**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1806班

学 号： U201814657

姓 名： 杨峰

成 绩：

指导教师： 邹复好

**完成日期： 2020年 7月 1日**

# 基于贝叶斯分类器的语音性别识别

## 一、实验题目：基于贝叶斯分类器的语音性别识别

## 二、实验要求

### 2.1数据集

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。

### 2.2任务描述

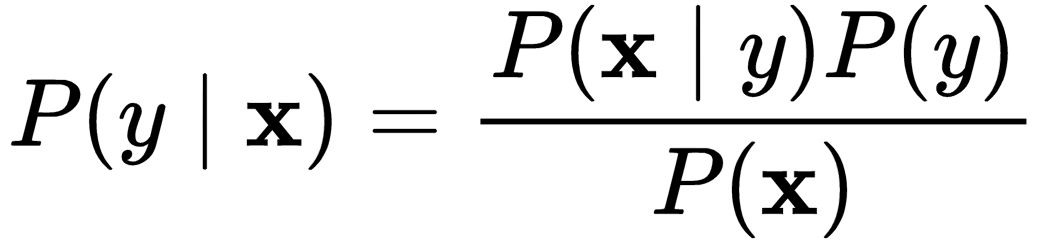
通过朴素贝叶斯方法，可以先对所有特征值做统计，并且通过连续性参数估计（高斯分布）方法得到参数。之后使用预测函数预测测试集。

## 三、算法设计

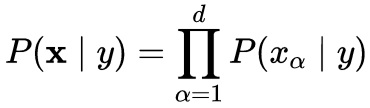
朴素贝叶斯算法描述：

朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征独立学习输入/输出的联合概率分布；然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。

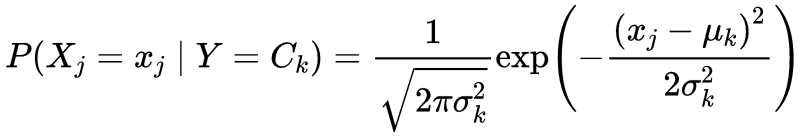
对于给定的输入x，先计算特定后验概率，再将后验概率最大的对应标签作为输入x的输出分类标签。

 (1)

估计P(x|y)需要对其做一定的假设，称其为朴素贝叶斯假设：

 (2)

其中=是特征维度上的值。在本次项目中，数据特征表现为连续特征，即，所以使用高斯分布来表征

(3)

P(x)在公式（1）中作为正则项出现，而P(y)由数据集可以很容易得到，所以由训练数据基于高斯朴素贝叶斯假设可得到后验概率。

## 四、实验环境与平台

实验环境：基于Anaconda的python 3.6.2

实验平台：jupyter notebook

## 五、程序实现

完整程序代码见附录

主要使用第三方开源机器学习框架sklearn对数据进行处理，并使用高斯贝叶斯模型对数据拟合：

1. 数据加载：使用pandas读取csv文件



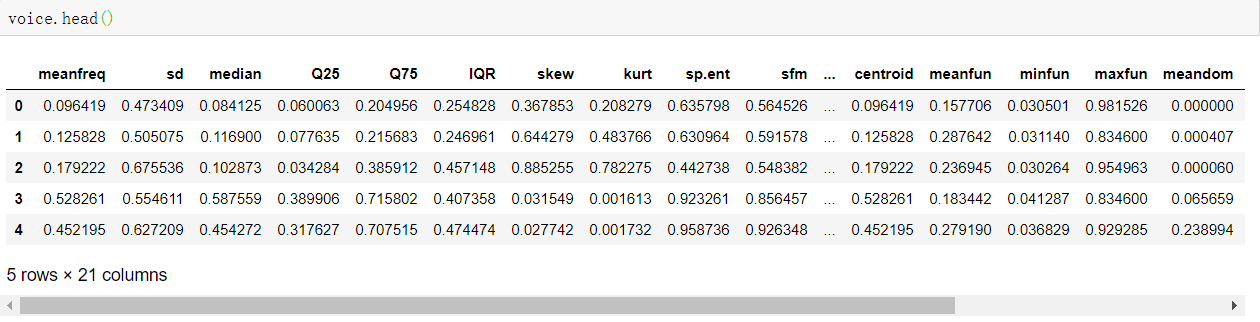


图1 使用pandas加载csv文件

2. 数据预处理：主要是替换文本并对数据特征数据进行归一化以及数据集的分割，按照7：3的比例进行分割。 这里我调用了sklearn库中的StanderdScaler、MinMaxScaler两个类和train\_test\_split函数：

