Gender Detection

Selen Akkaya s289332

January 2023

Abstract

In this paper, 'synthetic speaker embeddings that represent the acoustic characteristics of a spoken utterance' is analyzed and a gender classification task is applied by building commonly used machine learning algorithms. Moreover, the performances of applied machine learning models and the comparison of models are analyzed.

1 Introduction

1.1 Problem Overview

The data-set contains synthetic speaker embeddings which represent the acoustic characteristics of a spoken utterance. Each row corresponds to a different speaker and contains **12 features** followed by the gender label:
1: female,
0: male
The features do not have any particular interpretation. Speakers belong to four different age groups. The age information, however, is not available.
The **training set** consists of **3000** samples for each class, whereas the **test set** contains **2000** samples for each class.

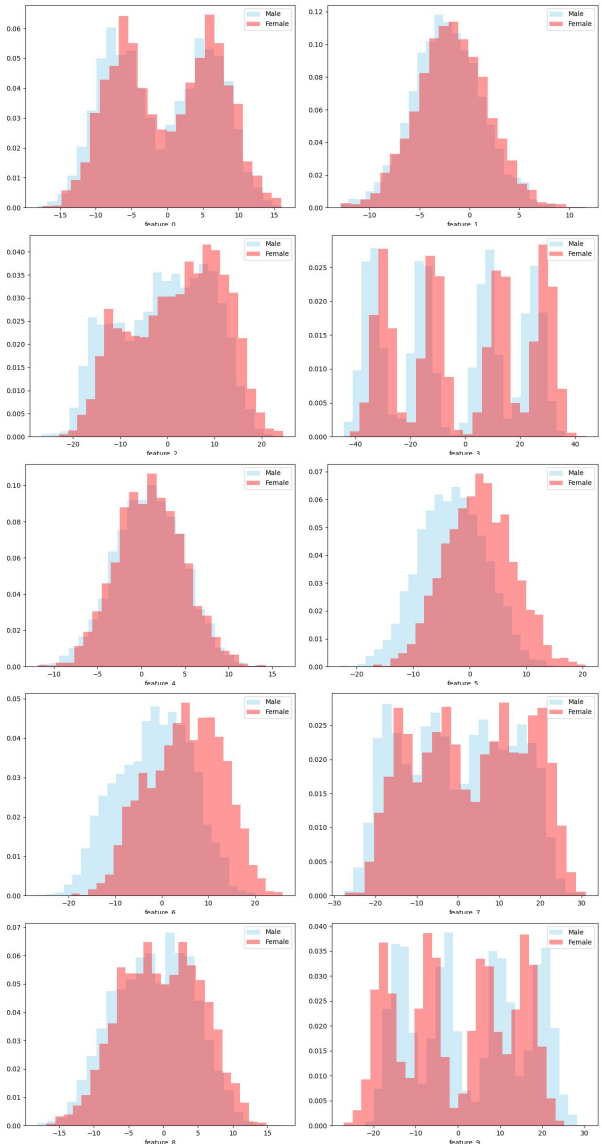
1.2 Exploratory Data Analysis

The 12 features are in a scale that have considerably similar means and variances, so it does not worth to apply Z-normalization which is basically centering every feature to its mean and scaling to unit variance xi = (xi-μ) / σ

μ : [-0.40439904, -1.98045219, 0.84747715, -2.37863374, 0.97348671, -0.72096827, 1.684338, 1.49200716, -0.8046595, 1.31572434, -0.07712583, 1.00468738]

σ : [7.09209235, 3.52880203, 9.8027367, 23.02600239, 3.85825232, 6.35299195, 8.5832784, 13.35106596, 5.71561775, 13.29758557, 10.69372272, 6.69376688]

Histograms of each 12 features (raw data) are shown below as Figure 1. It is obvious that raw features have approximated Gaussian distribution.



