式中, 求得 μ_k, σ_k, π_k , 再代入 (5) 式计算 $r(z_k)$, 循环往复, 直至算法收敛。

三、实验结果与分析

采用数据: 男生平均身高 176, 标准差 5, 样本数 1500. 女生平均身高 164, 标准差 3, 样本数 500. 可以得到男女占比分别为 0.75、0.25.

设定初始条件: 男生平均身高 170, 标准差 10, 比例 0.7. 女生平均身高 160, 标准差 10, 比例 0.3.

经过 EM 算法的 100 次迭代, 得到男生平均身高 175.82, 标准差 5.309, 占比 0.769, 女生平均身高 163.63, 标准差 2.866, 占比 0.231. 改变初始条件和增加 EM 算法的迭代轮数, 结果无变化. 得到男女生身高的混合概率密度曲线如下图:



EM 算法的预测结果与真实数据总体比较接近，但有一定误差。

分析误差原因，可能是样本数量太少、采样数据并不严格满足正态分布。为了验证这一猜想，本人将男女生的样本量增加到 15000、5000 重新进行实验。经过 EM 算法的 100 次迭代，得到男生平均身高 176.00，标准差 5.047，占比 0.752，女生平均身高 163.92，标准差 2.918，占比 0.248。课件，在增大样本量后，EM 算法的预测结果更接近真实值。