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：CS1806

学　　号：U201814655

姓　　名：杨雨鑫

成　　绩：

指导教师：邹复好

完成日期：　2020 年　6 月　30　日

目录

1. 实验题目：基于贝叶斯分类器的语音性别识别

2. 实验要求

2.1 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

2.2 数据集

数据集可自行在 https://www.kaggle.com/primaryobjects/voicegender 下载或附件。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。

2.3 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。

2.4 评测标准

要求得到 2*2 预测情况

男声正确率	男声错误率
女声正确率	女声错误率

3. 算法设计

3.1 数据分析

打开 voice.csv 文件，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。

在 csv 文件中有些地方的缺失值被记作了 0。这里给我们的数据是小数形式，说明是连续型数据，我们在三种的朴素贝叶斯模型中选择针对连续型变量的高斯朴素贝叶斯模型。

3.2 算法原理

这里我们得到的数据是连续型的，因此我们选择擅长处理连续型数据的高斯朴素贝叶斯模型。

高斯朴素贝叶斯模型原理：首先我们假设每一个特征都是相互独立的，即相关系数为 0，并且每个特征的数据分布是符合高斯分布的。对于单个变量来说我们需要在知道这个变量的值的条件下判断它是男性的可能性，这里我们记该变量等于某个值为事件 x，性别为男性为事件 y_i。由贝叶斯公式可得：

$$P(y_i|x) = P(x|y_i)P(y_i)/P(x)$$

这个公式的实际含义就是利用先验概率来预测出后验概率，这里我们的 $P(x|y_i)$是通过高斯分布函数的概率分布求出的先验概率，$P(y_i)$为男性样本占总样本数的比例，P(x)为该特征等于该值的次数占总样本数的比例，对于高斯分布函数的 σ 和 μ 这两个参数，得计算出该特征所有数据的均值和方差来确定。

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

最后我们就可以通过该变量后验概率的值预测出是男性还是女性了。

4. 实验环境与平台

实验设备：windows 10，i7-10750H，6 核
实验环境：基于 miniconda 的 python 2.7.16
实验平台：vscode

5. 程序实现

5.1 数据预处理

使用 pandas 库加载 csv 文件，把标签为男性的替换成 1.0，标签为女性的替换成 0.0。

```
#读取数据
voice = pd.read_csv('./input/voice.csv')
#文本替换成数字
le = preprocessing.LabelEncoder()
voice["label"] = le.fit_transform(voice["label"])

voice[:]=preprocessing.MinMaxScaler().fit_transform(voice)
```

图 1 加载 csv 文件并替换标签

我们在 excel 中统计分析每一条特征，发现特征分布中会出现偏差特别大的点，为了在不改变特征的数据分布的前提下把数据都统一在同一个数量级中，我使用了 sklearn 中的归一化函数，把每个特征的变量都投影到 0~1 之间，归一化函数如下：

$$x_s = \frac{x - x_{min}}{x_{max} - x_{min}}$$

图 2 归一化转换函数

最后得到的新的 csv 文件存储在 first_step.csv 文件中。

根据题目要求，我们对数据进行 7：3 比例划分训练集和测试集，这里我调用了 sklearn 库中的 train_test_split 函数：

```
train, test = train_test_split(voice, test_size=0.3)        #随机划分训练集和数据集，这里使用7：3划分
```

图 3 按照 7：3 比例划分数据集

最后我得到了 train 和 test 两个集合分别用来做训练和测试。

5.2 训练并输出测试结果

根据我们数据处理后得到的两个集合，我们直接调用 sklearn 库中的高斯朴素贝叶斯分类器来进行训练，最后调用 report 库直接输出训练结果：

```
#使用所有的变量来训练
x_train = train.iloc[:,:-1]
y_train = train["label"]
x_test = test.iloc[:,:-1]
y_test = test["label"]

```

图 4 所有特征进行训练

```
def classify(model,x_train,y_train,x_test,y_test):
    from sklearn.metrics import classification_report
    target_names = ['female', 'male']
    model.fit(x_train,y_train)              #训练数据
    y_pred=model.predict(x_test)
    print(classification_report(y_test, y_pred, target_names=target_names, digits=6))
```

