Immagine che contiene testo, Elementi grafici, Carattere, design

Descrizione generata automaticamente

Gender Identification Project

# Master of Science in Computer Engineering

# Machine learning and pattern recognition exam

# 2022/2023

Gelsi Alessandro

S303525

# Abstract

This report is dedicated to the analysis of a dataset consisting of low-level images depicting males and females, employing a range made of different Machine Learning (ML) algorithms. The goal is to determine the models that achieve the highest classification performance. Initially it was conducted an examination of the dataset's features, followed by the exploration of various classifiers, including Multivariate Gaussian Models (MVG), Logistic Regression (both linear and quadratic), Support Vector Machine (linear, RBF, and quadratic), Gaussian Mixture Models, and Fusion. A validation dataset is derived from the training dataset using K-fold cross-validation, in order to find the best hyperparameters for each model. The performances are evaluated at first taking into account minimum detection cost function(min DCF), followed then also by considerations of actual DCF and score calibration. Finally, the models equipped with the chosen hyperparameters were tested on evaluation set, made of unseen data.

It will be demonstrated that all classifiers yield good results on the provided dataset. As a best model it was selected SCRIVERE QUA IL BEST MODEL

# Index

[Master of Science in Computer Engineering 1](#_Toc144297980)

[Machine learning and pattern recognition exam 1](#_Toc144297981)

[2022/2023 1](#_Toc144297982)

[Abstract 2](#_Toc144297983)

[Index 3](#_Toc144297984)

# Dataset overview

The training dataset consists of 2,400 samples, comprising 720 males and 1,680 females, made extracting speaker embeddings from face images. A speaker embedding is a small-dimensional, fixed sized representation of an image, where features are continuous values that represent a point in the m-dimensional embedding space. The dataset results in a substantial bias towards females, accounting for 70% of the dataset. Each sample is made of 12 features that do not have a physical interpretation. It is also known that the samples belong to three distinct age groups, each characterized by a different distribution of embeddings. however, no age information is available. The test dataset instead is characterized by 6,000 samples, with 4,200 males and 1,800 females. As such, dataset are imbalanced, with the training set that has significantly more female samples, whereas the test set has significantly more male samples.

The mean (μ) and standard deviation (σ) of each feature for the training datasetare:

[  8.0, 10.6,  6.8,  4.5,  8.5,  9.4,  6.5,  6.5,  10.4,  4.3,  6.2 ]

It can be observed that features of the data set have different scales, they have large differences between their ranges. So, in this case, **Z-normalization** on the data-set to bring all the features on the same scale could be useful. Z-normalization centers the feature columns at mean 0 with standard deviation 1. Thus before applying any operation each sample of the training set has been transformed through the expression:

# Dataset Analysis

Before involve any classifier it is necessary first to perform an analysis of the dataset features. The mean (μ) and standard deviation (σ) of each feature for the training datasetare:

