EM 算法拟合身高数据

窦嘉祺 SY2342102

Nov.23 2023

1 引入

EM 算法是一种迭代算法，其用于对存在隐变量的概率模型进行参数估计。由于之前没有接触过，以上课秦老师所讲大概叙述一下我现阶段对算法核心思想的理解：模型具有观测变量和隐变量，隐变量的值是未知的，因此若想极大化观测数据的似然函数，则需要对隐变量的值有一个合理的预测，这个预测就是隐变量取值期望的形式，而隐变量出现在观测变量的似然函数中，因此问题转换成极大化似然函数在隐变量上的条件期望取值，这里的条件指的是第 k 次迭代确定的参数值和观测数据。·

2 重要公式

上面提到我们要求条件期望，就必须知道隐变量的条件概率 $P(z \mid x, \theta)$，根据贝叶斯公式我们有：

$$\begin{aligned} P(z \mid x, \theta) &= \frac{P(z, x \mid \theta)}{P(x \mid \theta)} \\ &= \frac{P(x \mid z, \theta) \cdot P(z \mid \theta)}{\sum_z P(x \mid z, \theta) \cdot P(z \mid \theta)} \end{aligned} \tag{1}$$

对应于本次实验两个高斯分布的混合，可以设

$$z = \begin{cases} 1 & x_i \in M \\ 0 & x_i \in F \end{cases} \tag{2}$$

对于模型中的参数，我们用 α 表示样本来自男生的概率，μ_1, σ_1 表示男生身高对应高斯分布的均值和标准差，μ_2, σ_2 表示女生身高对应高斯分布的均值和标准差。

那么公式 1可以具体为

$$P(z = 1 \mid x, \theta) = \frac{f(x \mid \mu_1, \sigma_1) \cdot \alpha}{f(x \mid \mu_1, \sigma_1) \cdot \alpha + f(x \mid \mu_2, \sigma_2) \cdot (1 - \alpha)} \tag{3}$$

其中 $f(x \mid \mu, \sigma)$ 表示参数为 (μ, σ) 的高斯分布的概率密度函数。

3 数据预分析

表 1展示了原始数据的基本信息，包含男女生的样本数、均值、标准差。

4 EM 算法表达式推导（手写）

该推导部分附在文章结尾，包含概念式推导与更正式一些的推导。

表 1: 身高-性别数据信息

	样本数	均值	标准差
男生	78(0.82)	178.33	5.51
女生	17(0.18)	167.35	5.57
整体	95	176.37	6.96

5 实验结果

5.1 初值设置

利用前面第二章提到的模型参数表示方式，我们设置初值：$\alpha = 0.6, \mu_1 = 170, \sigma_1 = 15, \mu_2 = 160, \sigma_2 = 10, T = 200$，T 为算法迭代次数。

5.2 结果分析

- 首先，通过一张散点图 1直观感受一下原始数据分布
- 接着，我们拟合数据，经过 200 次对模型参数的更新，我们得到表 2所展示的结果，模型拟合的可视化效

表 2: 算法结果

	预测	原始
μ_1	178.02	178.33
σ_1	5.48	5.51
μ_2	163.82	167.35
σ_2	2.46	5.57
α	0.88	0.82

果如图 2所示，与真实数据进行对比，我们发现男生数据分布的拟合程度比女生要好，这可能是因为女生样本量过少的原因 (在北航不得不考虑的一大问题)。

- 而后，我对算法近似逼近的似然函数值变化作了可视化，算法通过参数迭代近似极大化观测数据的对数似然函数 $\log L(\theta)$

$$\begin{aligned} \log L(\theta) &= \log P(x \mid \theta) \\ &= \sum_z P(x, z \mid \theta) \\ &= \sum_z P(x \mid z, \theta) P(z \mid \theta) \end{aligned} \tag{4}$$

其中 $P(x \mid z, \theta)$ 即代表男生或女生的高斯概率密度函数，$P(z \mid \theta)$ 即代表样本是男生或女生的概率。将所有样本的对数似然值求和取平均再加一个负号便得到了图 3中 y 轴代表的损失函数。

