模式识别第三次作业

一、问题分析

通过给定均值与标准差生成虚拟的大学男女生身高数据共 N 个：\mu_M=176, \sigma_M=8, \mu_F=164, \sigma_F=6， 其中男女比例为 3:2。

（1）用混合高斯模型对大学学生身高进行建模，并推导利用 EM 算法求解的公式（手写）

（2）编程实现 EM 算法对于以上 5 个参数的估计并对比正确结果并讨论 EM 算法的优缺点。

二、推导公式

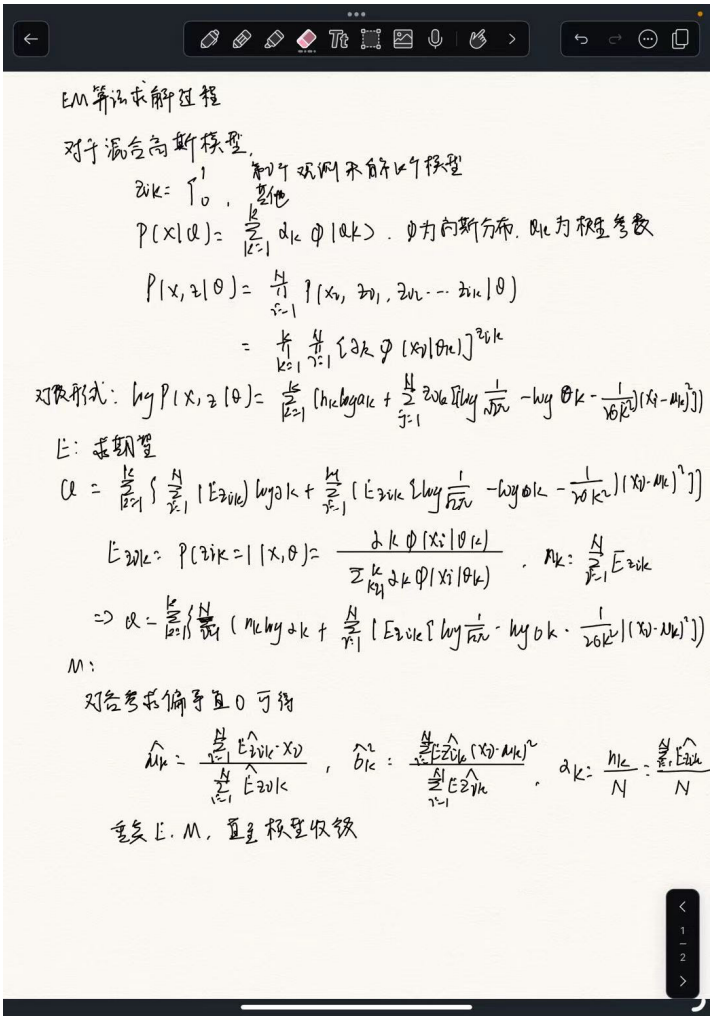

图 1 推导过程

三、数据生成

根据作业要求提供的各参数，生成符合要求的大学生男女身高数据。

将生成的实验数据进行可视化处理，示意图如下所示：

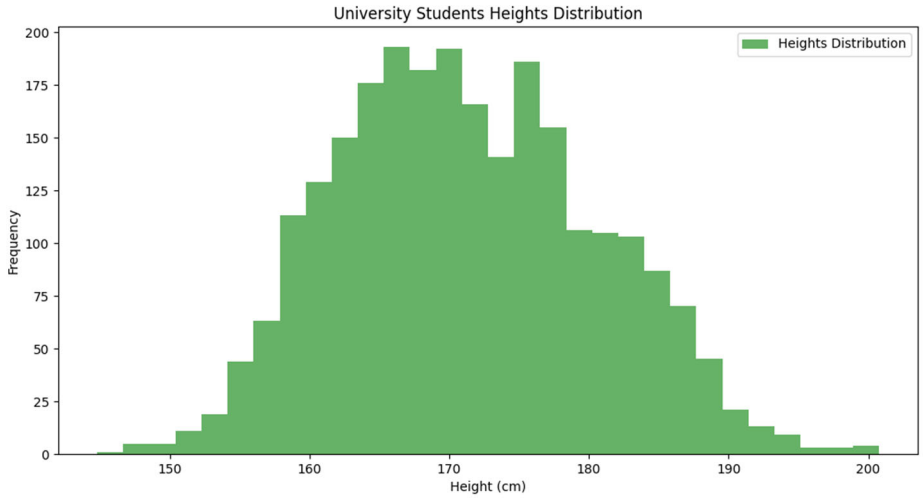

图 2 原始数据示意图

对应的生成数据所需要的参数如下：

N = 2500

mu_M = 176

sigma_M = 8

mu_F = 164

sigma_F = 6

接下来，利用给定的参数得到的男女大学生身高数据进行分析。

四、算法实现

EM 算法代码如下所示：

```
1.     def em_algorithm(heights, max_iter=100, tol=1e-6):
2.     N = len(heights)
3.
4.  # 初始化参数
 5.     pi = 0.6  # 初始男生比例
 6.     mu_M = np.mean(heights) + 5
 7.     sigma_M = np.std(heights)
 8.     mu_F = np.mean(heights) - 5
 9.     sigma_F = np.std(heights)
10.
11.     for _ in range(max_iter):
12.         # E 步：计算后验概率
```

```
        gamma_M = pi * gaussian(heights, mu_M, sigma_M)
        gamma_F = (1 - pi) * gaussian(heights, mu_F, sigma_F)
        gamma_sum = gamma_M + gamma_F
        gamma_M /= gamma_sum
        gamma_F /= gamma_sum

        # M 步：更新参数
