**华中科技大学计算机科学与技术学院**

**机器学习报告**



专 业： 计算机科学与技术

班 级： 计算机1807班

学 号： U201814685

姓 名： 张嘉洋

成 绩：

指导教师： 何琨

**完成日期： 2020年 6 月 25 日**

# 机器学习结课项目

## 一、实验题目：基于朴素贝叶斯分类器的语音性别识别

## 二、实验要求

* 题目背景

用朴素贝叶斯分类器进行数字手写体识别(基于MINIST数据集)，因此在这里用朴素贝叶斯在语音上做一个小应用——分辨声音是男性还是女性。具体题目可以参考<https://www.kaggle.com/primaryobjects/voicegender>

* 数据集

数据集可自行在<https://www.kaggle.com/primaryobjects/voicegender>下载或[附件](voicegender.zip)。这个数据集是基于对男女语音段进行合理的声音预处理而得到的语音特征(并不包含原始语音段)。集合中共有3168条数据，男女各1584条，每条数据可视作一个长度为21的一维数组。其中前20个数值是这条语音的20个特征值，这些特征值包括了语音信号的长度、基频、标准差、频带中值点/一分位频率/三分位频率等；最后一个数值是性别标记。元数据集中直接以字符串,即male和female进行标注。使用7：3划分数据集。

* 任务描述

通过朴素贝叶斯方法，可以先对所有特征值做统计，并且通过连续性参数估计（高斯分布）方法得到参数。之后使用预测函数预测测试集。

## 三、算法设计

本项目整体来说可分为加载数据，基于朴素贝叶斯进行数据处理两部分。其中朴素贝叶斯分类器可自己实现，也可使用使用sklearn库中的naïve\_bayes模块来进行实现。下面对这几部分分别进行介绍。

