单位代码

学 号ZY2203811

分类号

****

**EM 算法解决身高数据的参数估计**

**深度学习与自然语言处理（NLP）第二次课后作业**

|  |  |
| --- | --- |
| 院（系）名称 | 自动化科学与电气工程学院 |
| 专业名称 | 自动化 |
| 学生姓名 | 许铁 |

2023年04月

### 目录

[目录 2](#_Toc131695081)

[1内容介绍 2](#_Toc131695082)

[2实验原理 2](#_Toc131695083)

[4实验过程 3](#_Toc131695084)

[5实验结果分析 6](#_Toc131695085)

[6总结 7](#_Toc131695086)

[7参考文献 7](#_Toc131695087)

### 1内容介绍

题目：请使用身高数据，需要使用EM算法来估计高斯混合模型的参数，并使用这些参数来进行预测。你需要对模型进行评估，并解释模型的性能。

### 2实验原理

EM算法是一种在存在潜伏变量的情况下进行最大似然估计的方法。它通过首先估计潜在变量的值，然后优化模型，然后重复这两个步骤直到收敛。这是一种有效的通用方法，最常用于有缺失数据的密度估计，例如高斯混合模型、伯努利混合模型等聚类算法。对于服从混合分布的数据Y，数据集由许多点组成，这些点恰好由两个不同的分布产生。每个点都属于一个概率分布，但数据是组合在一起的，分布足够相似，以至于某个点可能属于哪个分布并不明显。可以使用两个分布和表示Y的概率密度，其中参数为，则可以将Y的概率密度表示如下：

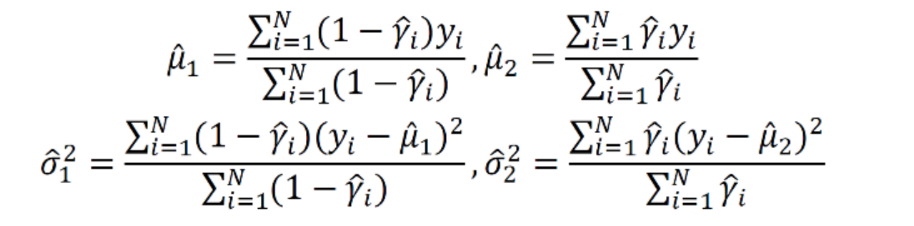
其中参数为，为混合比例。则参数估计的对数似然函数可以表示如下：

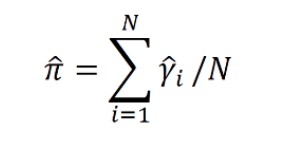
其中N为训练数据的个数。通过对上述对数似然函数进行变形，可得到如下表示：

其中为隐变量，当取值为0时表示来自于模型一，当取值为1时表示来自于模型二。在这种情况里，的极大似然估计是的数据样本均值和方差，的极大似然估计是的数据样本均值和方差。由于的实际值未知，我们使用迭代的方式进行处理，用期望代替，即：

下面介绍二分量参数估计的EM算法的具体步骤。首先初始化模型参数,,。之后设置E步算法如下：

该期望表示数据属于的概率，可以看到越接近0表示越可能属于模型一，反之，可以看到越接近1表示越可能属于模型二。最后设置M步算法如下所示：





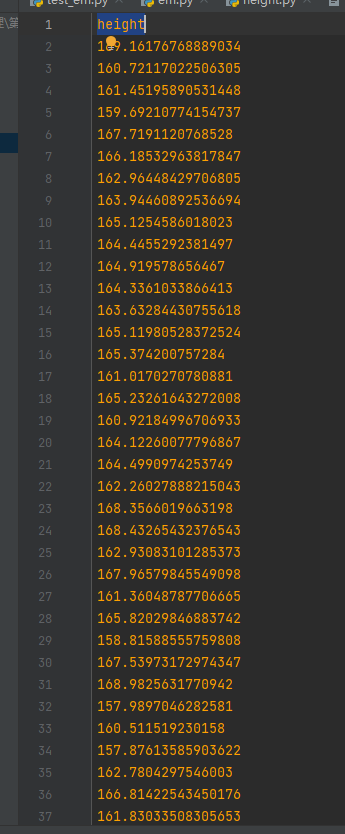
在EM算法中，估计步骤将为每个数据点的过程潜变量估计一个值，而最大化步骤将优化概率分布的参数，以试图最好地捕捉数据的密度。这个过程不断重复，直到有一组好的潜变量值和最大似然值符合数据。E步骤。估计每个潜变量的期望值。M步骤。使用最大似然法优化分布的参数。

