Machine Learning and Pattern Recognition Project Report

Paolo Magliano s314867

July 8, 2024

Abstract

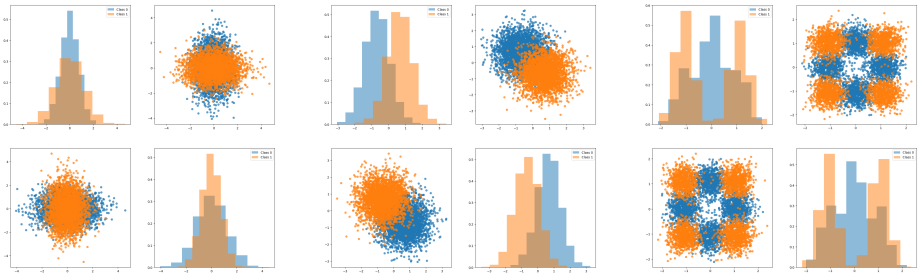
This report describes the work done on Machine Learning and Pattern Recognition project. The report follows the structure of course laboratories, which studies and analyses the dataset and the models:

- Multivariate Gaussian Model
- Logistic Regression
- Support Vector Machines
- Gaussian Mixture Model

1 Lab 2: Dataset

The samples are computed by a feature extractor that summarizes high-level characteristics of a fingerprint image. The data is 6-dimensional and it consists of labeled samples corresponding to the genuine (True, label 1) class and the fake (False, label 0) class.

The dataset can be summarized by the image in Figures 1. It shows the data distribution of features and their correlation with each other.



(a) Features 1 and 2. (b) Features 3 and 4. (c) Features 5 and 6.

Figure 1: The figure shows histograms of the projected data on each feature. The scatter plot shows the correlation between the two features.

The first two features are mostly overlapping, so it is difficult to separate the two classes. The histograms show that they can be probably well approximated by a Gaussian density function. Further more means are very close to each other and only the second feature has a slightly difference in variance.

The third and fourth features are more separated, but still overlapping. Also these features can be approximated by a Gaussian density function but the means are different and variance very similar. These features are more useful in a classification task as they are more discriminative.

The last two features are the most discriminative, they can't be approximated by a simple unimodal Gaussian density function because they seems to be bimodal. Thanks to this behavior they form 4 distinct clusters (two for each feature) that can be easily separated. They are probably the most useful features if the classifier is able to capture the complex structure.

2 Lab 3: PCA and LDA

2.1 Dimensionality Reduction

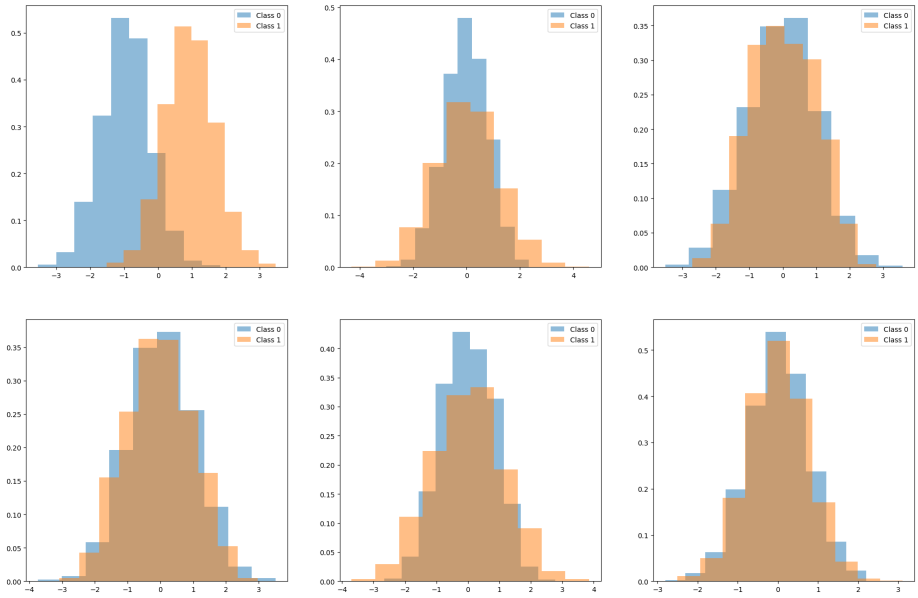


Figure 2: The figure shows the results of PCA on the dataset. The histograms show data distribution on each principal component. The top left one represents the first principal component, the bottom right is the last one.

Principal Component Analysis (PCA) is a dimensionality reduction technique that finds the directions of maximum variance in the data. It projects the data onto a lower-dimensional subspace while preserving as much variance

as possible. The Figure 2 shows the results of PCA on the dataset. The first principal component captures most of the variance and it separates decently the two classes. The other principal components overlap, so they are not very useful for classification.