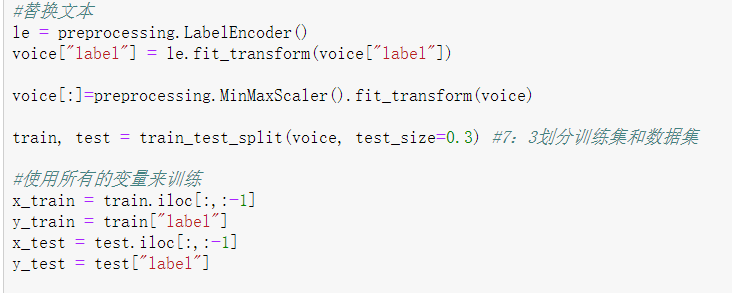


图2 数据预处理代码

3.训练及分类：根据我们数据处理后得到的两个集合，我们直接调用sklearn库中的GaussianNB分类器来进行训练，最后调用report库直接输出训练结果：

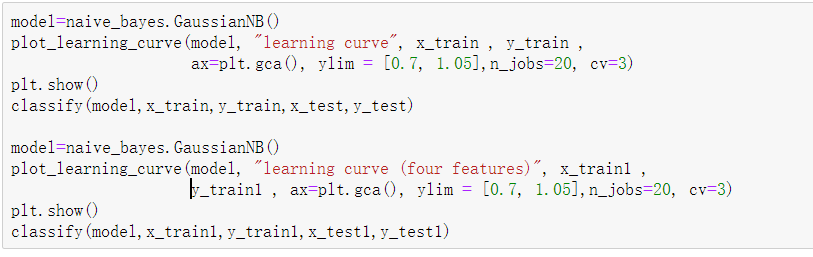


图3训练及分类

## 六、实验结果

实验结果如下图：

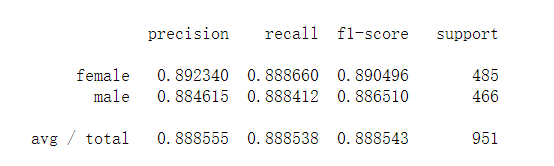


图4实验结果

使用交叉验证，绘制出学期曲线如下：

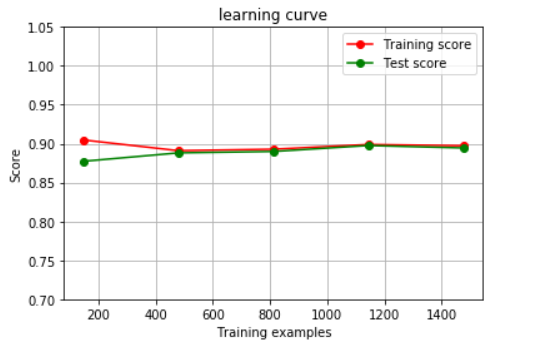


图5训练过程的学习曲线

## 七、结果分析

通过上面的实验结果，我可以总结发现，该模型对于预测男女性的性别正确率在90%左右，并且训练曲线随着样本数的增加快速收敛到89.5%附近，说明了高斯朴素贝叶斯模型具有快速收敛性并且能够保证较高的正确率。

