

成 绩

****

Expectation Maximum算法实现

|  |  |
| --- | --- |
| 院（系）名称 | 高等理工学院 |
| 专业名称 | 自动化 |
| 学生学号 | 18376251 |
| 学生姓名 | 乐祥立 |
| 指导教师 | 秦曾昌 |

2021年5月

Expectation Maximum算法实现

一、任务目标

首先使用Excel内置的NORM.INV()或者python中np.random.normal()功能生成男生和女生的身高数据，男生身高的正态分布为，即男生身高的均值为175，标准差为15；女生身高的正态分布为，即女生身高的均值为165，标准差为10。按照上述分布随机生成男生身高数据200个，女生身高数据100个，得到男女生的数量比为2：1。之后将数据进行混合，通过混合后的300个数据和EM算法估计男生和女生在数据集中的比例以及二者各自所服从的分布的形式。

二、EM算法原理

1. 高斯混合模型（GMM）

设有随机变量**X**，则高斯混合模型可以表示为：

其中称为混合模型中的第k个分量，分类数K满足：

可以认为，就是每个正态分布的权重。

2. GMM参数估计

引入一个K维的新随机变量，，其取值为0或者1。表示第k类被选中的概率为，表示第k类没有被选中的概率。

假设之间是独立同分布的，则的联合概率密度分布为：

因此对于服从正态分布的数据可以有：

根据贝叶斯定理，是先验概率，是似然概率，可以求出后验概率为：

EM算法需要第一步先求出参数的粗略值，之后第二步使用第一步的值最大化似然函数，需要先求似然函数。改写可有：

上式中，因此对这些独立同分布的样本连乘后取对数得到对数似然函数，再求导令导数为0可以得到极大似然函数为：

可以得到：

令，则有

同理可以得到：

若认为就是每个正态分布的权重，则可以很自然的得到：

3.EM算法流程：

a. 定义分类数量K，设置对应数量的，和的初始值，计算对数似然函数。

b. E step：

根据当前的，和值计算后验概率

c. M step：

根据E step中的计算新的，和

d. 计算对数似然函数

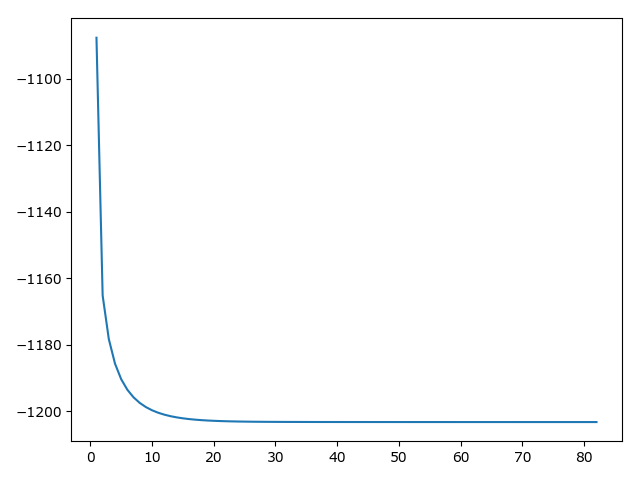
e. 检查参数是否收敛，不收敛则返回第二步；若收敛则结束程序返回估计值。

三、算法实现效果

1. 使用Excel生成的固定数据集

首先用NORM.INV函数生成对应分布和数量的男生以及女生的数据，将数据复制成为文本之后保存，得到固定的数据集，读入该数据，设定初值为男生比例0.6，均值180，标准差18，女生比例0.4，均值160，方差7，可以得到结果如下:

