机器学习纳米学位

声音性别识别

Fayren udacity

April 10th, 2018

一、问题的定义

1.1 项目概述

声纹性别识别是一个复杂的工作，作为人类本身可以很容易的听出来男生或者女生，但是机器如何能分辨男性声音和女性声音呢？

一方面，在机器学习技术风行之前，大多采用硬编码的方式来做声音性别识别，也就是在编写程序时，事先分别定义好男生女生的声音特点，再用写好的程序去匹配声音特点，这种传统的编写规则程序不仅区分准确率很难保证，规则编写更是非常复杂。

另一方面，随着机器学习技术的发展，使用机器学习方法判断一段音频是男性发出还是女性发出变得容易了许多。通过对声音数据的学习，程序能在大量的数据中统计学习得到女性男性的声音特点。再在习得的经验基础上再来做声音识别，这样不仅大大提高了识别的正确率，同时还不用编写复杂的规则，程序也拥有更灵活适应数据的特点。

本文旨在阐述一种使用机器学习方法来进行语音性别识别的方法。通过训练一个机器学习模型，能让程序自动区分一段声音是来自男性还是女性。

1.2 问题描述

声音性别识别也就是通过对声音文件的分析让程序自动识别。当然直观上，我们需要探索一种输入为声音文件，输出性别的方案。如下图所示：

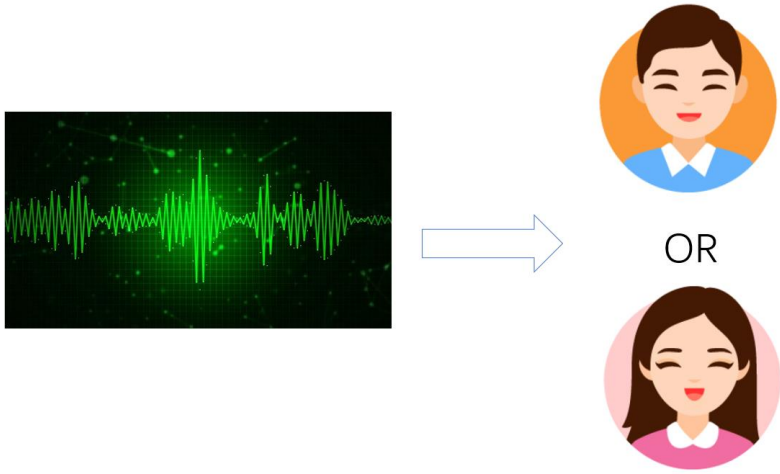
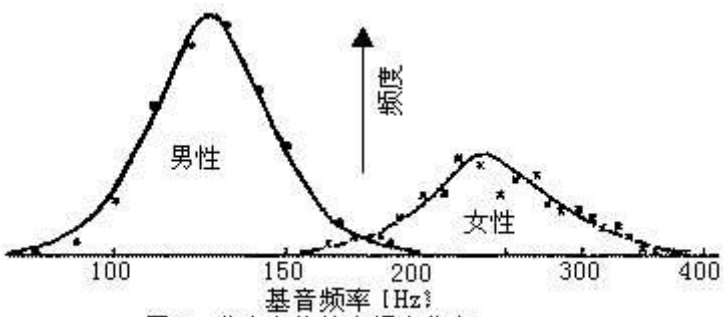


图 1.1 端到端的声纹识别

由于直接声音文件不容易直接使用作为机器学习算法的输入，这里还需要将声音文件预处理，从而获得声音的特征信息，如频率，响度等特征额。以频率为例，男性声音和女性声音的差异如下图所示：



图二 发声者的基音频率分布

可以看到男性的频度普遍分布在频率较低的区域，而女性则普遍分布在频率较高的区域。当然，仅凭一个频率判断声音是男性还是女性是不够精确的，如上图的分布交叉处。但是当我们有更多的声音特征时，从更多个方面去分析声音特点则能让我们更精确的判断声音的性别。