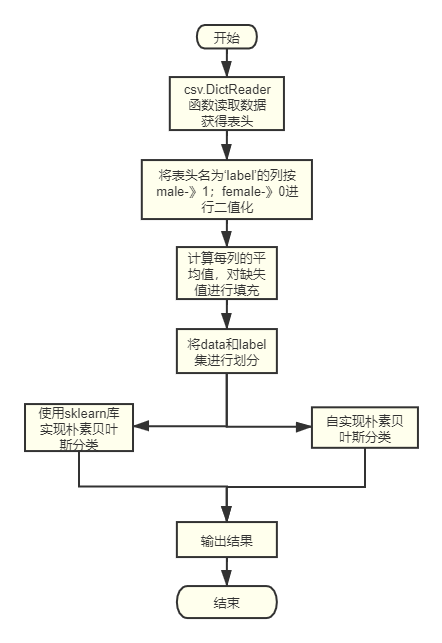
* 加载数据

本实验数据以.csv的文件格式储存，读取该文件借助csv库中的DictReader函数。



可从csv中得知该文件包含有语音信号的长度、基频、标准差、频带中值点/一分位频率/三分位频率等信息

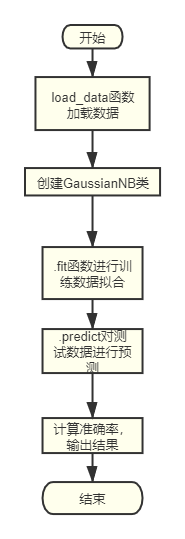
对上述信息的处理详细步骤如下：



**图3.1算法主体流程**

* 朴素贝叶斯分类器实现

接下来描述朴素贝叶斯的实现，在sklearn库中含有所要使用各函数，而自实现的过程即将这些函数逐个实现。



**图3.1朴素贝叶斯算法流程**

该过程基于贝叶斯公式

根据全概率公式，

展开分母，得到

也即已知特征x，求解yi，只需要知道先验概率P(yi)，和似然度P(x|yi)，即可求解后验概率P(yi|x)。

在本实验中，先验概率即为男女性别的比例。各特征值具有连续性，视为连续变量，因此对于似然度的估计应采用高斯分布，根据高斯分布的概率密度函数计算

所谓朴素贝叶斯，即将各特征值都视为独立变量，即为独立概率，则有

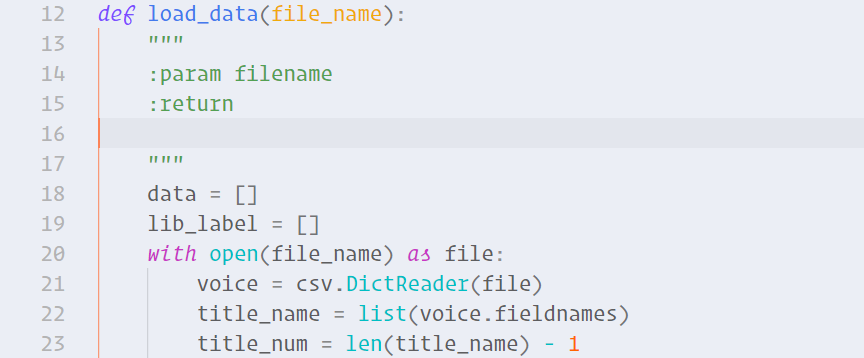
经过上述运算后，即可得到后验概率P(yi|x)，比较不同label下的概率值大小，即可做出预测。

## 四、实验环境与平台

Microsoft Windows 10.0.18363，16GB RAM，CPU @2.30GHz；Visual Studio Code 1.46；Python 3.8.2 32-bit；

## 程序实现

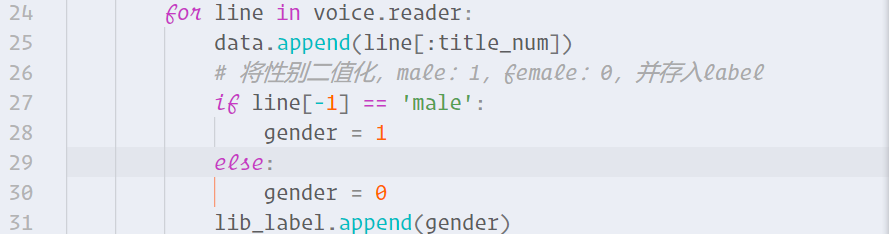
* **数据加载**



**图5.1**

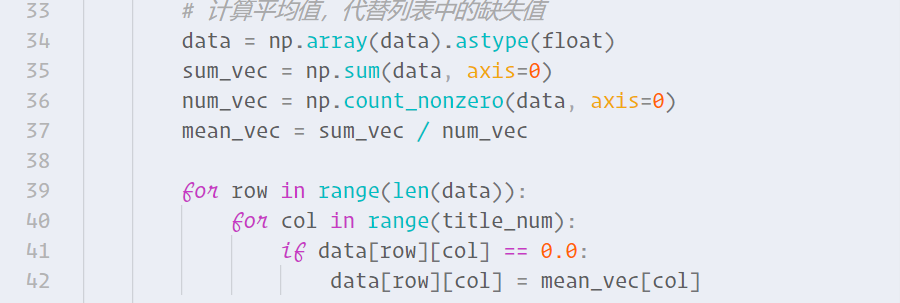
首先根据文件名读取csv文件，并获得所有表头的名称以及表头数量以用于后期分类和显示。此外直接查看csv文件，也可直观地观察上述信息，其中较为重要的是标签label位于最后。

