利用贝叶斯分类器实现基于身高的性别分类

1. 算法简介（算法原理及结构图）

（1）本实验采用基于贝叶斯理论的分类器，算法原理如下：

设样本空间S划分为，A为某一事件，则在A发生的条件下发生的概率为：

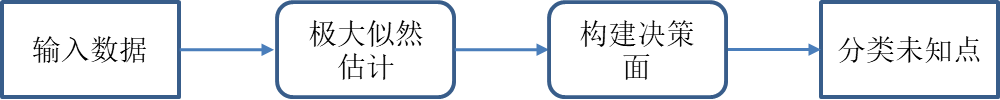
上式即为贝叶斯公式，其中为先验概率，为后验概率，为类条件概率密度。先验概率通过各类别数量计算，类条件概率密度通过估计分布与参数计算。类条件概率密度函数的参数估计需要利用样本集估计得到，期望当时，可以收敛于，估计方法包括极大似然估计和贝叶斯估计。

当样本为高斯分布时，参数的估计为：

决策面方程即当两类区域相邻，可表示为，其中，可进行对数变化得到：

决策面方程为：.

（2）贝叶斯分类器算法结构图如下：



1. 实验设置
   1. 数据集介绍

一共有两份数据集，traindata.csv和testdata.csv。

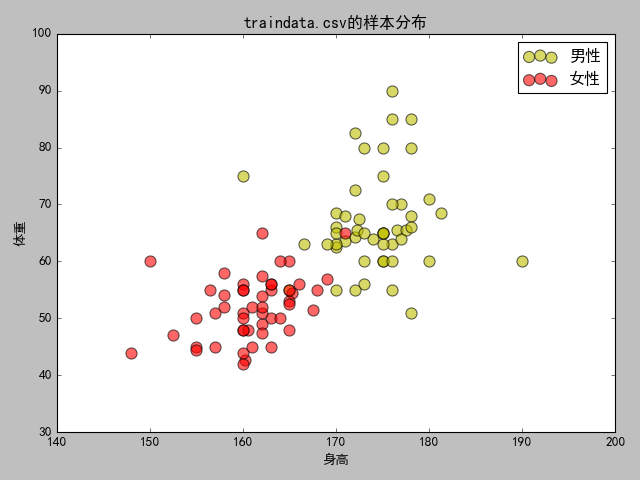
数据集traindata.csv和testdata.csv各包含了100份数据样本，数据的构成如下所示：

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | gender | height | weight |
| Sample1 | Male | 174 | 96 |
| Sample2 | Female | 185 | 110 |

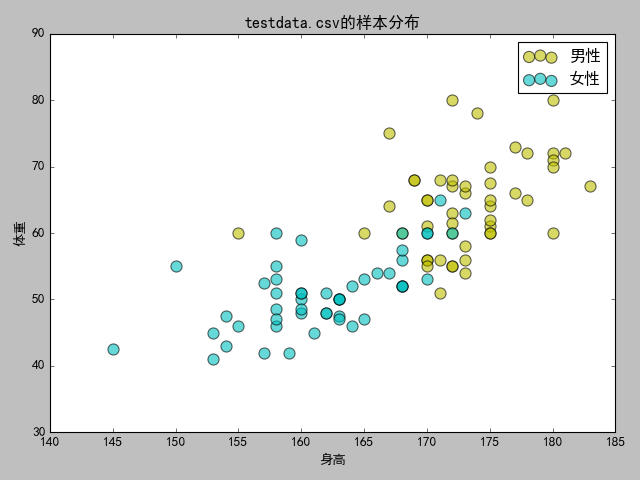
其中第一列代表性别：Male/Female，第二列表示身高height，第三列表示体重weight，数据间用‘，’分隔。其中第一行是列标题gender,height,weight，数据样本前50行类别是Male，后50行类别是Female。

数据集用身高和体重两个属性判断样本的类别是男性还是女性。

数据集traindata.csv的样本分布图如下：



数据集testdata.csv的样本分布图如下：



* 1. 评价指标介绍

本实验用的指标是**错误率**，公式如下：

用训练集训练贝叶斯分类模型，利用测试集对模型进行测试，并计算性别分类错误的样本占总样本的比例即为错误率。

* 1. 实验环境(开发工具、调用库等)

开发工具：pycharm

调用库：numpy、matplotlib.pyplot、math、pandas

* 1. 使用算法具体结构和参数

1. 读取数据：traindata = pd.read\_csv(**"traindata.csv"**)，traindata = pd.read\_csv(**"testdata.csv"**)
2. 绘制数据集样本分布图：plt.scatter()
3. 极大似然估计：loc = data.mean()，scale = np.sqrt(((data - loc)\*\*2).mean())
4. 计算先验概率：P\_male = len(male)/(len(male) + len(female))
5. 用训练集构建决策面
6. 用测试集计算错误率
7. 输入待预测数据并分类：plt.scatter(180, 120, norm = 2, c = **'red'**, s=100,marker=**'s'**)
8. 结果分析
9. 定量分析