图 5 进行训练并输出测试结果

6. 实验结果

训练之后生成的结果报告如下：

	precision	recall	f1-score	support
female	0.894737	0.885033	0.889858	461
male	0.892929	0.902041	0.897462	490
micro avg	0.893796	0.893796	0.893796	951
macro avg	0.893833	0.893537	0.893660	951
weighted avg	0.893806	0.893796	0.893776	951

图 6 高斯朴素贝叶斯训练结果

绘制出的训练拟合曲线如下（这里我选取了三倍交叉验证）：

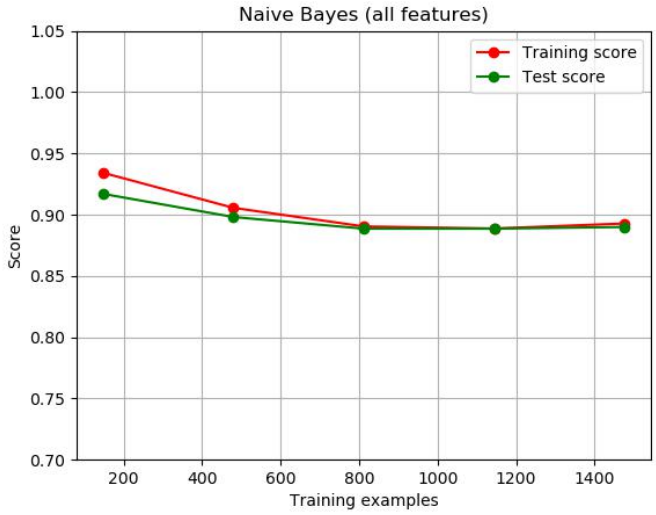


图 7 高斯朴素贝叶斯训练过程曲线

两种性别预测的正确率和错误率见表 1：

表 1 高斯朴素贝叶斯训练结果

男声正确率：89.3%	男声错误率：10.7%
女声正确率：89.5%	女声错误率：10.5%

7. 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。

当然如果我没有进行归一化操作，高斯贝叶斯模型的正确率只有 60%左右，这也进一步说明了归一化的必要性。

对于以上的模型所得到的正确率并不能满足很高的准确率，于是我开始思考算法模型的优化方式，最后我决定还是从最开始的数据分析开始优化。

8. 算法优化

8.1 数据可视化分析

仔细思考过高斯朴素贝叶斯的算法原理之后，我发现高斯朴素贝叶斯的前提要求是比较苛刻的，要求连续型变量分布满足正态分布并且每一个特征都是互相独立的，于是我决定借助 seaborn 图形库对每个特征的数据进行分析：

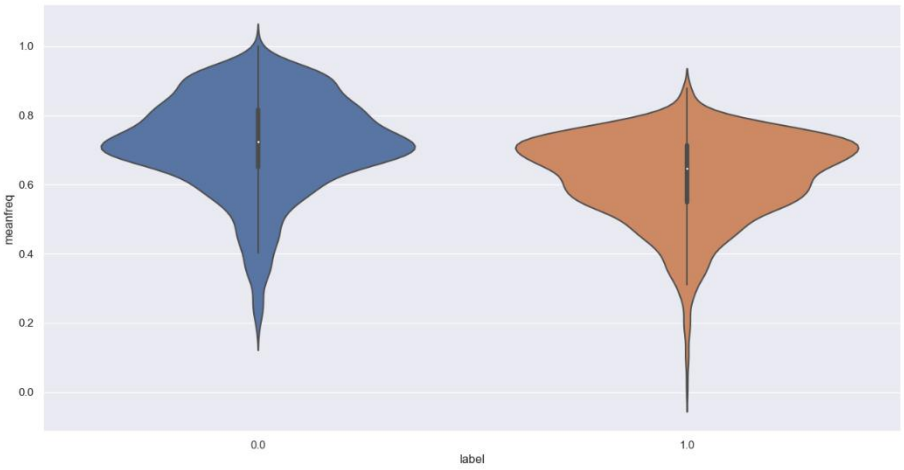


图 8 meanfreq 特征的数据分析

这里我就把 meanfreq 这个特征作为一个例子：首先我们的 label 轴代表性别，1.0 代表男性，0.0 代表女性。Y 轴代表归一化后的值。图的宽度代表了数据量分布的多少，中间的白点代表了该组数据的中位数。该图很好的反映了特征的分布情况。

