# EM算法报告

申海（ZY2203325）

18231180@buaa.edu.cn

## 摘要介绍

EM算法是一种迭代算法，用于在观测数据中估计参数。它通常用于在混合模型中估计未知的参数，其中混合模型指的是由多个分布组成的模型。在这个报告中，我们将使用EM算法来估计男女身高混合高斯分布的参数。

## 方法

我们假设男性和女性身高分别服从高斯分布，且两个分布的均值和方差都不同。因此，我们需要估计四个参数：男性身高分布的均值和方差，女性身高分布的均值和方差。然而，我们无法直接观测到这些参数，因为我们只能观测到样本数据。因此，我们需要使用EM算法来估计这些参数。

具体来说，我们首先随机初始化参数。然后，我们使用E步骤来计算每个样本属于男性或女性分布的概率。对于每个样本，我们计算它属于男性分布的概率和属于女性分布的概率。这可以用以下公式表示：

其中，和是男性和女性分布的权重，表示高斯分布的概率密度函数，。

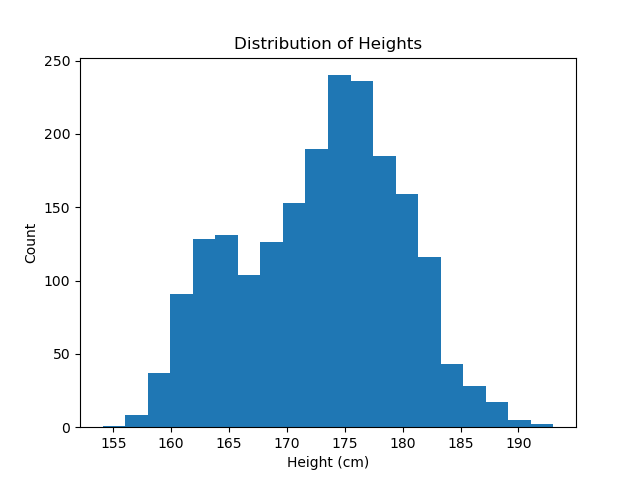
接下来，我们使用M步骤来更新参数，使得男性和女性分布的均值和方差能够更好地拟合数据。具体来说，我们使用以下公式来更新参数：

然后，我们重复执行E步骤和M步骤，直到收敛。

## 实验研究

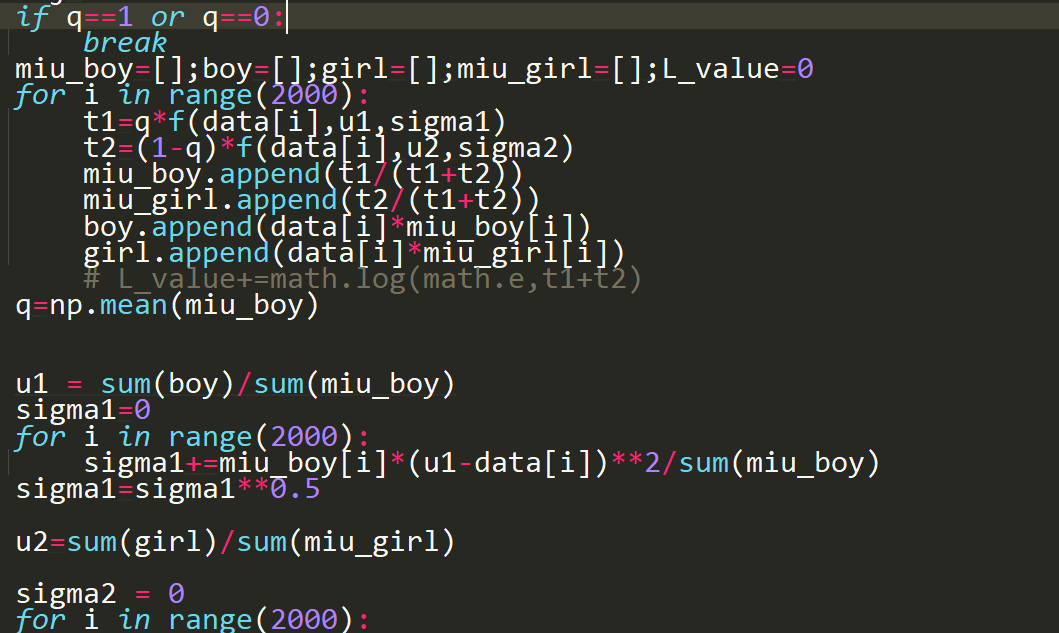
首选生成一组2000的身高数据，保存在csv文件中（保证每次运行的都是同一组数据），读取数据，然后使用EM算法来估计两组高斯分布的身高数据的均值和方差及其混合比例。