        N_M = np.sum(gamma_M)
        N_F = np.sum(gamma_F)

        mu_M_new = np.sum(gamma_M * heights) / N_M
        sigma_M_new = np.sqrt(np.sum(gamma_M * (heights - mu_M_new) ** 2) / N_M)

        mu_F_new = np.sum(gamma_F * heights) / N_F
        sigma_F_new = np.sqrt(np.sum(gamma_F * (heights - mu_F_new) ** 2) / N_F)

        pi_new = N_M / N

        # 检查收敛性
        if np.abs(mu_M - mu_M_new) < tol and np.abs(sigma_M - sigma_M_new) < tol and \
                np.abs(mu_F - mu_F_new) < tol and np.abs(sigma_F - sigma_F_new) < tol and \
                np.abs(pi - pi_new) < tol:
            break

        mu_M, sigma_M, mu_F, sigma_F, pi = mu_M_new, sigma_M_new, mu_F_new, sigma_F_new, pi_new

    return pi, mu_M, sigma_M, mu_F, sigma_F
```

根据所编写代码，也可以印证第二部分中手写推导公式的正确性，加深对于 EM 算法的理解。

详细代码见代码文件。

五、结果分析

根据不同的 EM 算法参数设置，将每次的结果绘制成直方图的形式，以方便观察与总结，示意图如下图所示：

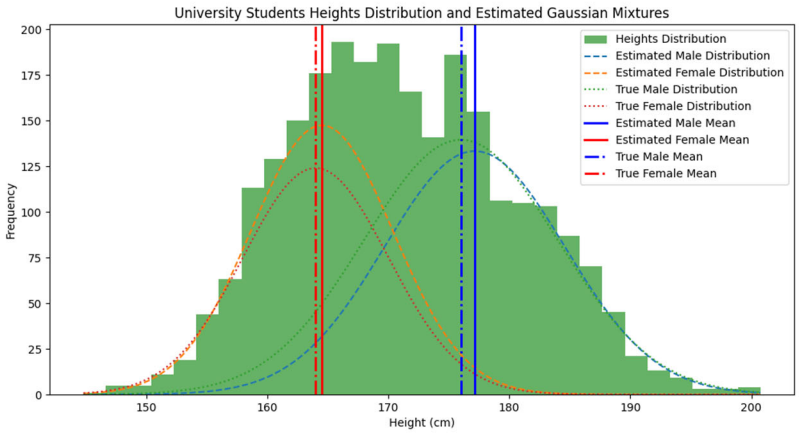

图 3 结果示意图

图中对应的迭代次数为 500，参数设置如下：

pi = 0.6 # 初始男生比例

mu_M = np.mean(heights) + 5

sigma_M = np.std(heights)

mu_F = np.mean(heights) - 5

sigma_F = np.std(heights)