但是所得到的正确率并不能满足很高的准确率，于是选择对算法进行优化。

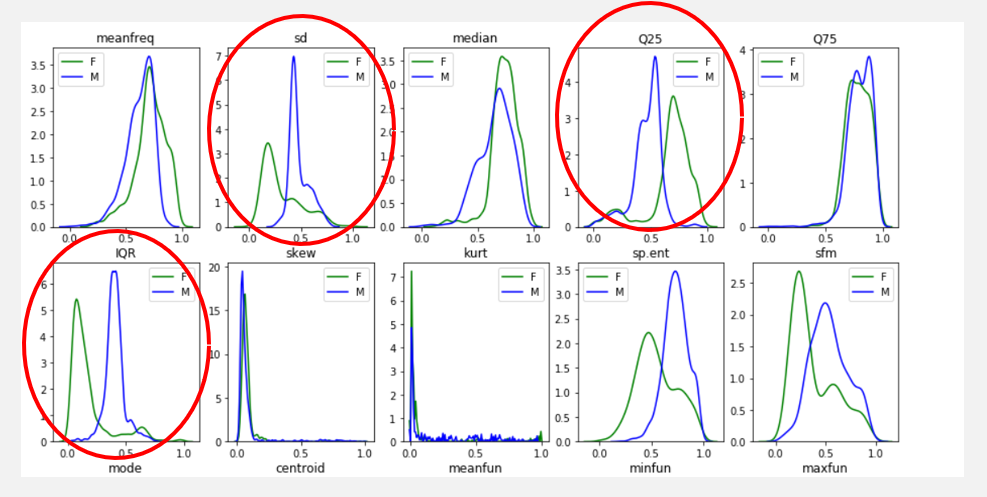
算法优化：

我们可以看到voice.csv中有一些都有为0 的值，说明数据可能有缺失，我们在处理数据集时要考虑到数据缺失的问题。

一共有三个特征列出现了缺失：mode , dfrange , modindx。将这三个特征舍去。

同时我们还能发现有些特征男女性别区分度更高一些，因此我决定通过各个特征的KDE曲线对所有特征进行区分度的分析。

各个特征KDE曲线如下：



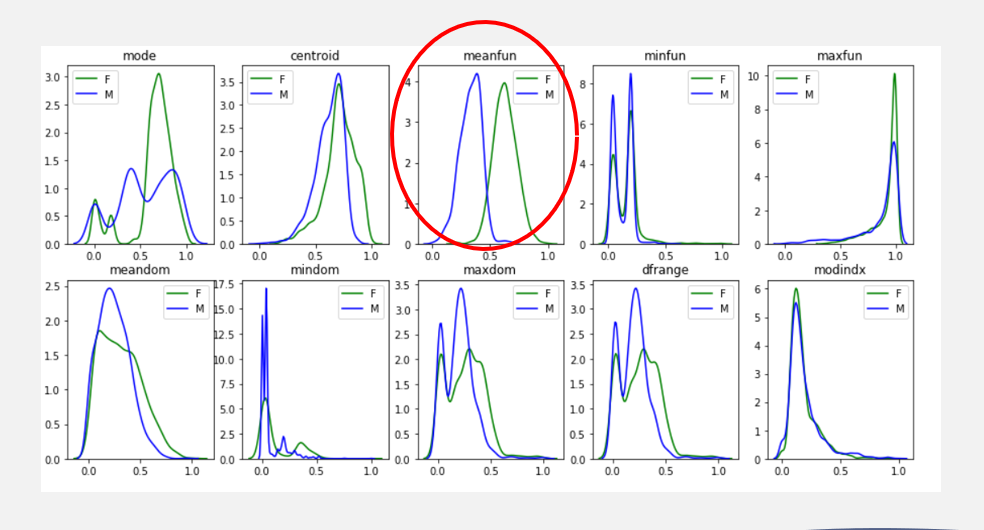


图6各个特征的KDE曲线

通过各个特征男女区分度的图中可以看到，sd、Q25、IQR、meanfun这四个特征的区分度较为明显，于是只考虑这四个特征来进行识别，

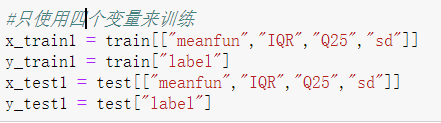


图7使用四个特征进行训练

其结果如下：

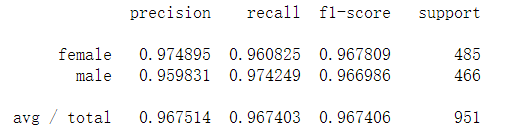


图8只使用四个特征的实验结果

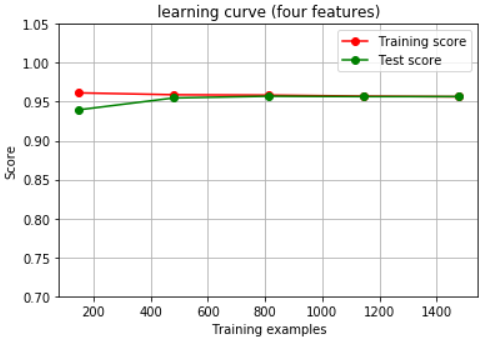


图9只使用四个特征的学习曲线

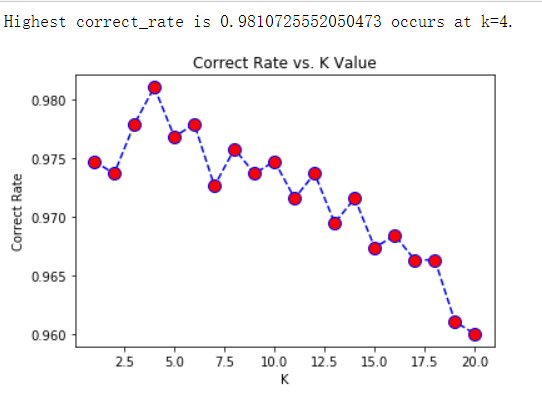
通过对两次实验结果进行对比，我们发现男性和女性的准确率获得了很大的提升，都达到了96%以上，相比之前有很大的提升。

通过资料查询发现，sd、Q25、IQR、meanfun这四种特征分别表示频率标准差、第一分位数、分位数范围和基频平均值。其中基频平均值代表了人声音的高低，而男性女性的声音高低有很大区别，这也非常符合我们日常生活中的常识。