5.3 模型预测

通过算法计算出的模型来完成课堂上的小预测，预测接下来进教室的十个人的身高，我们得到表 3

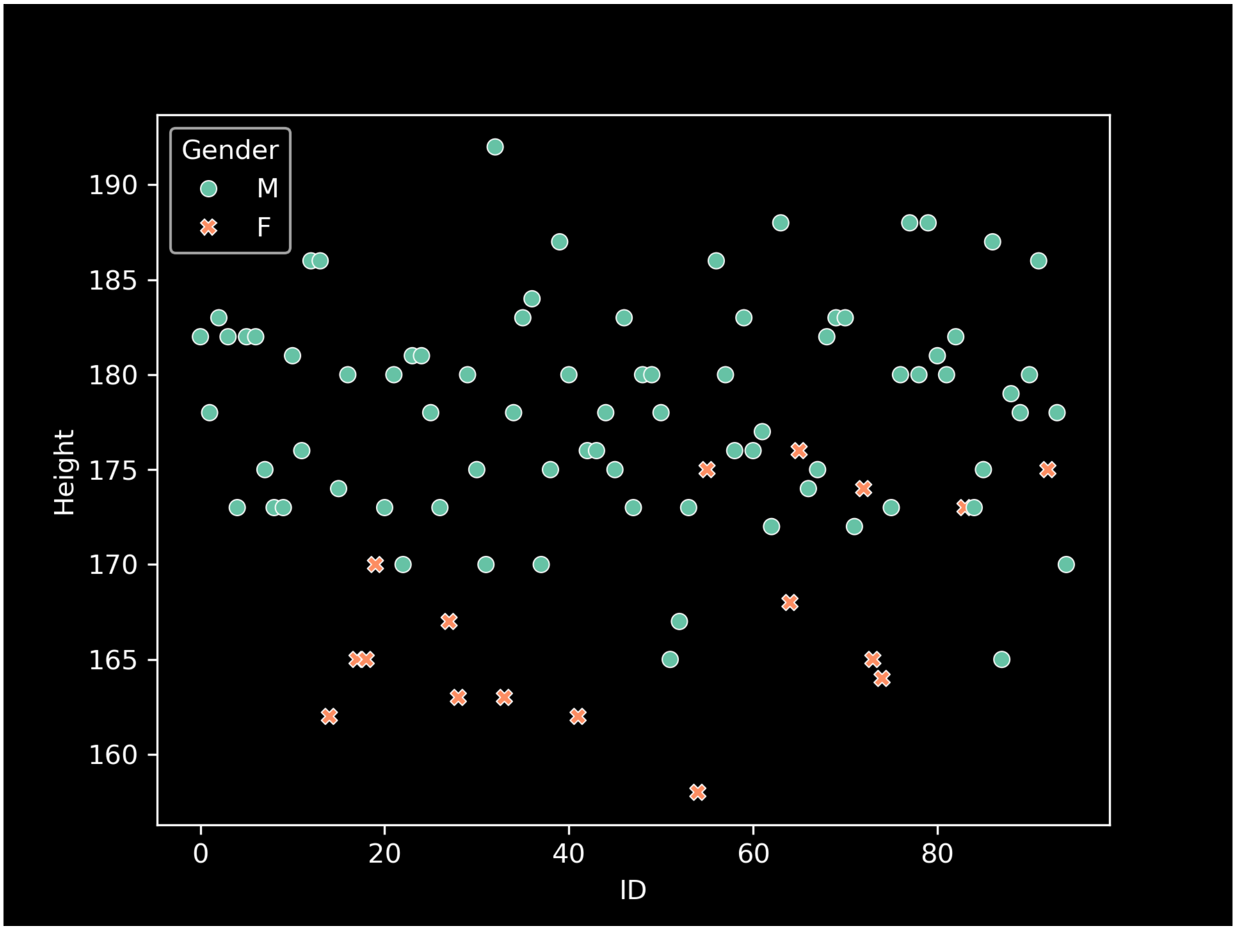


图 1: 数据分布散点图

表 3: 预测结果

	1	2	3	4	5	6	7	8	9	10
身高	182.33	175.73	163.79	175.50	176.03	163.29	173.78	179.93	183.41	175.20