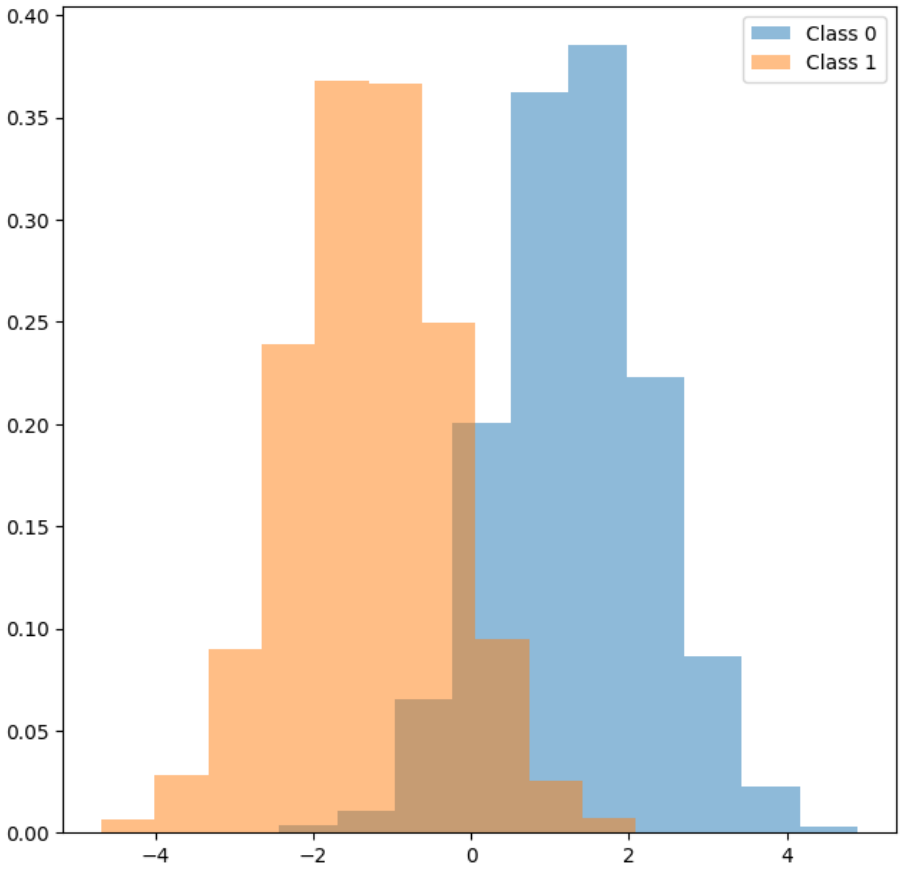


Figure 3: The figure shows the results of LDA on the dataset. The histogram shows data distribution on the only linear discriminant found.

Linear Discriminant Analysis (LDA) is a supervised dimensionality reduction technique that finds the directions that maximize the separation between classes. The Figure 3 shows the results of LDA on the dataset. Only one linear discriminant is found because the number of classes is 2. The linear discriminant separates the two classes partially. It is comparable to the first principal component of PCA.

Both cases simplify the dataset as much that the well-distinctive cluster of the last two features is lost. So, the preprocessing techniques cloud be useful to speed up the computation but they could also lose important information. In

very simple dataset ad this one, with only 6 dimensions, presumably the best choice is to use the original dataset.

2.2 Classification

The project analyses the performance of LDA method as a classifier. In order to get reliable results, the experiments' performance is evaluated with k-fold cross-validation with k equal to 10. It is also inspect the improvement of the classifier by using PCA as a preprocessing technique.

The Table 1 shows the performance results as error rate, Detection Cost Function (DCF) and Minimum DCF. The last one is useful to understand the gains of changing the threshold to optimal value.

PCA	**Error Rate (%)**	**DCF**	**Min DCF**
None	9.58	0.192	0.176
1 PC	9.68	0.194	0.177
2 PC	9.63	0.193	0.178
3 PC	9.43	0.189	0.176
4 PC	9.57	0.192	0.175
5 PC	9.58	0.192	0.175
6 PC	9.62	0.193	0.176

Table 1: Performance metrics for LDA models with various dimensionality reduction techniques.

As intuitively retrieved from PCA and LDA graphs, the capabilities of both methods are similar. In fact, the performance of the classifier is not improved by using PCA as a preprocessing technique.

In general the performance are terrible considering the simplicity of the dataset. Further more the threshold selection is not optimal, so the performance could be improved by tuning it a little bit.

3 Lab 4: Multivariate Gaussian density

Describe the methods used in your project. Include algorithms, data preprocessing steps, feature selection, model selection, and any other relevant details.

3.1 Data Collection and Preprocessing

Detail the data sources, collection methods, and preprocessing steps.

3.2 Algorithm and Model Selection

Discuss the algorithms and models used. Include any hyperparameter tuning and model selection processes.

3.3 Evaluation Metrics

Explain the metrics used to evaluate the performance of your models.

4 Experiments

Describe the experiments conducted to validate your models. Include details such as experimental setup, datasets used, and any specific configurations.

5 Results

Present the results of your experiments. Use tables, figures, and charts to illustrate the performance of your models.

Model	Accuracy	F1 Score
Model 1	0.90	0.88
Model 2	0.85	0.84
Model 3	0.92	0.90

Table 2: Model performance comparison

6 Discussion

Interpret the results. Discuss any patterns, insights, or anomalies observed. Compare the performance of different models and explain any differences.

7 Conclusion

Summarize the findings of your project. Discuss the implications of your results, any limitations of your study, and possible future work.