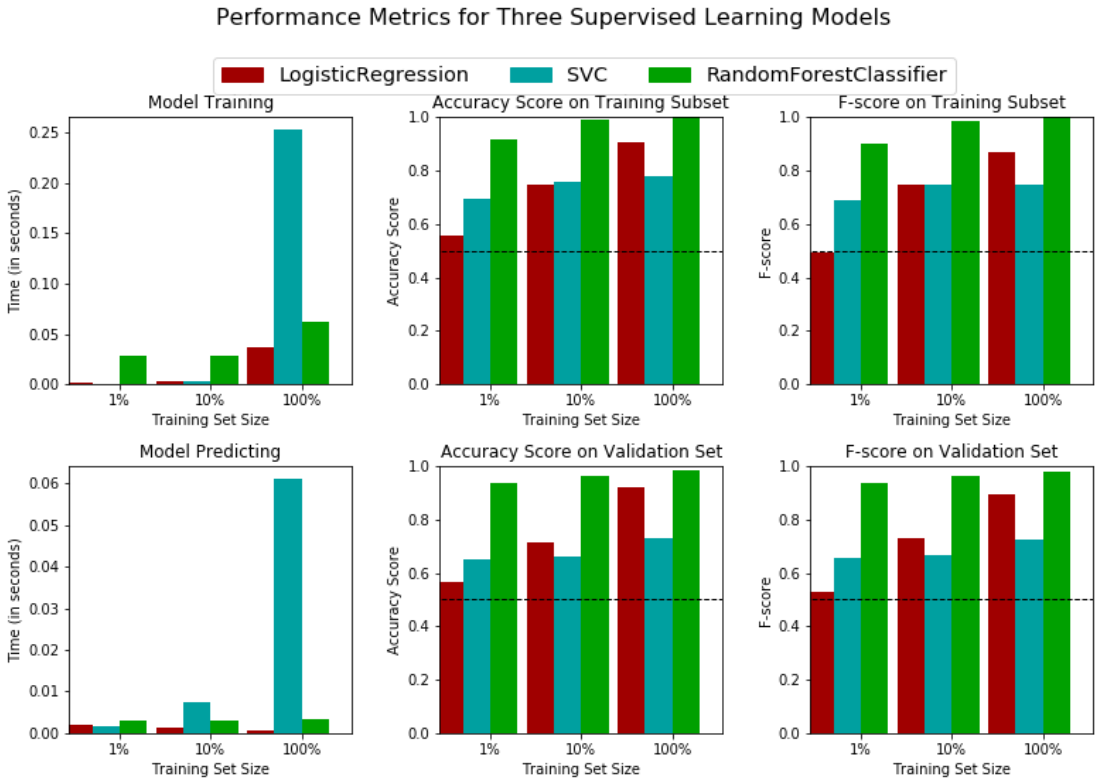
1.3 评价指标

本文主要涉及四种模型，评估上主要以训练集正确率和验证集正确率来评估模型性能。评估公式：

$$accuracy = T/(T+F)$$

其中 T 表示判断正确的样本数量，F 表示判断错误的样本数量。

最后输出如下图所示的正确率图：



二、分析

2.1 数据探索

本文将根据以下特征来研究如何区分男性声音和女性声音：

- meanfreq: 频率平均值 (in kHz)
- sd: 频率标准差
- median: 频率中位数 (in kHz)
- Q25: 频率第一四分位数 (in kHz)
- Q75: 频率第三四分位数 (in kHz)
- IQR: 频率四分位数间距 (in kHz)
- skew: 频谱偏度
- kurt: 频谱峰度
- sp.ent: 频谱熵
- sfm: 频谱平坦度
- mode: 频率众数
- centroid: 频谱质心
- peakf: 峰值频率
- meanfun: 平均基音频率
- minfun: 最小基音频率
- maxfun: 最大基音频率
- meandom: 平均主频
- mindom: 最小主频
- maxdom: 最大主频
- dfrange: 主频范围
- modindx: 累积相邻两帧绝对基频频差除以频率范围