### 4实验过程

按照EM算法的流程，使用Python语言编写代码。

在test\_em.py中，实现了Solution、EM4GMM两个类，其中Solution是问题的解，EM4GMM只有一个静态方法solve可以调用。

test\_em.py是使用生成的身高数据来运行EM算法的脚本文件。在脚本中，读取了男女生共2000个样本数据，其实际参数未知。

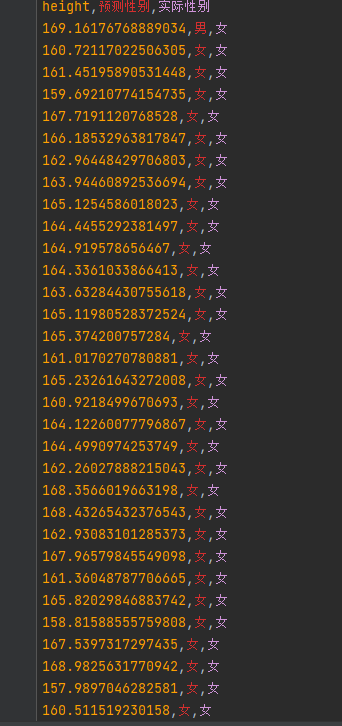


调用EM算法时，给定初始参数

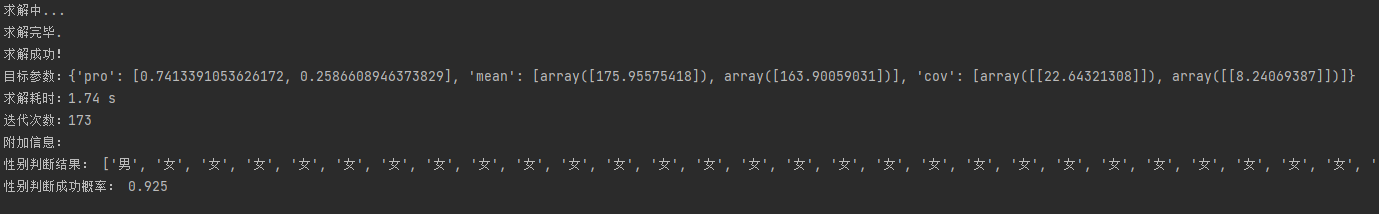
该初始参数是基于如下先验信息给出的：

1. 男生一般比女生高；
2. 北航男生比女生多。

运行程序，根据条件概率计算某个样本属于哪一类

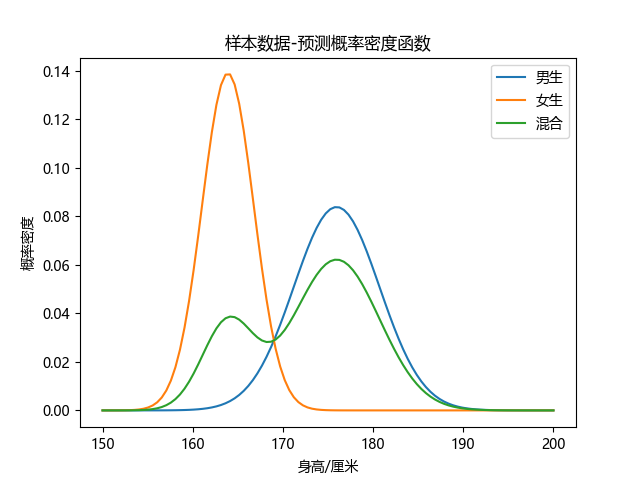


算法的运行结果如下所示。



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 样本数量 | 是否有解 | 求解结果 | 迭代次数 | 求解耗时/秒 | 预测成功率 |
| 2000 | 是 |  | 173 | 1.69 | 0.925 |

绘制在预测参数下的概率密度曲线，如下图所示。



##### 5实验结果分析

结果表明，算法能在可接受的时间内收敛，并且对于身高数据的预测有着较高的正确率。根据EM算法收敛性分析的过程，可知EM算法找到的是一个局部极值点，而不一定是全局最值点。因此从理论上来说，当在M-Step中优化的目标函数不是一个凸函数时，给定不同的初始参数值，可能会收敛到不同的结果。

下面给定了不同的初始参数来训练模型。在随机生成数据测试EM算法的过程中，发现给定不同初始参数，EM算法得到的结果将有一定的差别，如下表所示。可以看出EM算法对初始身高参数敏感，容易陷入局部最优解而导致参数预测不准确。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 初始参数 | 是否有解 | 求解结果 | 迭代次数 | 求解耗时/秒 | 预测成功率 |
|  | 是 |  | 122 | 1.69 | 92.5% |
|  | 是 |  | 211 | 2.05 | 92.5% |
|  | 是 |  | 188 | 1.84 | 92.5% |
|  | 是 |  | 262 | 2.55 | 7.5% |

### 6总结

由于EM算法容易陷入局部最优，本文探究了参数不同的初始状态对最终结果的影响，进行了对最终收敛结果的影响进行分析。同时研究发现，收敛结果对于均值初始化非常敏感。

传统EM算法对初始值敏感，聚类结果随不同的[初始值](https://www.zhihu.com/search?q=%E5%88%9D%E5%A7%8B%E5%80%BC&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7b)而波动较大。总的来说，EM算法收敛的优劣很大程度上取决于其初始参数，其中均值对于收敛结果影响较大。本次实验增强了我对于EM算法的掌握，加深了对于聚类算法的理解，提升了对于问题的分析能力与实现程序过程中的设计能力与完成能力。

### 7参考文献

[1] https://zhuanlan.zhihu.com/p/30483076

[2]https://zhuanlan.zhihu.com/p/93513123

[3]https://baike.baidu.com/item/%E4%BC%AF%E5%8A%AA%E5%88%A9%E5%88%86%E5%B8%83/7167021?fr=aladdin