EM 算法预测男生女生身高分布

郭润堂
guoruntang@buaa.edu.cn

摘要

本次作业的目标是基于给定的 1500 名男生和 500 名女生的身高数据，使用 EM（Expectation-Maximization）算法进行分析，预测男性和女性各自身高的均值和标准差。其中，男性身高的均值为 176cm，标准差为 5cm，女性身高的均值为 164cm，标准差为 3cm。考虑到数据集中存在男女身高数据混合的情况，我们采用混合高斯模型对数据进行建模，从而更好地描述身高数据的特征。通过迭代求解，在本次作业当中，经过 500 次迭代，得到男生身高的均值 175.82cm、标准差 5.06cm，女生身高均值 163.86cm、标准差 2.82cm，男生占 2000 人总人数的 76%，女生占总人数的 24%

Introduction

1. 混合高斯模型

混合高斯模型是一种常用的概率密度估计模型，也被称为高斯混合模型。它的基本思想是将数据看作由多个高斯分布组合而成的混合物生成。每个高斯分布对应着一个类别或簇，而数据集中同一类别的样本数据服从相同的高斯分布，不同类别之间的样本则服从不同均值和方差的高斯分布。

具体来说，混合高斯模型假设数据集由 k 个高斯分布组成，每个高斯分布都由三个参数决定：均值μi、协方差矩阵 Σi 和权重系数 wi ，其中 i∈{1,2,⋯ ,k}。给定一个观测值 xi ，它属于第 j 个高斯分布的概率可以通过下面的公式计算得到：

$$P(j|x_i) = \frac{w_j N(x_i|\mu_j, \Sigma_j)}{\sum_{l=1}^{k} w_l N(x_i|\mu_l, \Sigma_l)}$$

其中，N(xi｜μj,Σj) 表示第 j 个高斯分布的概率密度函数，即：

$$N(x_i|\mu_j,\Sigma_j)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_j|^{\frac{1}{2}}}exp\left(-\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)\right)$$

其中，d 表示数据集的维度，｜Σj｜ 表示第 j 个高斯分布的协方差矩阵的行列式。这个公式可以理解为给出一个观测值 xi，计算它属于每个高斯分布的概率。

混合高斯模型的主要优点在于它可以应用于各种数据集，而且对随机噪声、离群值等具有一定的鲁棒性。在实际应用中，混合高斯模型主要用于聚类分析、异常检测、图像分割、建模等领域的数据分析和处理任务。

在本次作业模型当中对于每个身高样本，我们将其看做由混合模型中的多个高斯分布组成的混合物生成，每个高斯分布对应一个类别，即男性或女性。在此基础上，我们运用 EM 算法进行参数估计，包括估计男性和女性各自的高斯分布的均值和标准差，以及每个高斯分布的权重，通过迭代求解。

2. EM 算法

EM 算法 (Expectation-Maximization Algorithm) 是一种迭代优化算法，它通常用于最大化含有隐变量（latent variable）的概率模型的对数似然函数。这些概率模型通常是高斯混合模型、隐马尔可夫模型等，而隐变量则是指在训练过程中无法直接观测到的变量。

EM 算法的基本思想是通过不断计算数据的期望值和最大化似然函数来求解参数。具体而言，EM 算法分为两个步骤：E 步和 M 步。在 E 步中，通过已知参数计算隐变量的后验概率分布；在 M 步中，通过已知隐变量的后验概率分布最大化完全数据的对数似然函数。这两步迭代交替进行，直到收敛为止。

下面以高斯混合模型为例，详细介绍 EM 算法的步骤：

初始化参数： 随机选择混合高斯模型的均值、协方差矩阵和权重系数作为初始参数；

E 步： 计算当前参数下每个观测样本属于每个高斯分布的概率，即计算隐变量的后验概率分布。这个步骤可以通过贝叶斯公式得到：

$$P(z=k|x_i,\theta^{(t)})=\frac{w_k^{(t)}\mathcal{N}(x_i|\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{j=1}^{K}w_j^{(t)}\mathcal{N}(x_i|\mu_j^{(t)},\Sigma_j^{(t)})}$$

其中，z 表示隐变量，即数据点 xi 属于高斯分布的哪一类；wk(t)、μk(t)和 $ \Sigma_{k}^{(t)}$表示第 k 个高斯分布在第 t 次迭代时对应的权重系数、均值和协方差矩阵；N(xi｜μk(t),Σk(t)) 表示多维高斯分布的概率密度函数。在 E 步中，计算每个样本观测值属于每个高斯分布的概率，为后面的 M 步提供数据基础；3. M 步：在已知隐变量的后验概率分布的基础上，最大化完全数据的对数似然函数，寻找当前参数的最优解，即：

$$\max_{\theta} \sum_{i=1}^{N} \sum_{k=1}^{K} P(z=k|x_i,\theta^{(t)}) \ln P(x_i, z=k|\theta)$$

