深度学习与自然语言处理第二次作业

姓名：郑梓岳　　　学号：19231191

一、 作业描述

请使用链接中的代码身高数据，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。

作业提交要求：1）EM 算法代码文件 2）结果报告文件（可以是 Jupyter Notebook、PDF、Word 等格式。

二、 EM 算法原理

若总体 X 为离散型，其概率分布列为

$$P\{X=x\}=p(x;\theta)$$

其中θ为未知参数，设$(X_1,X_2,\ldots,X_n)$是取自总体的样本容量为 n 的样本，则$(X_1,X_2,\ldots,X_n)$的联合分布律为$\prod_{i=1}^{n}p(x_i;\theta)$。 又设$(X_1,X_2,\ldots,X_n)$的一组观测值为$(x_1,x_2,\ldots,x_n)$， 易知样本$X_1,X_2,\ldots,X_n$取到$x_1,x_2,\ldots,x_n$的概率为：

$$L(\theta)=L(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n}p(x_i;\theta)$$

这一概率随θ的取值而变化，它是θ的函数，称$L(\theta)$为样本的似然函数。

对于 n 个样本观察数据$x=(x_1,x_2,\ldots,x_n)$，找出样本的模型参数θ，极大化模型分布的对数似然函数如下：

$$\boldsymbol{\hat{\theta}=argmax\sum_{i=1}^{n}logp(x_i;\theta)}$$

如果我们得到的观察数据有未观察到的隐含数据$z=(z_1,z_2,\ldots,z_n)$，即上文中每个样本属于哪个分布是未知的，此时我们极大化模型分布的对数似然函数如下：

$$\boldsymbol{\hat{\theta}=argmax\sum_{i=1}^{n}logp(x_i;\theta)=argmax\sum_{i=1}^{n}log\sum_{z_i}p(x_i,z_i;\theta)}$$

上面这个式子是根据x_i的边缘概率计算得来，没有办法直接求出θ。因此需要一些特殊的技巧，使用 Jensen 不等式对这个式子进行缩放如下：

$$\boldsymbol{\sum_{i=1}^{n}log\sum_{z_i}p(x_i,z_i;\theta)=\sum_{i=1}^{n}log\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}}\quad\textbf{(1)}$$

$$\boldsymbol{\geq\sum_{i=1}^{n}\sum_{z_i}Q_i(z_i)log\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}}\quad\textbf{(2)}$$

(1)式是引入了一个未知的新的分布$Q_i(z_i)$，分子分母同时乘以它得到的

(2)式是由(1)式根据 Jensen 不等式得到的。由于$\sum_{z_i}Q_i(z_i)log\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}$为$\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}$的期望，

且$log(x)$为凹函数，根据 Jensen 不等式可由(1)式得到(2)式。

上述过程可以看作是对 $logl(\theta)$求了下界。对于$Q_i(z_i)$我们如何选择呢？假设θ已经给定，那么$logl(\theta)$的值取决于$Q_i(z_i)$和$p(x_i,z_i)$。我们可以通过调整这两个概率使(2)式下界不断上升，来逼近$logl(\theta)$的真实值。那么如何算是调整好呢？当不等式变成等式时，说明我们调整后的概率能够等价于$logl(\theta)$了。按照这个思路，我们要找到等式成立的条件。

如果要满足 Jensen 不等式的等号，则有：

$$\frac{\boldsymbol{p(x_i,z_i;\theta)}}{\boldsymbol{Q_i(z_i)}}=\boldsymbol{c}, \qquad \boldsymbol{c}\text{为常数}$$

由于$\mathrm{Q_i(z_i)}$是一个分布，所以满足：$\sum_{z_i}Q_i(z_i)=1$， 则$\sum_{z_i}p(x_i,z_i;\theta)=c$

由上面两个式子，我们可以得到：

$$\boldsymbol{Q_i(z_i)=\frac{p(x_i,z_i;\theta)}{\sum_z p(x_i,z_i;\theta)}=\frac{p(x_i,z_i;\theta)}{p(x_i;\theta)}=p(z_i|x_i;\theta)}$$

至此，我们推出了在固定其他参数 θ 后，$Q_i(z_i)$的计算公式就是后验概率，解决了$Q_i(\mathrm{z_i})$如何选择的问题。

如果$Q_i(z_i)=p(z_i|x_i;\theta)$则(2)式是我们包含隐藏数据的对数似然函数的一个下界。如果我们能最大化(2)式这个下界，则也是在极大化我们的对数似然函数。即我们需要最大化下式：

$$\mathbf{argmax}\sum_{i=1}^{n}\sum_{z_i}\boldsymbol{Q_i(z_i)log\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}}$$

当完成了$\mathrm{Q_i(z_i)}$的选择，那么我们就完成了 M 步。

三、 使用 EM 算法解决该作业

3.1 创建数据

作业要求自行生成数据，使用给出的代码可以生成 2000 个学生的身高数据，其中男生有 1500 人，女生有 500 人，男生身高服从均值为 176，方差为 5 的高斯分布，女生身高服从均值为 164，方差为 3 的高斯分布