8.2 优化策略及结果

我们为了满足朴素贝叶斯的假设前提，要挑选出大致符合正态分布的特征出来。此外，为了保证精度能有提高，我们还得尽量选取那些中位数相差较大置信区间重合小，分布集中的特征来进行分类，这样才能很好的体现两个性别在该特征上的差异性。

最后我筛选出了四个特征出来，分别是 meanfun,Q25,IQR,sd：

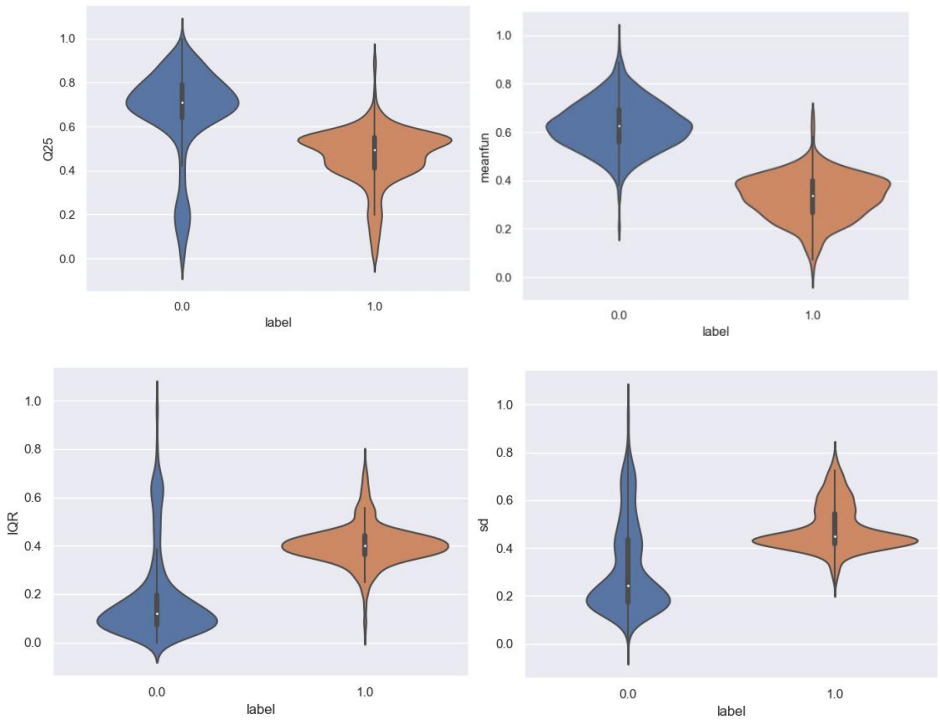


图 9 筛选出的四个声音特征的数据分布

对于筛选出来的这四个声音特征，我专门查阅了有关资料，了解到这里的四个特征的具体含义如下：

表 2 筛选出的四个特征的含义

声音特征	含义	形象化描述
Q25	第一分位数	代表了处于总数 1/4 位置的频率
meanfun	基频平均值	相当于代表了人声音的高低
IQR	分位数范围	代表了 1/4 和 3/4 分位点的频率差值
sd	频率标准差	代表数据在分布在均值附近的集中程度

从这里可以看出，我选择出来的四个特征，男性和女性的分布区别度都相差较大，具有比其他特征更加良好的区分度。我们改用这四个变量作为训练数据对模型进行再次的训练，得到训练结果如下图：

	precision	recall	f1-score	support
female	0.962963	0.958785	0.960870	461
male	0.961382	0.965306	0.963340	490
micro avg	0.962145	0.962145	0.962145	951
macro avg	0.962173	0.962046	0.962105	951
weighted avg	0.962148	0.962145	0.962143	951

图 10 四特征的改进贝叶斯算法结果

通过对生成的结果报告分析，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%左右，相比于之前的 89%的准确率，提升明显。为了查看训练过程拟合情况，我接着做出了四特征的学习曲线：