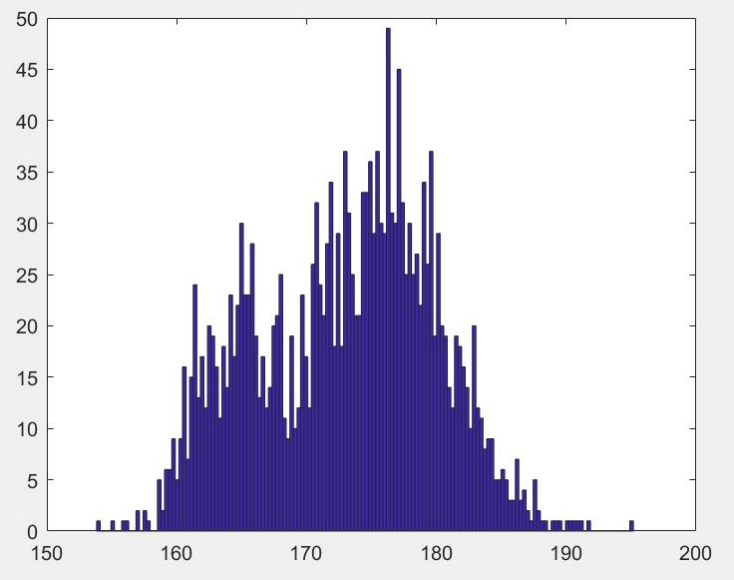
其中，θ表示模型参数，包括每个高斯分布的权重系数、均值和协方差矩阵。对于混合高斯模型，可以使用最大化对数似然函数的方式来求解模型参数。这个过程一般需要使用优化算法，比如牛顿法、梯度下降等；4.更新参数：将求得的参数作为新的参数，返回第 2 步，直到达到收敛条件。

需要注意的是，EM 算法有可能收敛于局部最优解，因此需要使用多组不同的初始参数来寻找全局最优解。另外，在实际应用中，EM 算法还存在一些问题，比如计算量大、初始化困难等。因此，在具体问题中需要结合实际情况进行一些改进和优化。

本次作业当中将混合高斯模型中每个高斯分布的均值、标准差和权重系数作为参数，计算出当前参数下每个观测样本属于每个高斯分布的概率，称为 E 步；然后在最大化这些概率的同时调整均值、标准差和权重系数等参数，称为 M 步。通过迭代交替进行 E 步和 M 步操作，最终可以得到模型收敛于局部最优解的模型参数。

Methodology

首先使用提供的 python 代码，生成均值为 176cm、标准差为 5cm 的男生身高数据 1500 个，生成均值为 164cm、标准差为 3cm 的女生身高数据 500 个，数据保存于 student_height.csv 文件中。身高分布图如下：



在 matlab 当中读取 student_height.csv 文件，获得 2000 个身高。并设定需要估计的男生身高、标准差、比例，女生的身高、标准差、比例 6 个参数的初值。

```
mu1_first=180;sigmal_first=8;w1_first=0.60;
mu2_first=160;sigma2_first=10;w2_first=0.4;
```

```
%Step 1.首先根据经验来分别对男女生的均值、方差和权值进行初始化
 mu1_first=180;sigmal_first=8;w1_first=0.60;
 mu2_first=160;sigma2_first=10;w2_first=0.4;

 iteration=500;%设置迭代次数
 outcome=zeros(iteration,6);%定义一个数组来存储每次的迭代结果
 outcome(1,1)=mu1_first;outcome(1,4)=mu2_first;
 outcome(1,2)=sigmal_first;outcome(1,5)=sigma2_first;
 outcome(1,3)=w1_first;outcome(1,6)=w2_first;%将第一列存储初始值
```

EM 算法部分：

E 步骤计算每个点属于哪一个高斯分布的概率

```
for i=1:N
    p1=w1_first*pdf('norm',h(i),mu1_first,sigmal_first);
    p2=w2_first*pdf('norm',h(i),mu2_first,sigma2_first);
    %p1,p2权重*男女生的后验概率
   R1i(i)=p1/(p1+p2);
   R2i(i)=p2/(p1+p2);
end
```