3.2 E 步

首先进行 E 步，假设x为某一学生的身高，$w1$和$w2$表示此时男生和女生各自所占的比例，μ_1和μ_2分别表示男生和女生身高的均值，σ_1和σ_2分别表示男生和女生身高的标准差，分别用$p1$和$p2$表示该学生属于男生或女生的概率。

$$p1(x)=\frac{w1*\mathcal{N}(x|\mu_1,\sigma_1^2)}{w1*\mathcal{N}(x|\mu_1,\sigma_1^2)+w2*\mathcal{N}(x|\mu_2,\sigma_2^2)}$$

$$p2(x)=\frac{w2*\mathcal{N}(x|\mu_2,\sigma_2^2)}{w1*\mathcal{N}(x|\mu_1,\sigma_1^2)+w2*\mathcal{N}(x|\mu_2,\sigma_2^2)}$$

其中$\mathcal{N}(x|\mu,\sigma^2)$表示均值为$\mu$，标准差为$\sigma$的高斯分布在$x$处的概率密度。

3.2 M 步

在 M 步中需要对参数进行一次极大似然估计，实现参数迭代。

在 E 步中将所有身高数据分为男生和女生两类后，计算得出新的参数，包括$w1$，$w2$，μ_1，μ_2，σ_1和σ_2

3.3 设定初始参数，开始 EM 算法迭代

设初始参数$w1=0.5$，$w2=0.5$，$\mu_1=180$，$\mu_2=165$，σ_1和σ_2均直接采用 2000 个数据的标准差，让 EM 算法开始迭代，迭代 100 次。