然后将每一列的特征值作为列表加入数据集data中，对最后的标签进行二值化，并将二值化后的结果加入标签集label。



**图5.2**

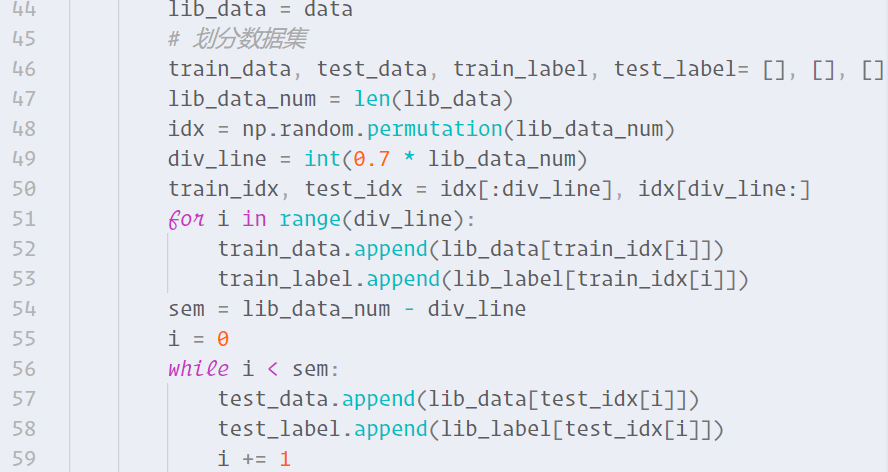
观察csv文件时可发现有些数据项为0，若不做处理会影响后序的计算。这里的处理方法是对该特征项的所有数据求特征值，并将该值填充在空白处。



**图5.3**

填充空白值之后，数据集就成功构建了，接下来要对数据集和标签集进行划分，划分为训练集和测试集。根据实验要求划分比例为7:3 。

划分方法为首先将构建从0到数据集总个数大小的随机序号序列，然后将该序列按7:3划分，前一部分序号对应的数据集内容加入到训练数据集train\_data中，标签集内容加入到训练标签集train\_label中；后一部分序号对应的数据集内容加入到测试数据集test\_data中，标签集内容加入到测试标签集test\_label中。



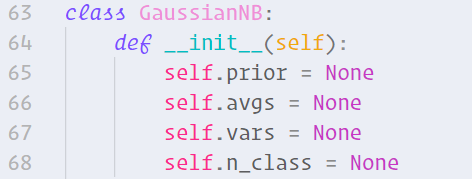
**图5.4**

之后返回训练数据集train\_data、训练标签集train\_label、测试数据机test\_data、测试标签集test\_label即可。

* **朴素贝叶斯分类器实现**

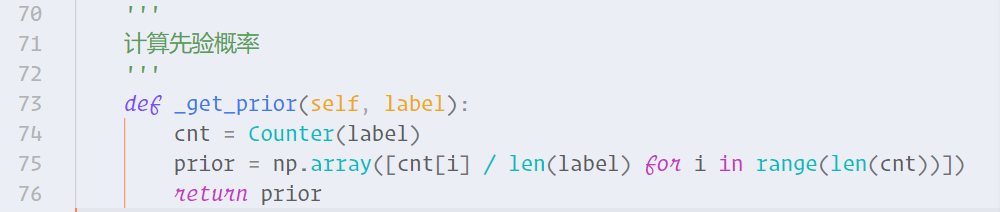
自己实现的朴素贝叶斯也即对sklearn库中将要用到的各种函数进行实现，原理过程更为详细，在此先介绍自己实现的朴素贝叶斯分类器。