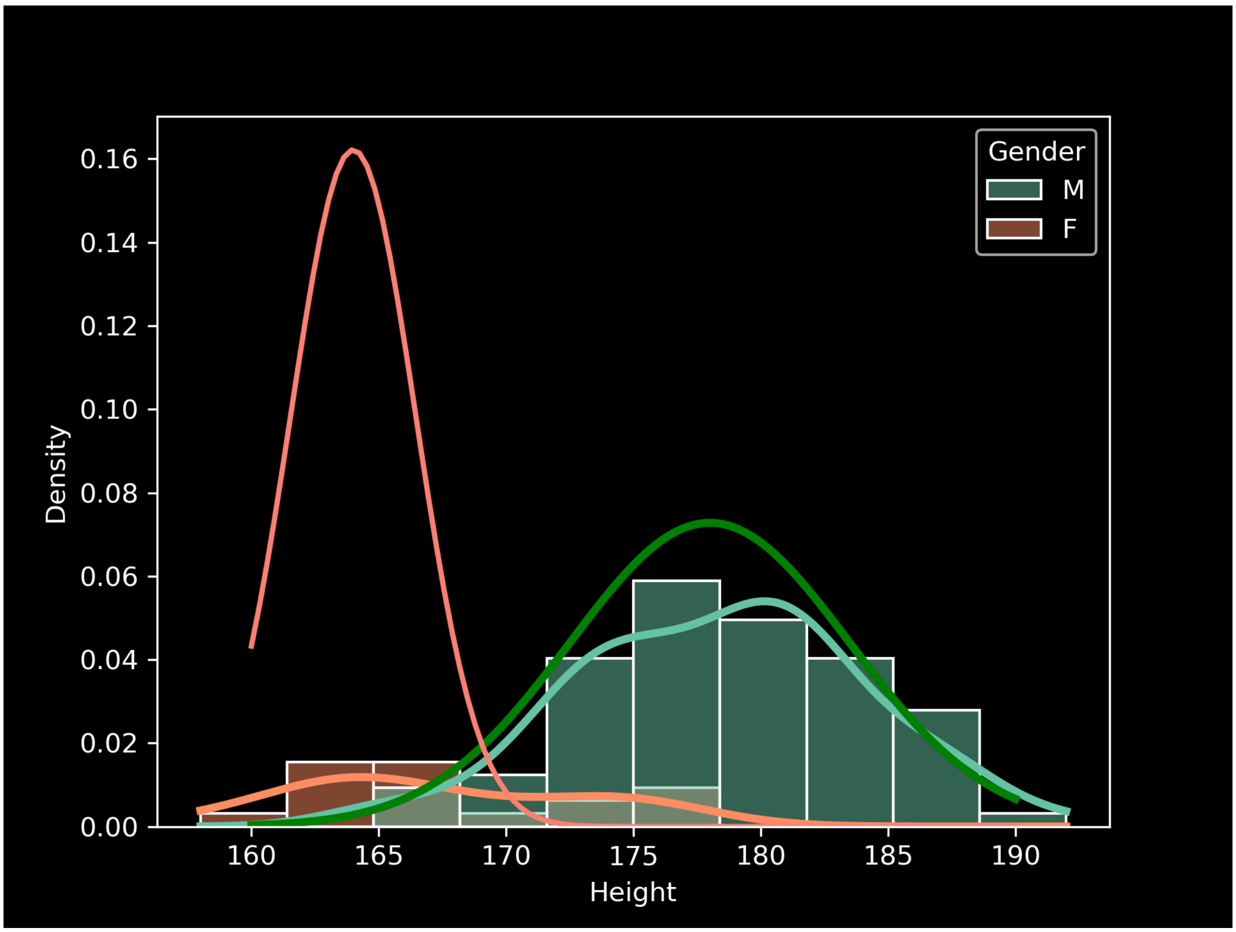


图 2: EM 算法拟合结果，图中较深的曲线代表算法拟合出的男女生身高高斯分布，图中较浅的曲线为真实数据的核密度函数估计

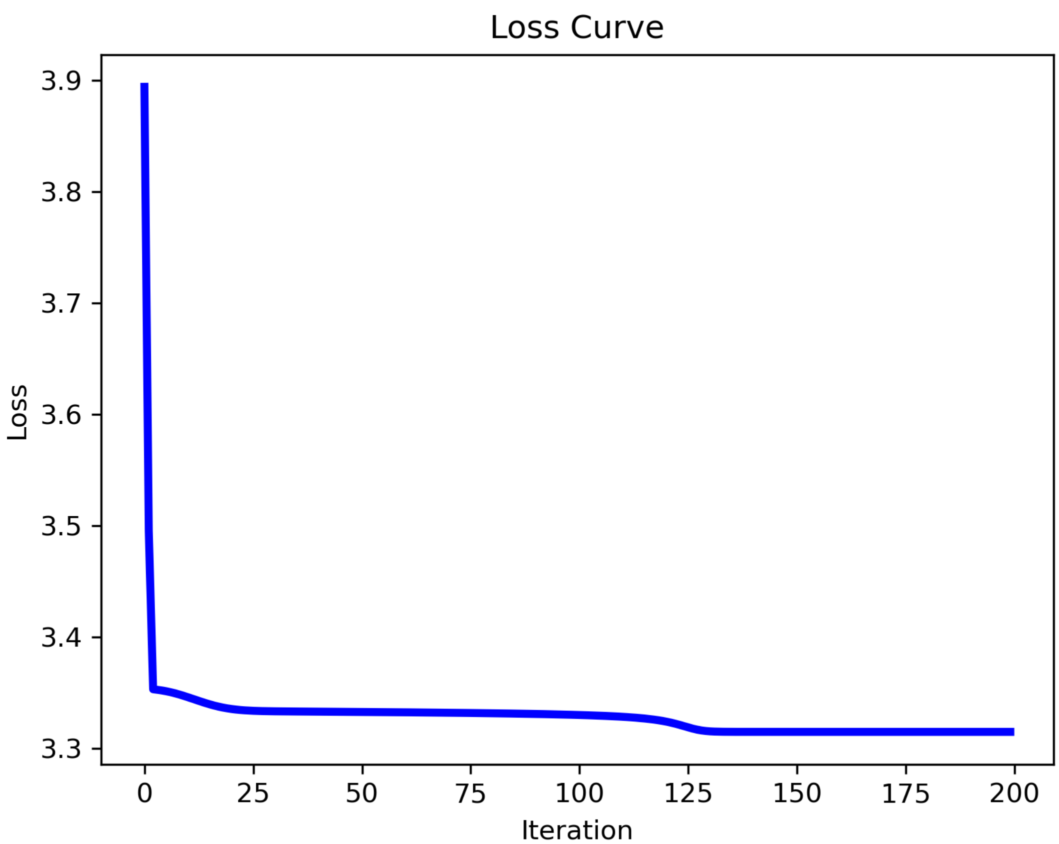

Loss Curve
3.9
3.8
3.7
3.6
3.5
3.4
3.3
Loss
0
25
50
75
100
125
150
175
200
Iteration

图 3: 损失函数变化

6 总结

通过本次实验，我们看到 EM 算法以迭代的方式能够对混合高斯模型进行求解，通过一定初值的选择，模型最终对数据的表达效果还不错，区分出了两个不同的高斯分布，但需要注意的是 EM 算法是初值敏感的，不同初值会导致算法得到不同的模型参数值。一个简单的例子，当初值设置的是 $\mu_1 = \mu_2, \sigma_1 = \sigma_2$，那么会导致每步计算的参数值都是一样的，这样最终男女生的高斯分布没有差异，决定样本属于男女生的参数 α 随即没有效果，这显然不是好的结果，模型的预测能力也会减弱，因此针对 EM 算法而言，开始时需要设置合理的初值。本次实验并没有具体分析及可视化初值对模型的影响，这是可以补充的一方面。

