Machine Learning and Pattern Recognition Gender Identification Project

Nicolò Caradonna - s316993

September 2023

1 Introduction

1.1 Abstract

The objective of this project is to create a classifier that can determine the gender of individuals based on high-level characteristics extracted from facial images.

Our dataset is comprised of image embeddings, which are compact representations of images achieved by mapping facial images to a shared, low-dimensional space, typically with just a few hundred dimensions. To maintain the feasibility of our model, we are working with synthetic data, resulting in embeddings with considerably lower dimensions compared to real-world scenarios.

1.2 Consideration on Dataset

The dataset contains **12-dimensional vectors**, consisting of continuous values, belonging to either the **male** category (labeled as 0) or the **female** category (labeled as 1). It's important to note that the individual components of these 12-dimensional vectors do not carry any specific real-world meaning so they don't have any physical interpretation.

In the file named **"Train.txt"**, you'll find the embeddings that can be used to construct the classification model, which serves as our training set. On the other hand, in the file labeled **"Test.txt"** you will discover the embeddings used for evaluation purposes, constituting our evaluation set.

Each row in both files represents a single sample.The features of each sample and their corresponding labels are separated by commas within each row. Specifically, the first 12 columns of each file contain the sample components, while the last column holds the corresponding label.

- The training set contains **2400** samples divided in **720** male and **1680** female. We can notice that the training set is very unbalanced with 70% of samples belonging to female class. The samples belong to 3 different age groups. Each age group may be characterized by different distributions for the embeddings, unfortunately the age information is not available.
- The test set contains **6000** samples divided in **4200** male and **1800** female. We can notice that the test set is very unbalanced with 70% of samples belonging to male class, therefore the proportion of samples in training and test sets is very different.

1.3 Feature Analysis

In the following section, some preliminary feature analysis are presented so that we can understand which models might be most suited for our problem.

1.3.1 Histograms and Scatter Plots

Plot of distributions and scatter plots are shown here. Note that before plotting the distributions, data have been centered.

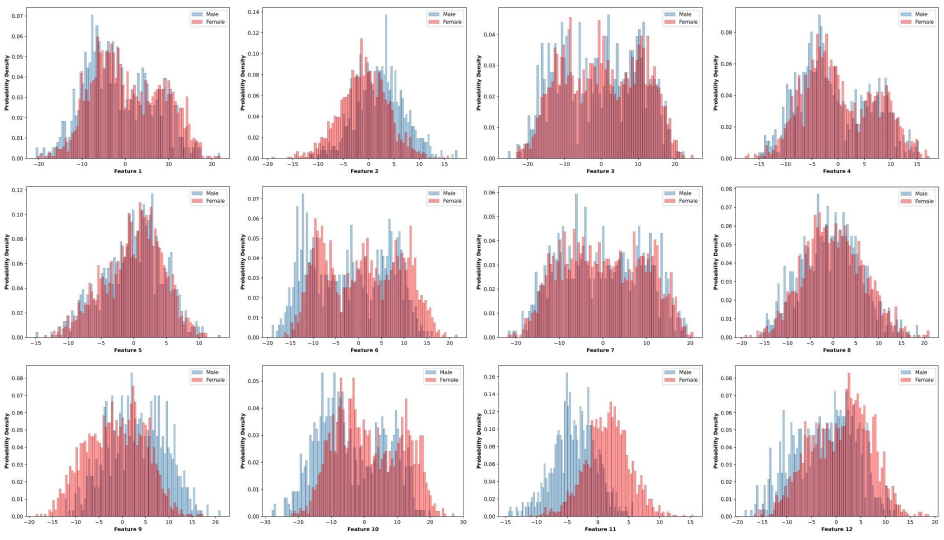


Figure 1: Feature Distributions

From an initial analysis, it is evident that some of these features can be accurately described by Gaussian distributions, particularly features 2-5-8-9-11-12 have a distribution resembling a one-dimensional Gaussian. Other features, however, appear to be distributed as a composition of multiple Gaussians, This is evident with features 3-6-10 where three Gaussian distributions appear. Probably the reason for this is that there are samples from three different age groups in the dataset and thus we find three different clusters in the distributions. In contrast, there does not seem to be any difference between the distributions of the two classes (blue-male/red-female).

