北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理

学 院 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	张一凡
指 导 教 师	秦曾昌

2023 年 4 月

深度学习与自然语言处理第二次大作业-EM 算法估计身高

一、实验内容

参考给出的链接中代码来生成 2000 分身高数据，并通过 EM 算法来估计高斯混合模型的参数，使用这些参数来进行预测，最后需要对模型进行评估，并解释模型的性能。

二、实验数据

Height 文件

三、实验算法设计

EM 算法是一种基于最大似然估计的参数估计算法，适用于高斯混合模型的参数估计。在本实验中，我们有 2000 份身高数据，可以使用 EM 算法来估计高斯混合模型的参数，以进行预测和评估。

首先，需要对高斯混合模型进行建模，假设身高数据是由两个高斯分布混合而成的，即：

$$p(x) = \pi_1 \cdot \mathcal{N}(x|\ \mu_1, \sigma_1^2) + \pi_2 \cdot \mathcal{N}(x|\ \mu_2, \sigma_2^2)$$

其中，π1 和π2 是两个高斯分布的混合系数，满足π1+π2=1，$\mathcal{N}(x|\ \mu_1, \sigma_1^2)$ 和 $\mathcal{N}(x|\ \mu_2, \sigma_2^2)$ 表示第 i 个高斯分布的概率密度函数，mu 和 sigama 是对应的均值和方差。

然后，可以使用 EM 算法来估计高斯混合模型的参数，具体步骤如下：

1. 初始化高斯混合模型的参数，包括混合系数π1 和π2，以及两个高斯分布的均值和方差 mu1，sigma1，mu2，sigma2。
2. E 步：根据当前模型的参数，计算每个样本属于第一个高斯分布和第二个高斯分布的概率，即：

$$\gamma_{ij} = \frac{\pi_j \cdot N(x_i|\ \mu_j, \sigma_j^2)}{\sum_{k=1}^{2} \pi_k \cdot N(x_i|\ \mu_k, \sigma_k^2)}$$

   其中，gammaij 表示第 i 个样本属于第 j 个高斯分布的概率。
3. M 步：根据当前模型的参数和 E 步得到的结果，重新估计高斯混合模型的参数，即：

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ij}$$

$$\mu_j = \frac{\sum_{i=1}^{N} \gamma_{ij} x_i}{\sum_{i=1}^{N} \gamma_{ij}}$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{N} \gamma_{ij} (x_i - \mu_j)^2}{\sum_{i=1}^{N} \gamma_{ij}}$$

   其中，N 表示样本数量。
4. 重复 E 步和 M 步，直到模型收敛或达到最大迭代次数。
5. 使用估计得到的模型参数进行预测和评估，可以使用均方误差、对数似然或者其他指标来评估模型的性能。

对于预测，可以根据估计得到的模型参数，计算每个样本属于第一个高斯分布和第二个

高斯分布的概率，选择概率较大的一个作为样本所属的类别，即：

$$y_i = \begin{cases} 1, & \text{if} p(x_i|\ z=1) > p(x_i|\ z=2) \\ 2, & \text{if} p(x_i|\ z=1) \le p(x_i|\ z=2) \end{cases}$$

其中，yi 表示第 i 个样本的预测类别，z 表示高斯分布的类别，即 z=1 表示第一个高斯分布，z=2 表示第二个高斯分布。

对于评估，可以使用均方误差和对数似然等指标来评估模型的性能，均方误差表示预测值与真实值之间的平均误差，可以用来衡量模型的精度，计算公式为：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2$$

其中，yi 表示真实值，y 表示预测值，N 表示样本数量。

对数似然是一种衡量模型拟合数据的好坏程度的指标，对于高斯混合模型，可以使用对数似然来评估模型的拟合效果，计算公式为：

$$\log L(\theta) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{2} \pi_j \cdot N(x_i|\ \mu_j, \sigma_j^2)\right)$$

其中，θ表示模型的参数，N 表示样本数量，$\mathcal{N}(x|\ \mu_1, \sigma_1^2)$ 表示高斯分布的概率密度函数。

通过计算均方误差和对数似然等指标，可以评估高斯混合模型的性能，从而确定模型是否合适并进行进一步的优化和改进。

四、实验结果及分析

```
pi1:  0.1686441892456131
pi2:  0.8313558107543869
mu1:  1.6309621732525108
mu2:  1.7509282325243092
sigma1:  0.02413109750011217
sigma2:  0.057327127261475125
MSE:  0.07737244883060179
Log-likelihood:  2579.281771312926
```

结果如图所示