概念化推导:

If we assume to know which distribution generates the results

For example

$N=10$

$x^{(2)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$
181	172	175	186	162	168	170
181	172	175	186	162	168	170

$x^{(8)}$	$x^{(9)}$	$x^{(10)}$
169	174	179
169	174	179

$f(x)$为正态分布的概率密度函数

$$f(x|\theta) = \alpha f_1(x|\mu_1,\sigma_1) + (1-\alpha) f_2(x|\mu_2,\sigma_2)$$

f_1代表男生的概密，f_2代表女生

We use γ to represent the probability of x coming from male given current parameter

$\theta = (\alpha; \mu_1, \sigma_1; \mu_2, \sigma_2)$

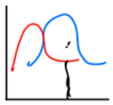
$$\gamma = \frac{\alpha f_1(x|\mu_1,\sigma_1)}{\alpha f_1(x|\mu_1,\sigma_1) + (1-\alpha) f_2(x|\mu_2,\sigma_2)} = P(z^M|x,\theta)$$

Given a sequence of Heights

E-step: $$\gamma^{(i)} = \frac{\alpha f_1(x^{(i)}|\mu_1,\sigma_1)}{\alpha f_1(x^{(i)}|\mu_1,\sigma_1) + (1-\alpha) f_2(x^{(i)}|\mu_2,\sigma_2)}$$

We now have a probability estimation of being male or female.

181 172 175 186 162 168 172 169 174 179

综上所述

M-step:

$$\Theta_1: \hat{\alpha} = \frac{\sum_{i=1}^{N} \gamma(i)}{N}$$

对于一个单变量正态分布，若有样本 $x_1, x_2, \cdots, x_n$

则该正态分布参数的极大似然估计为：

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

那么类比上此结论，混合高斯模型的极大似然估计：

$$\Theta_2: \hat{\mu}_1 = \frac{1}{\sum_{i=1}^{N}\gamma(i)} \sum_{i=1}^{N} \gamma(i) \cdot x^{(i)}$$

$$\Theta_3: \hat{\sigma}_1^2 = \frac{1}{\sum_{i=1}^{N}\gamma(i)} \sum_{i=1}^{N} \gamma(i) (x^{(i)} - \hat{\mu}_1)^2$$

$$\Theta_4: \hat{\mu}_2 = \frac{1}{\sum_{i=1}^{N} 1-\gamma(i)} \sum_{i=1}^{N} (1-\gamma(i)) \cdot x^{(i)}$$

$$\Theta_5: \hat{\sigma}_2^2 = \frac{1}{\sum_{i=1}^{N} 1-\gamma(i)} \sum_{i=1}^{N} (1-\gamma^{(i)})(x^{(i)} - \hat{\mu}_2)^2$$

更正式一点的推导：

写出观测变量和隐变量的联合分布概率密度函数，其中隐变量用 r_j 表示，r_j 意味样本 j 来自男生或女生

$$r_j = \begin{cases} 1, & \text{样本 } j \text{ 来自男生} \\ 0, & \text{------ 女生} \end{cases}$$

$$f(x, r \mid \theta) = \prod_{j=1}^{N} f(x_j, r_j \mid \theta)$$

$$= \prod_{j=1}^{N} [\alpha f(x_j \mid \theta_M)]^{r_j} \cdot [(1-\alpha) f(x_j \mid \theta_F)]^{1-r_j}$$

$$= \alpha^{n_m} \prod_{j=1}^{N} f(x_j \mid \theta_M)^{r_j} \cdot (1-\alpha)^{n-n_m} \prod_{j=1}^{N} f(x_j \mid \theta_F)^{1-r_j}$$