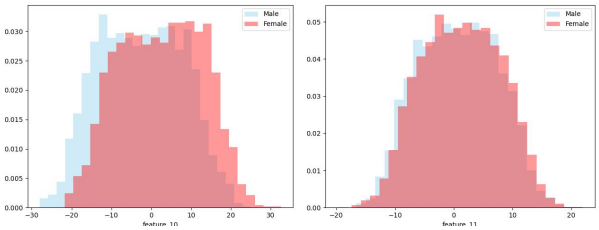


Figure 1: Raw Features

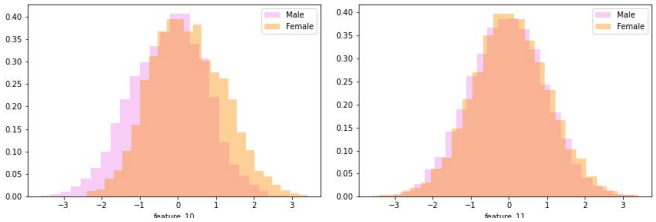
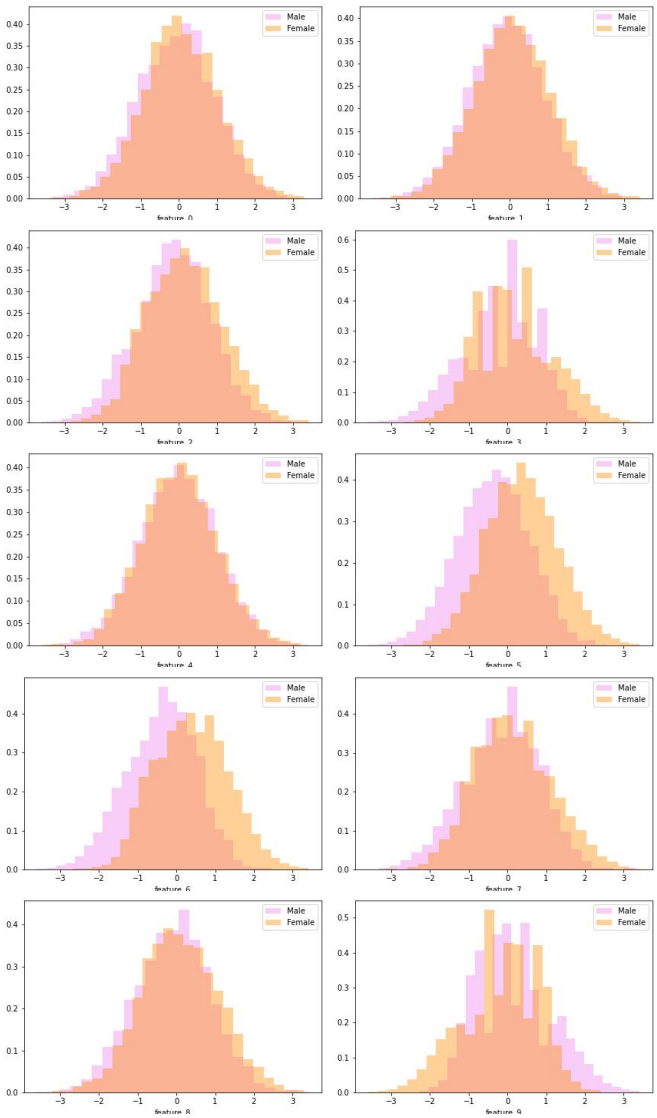


Figure 2: Gaussianized Features

However, to approve this idea, histograms of Gaussianized features are plotted to demonstrate. Every 12 features with Gaussianization as it is shown in Figure 2.



The gaussianization did not improve the histograms. Using raw data is better in this case. Especially feature-3 and feature-9 show how gaussianization worsens the result compared to the raw data.

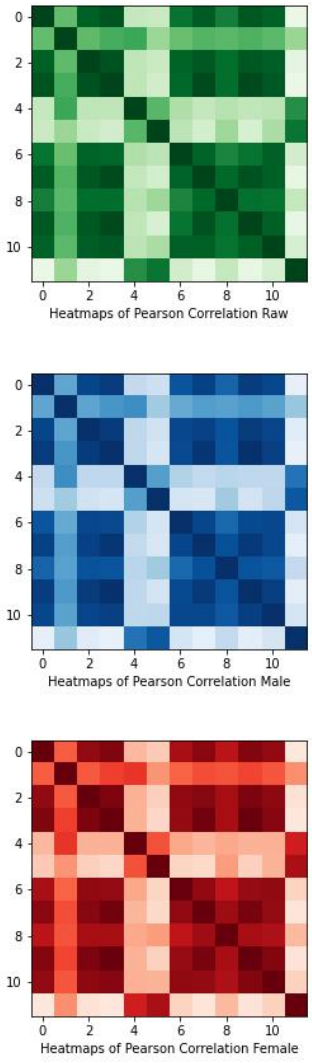


Figure 3: : Heatmap - Pearson Correlation

Pearson Correlation Heatmap shows that there are strongly corrolated features for instance, feature 3 is highly correlated to the 0, 2, 7 and 9. Also feature 10 is corraleted with 0, 2, 3, 6, 7 and 9. Therefore, we can benefit of PCA to reduce dimention and map data to less correlated features.