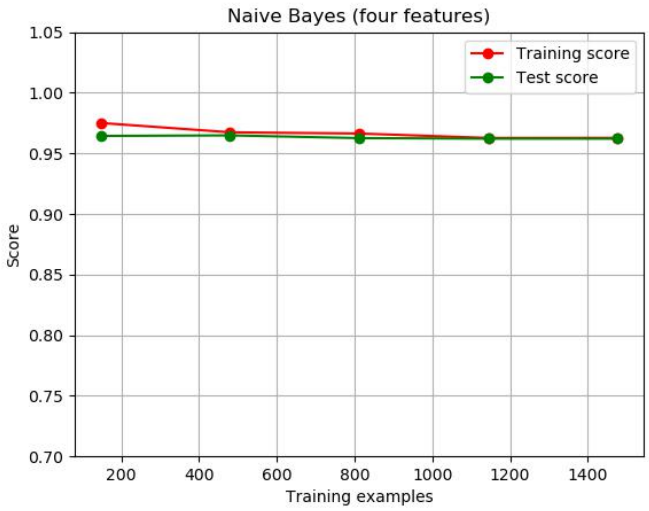


图 11 四特征的改进朴素贝叶斯模型学习曲线

对于选择的这四个变量，我分别由进行了探究，发现 meanfun 这个特征对于结果的影响是最大的，如果只是单一的选取 meanfun 这个特征进行训练，也可以达到 93%左右的准确率。结合我的现实生活也证明了确实分辨男性和女性声音的主要判断依据是声音高低。男性声音相对更低沉，基频 meanfun 更低，女性声音相对比较高，基频 meanfun 更高。

表 3 优化高斯朴素贝叶斯训练结果

男声正确率：96.1%	男声错误率：3.9%
女声正确率：96.3%	女声错误率：3.7%

9 拓展延伸

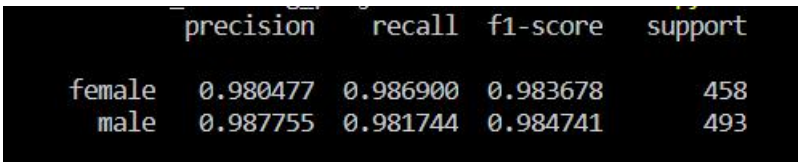
9.1 模型对比

对于上述朴素贝叶斯模型机器算法优化得到的准确率我感觉仍有提升的空间，于是决定尝试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

KNN 算法原理比较简单，直接通过计算距离最近的 K 的点来做出预测。

KNN 算法由于没有训练过程，我们这里还是选取 7：3 切分训练集和数据集，K=3 时生成的训练报告见下：



	precision	recall	f1-score	support
female	0.980477	0.986900	0.983678	458
male	0.987755	0.981744	0.984741	493

