 成 绩

****

**模式识别与机器学习**

**（EM算法的GMM应用）**

院（系）名 称 高 等 理 工 学 院

专 业 名 称 自动化

学 号 18231053

姓 名 则坤睿

2021年5月21日

1. 问题重述：

现有男生400人，女生200，男女生的身高均符合正态分布，我们可以获得男女生的全部身高，一共有600个数据，但是并不知道任何一个身高数据是男生的还是女生的。这时这600个身高数据符合一个混合高斯分布（GMM），我们使用EM算法希望还原出：

1：男女生比例（w1,(1-w1)）；

2：男生身高正态分布平均值mu1;

3：男生身高正态分布标准差sigma1;

4：男生身高正态分布平均值mu2;

5：男生身高正态分布标准差sigma2;

1. 方法原理：
2. 数据集的构建：

分别使用python numpy相关的函数生成400个男生身高数据，数据符合正态分布（）再生成200个女生身高数据，数据也符合正太分布（），固定随机数种子，以确保每次运行是生成的600个身高数据是相同的。

对于400个男生身高数据和200个女生身高数据，我设置了基于随机数的混合逻辑，每次运行都会将600个固定的身高数据（其中男生400个，女生200个）随机混合。混合之后可以得到要求的待测试数据集。

（二）EM算法：

如果一个概率模型的变量都是观测变量，那么给定数据，可以直接利用极大似然估计法估计模型参数。但是当模型参数中含有隐变量时，就不能使用常规的极大似然法了。最大期望算法（EM算法）是一种迭代算法，可以用来处理具有隐变量的概率模型，大致思路如下：