2.2 探索可视化

本文在研究声纹特征时发现一般声纹特征分布如下：

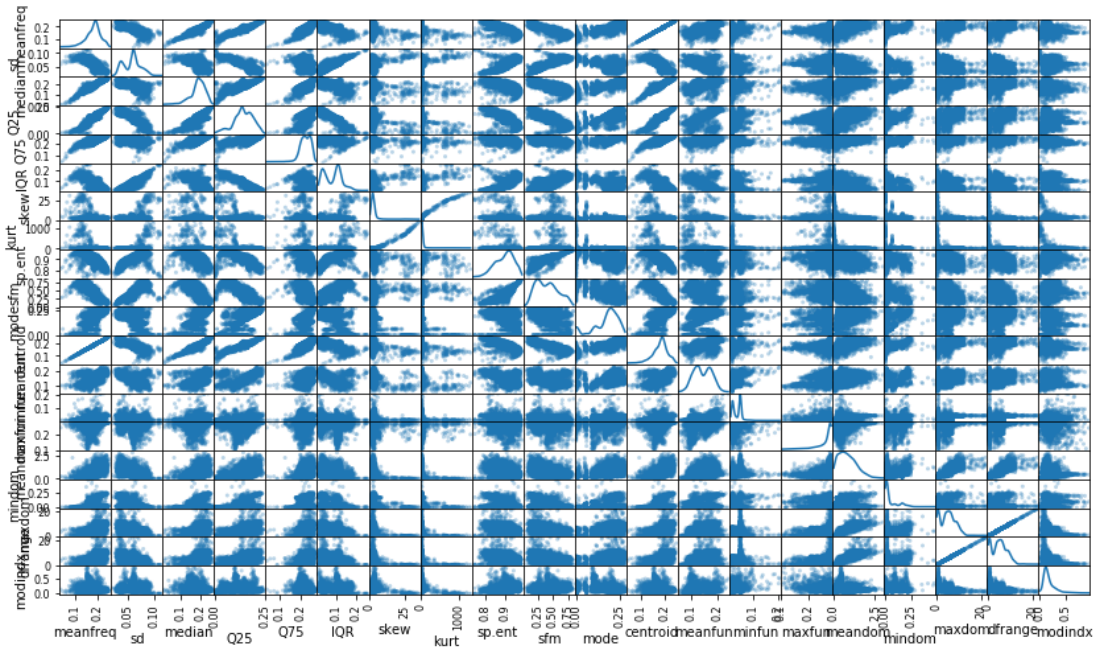


图 2.1 特征分布

如上图所示，在各个特征空间离，特征呈现正态分布的特点。

在背景介绍中，我们知道这里同样可以使用传统的编写规则的方式实现声音性别的区分，例如一般女生的声音音调比男生高，也就是频率会很高，对主频范围设置一个值，大于某个值的认为是女生。当然这么简单的规则肯定无法收获很高的正确率。虽然这里可以通过复杂一点的规则提高区分的正确率，但是编写规则将会是异常麻烦的。

所以，机器学习方法在这里就表现出它得天独厚的优势。通过对样本的分析，让程序习得男生女生的声音各个特征的一般分布，从而分析得到是男生还是女生的声音，避免了规则的编写，同时提高了程序的灵活性。

2.3 算法和技术

如前两章介绍，本文拟准备使用机器学习的方法来识别声音的性别。如下四种模型非常适合做对比试验。其中，随机森林，支持向量机，LR 回归为传统的机器学习模型，同时本文拟采用 DNN 神经网络试验。

- 随机森林（Random Forest)
- 支撑向量机（SVM)
- Logistic 回归（LogisticRegression）
- 深度神经网络（DNN）

其中 LR 和支持向量机可以直接使用，随机森林模型采用 GridSearch 方法找到最合适的决策树数量。DNN 神经网络拟采用如下的结构：

Layer (type)	Output Shape	Param #
the_input (InputLayer)	(None, 20)	0
dense1 (Dense)	(None, 128)	2688
activation_21 (Activation)	(None, 128)	0
dropout_19 (Dropout)	(None, 128)	0
dense2 (Dense)	(None, 256)	33024
activation_22 (Activation)	(None, 256)	0
dropout_20 (Dropout)	(None, 256)	0
dense3 (Dense)	(None, 256)	65792
activation_23 (Activation)	(None, 256)	0
dropout_21 (Dropout)	(None, 256)	0
dense4 (Dense)	(None, 128)	32896
activation_24 (Activation)	(None, 128)	0
dropout_22 (Dropout)	(None, 128)	0
dense5 (Dense)	(None, 128)	16512
activation_25 (Activation)	(None, 128)	0
dropout_23 (Dropout)	(None, 128)	0
dense-final (Dense)	(None, 1)	129
sigmoid (Activation)	(None, 1)	0