本实验利用贝叶斯分类器，根据身高和体重预测性别，在读取数据集后首先进行最大似然估计，即分男性和女性计算训练集中身高和体重的均值和方差，结果如下：

最大似然结果:

身高的参数:男性均值174.13400000000001,方差4.52723359238288,女性均值161.03,方差4.434151553566928

体重的参数:男性均值66.60799999999999,方差8.400377134391052,女性均值51.956,方差5.402412794298488

同时尝试用贝叶斯估计，贝叶斯估计结果如下:

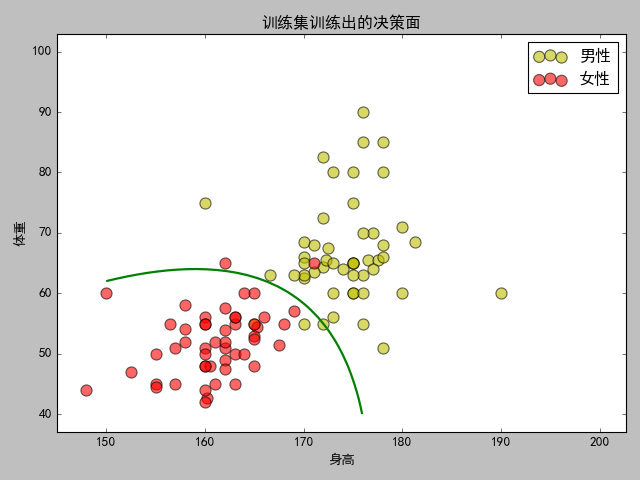
身高的参数:男性均值123.50657153633057,女性均值115.58001024953455

由常识可知，最大似然估计更符合现实。

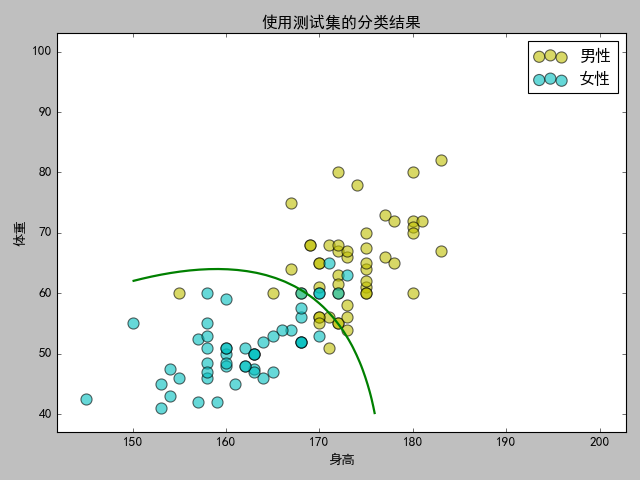
然后构建决策面并输入测试数据集进行测试，计算分类错误率为10%。

1. 定性分析

首先利用训练数据集构建决策面，如下图所示：

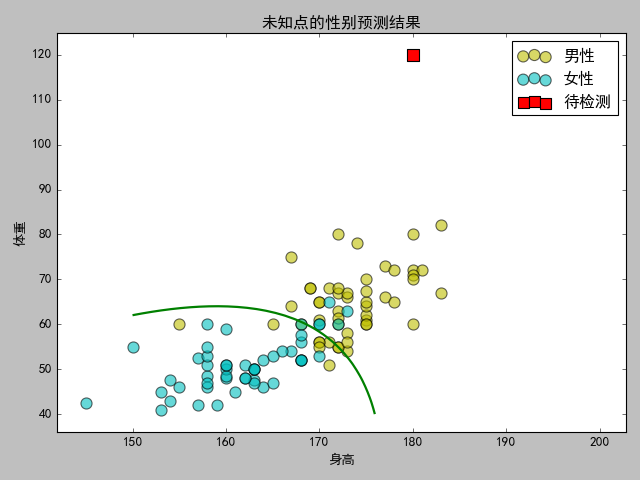


然后，在此决策面的基础上，输入测试集进行测试，如下图所示：



分析上图可见，在100个样本点中，有10个点分类错误，错误率为10%，模型分类效果较好。

最后，在测试集样本决策面上输入待预测点，如下图所示：



由上图可知，经贝叶斯分类器预测，身高为180，体重的120的人的性别为男性。

1. 总结与展望

本实验通过建立贝叶斯分类器，运用极大似然估计，实现了根据身高和体重对人的性别进行分类和预测。由本文的实验结果可知，贝叶斯分类器的决策面可直观看出所预测数据的所属类别，此外，贝叶斯算法还有方法简单、学习效率高的优点。然而，由于算法以自变量间的独立性和连续变量的正态性假设为前提，会导致算法精度在某种程度上受影响。本实验未能实现自动计算错误率的功能，在接下来的学习中，会尝试实现它。