* **自己实现的朴素贝叶斯分类器**



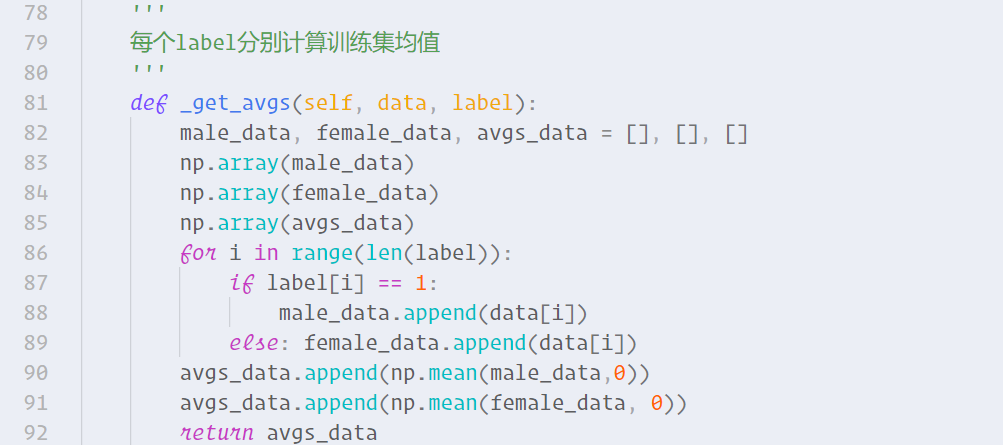
**图5.5**

首先创建一数据类，包含有先验概率prior、平均值avgs、方差vars、label数n\_class。



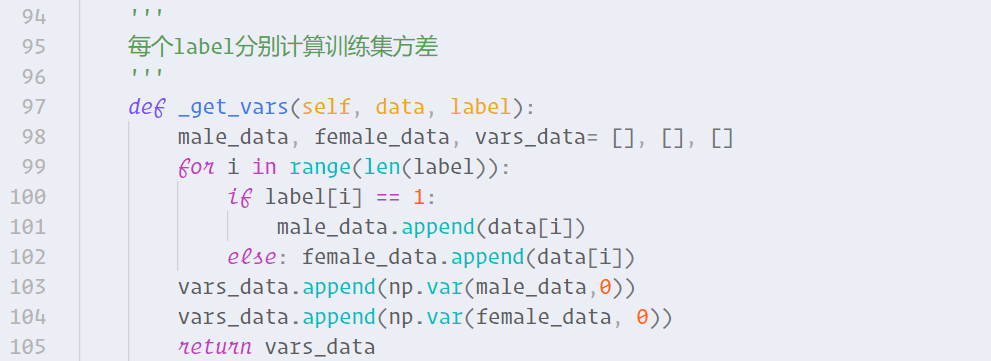
**图5.6**

然后为计算先验概率函数，即各标签值在训练数据集中所占的比例。可使用系统函数Counter进行计数，然后求比值即可。



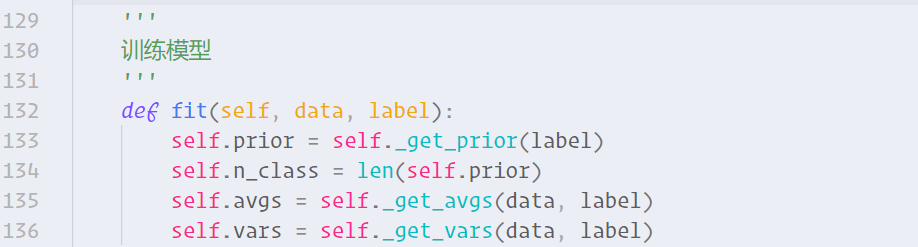
**图5.7**

接着为求训练集均值函数，这里的均值为各标签值下的均值，因此先对数据集按标签值进行分类，再对每一类利用mean函数求均值即可。



**图5.8**

求方差后函数与求均值函数相似，不过这里使用的函数是var函数。



**图5.9**

训练模型的过程即利用fit函数求得数据类中的各项值，其中调用了之前表述的各种计算函数。

上述函数即完成了对训练集的处理，接下来即利用训练数据得出的结果，对测试数据进行求解后验概率，并计算其准确度。



**5.10**

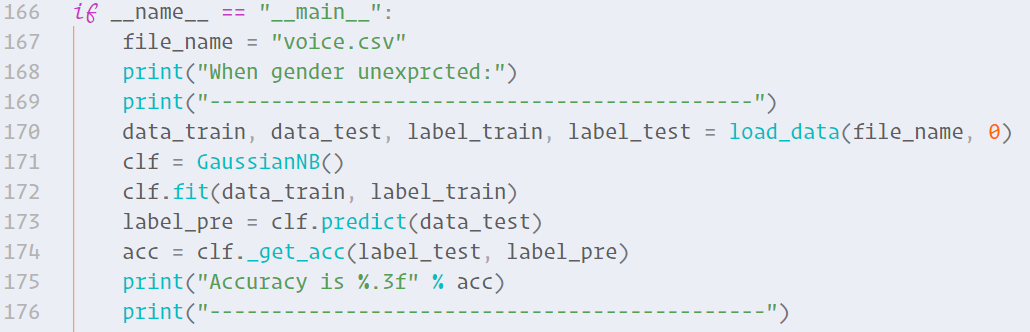


**图5.11**

计算似然度是对数据集中的每一行数据而言的。按照公式计算各特征值的概率，再利用.prod函数将各概率相乘，得到似然度。

对每一个标签来说，使用先验概率乘以似然度。分别将上述值作为分子，将各标签值对应的值之和作为分母，即可得到该标签下的预测值。

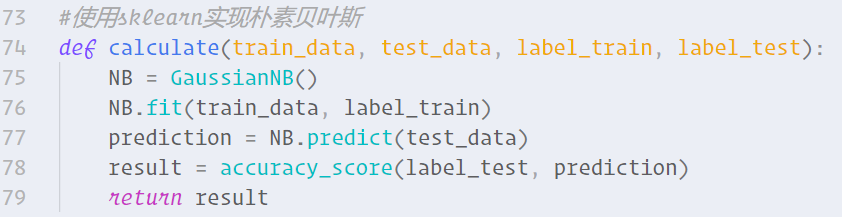
以上为各计算函数，该分类器的主函数如下：



**图5.12**

* **sklearn库实现朴素贝叶斯分类**