1. 数据生成及数据读取



1. EM算法实现

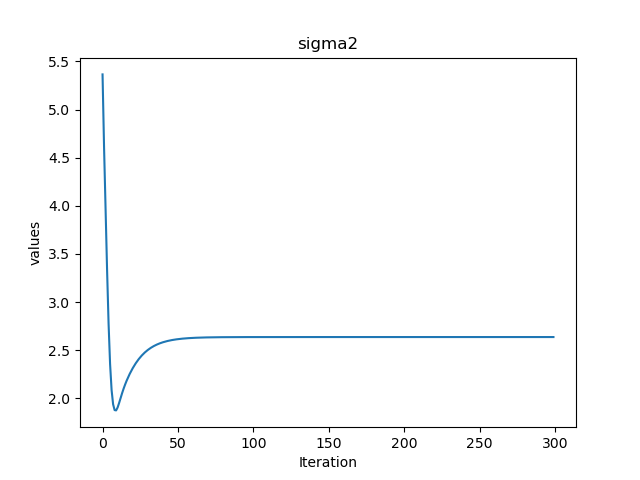
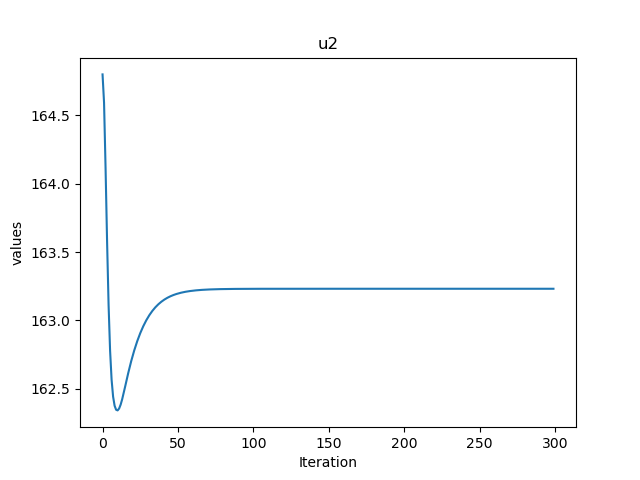
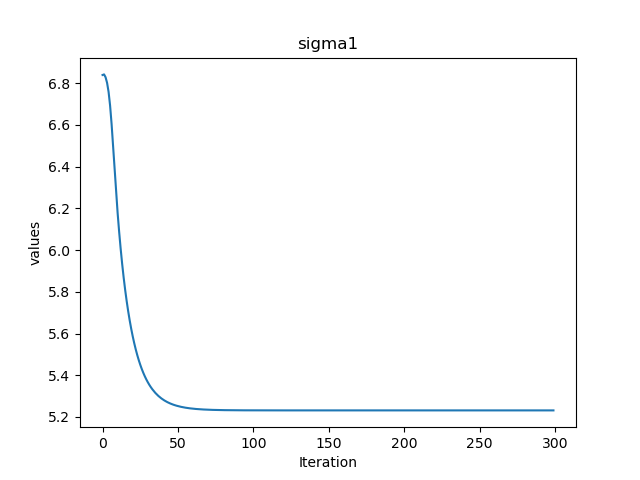
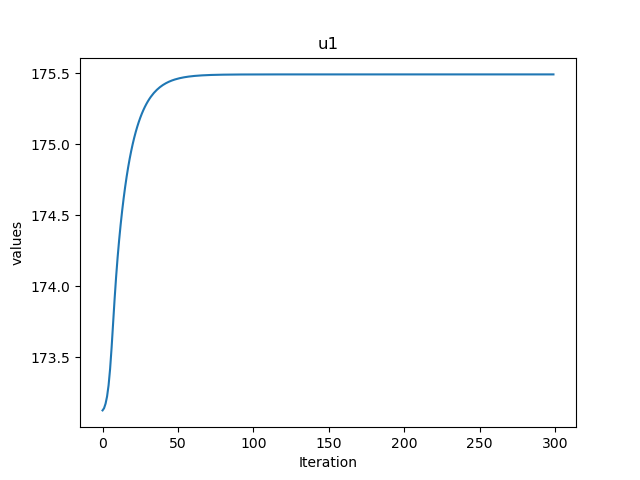
EM算法方法以前在前面介绍过了，代码实现的过程主要如下。