下图为极大似然函数随训练次数变化的曲线图，可以看到最终基本收敛。设定的程序结束条件为，前后两次极大似然函数值相差不超过0.0001或者训练次数超过2000次。



最终计算82次之后收敛，得到的参数为：

男生比例：0.6765，男生平均身高：174.09，男生身高标准差：15.83

女生比例：0.3235，女生平均身高：164.52，女生身高标准差：7.30

（具体的数据和训练过程见附录）

可以看到，最终得到的结果除了女生的标准差有较大误差以外，其余的结果较为准确。

2. 使用python随机生成数据集

和使用固定数据集不同之初在于，使用固定数据集进行训练时如果初始值改变的幅度不大或者在合理的范围之内，估计的参数值总是和上述的参数值相差不大。如果使用相同的初始值则一定对得到相同的结果。但是如果使用python随机生成的数据集进行训练，每次的结果并不会收敛到同一组结果附近，即使使用相同的初始值也不会得到相同的结果，这和数据的采样有一定的关系，采样不同，采样出的分布和原本分本的偏差也会有所不同。下面选择一组效果较好的结果进行展示：

设定初始值为男生比例0.5，男生均值180，男生标准差20，女生比例同样为0.5，均值160，女生标准差7。可以得到结果如下：

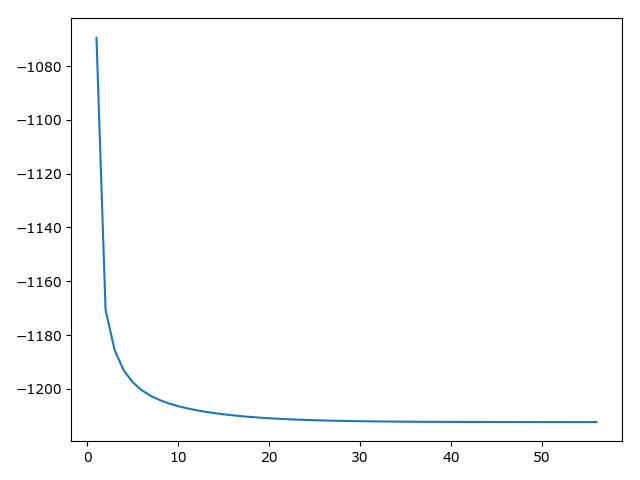
最终计算56次之后收敛，得到的参数为：

男生比例：0.6682，男生平均身高：174.35，男生身高标准差：15.36

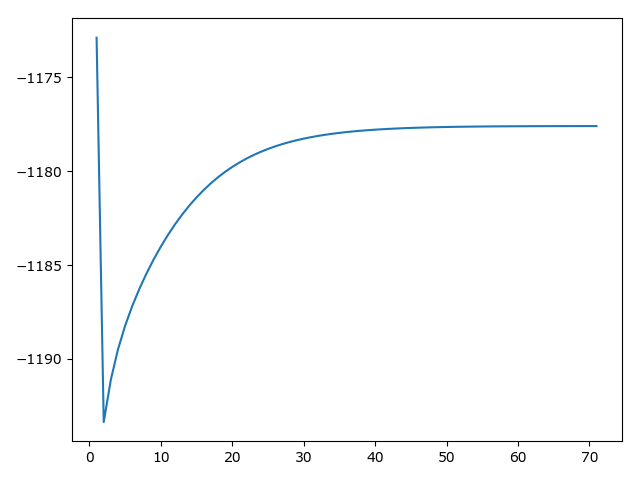
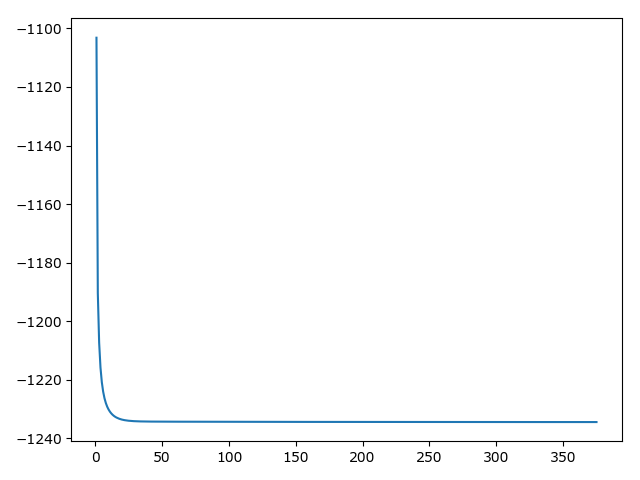
女生比例：0.3318，女生平均身高：165.16，女生身高标准差：9.25