其中 $n_m = \sum_{j=1}^{N} r_j$

这个密度概率函数的对数似然函数为：

$$\log f(x, r \mid \theta) = \log \alpha^{n_m} \prod_{j=1}^{N} f(x_j \mid \theta_M)^{r_j} + \log (1-\alpha)^{n-n_m} \prod_{j=1}^{N} f(x_j \mid \theta_F)^{1-r_j}$$

$$= n_m \log \alpha + \sum_{j=1}^{N} r_j \log f(x_j \mid \theta_M) + (n-n_m) \log(1-\alpha) + \sum_{j=1}^{N} (1-r_j) \log f(x_j \mid \theta_F)$$

E-step：

EM 算法中的 E 步实则计算 $E_r[\log f(x, r \mid \theta) \mid x, \theta]$

$$E_r[\log f(x, r \mid \theta) \mid x, \theta]$$

$$= E\left\{ n_m \log \alpha + \sum_{j=1}^{N} r_j \log f(x_j \mid \theta_M) + (n-n_m) \log(1-\alpha) + \sum_{j=1}^{N} (1-r_j) \log f(x_j \mid \theta_F) \right\}$$

$$= \sum_{j=1}^{N} (E r_j) \log \alpha + \sum_{j=1}^{N} (E r_j) \log f(x_j \mid \theta_M) + \sum_{j=1}^{N} [E(1-r_j)] \log(1-\alpha) + \sum_{j=1}^{N} [E(1-r_j)] \log f(x_j \mid \theta_F)$$

其中 $E r_j = P(r_j = 1 \mid x_j, \theta)$ (就是上一篇推导中的 $\gamma(j)$)

$$= \frac{P(r_j = 1, x_j \mid \theta)}{P(x_j \mid \theta)}$$

$$= \frac{\alpha f(x_j \mid \theta_M)}{\alpha f(x_j \mid \theta_M) + (1-\alpha) f(x_j \mid \theta_F)}$$

$$E(1-r_j) = \frac{(1-\alpha) f(x_j \mid \theta_F)}{\alpha f(x_j \mid \theta_M) + (1-\alpha) f(x_j \mid \theta_F)}$$

将 $\gamma^{(j)} = E\gamma_j$，$1-\gamma^{(j)} = E(1-\gamma_j)$ 代入上式中

$$= \sum_{j=1}^{N} \gamma^{(j)} \log\alpha + \sum_{j=1}^{N} \gamma^{(j)} \log f(x_j|\theta_M) + \sum_{j=1}^{N} (1-\gamma^{(j)}) \log(1-\alpha) + \sum_{j=1}^{N} (1-\gamma^{(j)}) \log f(x_j|\theta_F)$$

M-Step:

将上式记为 $Q(\alpha, \theta_M, \theta_F)$，求解 $\arg\max Q(\alpha, \theta_M, \theta_F)$

则计算 $\begin{cases} \dfrac{\partial Q}{\partial \alpha} = 0 \\ \dfrac{\partial Q}{\partial \theta_M} = 0 \\ \dfrac{\partial Q}{\partial \theta_F} = 0 \end{cases}$

以 α 为例：

$$\frac{\partial Q}{\partial \alpha} = \frac{1}{\alpha} \sum_{j=1}^{N} \gamma^{(j)} - \frac{1}{1-\alpha} \sum_{j=1}^{N} (1-\gamma^{(j)})$$

$$0 = \frac{1}{\alpha} \sum_{j=1}^{N} \gamma^{(j)} + \frac{1}{1-\alpha} \sum_{j=1}^{N} \gamma^{(j)} - \frac{N}{1-\alpha}$$

$$N = \left(\frac{1-\alpha}{\alpha} + 1\right) \sum_{j=1}^{N} \gamma^{(j)}$$

$$\frac{N}{\sum_{j=1}^{N} \gamma^{(j)}} = \frac{1}{\alpha}$$

求得 $\alpha = \dfrac{\sum_{j=1}^{N} \gamma^{(j)}}{N}$

同理 $\mu_M = \dfrac{\sum_{j=1}^{N} \gamma^{(j)} \cdot x_j}{\sum_{j=1}^{N} \gamma^{(j)}}$ . . . .