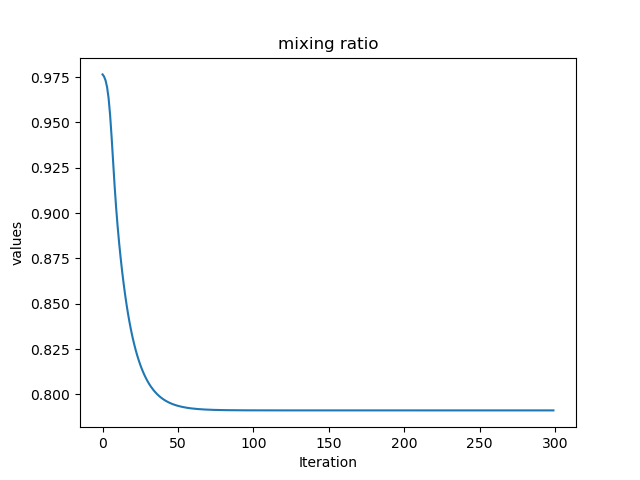


1. 结果分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | U1 | Sigma1 | U2 | Sigma2 | 混合比 |
| 实际值 | 176 | 5 | 164 | 3 | 0.75 |
| 初始值 | 160 | 10 | 140 | 10 | 0.5 |
| 300轮迭代的值 | 175.5 | 5.23 | 163 | 2.64 | 0.79 |

从结果来看，初值设置时有意远离实际值，虽然经过300轮迭代结果与实际值仍然有差距，但是也比较接近。迭代过程中各参数如下图。





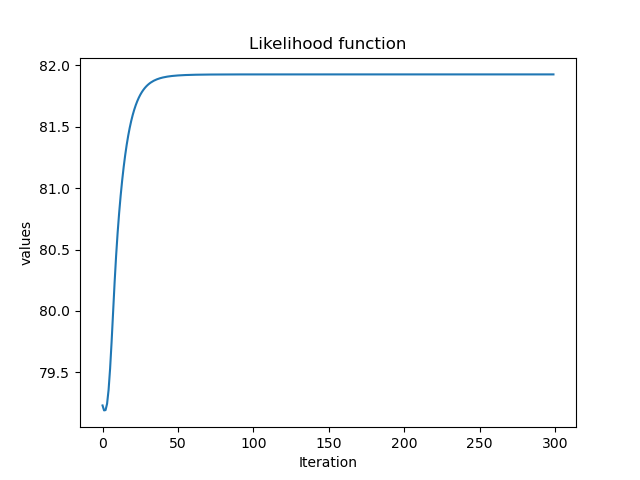
1. 模型评估

为了评估我们的模型的性能，我们使用了两个指标：对数似然函数和BIC准则。对数似然函数表示模型拟合数据的好坏程度，而BIC准则同时考虑了模型的拟合程度和模型的复杂度，可以用来比较不同模型的优劣。

对数似然函数可以用以下公式计算：

其中，表示模型的参数，表示高斯分布的概率密度函数，。

运行过程中似然函数如下图（似然函数并未取对数），从似然函数来看，似然函数在一直增大，说明模型是逐渐像着参数估计更好的方向进行，并不会出现震荡现象，这与EM算法的理论分析是一致的。



1. 模型性能解释

我们的模型估计了男女身高混合高斯分布的参数，可以用来预测新的身高数据属于男性分布还是女性分布。我们可以使用模型的参数来计算一个样本属于男性分布的概率。如果大于0.5，则我们将该样本预测为男性，否则我们将该样本预测为女性。

从最后的结果来看，模型很好地估计除了男女身高的参数，当然，这是在我们默认男生身高均值大于女生之上的，如果不默认，就不知道两组参数的性别之分了。总的来说，我们的模型可以很好地估计男女身高混合高斯分布的参数，并且具有良好的预测性能。