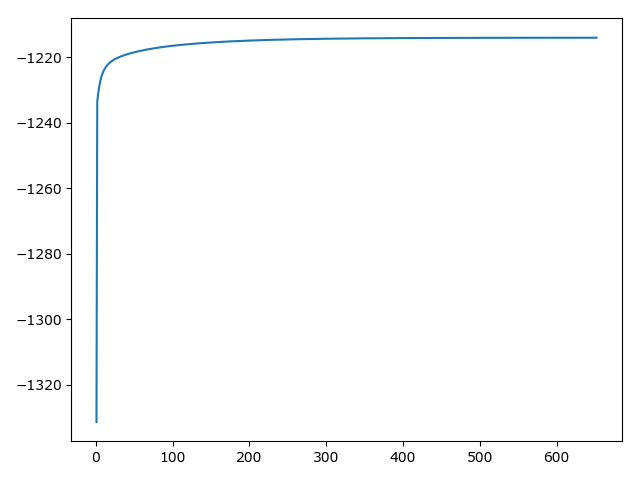
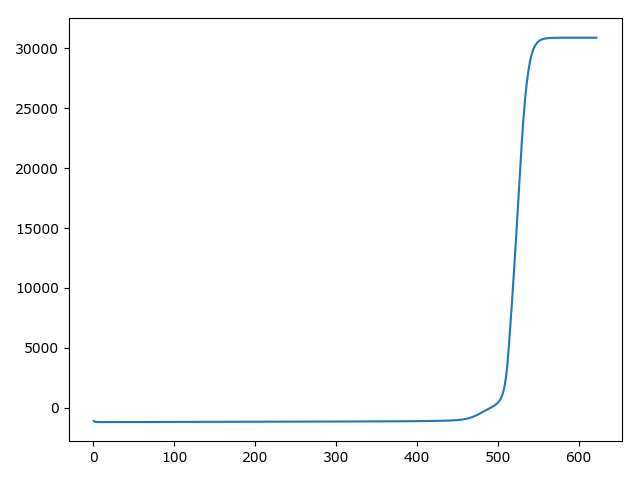
（具体的数据和训练过程见附录）

下图为极大似然函数随训练次数变化的曲线图，可以看到最终基本收敛。



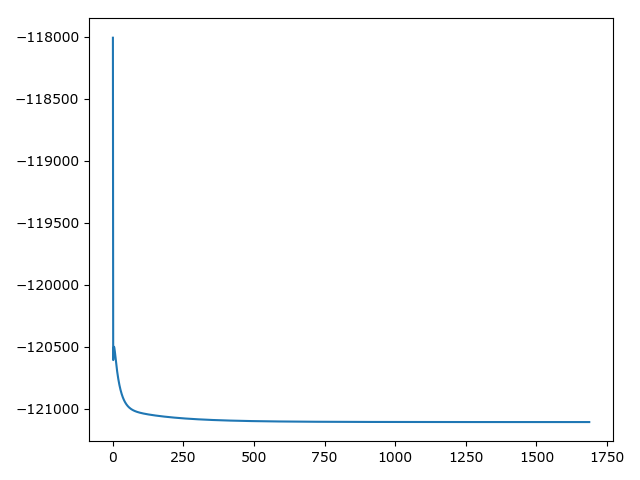
3. 收敛结果和初值的关系





在使用随机采样得到的数据时，数据的收敛情况收到初值和采样的双重影响，但是若较为明显的改动初值，比如使得男生的平均身高小于女生，则会使得算法得到的结果收敛到一些不同的局部最优解。同时也会改变极大似然函数的变化形式，如上所示。