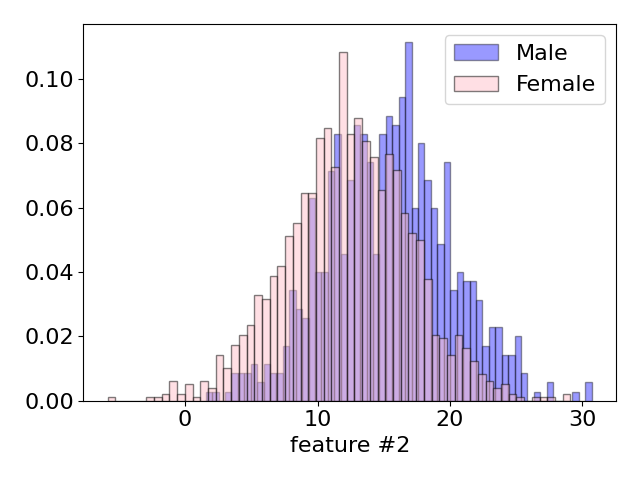
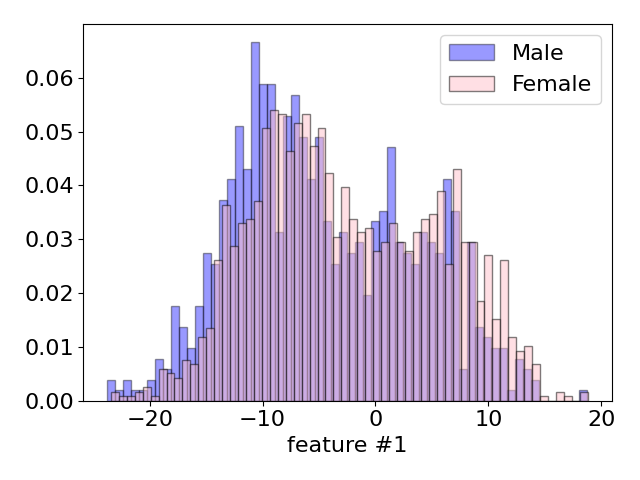
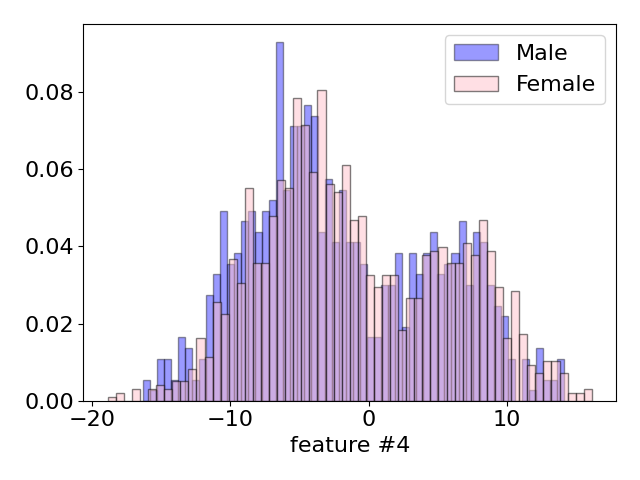
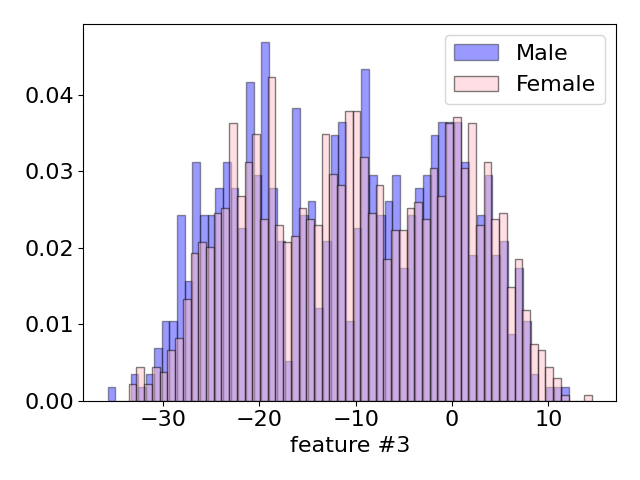
[  8.0, 10.6,  6.8,  4.5,  8.5,  9.4,  6.5,  6.5,  10.4,  4.3,  6.2 ]

The features do not exhibit significantly different scales., there is not a large differences between their ranges. A technique called **Z-normalization** generally is useful to bring all the features on the same scale, by centering the feature columns at mean 0 and with standard deviation 1. It is possible to consider it inside this study, despite not expecting substantial improvements. So, in parallel with the analysis of the raw features, an analysis of the normalized features was conducted by normalizing before applying any operation. Each sample of the training set has been transformed through the expression:

Where 𝐱′ is the sample after the Z-score normalization, while 𝐱 is the original sample in the data set.

Histograms

The initial step involves plotting histograms of each dataset feature to examine their distributions, after we normalized the dataset.



We can see that there are some distributions, such as the ones relative to features 5, 8, 2, that recall directly to a gaussian density. It is also important to observe that some plots, like for example plot number 3, 6 and 7, resemble a distribution made of three gaussian. This can be associable to the 3 groups of ages from where the features are extracted of.

Altogether, the distribution of individual features is consistent across both classes, but there are certain distributions that enable us to differentiate between classes in an easier way. This happens for example in figure 11, where it is possible to observe the most distinguishable feature distribution in our dataset.

Immagine che contiene schermata, testo, Diagramma, diagramma

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Diagramma, diagramma

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Diagramma, diagramma

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Diagramma, diagramma

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Diagramma, diagramma

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Diagramma, linea

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Diagramma, linea

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Diagramma, diagramma

Descrizione generata automaticamente

LDA

Scatter plots

The analysis continues by leveraging scatter plots, which are particularly useful to visualize the relationship between two continuous variables. They can help identify patterns, trends, correlations, or clusters within the data. By examining the distribution and dispersion of the dots, you can gain insights into how the variables interact with each other. For our dataset, these plots are aligned to the one present in the gaussian model, and for this reason, we expect that gaussian model are able to perform well on this kind of data.

The following images contains some of the most significant plot. For example in the scatter plot relative to feature #6 it is evident the presence of three clusters. It can be associable again to a gaussian distribution with more components, and directly related to the three groups of age from where the dataset sample are taken.