图 12 KNN 算法训练结果

这里我们做出 K 取不同值的时候的模型准确率如下图：

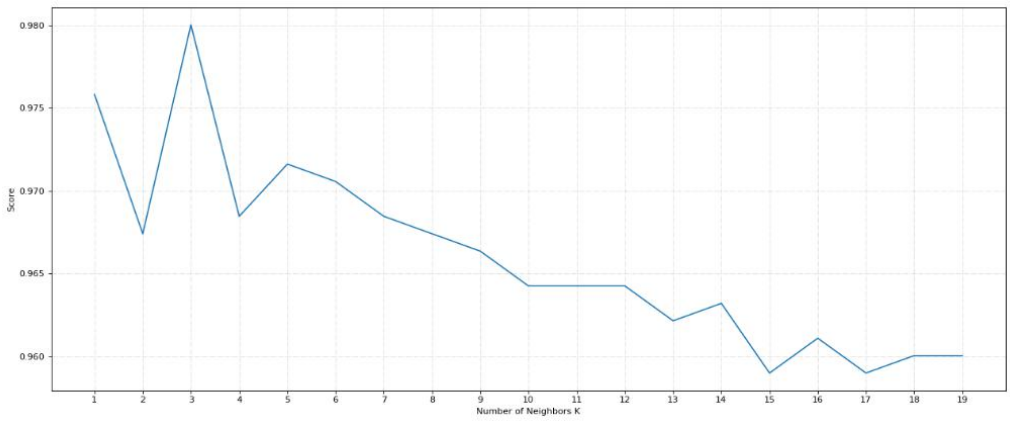


图 13 不同 K 值的模型准确率变化曲线

这里我们可以发现 K 取 3 的时候，模型准确率最高达到了 98%左右，可以发现相比之前的朴素贝叶斯有了很大提升。

（2） Logistic Regression

逻辑回归模型引入了 sigmoid 函数，对于数据中的一些小噪声产生的影响更小，训练速度较快，这里我生成的训练报告如下：

	precision	recall	f1-score	support
female	0.976242	0.969957	0.973089	466
male	0.971311	0.977320	0.974306	485

图 14 Logistic Regression 算法训练结果

对于逻辑回归模型，我这里同样也绘制出它的学习曲线如下图：

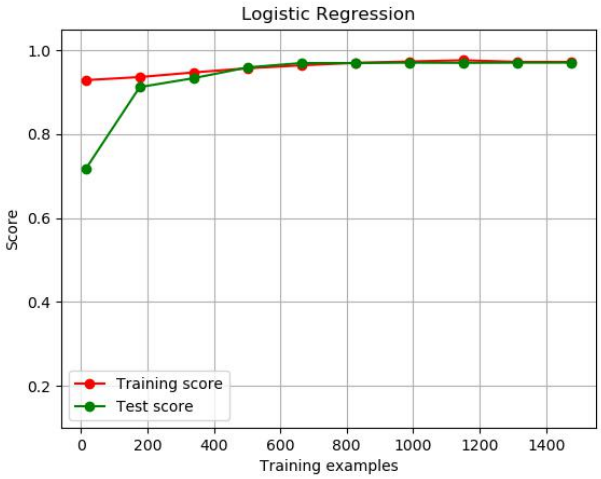


图 15 Logistic Regression 算法学习曲线

这里我可以发现逻辑回归模型也是快速收敛最后趋于稳定下来，证明拟合效果比较理想。

（3） Random Forest

随机森林模型相当于是基于决策树模型的一个优化，通过生成的多个决策树，对测试数据进行预测，最后统计整个森林中所有树的预测值选取预测值较多的作

为预测结果。对于一些非线性数据效果不错，并且可以改变训练森林中树的个数来减小一些噪声的影响，达到提升准确率的效果。

随机森林模型方便我处理高维数据并且不需要降维，对于这里的 20 维特征数据比较适合，这里我们选取的 n_estimate 参数选择的是 20，因为即使再增加树的颗树，由于这里只有 20 维特征，即使再提升 n 的值，对于准确率的提升并没有十分的明显，这里我生成的训练报告如下：

	precision	recall	f1-score	support
female	0.981013	0.976891	0.978947	476
male	0.976939	0.981053	0.978992	475

图 16 Random Forest 算法训练结果

对于随机森林，我这里同样也绘制出它的学习曲线：

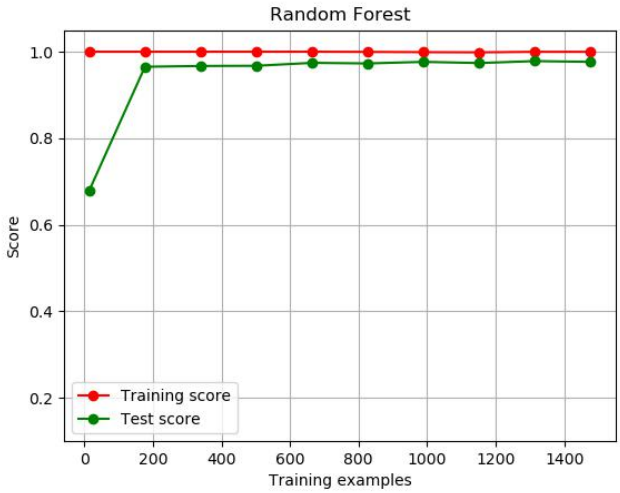


图 17 Random Forest 算法学习曲线

这里我可以看到随机森林算法的训练得分一直在 1.0，后来我查阅有关资料得知随机森林算法由于其特殊性，训练分数一直会保持在 1.0，因为在这个过程中每颗决策树的参数会随着训练数据调整权重而满足该训练数据。当然这里的学习曲线还是收敛到 0.98 附近，保持稳定，证明拟合效果比较好。