经过尝试可以发现解决的办法是增大数据集的量，即增大采样率。当有足够多的数据时，数据会有更大的概率收敛到真实的情况的附近。



观察训练结果可以发现，各个参数很少的迭代次数之内向真实的方向开始收敛。观察极大似然函数曲线也可以发现，曲线前期变化很大，在较少的代数之内，逐渐稳定，之后的变化不大。

四、总结和收获

通过本次作业，我了解了EM算法以及混合高斯分布（GMM）的概念，并通过理解原理部分学习了如何估计对正态分布进行极大似然函数的参数估计以及EM算法的迭代步骤。在编写代码的过程中，发现了EM算法受到初值的影响很大，通过多次实验可以找到收敛较好的结果，以及容易出现较好收敛结果的初值。此外，还发现，如果增大数据集的量，即增大采样率，无论怎样更改初值，算法都会更容易收敛到真实的情况。

附录

1. 使用Excel生成的固定数据集

male: [166.15000043825032, 190.30523186648455, 163.50220761119732, 163.62426392277925, 187.96912658225216, 169.11041430362613, 186.88888403989557, 175.1233771031881, 164.21854637683705, 196.5570748268579, 150.10929088162948, 172.66794931831188, 166.98670016886095, 149.93649904636231, 174.51292818114416, 177.50956846990562, 180.77637761082585, 170.54760066581753, 171.75867514915538, 162.80111962689801, 120.00466880999016, 165.98294510507068, 145.8583188680164, 164.73886081687695, 164.66086655537947, 180.0385446150517, 167.4179492155944, 196.7006482650286, 169.5855518229282, 200.90729213484985, 172.66244553154812, 161.08653006779966, 185.212724842339, 176.10880736947283, 165.0293786678737, 175.5283699715323, 187.78659706477265, 144.09759841929818, 169.89041691755492, 168.55051503589794, 209.57200578573512, 162.7624480884757, 179.69747364785624, 173.92485529278687, 166.31444300179038, 155.27880716103346, 174.7129645616101, 155.7177259222184, 192.3310678943784, 172.12973413578084, 187.81864808364173, 183.31043593929076, 162.52405336004668, 196.897831842528, 168.6754771290935, 183.06553261299283, 173.31042947950593, 173.65854304879883, 188.1633766841651, 156.23069011293703, 162.59077279188566, 178.97808170009748, 169.19407827909166, 179.68729339650167, 198.31574299073264, 162.19400917766174, 156.56176153440725, 148.4464735951712, 154.92308654675531, 156.45397737583772, 186.67100122665474, 167.85183002475983, 171.74178408877307, 208.52070917370688, 177.9010425411354, 164.20870241033975, 176.6151772965798, 145.37776265835825, 143.98187556862888, 186.6780500345163, 171.35077873426798, 186.5753099221021, 172.42517320139225, 165.62802211032383, 181.34294914523977, 151.19341923424432, 193.57472355673454, 189.8639958066358, 197.30422741693113, 170.7180218917443, 207.9360063654125, 220.01020995008628, 172.65718380905048, 134.8180739890339, 170.37843707246398, 187.81923732665206, 186.58722390879197, 150.14526582029333, 160.40771967653734, 163.9386168520724, 181.0082631684281, 198.01353733432765, 179.59402043138945, 185.73056317029506, 185.5807685331131, 172.03365016773583, 158.04949000433243, 158.645908061195, 164.57477777202325, 196.22847352994472, 193.37098281138046, 174.18641095470812, 198.71918224112062, 152.09468307170164, 183.50164544808808, 198.28476886452012, 206.76540969493735, 171.02718118784364, 169.57000331063006, 151.10876007296517, 155.54316016445702, 194.70176987991425, 176.3998952034989, 163.0255936089669, 143.2198648625717, 183.29309650236158, 154.26385170726613, 190.74036107187436, 203.81638859504332, 190.08743803742547, 166.35908141106648, 175.35816017544886, 183.74617514359102, 191.04511413621023, 196.75364013240707, 155.49502553691366, 164.27808178370165, 180.37748029565017, 150.8737375164919, 191.8850635749946, 179.0522240546126, 182.82585966566887, 183.5721996373926, 178.31788132018497, 172.08858917727963, 179.02195512445093, 173.35680687922212, 179.735443823012, 174.13837736041202, 179.2205090686048, 160.76069222199035, 157.403820427102, 154.93236538472476, 179.53976223614768, 191.9594022854748, 175.13798070107734, 187.92596216045584, 177.36217889243264, 182.05062633160986, 200.69719433392282, 152.56066857685386, 187.38967345534206, 163.16029344694593, 165.53941306035838, 162.24761822446246, 165.11953745338565, 169.51872295926339, 173.0734936695194, 157.097790305611, 173.96950007725505, 193.6625661571887, 169.50132299047928, 192.15553780717943, 195.68995822886032, 169.5257185214152, 183.0107369737227, 142.93928634073566, 166.50908922674586, 161.26831346985315, 166.58160295622775, 168.86552932322542, 173.04255066552247, 155.23247826862914, 166.05117078697538, 173.08581121521303, 174.51851490494226, 172.32305792539066, 167.6237674231675, 168.03462131644207, 165.64659223954055, 189.36870918545776, 161.18899415546926, 194.1825236449743, 178.04438686169593, 183.9440892245, 193.42476840166427, 182.807156524834, 169.54732551592497, 166.95042971364944, 156.54902674943025]