Immagine che contiene testo, schermata, diagramma, Policromia

Descrizione generata automaticamente Immagine che contiene testo, schermata, diagramma, Policromia

Descrizione generata automaticamente Immagine che contiene testo, schermata, diagramma

Descrizione generata automaticamente

Correlation

A way to analyze features interaction is to compute the **correlation** of features. This is useful also to understand if PCA could be useful and how many features can be discarded. **Pearson correlation coefficient** can be used to measure correlation between two features and it can be computed as:

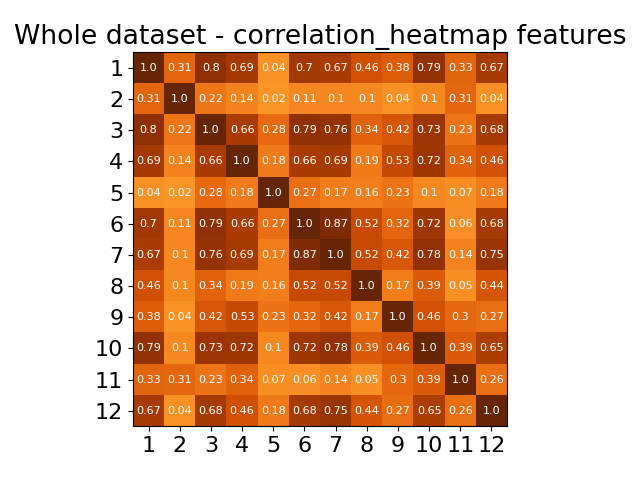
𝐶𝑜𝑣(𝑋,𝑌)√𝑉𝑎𝑟(𝑥)√𝑉𝑎𝑟(𝑦)

Actually absolute value of Pearson correlation is considered because we are only interested to understand if there is correlation or not:

‖𝐶𝑜𝑣(𝑋,𝑌)√𝑉𝑎𝑟(𝑥)√𝑉𝑎𝑟(𝑦)‖

The absolute value of Pearson correlation coefficient can take value between 0 and 1. If 0 it means that the two considered features are uncorrelated, while 1 means that the features are completely correlated (one feature is the scaled version of the other).

To visualize the correlation between features we employed heatmaps. In the following heatmaps, darker colors indicate a strong correlation between two features, while lighter colors suggest a weaker correlation between the two features.

Immagine che contiene testo, schermata, quadrato, Rettangolo

Descrizione generata automaticamente

Immagine che contiene testo, schermata, quadrato, modello

Descrizione generata automaticamente

# Validation approach

Dasdasda

# Multivariate Gaussian classifier (MVG)

Dsdafsafafascaca

*Table 1: Min DCF results for MVG with K-fold cross validation (K=5)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RAW | | | Z-score | | |
| Model | **π= 0.5** | **π= 0.1** | **π= 0.9** | **π= 0.5** | **π= 0.1** | **π= 0.9** |
| Full-cov | 0.113 | 0.297 | 0.350 | 0.113 | 0.297 | 0.350 |
| Diag-cov | 0.463 | 0.770 | 0.777 | 0.463 | 0.770 | 0.777 |
| Tied full-cov | 0.109 | 0.299 | 0.341 | 0.109 | 0.299 | 0.341 |
| Tied diag-cov | 0.457 | 0.769 | 0.780 | 0.457 | 0.769 | 0.780 |

*Table 2: Min DCF results for MVG with K-fold cross validation (K=5) and PCA(m=11)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RAW- PCA(11) | | | Z-score -PCA(11) | | |
| Model | **π= 0.5** | **π= 0.1** | **π= 0.9** | **π= 0.5** | **π= 0.1** | **π= 0.9** |
| Full-cov | 0.117 | 0.308 | 0.349 | 0.121 | 0.311 | 0.358 |
| Diag-cov | 0.126 | 0.324 | 0.370 | 0.124 | 0.311 | 0.348 |
| Tied full-cov | 0.118 | 0.288 | 0.355 | 0.118 | 0.298 | 0.355 |
| Tied diag-cov | 0.124 | 0.302 | 0.360 | 0.123 | 0.294 | 0.354 |

*Table 3: Min DCF results for MVG with K-fold cross validation (K=5) and PCA(m=10)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RAW- PCA(10) | | | Z-score -PCA(10) | | |
| Model | **π= 0.5** | **π= 0.1** | **π= 0.9** | **π= 0.5** | **π= 0.1** | **π= 0.9** |
| Full-cov | 0.163 | 0.401 | 0.492 | 0.187 | 0.406 | 0.537 |
| Diag-cov | 0.168 | 0.448 | 0.468 | 0.184 | 0.434 | 0.545 |
| Tied full-cov | 0.161 | 0.392 | 0.474 | 0.183 | 0.427 | 0.535 |
| Tied diag-cov | 0.170 | 0.396 | 0.479 | 0.182 | 0.421 | 0.543 |

# Logistic Regression

Linear Logistic Regression

Teoriaaadasdadsfasfvasva

Immagine che contiene testo, diagramma, linea, Diagramma

Descrizione generata automaticamenteImmagine che contiene testo, diagramma, linea, schermata

Descrizione generata automaticamente

Figure 1: minDCFwrt Lambda RAW and on the right is ZSCORE

Linear lambda raw= 10 alla meno 6 DA RICONTROLLARE!