## 八、拓展延伸

除了使用上述朴素贝叶斯模型机器算法优化得到的准确率外，我还尝试了本课程中所讲的其他的几个机器学习算法对这些特征再进行训练。

1.KNN



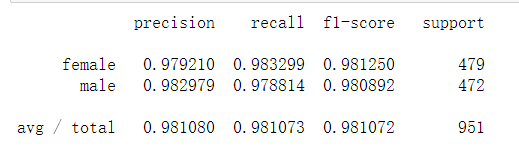
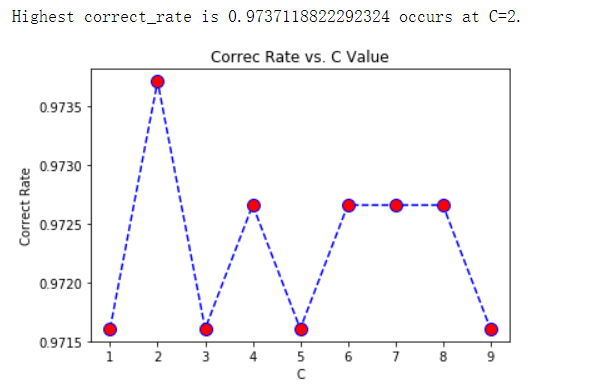


图10 KNN实现语音识别

1. CVM



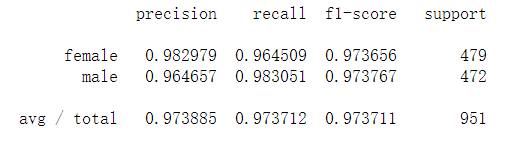


图11 SVM实现语音识别

1. 决策树

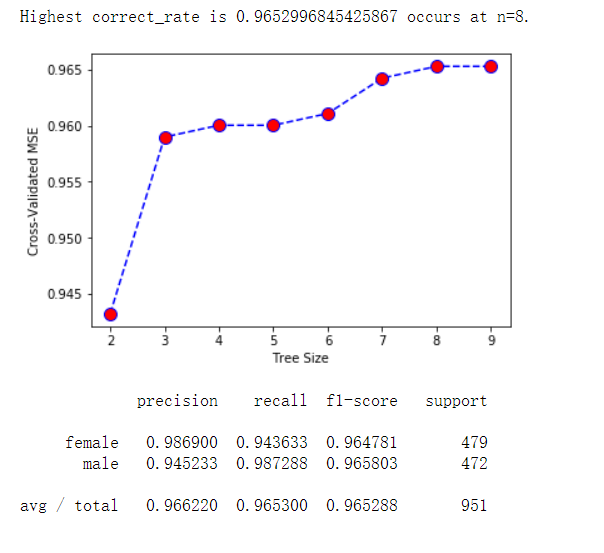


图12决策树实现语音识别

模型差异比较：

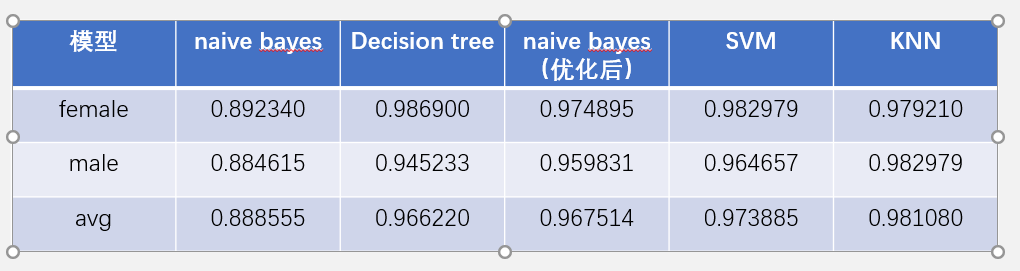


图13模型正确率比较

图中可以看出其他几个模型的正确率都略高于朴素贝叶斯，可能原因为特征之间的相关系数并不为0，代表特征之间并不是独立，因此朴素贝叶斯不能达到真正的分类效果，从而正确率相较于其他几个模型较低。

## 九、小结

总结与收获：由于python语言掌握的还不够熟练，以及各个算法了解程度比较粗浅，这次项目的过程还是十分曲折的，不过好在通过同学的帮助、查阅相关的资料，最后完成的还算成功。通过这次课程项目以及一个学期的机器学习课程的学习，收获了很多东西，对很多算法模型都有了一定的了解，同时也学会了python这门新的语言，了解到了sklearn、matplotlib等非常实用的库。

最后感谢老师的教导以及同学的帮助。

、

# 参考文献

[1] https://blog.csdn.net/GoodShot/article/details/80373372

[2] https://www.kaggle.com/primaryobjects/voicegender

[3] 周志华.机器学习[M].北京:清华大学出版社

[4] 李航.统计学习方法[M].北京:清华大学出版社