female: [143.39038590797486, 150.07874544899857, 165.0288682532469, 179.09517208230423, 175.13080422699167, 167.8233993548947, 158.02812090292122, 161.67746854257365, 162.47609552805983, 158.098583193272, 152.6574139248449, 177.07825333187583, 156.59824086030955, 163.5166584095032, 153.76840515031182, 162.80616990499314, 156.73650260657718, 178.3298266278253, 160.60792540363119, 156.9465524481181, 170.77144629430921, 164.6338876115548, 156.92429224536448, 173.07406702926826, 188.69694189813103, 164.07184522914613, 165.3012148966906, 173.82859622119554, 169.4156889171395, 182.4855508231447, 167.8631697437071, 177.73509252956165, 167.88261231925517, 170.44265191135707, 166.6111148023016, 157.73269147638857, 172.4827096896167, 166.1319528344039, 149.11109102952562, 165.6824885468481, 176.12715701063354, 171.6539093008458, 168.70553758777845, 184.47850900969783, 160.46460116445436, 164.80513326141855, 162.7568024648385, 145.86959255713703, 171.9970660802402, 166.5028431470203, 177.2538713183825, 162.99686058763345, 170.4090023145845, 154.02876778473237, 162.9519667558881, 174.99148589073482, 161.43623213478304, 164.23592723036103, 168.81112557913804, 179.48152412821463, 169.31382845556237, 161.68373238096027, 195.52616210354424, 159.55550254030126, 168.94721599648616, 155.31779647899876, 156.24051397005482, 167.29722353090466, 162.65809002765081, 180.28374277560698, 168.68490783459205, 158.81413037645365, 146.24331018331011, 162.7647382660531, 158.58234506185744, 170.86467697125474, 158.22569368762052, 159.59197639542762, 163.4477246679743, 152.4825640256372, 161.76824909938534, 178.07961614979007, 162.61171839582266, 161.37415520854512, 159.6927797652289, 167.9606151391455, 158.99099819358213, 150.77309298098456, 150.29206808937764, 162.91248569944972, 174.17124745809042, 180.4520842855552, 157.7848500247845, 166.94713258797594, 160.4754272099528, 162.45871938066193, 164.18448454393413, 158.5532223063653, 165.67079573362957, 176.02818808798753]