得到的一组参数结果如下所示：

pi (male proportion): 0.578

mu_M (male mean height): 176.578

sigma_M (male height std): 7.847

mu_F (female mean height): 164.074

sigma_F (female height std): 5.972

可以看到，在这组参数设置的情况下，拟合的效果较好。

接下来分析 EM 算法参数改变对于结果的影响：首先我尝试的是迭代次数的改变，在经过若干次实验之后，在初始值设定为 0.6（男生比例）时，更改迭代次数得到的结果并没有太大差别，拟合效果都较好。下面着重分析初始值对于结果的影响。

改变男生比例的初始值为 0.3，得到的结果如下所示：

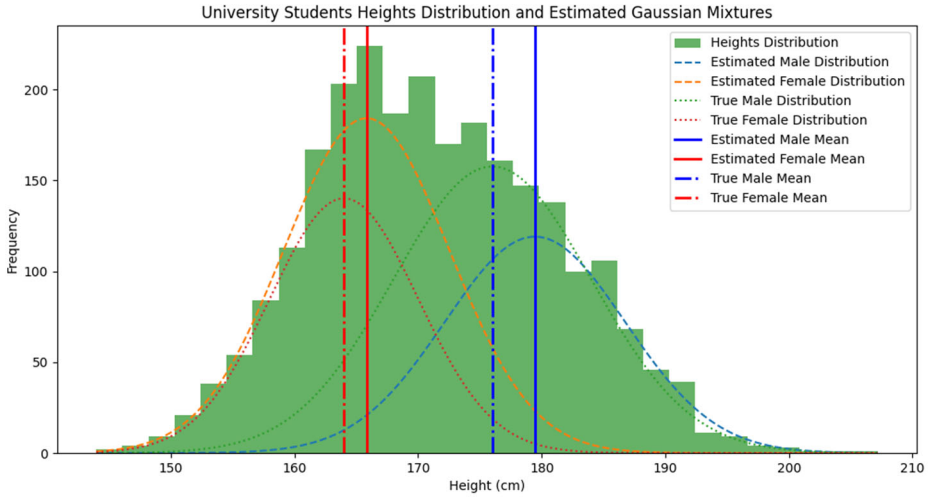

图 4 pi=0.3 对应示意图

五个参数对应为:

pi (male proportion): 0.413

mu_M (male mean height): 179.451

sigma_M (male height std): 7.291

mu_F (female mean height): 165.823

sigma_F (female height std): 6.700

可以看到，EM 算法所得到的结果有着较大的偏差，在接下来的实验过程中，继续对比例这一参数进行较大改动（偏离数据真实设定值），均可以看到，EM 算法得出的结果效果并不好，同时在增加迭代次数的情况下，这种现象也并没有改观。另外，我也尝试更改了几组均值，使其偏离理想设定值，得到上述相同结论，由此可以推断，EM 算法最终的效果十分依赖于我们初始参数的设定。

六、总结

通过理论学习以及实验验证，下面对 EM 算法的优缺点进行分析：

优点：适用于缺失数据： EM 算法它通过期望步骤（E 步）计算缺失数据的期望值，然后在最大化步骤（M 步）更新参数，使其在有缺失数据的情况下也能进行参数估计。

EM 算法可以有效地处理包含隐变量或缺失数据的情况。它通过迭代地估计隐变量和优化参数来最大化对数似然。

EM 算法可以应用于各种统计模型，如混合高斯模型、隐马尔可夫模型等。它在许多领域（如生物信息学、图像处理、机器学习）都有广泛应用。

缺点：首先最大的缺点便是 EM 算法对于初始值的依赖，初始值选择不当

时，结果的偏差可能会非常大。

另外，对于局部最优问题，通常需要多次运行 EM 算法，并选择对数似然值最高的结果。这增加了计算成本和复杂性。