Also from scatter plot analysis, it is possible to see the presence of features composed of only one cluster and features where there are three different clusters. Again the distribution of the

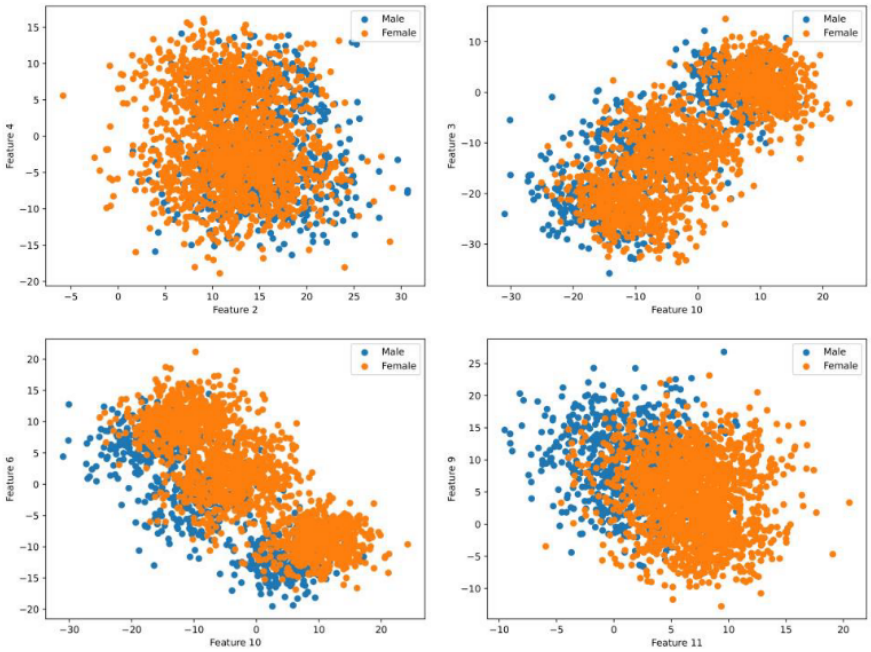


Figure 2: Scatter Plots

samples appears to be a Gaussian so we can assume that a Gaussian model will perform well to classify our data.

1.3.2 Linear Discriminant Analysis

Now we can focus our attention on a preprocessing transformation known as Linear Discriminant Analysis (LDA). The aim of LDA is to find the most discriminant direction that separates the classes of our dataset. Since LDA can find C-1 discriminant directions where C is the number of classes in the dataset, we will have only one discriminant direction. Looking at the graph below, it can be seen that by applying this transformation technique, the data become more linearly separable but nevertheless there is a region where classification errors would still be made.

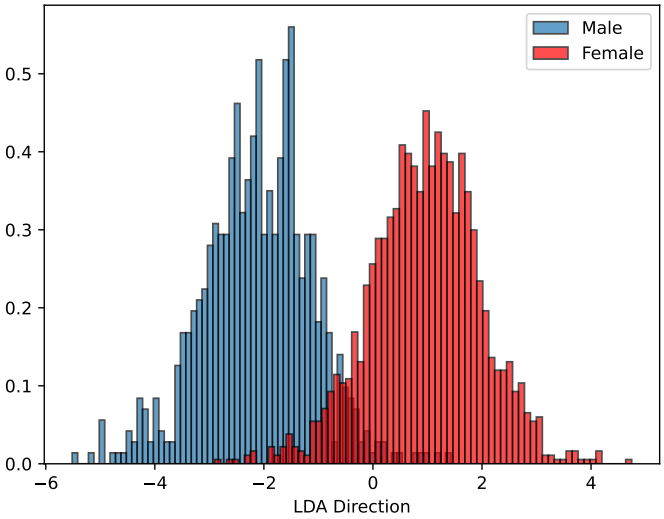
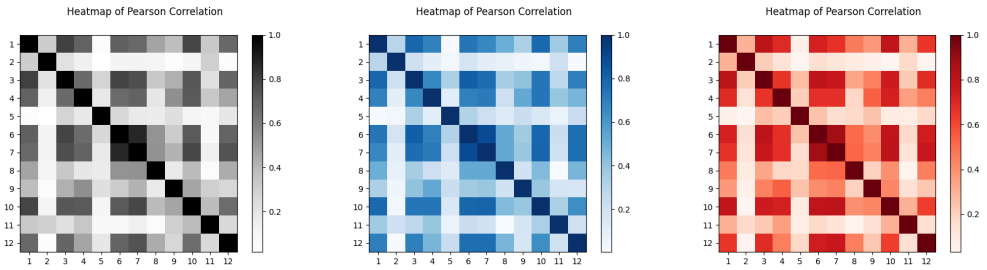