初始化的参数： [0.6, 180, 18, 0.4, 160, 7] %男生比例，男生均值，男生方差，女生比例，女生均值，女生方差

0 : -1087.74605445968

第1次迭代参数： [0.6105450430370014, 176.89807555556953, 14.677526368798011, 0.38945495696299826, 161.73987083046225, 7.297049866401867]

1 : -1165.2031496278562

第2次迭代参数： [0.6210967240791923, 176.31578321390293, 14.939228571371334, 0.37890327592080736, 162.27223710241847, 7.597652893158511]

2 : -1178.2842807349855

第3次迭代参数： [0.6273279942929415, 175.90924020448637, 15.14358728742082, 0.37267200570705866, 162.72176558200812, 7.749358810636641]

3 : -1185.6505153244343

第4次迭代参数： [0.6313684762889685, 175.60823155379956, 15.305181229711255, 0.3686315237110317, 163.09276917381146, 7.8210285357674545]

4 : -1190.3178231577092

第5次迭代参数： [0.6341842070825489, 175.3782303481999, 15.433152116060901, 0.3658157929174514, 163.3951698944031, 7.849124931806199]

5 : -1193.497547702393

……

……

……

第80次迭代参数： [0.6761037330946439, 174.09652791151598, 15.835738367324605, 0.3238962669053567, 164.51972619522454, 7.302107385367071]

80 : -1203.206513002817

第81次迭代参数： [0.6762896165387606, 174.0941700275096, 15.835218760647649, 0.32371038346123965, 164.51915297785024, 7.299551071238389]

81 : -1203.2066169591067

第82次迭代参数： [0.6764710804563987, 174.0918687930374, 15.834711319891104, 0.3235289195436009, 164.51859413589062, 7.297054099842708]