9.2 模型结果差异分析

首先，对于我这次项目采用的 5 个训练方法做出结果比对图：

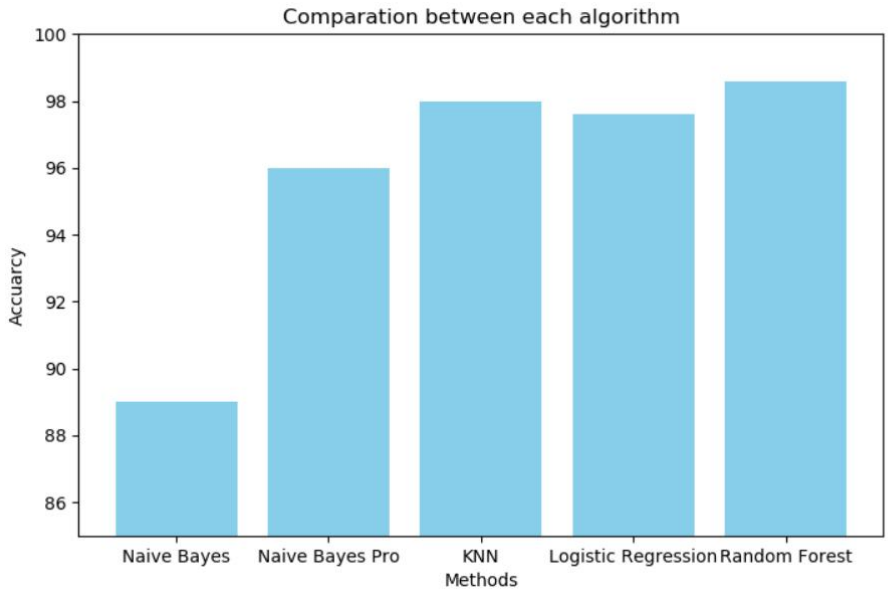


图 18 5 种实现准确率比较

从上面的柱状图我可以发现最开始的全特征朴素贝叶斯准确率是最低的，特征选择优化后的朴素贝叶斯效果比较理想，但是 K=3 的 KNN 和逻辑回归和随机森林明显要高出一些，针对这个情况，我进行了查阅了想到了朴素贝叶斯模型假设的另一个前提：特征独立性，于是我输出了 20 个特征之间的相关系数：