Quadratic lambda raw= 100

*Table 3: Min DCF results for Linear logreg with K-fold cross validation (K=5)*

|  |  |  |  |
| --- | --- | --- | --- |
|  | RAW | | |
| Model | **π= 0.5** | **π= 0.1** | **π= 0.9** |
| LR(λ=1e-6, πT=0.5 ) | 0.112 | 0.283 | 0.339 |
| LR(λ=1e-6, πT=0.1 ) | 0.120 | 0.299 | 0.377 |
| LR(λ=1e-6, πT=0.9 ) | 0.110 | 0.315 | 0.344 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | ZSCORE | | |
| Model | **π= 0.5** | **π= 0.1** | **π= 0.9** |
| LR(λ=1e-5, πT=0.5 ) | 0.112 | 0.283 | 0.343 |
| LR(λ=1e-5, πT=0.1 ) | 0.131 | 0.353 | 0.360 |
| LR(λ=1e-5, πT=0.9 ) | 0.135 | 0.323 | 0.313 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Quadratic LR-RAW | | |
| Model | **π= 0.5** | **π= 0.1** | **π= 0.9** |
| LR(λ=100, πT=0.5 ) | 0.121 | 0.299 | 0.363 |
| LR(λ=100, πT=0.1 ) | 0.125 | 0.313 | 0.406 |
| LR(λ=100, πT=0.9 ) | 0.122 | 0.331 | 0.356 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Quadratic LR-ZSCORE | | |
| Model | **π= 0.5** | **π= 0.1** | **π= 0.9** |
| LR(λ=1e-3, πT=0.5 ) | 0.125 | 0.334 | 0.332 |
| LR(λ=1e-3, πT=0.1 ) | 0.131 | 0.353 | 0.360 |
| LR(λ=1e-3, πT=0.9 ) | 0.135 | 0.323 | 0.313 |

# Support Vector Machines

LINEAR RAW

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

QUADRATIC RAW

Immagine che contiene testo, linea, diagramma, Diagramma

Descrizione generata automaticamente

RBF RAW

Immagine che contiene testo, linea, diagramma, Diagramma

Descrizione generata automaticamente

LINEAR ZSCORE

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

QUADRATIC ZSCORE

Immagine che contiene testo, schermata, linea, diagramma

Descrizione generata automaticamente

RBF ZSCORE

Immagine che contiene testo, diagramma, linea, Diagramma

Descrizione generata automaticamente

*Table 3: Min DCF results for Linear svm with K-fold cross validation (K=5)*

|  |  |  |  |
| --- | --- | --- | --- |
|  | RAW | | |
| Model | **π= 0.5** | **π= 0.1** | **π= 0.9** |
| LR(C=0.1, πT=0.5 ) | 0.115 | 0.294 | 0.351 |
| LR(C=0.1, πT=0.1 ) |  |  |  |
| LR(C=0.1, πT=0.9 ) |  |  | 0.339 |

*Table 3: Min DCF results for Quadratic svm with K-fold cross validation (K=5)*

|  |  |  |  |
| --- | --- | --- | --- |
|  | RAW | | |
| Model | **π= 0.5** | **π= 0.1** | **π= 0.9** |
| LR(C=0.1, πT=0.5 ) | 0.115 | 0.294 | 0.351 |
| LR(C=0.1, πT=0.1 ) |  |  |  |
| LR(C=0.1, πT=0.9 ) |  |  | 0.339 |

# Gaussian Mixture Model

# Best Model

GMM full 4components pi=0.5 raw: 0.0706

GMM full tied 8components pi=0.5 raw : 0.06

GMM FULL TIED 4 Component zscore is the best

# Score calibration

Prima su GMm ma non ce neanche bisogno

Poi si fa vedere che funziona su qualcosa che va male tipo logreg

# Fusion

Non ce bisogno di fare calibration

WORKS ON ZSCORE

GMM è il full tied 4 component

LR è logreg linear con piT=0.9 e lambda 10^-5

SVM è piT=0.9 RBF SVM no rebalancing -gamma =0.100000 -C=10.000000 - pi = 0.500000 -> minDCF = 0.090675

Da testare:

GMM+SVM+LR

GMM+LR ok

GMM+SVM ok

SVM+LR

DA FIXARE SCORE CALIBRATION E ROC QUANDO FA SVM.TRAIN ecc, ci sono errori