四、 结果及分析

4.1 实验结果展示

部分实验结果如下图所示：

```
第1次迭代的结果为，男生有1061人，男生身高的均值为177.23690258313908，标准差为5.010290531842007，女生有939人，女生身高的均值为168.1249023732507З，标准差为5.573300858115712
第2次迭代的结果为，男生有1069人，男生身高的均值为177.30049439160027，标准差为4.852178354383337，女生有931人，女生身高的均值为167.97255326019527，标准差为5.55196910331854
第3次迭代的结果为，男生有1075人，男生身高的均值为177.36462277237928，标准差为4.760531274410303，女生有925人，女生身高的均值为167.841273679569４，标准差为5.472744949965371
第4次迭代的结果为，男生有1081人，男生身高的均值为177.41514064731237，标准差为4.696519076168387，女生有919人，女生身高的均值为167.7223850697658，标准差为5.384541915746843
第5次迭代的结果为，男生有1087人，男生身高的均值为177.4493591788076，标准差为4.648670149132508，女生有913人，女生身高的均值为167.61239121120016，标准差为5.301023186420268
第6次迭代的结果为，男生有1095人，男生身高的均值为177.46842919952286，标准差为4.612955493153724，女生有905人，女生身高的均值为167.50934438230647，标准差为5.225144532204649５
第7次迭代的结果为，男生有1103人，男生身高的均值为177.47465281514897，标准差为4.587103266018385，女生有897人，女生身高的均值为167.41170441652218，标准差为5.156632962441437
第8次迭代的结果为，男生有1111人，男生身高的均值为177.47048594764064，标准差为4.569249587198577，女生有889人，女生身高的均值为167.31809332979975，标准差为5.094281751923499
第9次迭代的结果为，男生有1121人，男生身高的均值为177.45815373735854，标准差为4.557746845272738，女生有879人，女生身高的均值为167.22729833879526，标准差为5.036696371928309
第10次迭代的结果为，男生有1130人，男生身高的均值为177.43951531577213，标准差为4.551189582491337，女生有870人，女生身高的均值为167.13830088986163，标准差为4.982567756330918
第11次迭代的结果为，男生有1140人，男生身高的均值为177.41604502589269，标准差为4.5484353620077504，女生有860人，女生身高的均值为167.05028143814766，标准差为4.930773828367647
第12次迭代的结果为，男生有1150人，男生身高的均值为177.38886886527612，标准差为4.548592932854846，女生有850人，女生身高的均值为166.96260235317098，标准差为4.880403221239476
第13次迭代的结果为，男生有1161人，男生身高的均值为177.35882243660254，标准差为4.550989744204692，女生有839人，女生身高的均值为166.87478032534105，标准差为4.830741620841321
第14次迭代的结果为，男生有1171人，男生身高的均值为177.3265118233417，标准差为4.555132287192329，女生有829人，女生身高的均值为166.78645698117944，标准差为4.781242751704578
第15次迭代的结果为，男生有1182人，男生身高的均值为177.29236837914405，标准差为4.560667605556377，女生有818人，女生身高的均值为166.6973724424913，标准差为4.731496172068328
第16次迭代的结果为，男生有1193人，男生身高的均值为177.25669416170427，标准差为4.567349869428361，女生有807人，女生身高的均值为166.60734372225525，标准差为4.681197889125719
第17次迭代的结果为，男生有1204人，男生身高的均值为177.21969780072877，标准差为4.575013155261198，女生有796人，女生身高的均值为166.51624824296277，标准差为4.630126154628037
第18次迭代的结果为，男生有1215人，男生身高的均值为177.18152196328728，标准差为4.583550154853333，女生有785人，女生身高的均值为166.42401200725902，标准差为4.578122857999415
第19次迭代的结果为，男生有1226人，男生身高的均值为177.14226401315207，标准差为4.592895941710472，女生有774人，女生身高的均值为166.33060168241184，标准差为4.525080040018501
第20次迭代的结果为，男生有1237人，男生身高的均值为177.10199143150916，标准差为4.6030157743641515，女生有763人，女生身高的均值为166.23601983316098，标准差为4.4709307328206975
第21次迭代的结果为，男生有1249人，男生身高的均值为177.06075334699307，标准差为4.613895978396003，女生有751人，女生身高的均值为166.14030260771025，标准差为4.415643295042303
第22次迭代的结果为，男生有1260人，男生身高的均值为177.01858925288613，标准差为4.625537091585566，女生有740人，女生身高的均值为166.04351927260345，标准差为4.359218485678807
第23次迭代的结果为，男生有1272人，男生身高的均值为176.97553573138123，标准差为4.637948614963618，女生有728人，女生身高的均值为165.9457730674116６，标准差为4.301688618199754
第24次迭代的结果为，男生有1283人，男生身高的均值为176.9316317814847，标准差为4.651144858695936，女生有717人，女生身高的均值为165.84720289489687，标准差为4.243118215234398
第25次迭代的结果为，男生有1295人，男生身高的均值为176.88692316163255，标准差为4.665141497039176，女生有705人，女生身高的均值为165.74798537404175，标准差为4.183605626530473
```

```
第93次迭代的结果为，男生有1564人，男生身高的均值为175.58095549018748，标准差为5.324396312841466，女生有436人，女生身高的均值为163.5551742639307，标准差为2.8469421447034375
第94次迭代的结果为，男生有1564人，男生身高的均值为175.58024617501326，标准差为5.324861052639762，女生有436人，女生身高的均值为163.55436242647386，标准差为2.846493222083454
第95次迭代的结果为，男生有1564人，男生身高的均值为175.5795934046063６，标准差为5.325288803481044，女生有436人，女生身高的均值为163.55361559142344，标准差为2.846080257889123
第96次迭代的结果为，男生有1564人，男生身高的均值为175.57899268041638，标准差为5.325682499028835，女生有436人，女生身高的均值为163.55292854361534，标准差为2.8457003648787316
第97次迭代的结果为，男生有1564人，男生身高的均值为175.578439860366，标准差为5.326044841882234，女生有436人，女生身高的均值为163.5522964878744３，标准差为2.845350888708067
第98次迭代的结果为，男生有1564人，男生身高的均值为175.57793113082803，标准差为5.326378321521593，女生有436人，女生身高的均值为163.55171501495963，标准差为2.8450293889807328
第99次迭代的结果为，男生有1565人，男生身高的均值为175.57746298076975，标准差为5.326685230901051，女生有435人，女生身高的均值为163.55118007030416，标准差为2.844733621865075
第100次迭代的结果为，男生有1565人，男生身高的均值为175.57703217790205，标准差为5.326967681783208，女生有435人，女生身高的均值为163.5506879253159，标准差为2.84446152414451９
第101次迭代的结果为，男生有1565人，男生身高的均值为175.57663574668328，标准差为5.327227618905848，女生有435人，女生身高的均值为163.55023515102567，标准差为2.8442111985797847
第102次迭代的结果为，男生有1565人，男生身高的均值为175.576270948037６，标准差为5.3274668330648725，女生有435人，女生身高的均值为163.54981859388613，标准差为2.843980900472619
第103次迭代的结果为，男生有1565人，男生身高的均值为175.57593526065742，标准差为5.327686973192742，女生有435人，女生身高的均值为163.54943535354414，标准差为2.8437690253299674
第104次迭代的结果为，男生有1565人，男生身高的均值为175.57562636376974，标准差为5.32788955750609，女生有435人，女生身高的均值为163.54908276242253，标准差为2.843574097536603
第105次迭代的结果为，男生有1565人，男生身高的均值为175.5753421212534，标准差为5.328075983791566，女生有435人，女生身高的均值为163.54875836696132，标准差为2.843394759952148
```

4.2 实验结果分析

由上图可见，在经过 100 次迭代后，EM 算法基本收敛，将收敛后的数据参数与数据生成时所用的参数对比，虽然并不是完全相同，但是在误差允许范围内。因此，EM 算法完成了参数的估计。

在经过多次设置不同初始参数并运行后发现，EM 算法对初始值十分敏感，结果随不同的初始值而波动较大。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。

EM 算法可以保证收敛到一个稳定点，但是却不能保证收敛到全局的极大值点，因此它是局部最优的算法，