输入：观测变量数据为Y，隐变量数据为Z，联合分布为,条件分布。

输出：模型参数。

（1）选择参数的初值，开始迭代。

（2）E步：记为第i带迭代参数，在第i+1次迭代的E步，计算：



也就是说用当前的可观测变量参数估计隐变量参数。

（3）M步：求使得极大化的一组，作为第i+1次迭代的参数估计值：



也就是根据隐函数的参数，利用极大似然法确定可观测变量的参数。

重复迭代（2）、（3）直到收敛。

（三）混合高斯分布（GMM）：

高斯混合模型是指具有如下形式的概率分布模型：



其中一共有K个不同高斯分布构成，是每一个高斯分布的系数，且有：



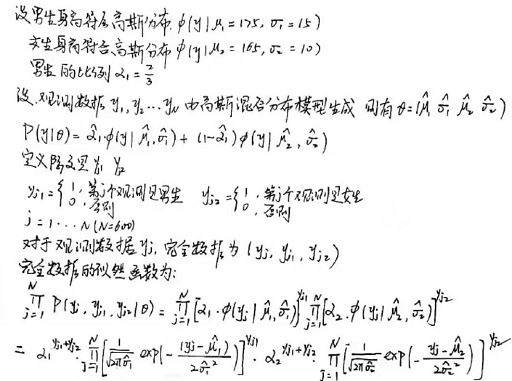
是一维高斯概率密度分布函数，，有：

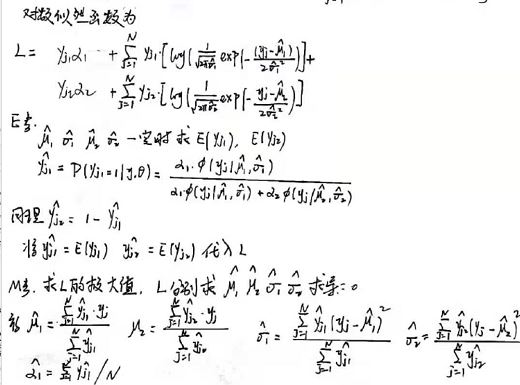


对于本问题，只有男生女生身高分布两个不同的高斯分布，则K=2。

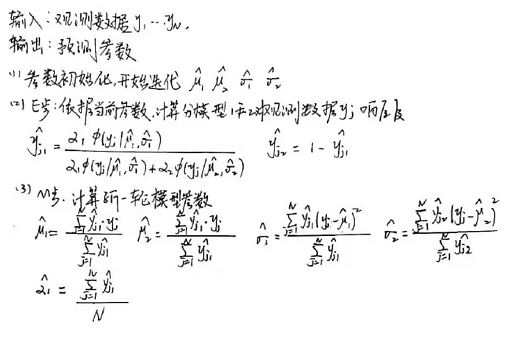
（四）混合高斯分布（两高斯分布混合）EM法推导：

EM推导为手写推导，参考文献《统计学习方法》





（五）EM算法实现



其中：

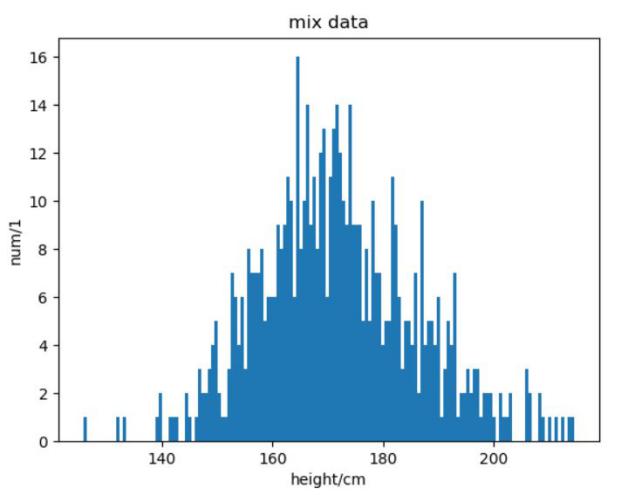
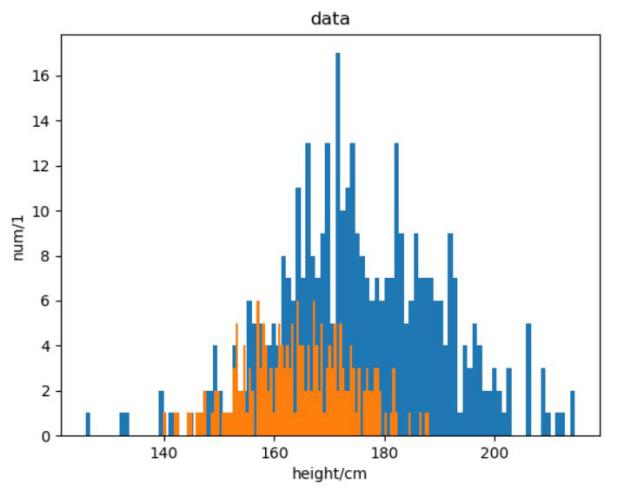


1. 测试结果：

使用固定的400个男生身高数据和200个女生身高数据，程序每次会随机组合数据以打乱顺序。

对打乱顺序的600个样本数据使用EM算法，预测混合高斯分布的参数以及男生女生比例，程序运行结果如下：

（1）原始数据展示

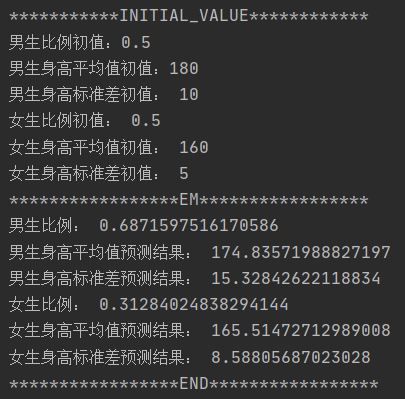


左侧图中橘黄色代表200个女生的身高样本数据，蓝色代表400个男生的身高样本数据。

右侧图中是男女生身高数据随机混合之后的结果。

1. EM预测结果

对于随机的初始值，EM算法预测的结果如下：



而实际上标准值为：

男生比例：0.66

男生身高平均值：175

男生身高标准差：15

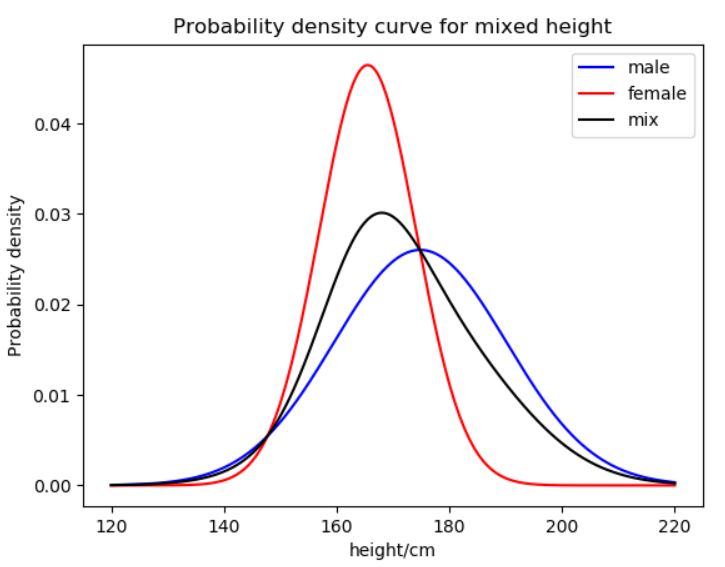
女生比例：0.33

女生身高平均值：165

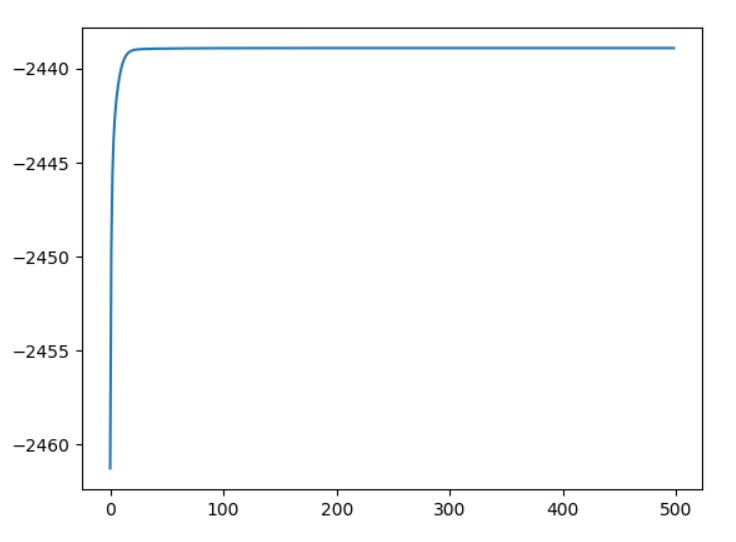
女生身高标准差：10

预测结果与实际结果一致的程度很好。

将预测的男生身高分布（图中蓝色实线）、女生身高分布（图中红色实线）和预测出的混合身高分布（图中黑色实线）绘制如下。



1. EM迭代过程中对数似然函数变化



对数似然函数增长。

1. 总结：

本文中利用EM方法求解混合高斯模型，程序使用Python编写，依赖库环境如下：

Numpy、Matplotlib、Scipy、random