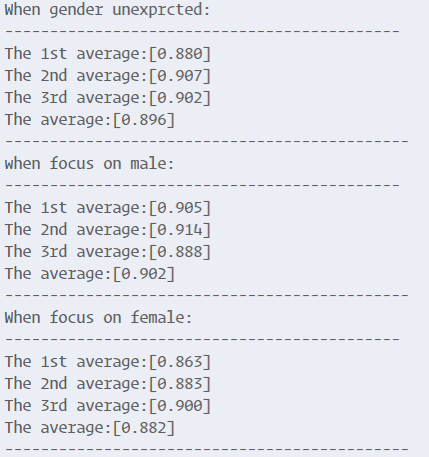
在sklearn库中封装有上述实现的各计算函数，直接调用即可。



**图5.13**

## 六、实验结果

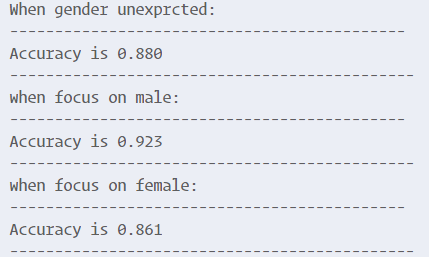
sklearn库实现朴素贝叶斯分类的运算结果如下：



**图6.1**

|  |  |
| --- | --- |
| 男声正确率 | 男声错误率 |
| 90.20% | 9.80% |
| 女声正确率 | 女声错误率 |
| 88.20% | 11.80% |

自实现朴素贝叶斯分类的运算结果如下：



**图6.2**

|  |  |
| --- | --- |
| 男声正确率 | 男声错误率 |
| 92.3% | 8.7% |
| 女声正确率 | 女声错误率 |
| 86.1% | 13.9% |

## 七、结果分析

根据上述实验结果可得知，使用sklearn库中的GuessianNB模块和自己实现的高斯朴素贝叶斯算法运算结果没有很大差别，自己设计的计算方法达到了sklearn库中相关函数的运算性能。

但在不考虑性别分组的情况下，两者的运算准确率均在90%左右，具有较高的错误率，还有很大提升空间。使用平均值替代缺失值这一操作，虽然方便了运算，但可能带来了较大的损失。

此外，两种运算方法的结果都体现出男性声音的识别准确率略高于女性声音，在声音特征的区分上，应当注重不同特征的权值分布。

# 参考文献

[1]何琨，机器学习内部讲义，2020,4,6;

[2]周志华，机器学习，清华大学出版社

[3]朴素贝叶斯算法 --python 实现，<https://www.cnblogs.com/yiyezhouming/p/7364688.html>