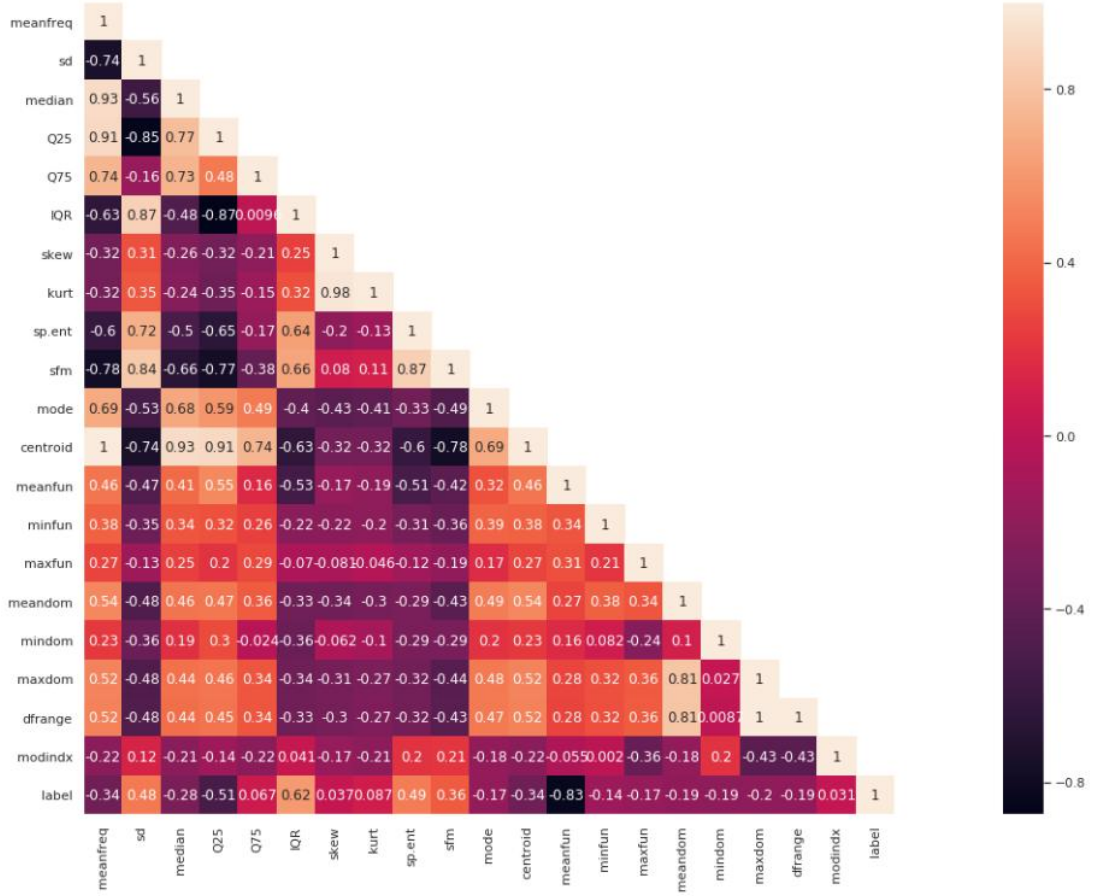


图 19 20 个特征之间的相关系数热图

很明显我们看到这里特征之间的相关系数并不为 0，代表特征之间并不是独立的这也是为什么我的朴素贝叶斯模型即使在优化过后仍然不能达到其余三个模型的准确率的原因。

10 总结

这次任务，我的探究过程比较曲折。首先我是根据实验要求直接使用高斯朴素贝叶斯算法训练准确率最初只有 60%左右，后来增加了标准化操作之后把准确率提升到了 89%左右。由于朴素贝叶斯模型原理简单，在优化模型的时候，我发现必须得从模型假设前提出发，对数据进行筛选处理，尽量挑选满足模型假设的数据，最后得到的训练模型准确率提升到了 96%左右。之后的模型对比我也是使用了本次课程中老师所讲的 KNN，逻辑回归，随机森林等算法，重新动手操作并深入理解了这些算法的原理，并且在模型对比分析之中，了解了不同模型对于不同数据类型的适用性。

除了学习 python 和算法模型外，我专门花了 2 天的时间学习了基于 matplotlib 的 seaborn 可视化库。发现了这个可视化库可以很好的帮助我做出清晰明了的数据分析图，帮助我在模型优化选取策略中一目了然的分析出每个特征的特点。此外结合 sklearn 库中的一些数据分析的函数，也可以帮助我快速生成训练报告和学习曲线，让我更快的了解训练结果和训练过程中的模型拟合效果。

这次任务，我最后还是去 cmu 的语料库中找到了每条语音数据的来源，自己听了之后，发现确实有些语音数据我也不太好区分是男性还是女性。对于每个声音特征的意义，我也进行了探究，懂得了每个特征的具体含义和作用，对数据集的理解更加深入了。

我这次任务的最大的感悟是机器学习，除了有合适的模型外，一定要有数据分析的能力，能够根据所选择的模型对数据进行针对化处理，只有这样才能够做出比较好的优化模型。

参考文献

[1] https://zhuanlan.zhihu.com/p/64498790
[2] https://blog.csdn.net/GoodShot/article/details/80373372
[3] https://blog.csdn.net/iModel/article/details/80042862
[4] https://www.kaggle.com/primaryobjects/voicegender
[5] 周志华.机器学习[M].北京:清华大学出版社
[6] 李航.统计学习方法[M].北京:清华大学出版社

项目代码见：https://github.com/yux20000304/Machine-Learning-Project