Total params: 151,041
Trainable params: 151,041
Non-trainable params: 0

图 2.2 DNN 网络结构图

2.4 基准模型

在实验时，本文着重使用随机森林模型，并为其设置正确率 98%的基准阈值。参考 kaggle 数据竞赛上的基准，如下表：

模型	正确率
Baseline(总是认为是男性)	50%
逻辑回归	97%
CART	97%
随机森林	98%
SVM	99%
XGBoost	99%

图 2.3

三、方法

3.1 数据预处理

本文实验中用到的数据集包含 3168 个样本，其中 50%为男性，50%为女性。样本数据中包含问题描述中提到的各个声音特征以及该声音对应的性别。
由于本文采用机器学习的方法来对声音性别进行识别，在训练之前的数据处理时，主要做数据做了一下操作：

1. 原数据将 label 从其中分离，分别得到训练数据和 label。
2. 将源数据去除事先随机预留了 20%的数据作为测试集，用来衡量模型的性能。
3. 其次剩下的 80%数据中的 20%数据用来做训练时的验证集，用来找出模型的最佳效果。剩下的 80%数据作为训练数据

3.2 执行过程

由于问题的复杂性，难以编写规则来直接分类声音，本文使用机器学习的方法来实现声音性别的分类。项目的总体训练流程如下：

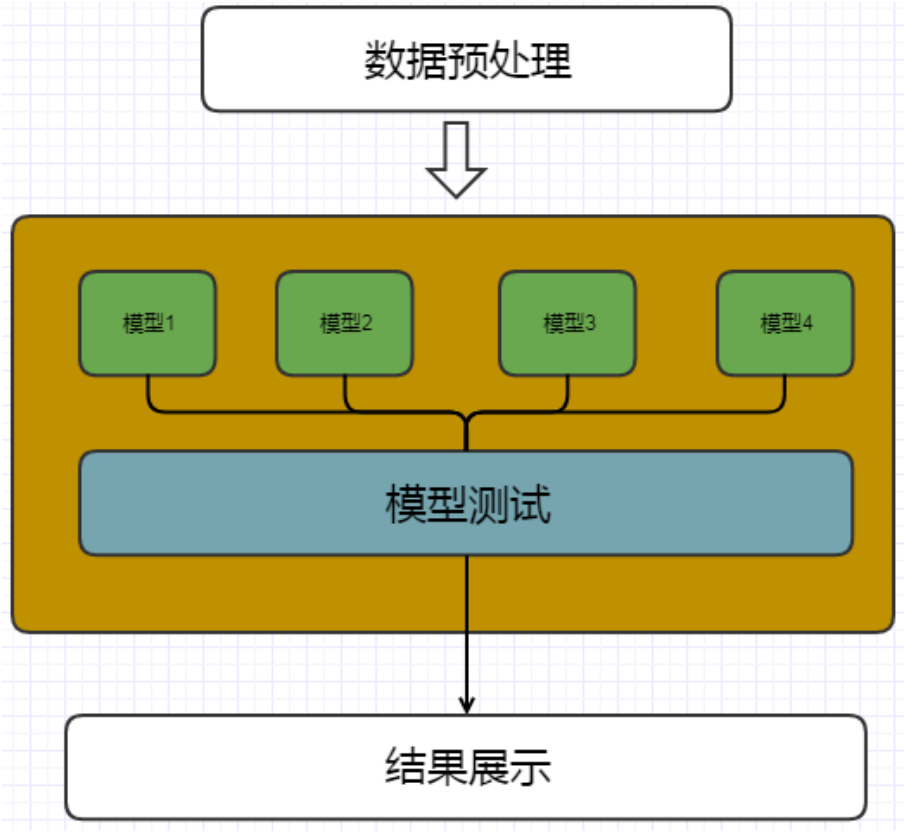


图 3.1 总体过程流程图

主要包含以下几个步骤：

1. 数据预处理，切割得到测试集，训练集和验证集。
2. 分别使用多个模型训练来拟合数据。
3. 评测各个模型的性能，对于每个模型选出最佳性能下的参数。

展示最终结果。

3.3 完善

值得注意的是，由于训练时随机森林算法的 n-trees 参数使用的固定值 5 进行的实验，这里使用 5 颗决策树不一定能得到最佳的识别效果，而使用 Grid-Search 方法则可以找到最佳的 n-trees 参数。

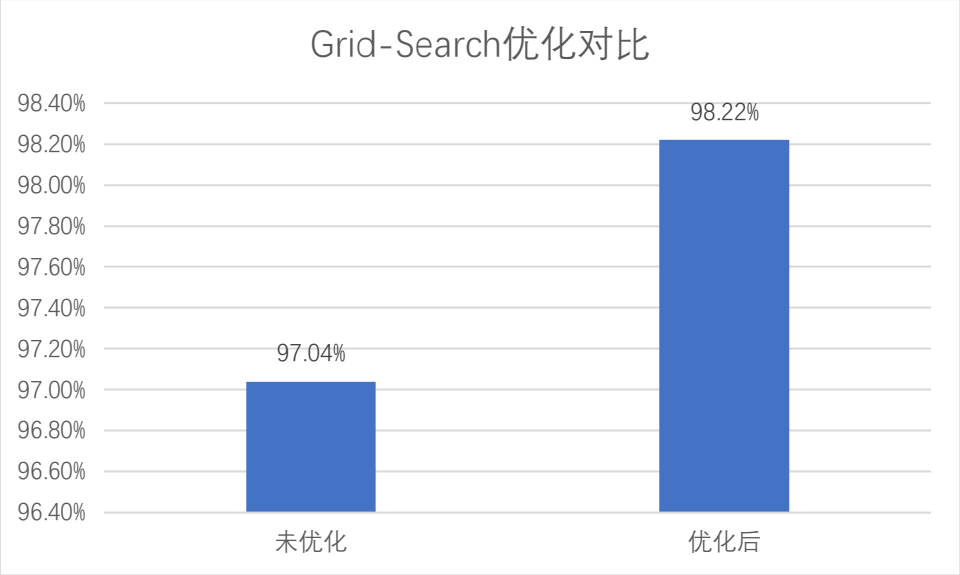


图 3.2

四、结果

4.1 模型评价与验证

本文主要采用随机森林模型做声音性别识别。在实验过程中，为了确保模型的有效性，分别尝试了 1%的数据量，10%的数据量以及 100%的训练集数据量做训练，具体表现如下图：