Figure 3: LDA Plot

1.3.3 Pearson Correlation

An additional analysis we can perform on our dataset concerns the correlation between features. To do this we exploit Pearson correlation, showing the results on a heatmap. To better understand the results of the heatmap we should note that the value of the correlation can range from 0 to 1. The higher the value, the higher the correlation between the corresponding features. This information will allow us to develop new hypotheses about some classifiers, which are influenced by feature correlation.



(a) All classes correlation (b) Male class correlation (c) Female class correlation

As we can see from the heatmaps, the dataset has slightly correlated features, in particular this correlation is more evident between some features such as between 1-3 , 1-10 and 7-6. In other cases the correlation is totally absent such as between features 1-5 or 2-5. Analyzing individually the correlations within the two classes we can see that the results are the same as in the dataset , in fact the remarkable information is that the heatmaps of the two classes are almost equal. This

fact is very important since having the classes correlated in the same way indicates that some classifiers such as MVG and Tied Gaussian will perform similarly. This data is also useful for making assumptions about models such as Naive Bayes where it is assumed that the features are correlated. It is clear from the heatmaps that a Naive Bayes type model may not perform so poorly since a slight correlation is present, however, it is necessary to further investigate this assumption by directly applying the algorithm.

1.3.4 PCA and Explained Variance

We now focus instead on PCA and in particular on Explained Variance, which allows us to understand for each component identified by PCA, how much variability in the dataset is preserved.

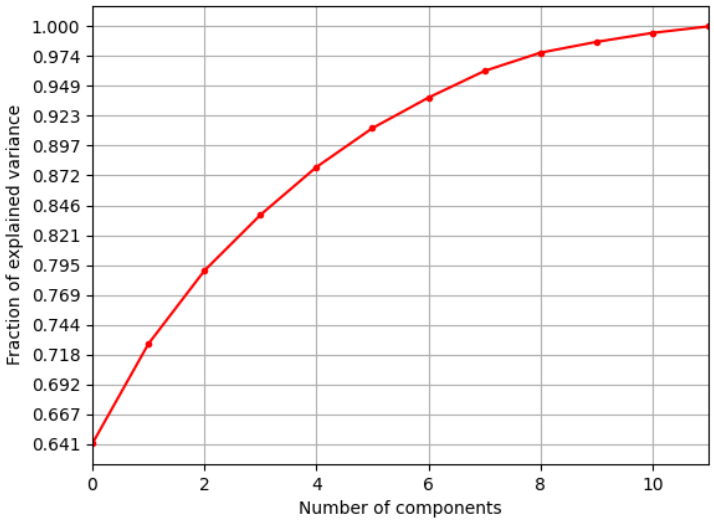


Figure 5: Explained Variance

From the graph we can see that by reducing the size of our dataset from 12 to 8 by PCA, the variance of the dataset still remains above 95% while going from 12 dimensions to 9 the variance of our dataset is above 97%. Decreasing to 9 we will get above 90%.