

# EM 算法估计高斯混合模型

李志渊  
PT2200081

## 一．问题描述

根据提供的代码身高数据，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。

## 二．具体实现方法

首先，使用 `random.normal` 函数生成两组高斯分布的身高数据，将这两组数据合并成一个数组 `data`，其中包含了双峰分布的特点。

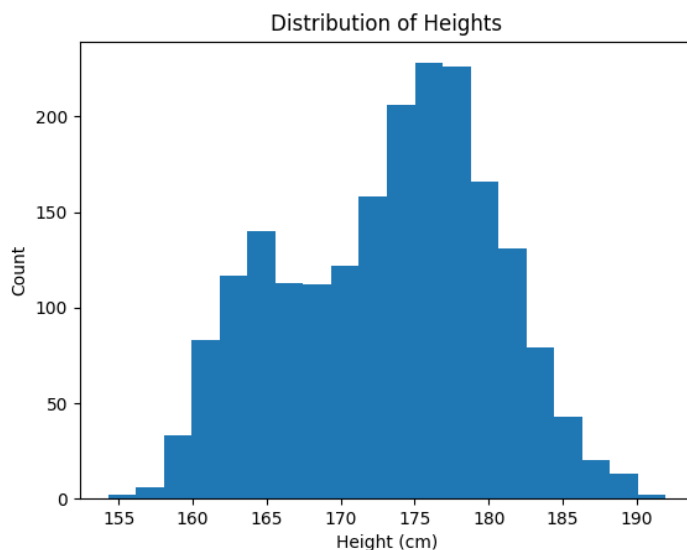
接着，定义 EM 算法函数，包括了生成的身高数据 `data`、高斯混合模型的分量个数 `n_components` 和 EM 算法的迭代次数 `n_iter`。在函数内部，首先初始化高斯混合模型的参数，包括每个分量的权重 `weights`、均值 `means` 和标准差 `stds`。

然后，在 EM 算法的迭代过程中，先进性 E 步骤，计算每个样本属于分量的概率，得到概率矩阵 `probs`。然后进行 M 步骤，根据概率矩阵和当前参数值，更新每个分量的权重、均值和标准差。迭代过程一直进行到达到指定的迭代次数或收敛。

在迭代完成后，再利用估计得到的高斯混合模型的参数，计算出给定身高 `mean1` 和 `mean2` 属于每个分量的概率 `p1` 和 `p2`，以及每个分量的权重、均值和标准差的值。

## 三．运行结果

直方图如下：



以下分别是迭代次数不同时得到的权重、均值和标准差的值：

25 次迭代：

```
lizhiyuan@naiyuandeMacBook-Pro W5大作业 % python -u "/Users/lizhiyuan/学习/深度学习与NLP/W5大作业/homework.py"
weights: [0.28323145 0.71676855]
means: [164.4731823 176.67910209]
stds: [3.17754299 4.67652756]
P(mean1, std1): [0.95780055 0.04219945]
P(mean2, std2): [8.15253686e-04 9.99184746e-01]
```

50 次迭代：

```
lizhiyuan@naiyuandeMacBook-Pro W5大作业 % python -u "/Users/lizhiyuan/学习/深度学习与NLP/W5大作业/em1.py"
weights: [0.24507739 0.75492261]
means: [163.87768383 176.25553263]
stds: [2.80275862 4.95160176]
P(mean1, std1): [0.92456731 0.07543269]
P(mean2, std2): [4.97719357e-05 9.99950228e-01]
```

75 次迭代：

```
lizhiyuan@naiyuandeMacBook-Pro W5大作业 % python -u "/Users/lizhiyuan/学习/深度学习与NLP/W5大作业/homework.py"
weights: [0.24216795 0.75783205]
means: [163.83631537 176.22123153]
stds: [2.77835067 4.97565486]
P(mean1, std1): [0.92104326 0.07895674]
P(mean2, std2): [3.94401293e-05 9.99960560e-01]
```

100 次迭代：

```
lizhiyuan@naiyuandeMacBook-Pro W5大作业 % python -u "/Users/lizhiyuan/学习/深度学习与NLP/W5大作业/em1.py"
weights: [0.24196728 0.75803272]
means: [163.83348532 176.21885631]
stds: [2.77668799 4.97732858]
P(mean1, std1): [0.92079408 0.07920592]
P(mean2, std2): [3.88092184e-05 9.99961191e-01]
```

500 次迭代：

```
lizhiyuan@naiyuandeMacBook-Pro W5大作业 % python -u "/Users/lizhiyuan/学习/深度学习与NLP/W5大作业/em1.py"
weights: [0.24195248 0.75804752]
means: [163.83327673 176.21868109]
stds: [2.77656547 4.97745209]
P(mean1, std1): [0.92077568 0.07922432]
P(mean2, std2): [3.87630753e-05 9.99961237e-01]
```

1000 次迭代：

```
lizhiyuan@naiyuandeMacBook-Pro W5大作业 % python -u "/Users/lizhiyuan/学习/深度学习与NLP/W5大作业/em1.py"
weights: [0.24195248 0.75804752]
means: [163.83327673 176.21868109]
stds: [2.77656547 4.97745209]
P(mean1, std1): [0.92077568 0.07922432]
P(mean2, std2): [3.87630753e-05 9.99961237e-01]
```

可以很明显看出，程序在 100 次迭代后得到的结果趋于稳定，得到的均值、标准差变化很小。

从直方图可以看出，模型拟合效果较好，能够很好的反映身高数据的分布情况。

通过高斯分量的均值和标准差，可以了解模型对数据的拟合情况，从输出结果中可以看出，这些参数能够很好的反映身高数据的分布情况。

## 四．总结

通过本次作业，使用了 EM 算法对高斯混合模型进行训练，学习了如何定义高斯混合模型，如何定义一个 EM 算法函数，并且通过调参得到不同的结果并进行判断。