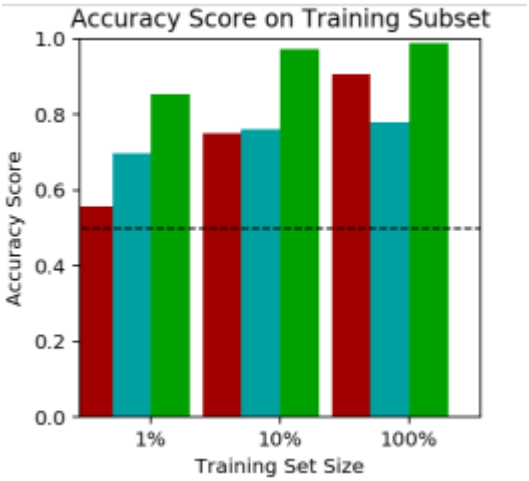


表 4.1

可以看出，随着训练数据集的增大，模型的表现越来越好，说明模型确实能从数据中习得数据规律。

另一方面，训练过程中设置了验证集，依此来验证模型是否能真正习得训练数据中的规律同时保证模型的鲁棒性。效果如下：

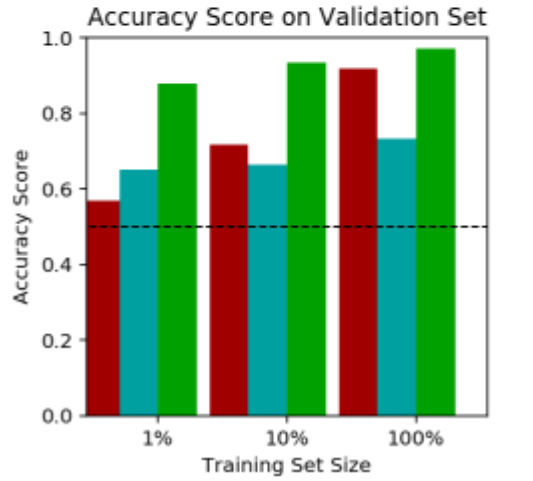


表 4.2

可以看出，训练集中的正确率和验证集中差不多，说明模型的鲁棒性没有问题。反之，则说明模型很可能过拟合或者欠拟合了。

4.2 合理性分析

从图 4.1 中分析对比可以看出，三个模型中随机森林模型的表现最为优异，其中 LR 逻辑回归模型存在明显的正确率不高的问题，对于声音性别识别来说 80% 多的正确率显得苍白无力。再看 SVM 模型明显存在训练复杂度高，耗时长的问题，同时在验证集的正确率较训练集存在明显的下滑，说明 SVM 出现了明显的过拟合现象。主要优劣对比如下表：

	逻辑回归（LR）	支持向量机（SVM）	随机森林（RF）
训练耗时	短	长	短
训练集正确率	90%	79%	99.1%
验证集正确率	90%	75%	98.2%
是否过拟合	否	是	否
增加训练数据是否有效	明显	不太明显	很明显

表 4.3

从图 4.2 中分析可以看出，当迭代次数达到 50 次之后模型开始收敛，之后的学习变得很缓慢，最终在 98%左右开始趋于平稳，考虑到更多的迭代只会让模型陷入过拟合，故在 300 轮迭代的时候终止了训练。

