Machine Learning Engineer Nanodegree

Capstone Proposal

Hanwei Zhu
April 19st, 2018

Proposal

Domain Background

随着人工智能的发展，人机语言交互已经成为重要的研究领域之一。Apple公司的Siri，Microsoft的Cortana，Google的Google Assistant和Amazon的Alexa等等开发的虚拟AI助手都已经投入商用并取得了巨大的成功。这都是以语音识别技术作为基石。可以想象未来基于语音识别的各种场景：对话的方式来完全地取代人工键盘输入；声控智能家居；声纹加密；在现有的技术支持下实现这些场景似乎已经指日可待。

现有的研究指出传统的语音交互有以下几种缺点：第一，交互距离要近；第二，发音必须标准；第三，环境必须安静；第四，人机不能持续对话[1]。随着算力的提升和人工智能研究的发展，现有的机器学习，深度学习技术等等都能够很好地改进以上几点。这也是本次项目探索的重点和方向。

Problem Statement

本次毕业项目将实现一个简单的语音识别器：对不同的声音样本进行性别识别。该项目基于Kaggle竞赛赛题“Gender Recognition by Voice”。根据Kaggle上竞赛举办方提供的声音数据集，搭建模型来识别采集到的声音是男性还是女性。针对这个问题，本次项目将采用现有的机器学习算法——特别是监督学习来解决。该问题本质上属于监督学习中基本的二元分类问题，即根据输入的一系列特征输出二元标签。在给出人工标签的数据集上训练模型，通过减小和标签差距的方式来找到模型的最佳参数。最后通过准备好的测试数据来衡量模型的好坏。

Datasets and Inputs

本次项目使用了Kaggle竞赛“Gender Recognition by Voice”提供的数据集进行实验。现有的数据集提供了`voice.csv`文件。`voice.csv`文件内一共包含了3168个样本，其中50%为男性，50%为女性。其中每条记录包括：

- 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
- label: 男性或者女性

以上特征都是一段语音进过筛选后的重要信息，已经通过了R语言脚本对原始的音频进行了预处理[2]。项目将基于这一数据集进行实验，同时将其拆分为训练集和测试集对模型进行训练以及验证。

Solution Statement

现有的许多机器学习算法（特别是监督学习）都能够为本次项目研究的问题提供解决方案。现有的大多数监督学习方法都是通过最大似然估计以及正则化等方式来最小化损失函数。针对已有的数据集，我们可以采用例如使用逻辑回归，随机森林，XGBoost等模型进行预测。

以逻辑回归为例：我们通过赋予声音信息的每个特征一个权重并将结果相加后，用sigmoid函数进行非线性变换。再以交叉熵作为损失函数对预测标签（模型预测是男性或女性）和真实标签（实际是男性或女性）作为损失函数，最后用梯度下降的方式不断调节各项权重直至损失函数收敛。训练完成后我们可以在测试集通过得到的各项指标（例如f1-score，accuray等等）来对模型进行评估。

Benchmark Model

实验采用naive分类器——总是预测输入为男性（或女性）作为baseline。这意味着naive分类器的accuracy能够达到50%（因为数据集中男性和女性的样本各占了50%）。如果我们的模型能够大于这一评估指标，说明模型至少比随机预测（50%）的效果要好。

Evaluation Metrics

现有多种方式能够衡量模型的好坏。常用的方法有f-score，混淆矩阵等等。通过这些衡量指标我们能够清楚地得到模型在各方面的表现。

Project Design

我将采用在纳米学位的课程中学习到的标准数据处理流程进行实验。具体实验过程如下：

- 数据预处理

  包括数据清洗，标签编码，数据标准化，将数据集拆分成训练集，验证集和测试集。

- 模型训练

  可选择的模型包括：逻辑回归，决策树，随机森林，支持向量机，神经网络，XGBoost等等。并使用交叉验证的方式进行参数选择。

- 模型评估

  使用f1-score，混淆矩阵等对模型的预测结果进行评估。

- 结果可视化

  使用可视化工具对结果进行可视化分析。

Reference

[1] http://www.eepw.com.cn/article/201704/358275.htm

[2] http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/