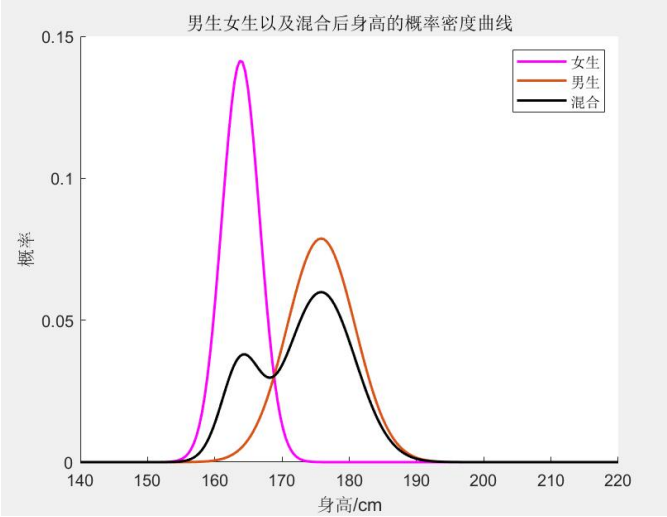
M 步骤根据每个点属于每一高斯分布的概率，重新计算男生、女生身高的均值、标准差和权重

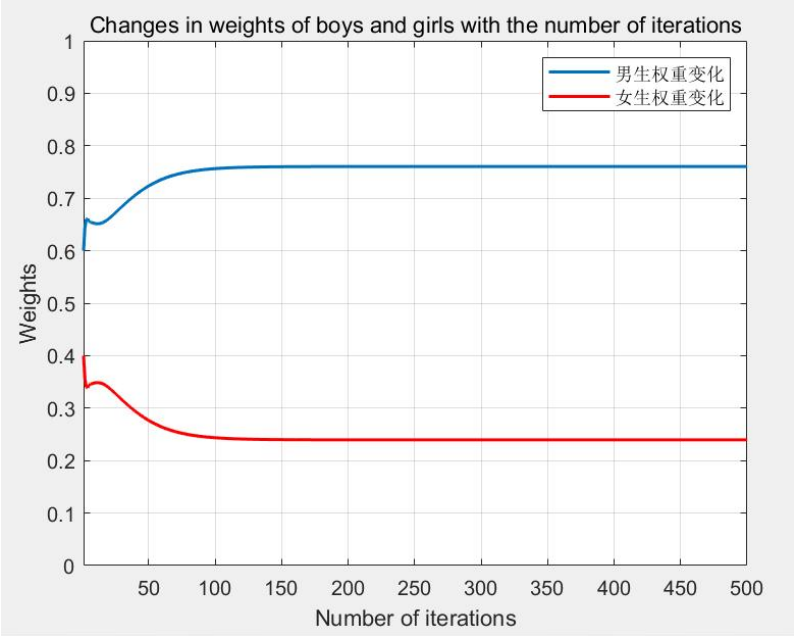
```
%更新男、女生身高分布的期望mu
s1=0;
s2=0;
for i=1:N
    s1=s1+R1i(i)*h(i);
    s2=s2+R2i(i)*h(i);
end
s11=sum(R1i);
s22=sum(R2i);
mu1_last=s1/s11;
mu2_last=s2/s22;
```

```
%更新男、女生身高分布的标准差sigma(一维)
t1=0;
t2=0;
for i=1:N
    t1=t1+R1i(i)*(h(i)-mu1_last)^2;
    t2=t2+R2i(i)*(h(i)-mu2_last)^2;
end
t11=sum(R1i);
t22=sum(R2i);
sigmal_last=sqrt(t1/t11);
sigma2_last=sqrt(t2/t22);
```

```
%更新权值
w1_last=s11/N;
w2_last=s22/N;
```

不断重复以上步骤，迭代 500 次，得到得到男生身高的均值 175.82cm、标准差 5.06cm，女生身高均值 163.86cm、标准差 2.82cm，男生占 2000 人总人数的 76%，女生占总人数的 24%。与真值男性身高的均值为 176cm，标准差为 5cm，女性身高的均值为 164cm，标准差为 3cm，男女比例 3：1 十分接近。





Conclusions

通过直方图可以清晰的看到，男女身高数据的混合近似于两个高斯分布的混合，适合使用混合高斯模型进行求解。本次作业通过设定两个不同的高斯分布初值与权重，建立混合高斯模型，使用 EM 最大化期望方法进行迭代，不断依次对数据点属于某一分布的概率，与隐变量两个高斯分布的均值、标准差、权重进行估计。在迭代 500 轮时获得与真实值十分相近的结果。