五、项目结论

5.1 结果可视化

本文在实验时前后选择了四个模型进行对比试验。三个传统的机器学习模型正确率如下图所示：

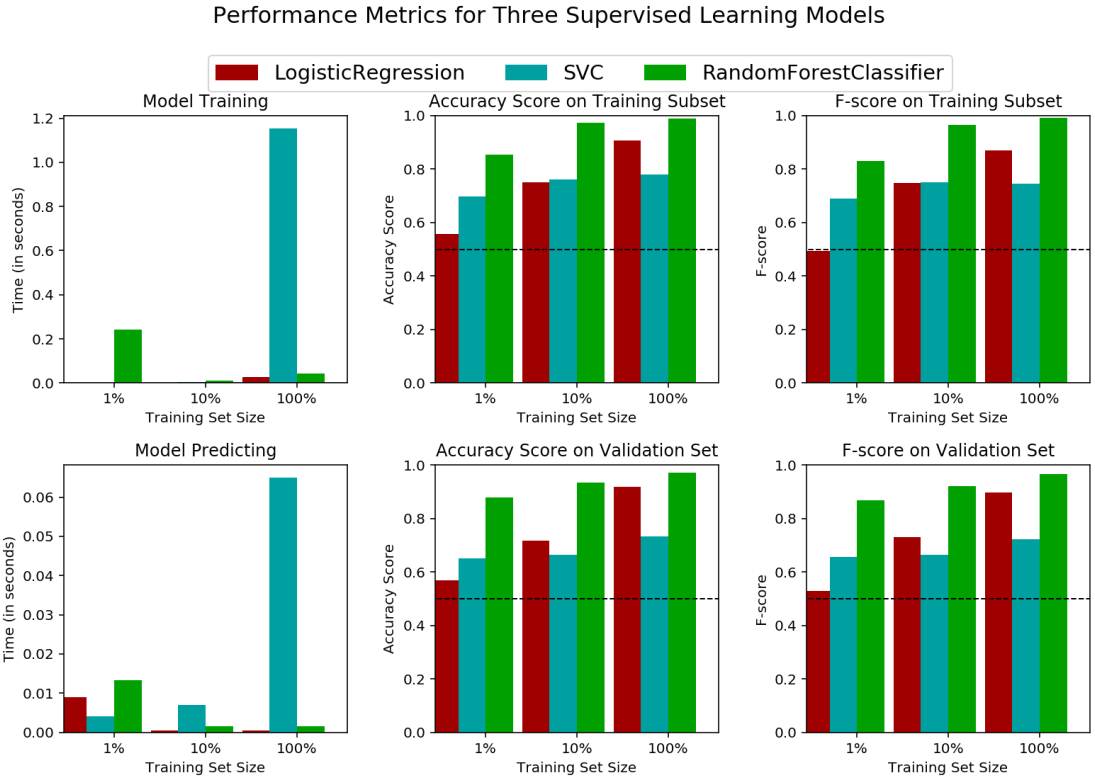


图 5.1

使用 DNN 神经网络的结果如下图所示：

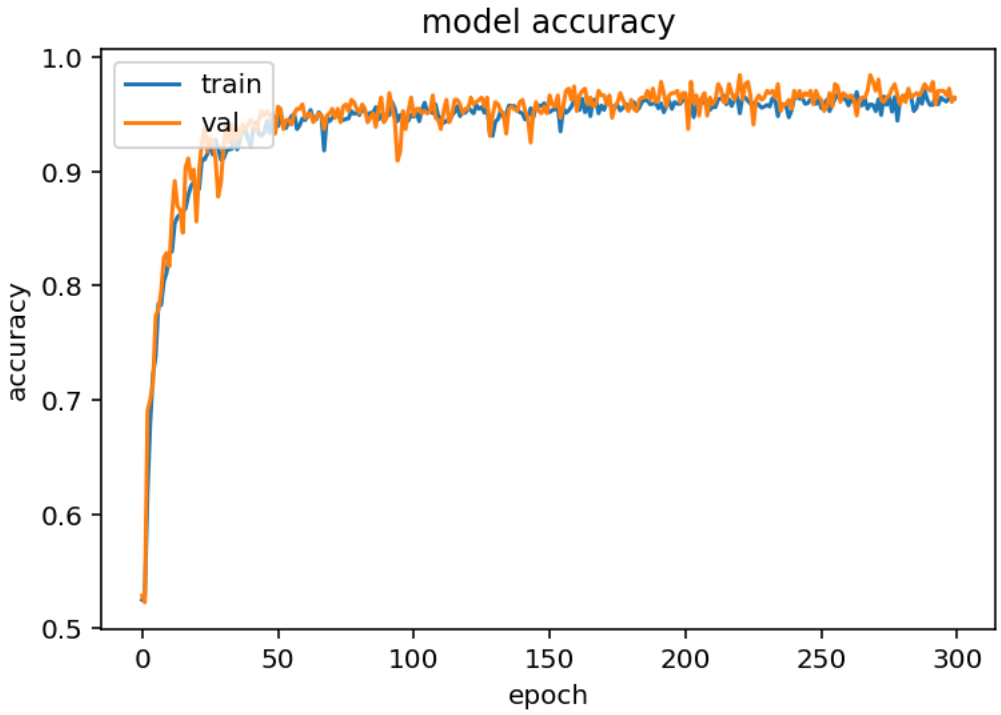


图 5.2

在经历了 300 个迭代之后，正确率开始稳定在 98.1%左右。

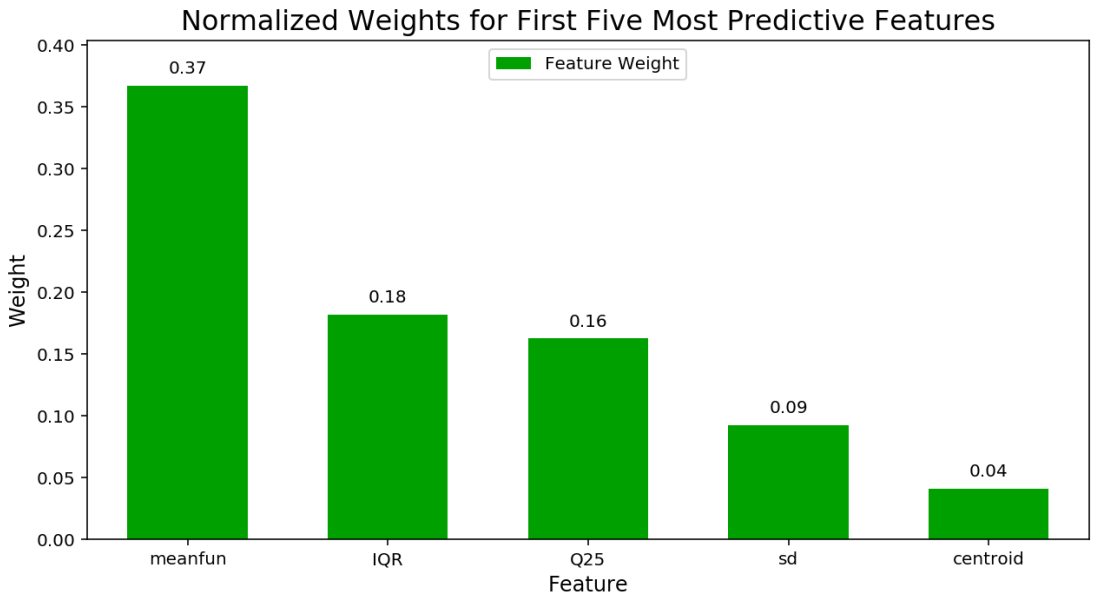


图 5.3 特征权重

通过将特征权重输出可以看到，meanfun,IQR,Q25,SD,centroid 这五个特征的权重最大，这五个特征对结果的影响力占到了 84%，可以看出比重非常之大。

5.2 对项目的思考

本文首先明确要解决的问题是实现一个能区分声音的性别的程序。其次，本文分析了实验中所获取到的数据，阐述了声音文件的特征抽取，这些特征对于性别的识别至关重要。紧接着是机器学习模型的选型，本文在实现上拟采用四种模型进行实验，分别是逻辑回归模型，SVM 支持向量机模型，随机森林模型和 DNN 深度神经网络模型。这些模型在已经获取到声音文件特征的情况下可以直接开始训练。最后，需要对本文所选择的模型优劣进行评估，针对前三种机器学习模型，使用传统的测试集正确率，验证集正确率和 Fscore 进行建成。

本文最有特色的是使用了多种模型进行对比试验，通过对比探索了各个模型在语音性别识别问题上的优势和劣势。

总的来说，模型的最终结果还是比较符合预期，98.22%的正确率可以达到使用的标准。

5.3 需要作出的改进