82 : -1203.206716835104

2. 使用python随机生成数据集

male: [166.71148854 160.37712716 156.00419853 192.34818711 184.8811979

179.21146327 151.02931671 200.89968172 185.0901872 166.04912775

175.91107909 170.48223998 174.44366184 179.56358191 180.99365954

188.16558632 179.22877146 187.35671327 207.68409567 185.67758794

190.40714509 183.88877986 167.10408222 157.15545985 161.02194761

183.60864495 211.15188216 192.43645738 157.0636827 190.95207858

177.64515624 173.29771977 176.39548344 175.89801228 167.68282977

163.53598428 169.0530466 162.34670727 209.54266333 172.59145665

179.15869102 169.77468781 176.44433015 180.44558352 164.00008848

133.3435682 168.81182001 173.09249441 155.26541183 171.6120357

189.45816619 197.40177879 186.62263032 162.49847744 139.96190553

187.80432341 175.56404939 195.90468052 177.38626439 188.73413801

163.69544841 176.13987382 203.64021526 184.24807534 174.69902229

201.14316815 179.49034198 188.36600525 201.71814511 162.41187587

156.961367 171.28426614 176.13793024 171.79424244 161.98169376

164.39508208 178.04749842 171.21052103 165.58013058 192.07979354

161.58055609 151.78949406 190.48776434 171.21861649 193.35432694

160.26736894 194.34004806 182.73771824 180.38751448 177.48035622

150.22783702 183.44532311 176.78857818 170.59265613 179.96456927

181.90747828 161.21649562 167.38776625 197.8317528 181.7885713

151.79932655 165.95685151 168.23217172 181.79675843 160.35209453

185.05211915 197.04068728 192.30192532 178.02627449 191.33249383

197.44159262 167.80549499 187.11542481 196.46009106 190.67586494

175.98328423 176.22178229 189.37317784 156.39868282 159.71473574

168.61831031 192.010337 160.91975326 199.16391502 157.64732814

168.9574624 173.97162316 170.13562021 167.91227205 149.70212836

153.75044661 145.14198708 184.28506818 178.11152304 164.77404672

126.74623812 172.66827871 142.39278471 182.41644499 170.83746769

195.44280525 171.18487816 173.89051052 185.97066868 175.79578435

158.87823529 165.289631 171.00737278 179.65868124 174.72819305

164.6295574 193.37760021 179.58324576 170.95417225 170.62783358

193.57153699 178.79468882 186.2207389 179.42913746 178.71206616

178.31594204 155.59934967 187.53896265 161.94842508 171.37857216

172.80196656 176.00882608 198.34720334 195.60961395 159.39472309

203.76978953 177.65902001 166.83055568 182.00623226 174.61476039

175.74480236 171.82683885 160.2557604 206.97146076 161.5444471

160.07558196 199.22950076 193.42027713 154.29652579 155.32147599

154.16185956 157.98481826 171.22951689 172.96055299 178.38885186

170.37523128 154.03087921 161.64248364 160.56203941 184.33021765

181.77849836 188.26228678 181.69430083 161.89831283 158.12482903]

female: [159.13765606 157.91639483 156.75142527 153.70959741 154.47816859

164.8674679 178.517941 157.36481572 156.77471198 163.75334253

154.14218045 172.62238428 144.33084578 174.64208381 154.35384234

173.20293909 165.55271454 166.21502427 162.69612439 165.49530461

150.35504291 141.88805081 153.22287799 159.15250107 179.95515813

172.38690073 172.75481148 172.4929444 165.28451851 173.25792874

163.19403479 165.95125713 163.82503747 156.59309834 164.25312319

159.0476187 162.67418429 162.23763657 148.46732613 172.50941168

169.75835201 181.53293599 170.7873792 173.68187288 177.45794299

158.07469511 156.90712882 173.98500653 165.19974195 153.98802789

157.51937566 159.64905195 156.40433758 170.93380337 171.00094438

173.2043337 179.73970873 171.45149905 172.41028217 167.85658315

155.71947068 151.15453233 153.81097989 152.33129102 158.00500186

173.48468479 142.64901075 166.87489841 167.7655802 157.79408285

166.18799259 144.67223458 147.47764791 174.32641115 151.66141409

161.25887834 167.47641884 152.67439247 181.71962934 168.61332943

144.60537706 172.40933083 171.7379806 159.9003299 166.32891598

169.25839994 164.52048204 171.7892804 164.77348749 166.24602552

160.12690414 177.00326247 172.89644899 168.27164165 155.09464125

176.01946688 163.35413331 155.67353822 154.84776515 170.26804324]

初始化的参数： [0.5, 180, 20, 0.5, 160, 7] %男生比例，男生均值，男生方差，女生比例，女生均值，女生方差

0 : -1069.4739868487045

第1次迭代参数： [0.5741767086029091, 178.02272557527573, 14.351036903085403, 0.4258232913970908, 162.2451040497167, 7.764296809121082]

1 : -1170.7111755033504

第2次迭代参数： [0.5942122662253567, 177.29177104835804, 14.40616800089958, 0.4057877337746433, 162.53646007378165, 8.449137143833381]

2 : -1185.641489997828

第3次迭代参数： [0.6050720547333569, 176.83748636908436, 14.472862260389707, 0.39492794526664343, 162.82672924602497, 8.879502666658118]

3 : -1193.1527639711444

第4次迭代参数： [0.6118157674329349, 176.51258331557545, 14.547552291340052, 0.38818423256706575, 163.09540654286573, 9.157542309385176]

4 : -1197.627552273905

第5次迭代参数： [0.6164405847143668, 176.2599612060791, 14.62617105849463, 0.3835594152856328, 163.33963084332677, 9.33786917088359]

5 : -1200.5827597462953

……

……

……

第55次迭代参数： [0.6676857803358431, 174.35773405635206, 15.363349221484317, 0.3323142196641561, 165.16917988436, 9.253997764002593]

55 : -1212.3623187033622

第56次迭代参数： [0.6681971639173372, 174.3528414561675, 15.36174411610135, 0.33180283608266253, 165.16487113826645, 9.249587702228279]

56 : -1212.3622715165677