本文探讨了使用随机森林模型对声音进行性别识别，从 kaggle 数据竞赛上的结果可以看到，随机森林在正确率上面是比不上 SVM 和 XGBoost。一方面对比试验要优化参照模型如 LR，SVM 的性能才能凸显本文主体模型随机森林模型的能力，同时在优化本文主体模型随机森林的模型的时候还可以考虑优化一下数据预处理，优化数据，筛选和剔除异常数据。另一方面，可以尝试性能更加优秀的模型如 XGBoost。

参考文献

[1] Thomas S. Data-driven Neural Network Based Feature Front-ends for Automatic Speech Recognition[Ph.D. dissertation], Johns Hopkins University, Baltimore, USA, 2012.

[2] Wu Wei-Lan, Cai Meng, Tian Yao, Yang Xiao-Hao, Chen Zhen-Feng, Liu Jia, Xia Shan-Hong. Bottleneck features and subspace Gaussian mixture models for low-resource speech recognition. Journal of University of Chinese Academy of Sciences, 2015, 32 (1):97-102(吴蔚澜，蔡猛，田垚，杨晓昊，陈振锋，刘加，夏善红. 低数据资源条件下基于 Bottleneck 特征与 SGMM 模型的语音识别系统. 中国科学院大学学报, 2015, 32 (1):97-102)

[3] Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors. Computer Science, 2012, 3 (4):212-223

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