Solution of the fourth exercise assignment

Lucas Nogueira Ribeiro - 373082

1 Introduction

Linear methods such as Least-squares estimators, Fisher's Linear Discriminant, Multiclass Logistic Regression, were employed to solve classification problems in this exercise assignment.

2 Methodology

Python scripts were written to solve this assignment. The sourcecode is available in the Jupyter notebook format at `https://github.com/lnribeiro/patternrecognition/blob/master/4th%20assignment/4th_assigment.ipynb`.

3 Results

First exercise

The obtained mean square error (MSE) and sum-of-square error are

```
MSE [train]: 0.0190868610825
MSE [test]: 0.0188001558568
Sum-of-squares error [test]: 9.40007792838
Sum-of-squares error [train]: 28.6302916237
```

Data from the test set and the decision boundary are plotted in Figure 1. The obtained classification rate is 100%.

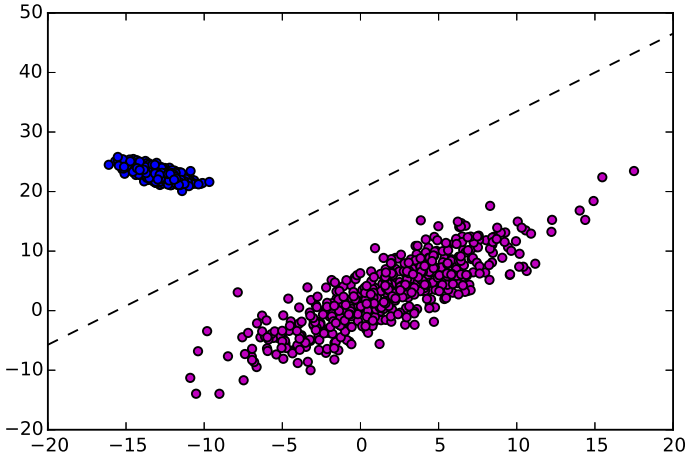

Figure 1: Test dataset and its decision boundary

Second exercise

The generated dataset and its decision boundary, obtained using Fisher's Linear Discriminant, are plotted in Figure 2. The obtained classification rate was 100%.

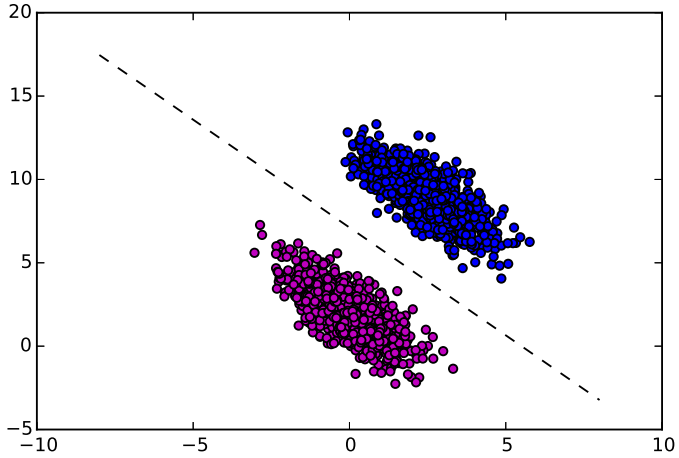

Figure 2: Generated dataset and its decision boundary, which was calculated using the Fisher's Linear Discriminant.

Third exercise

The decision boundaries obtained using a Perceptron in the data of the 1st and 2nd questions are depicted in Figures 3 and 4, respectively. Although the boundaries seem to be misplaced, the obtained classification rates for both data are 100%. The Perceptron algorithm is indeed capable of providing high classification rates given that the data is linearly separable. In practical scenarios, however, data seldom satisfies this condition. Multilayer perceptron classifiers are more adapted to real scenarios.

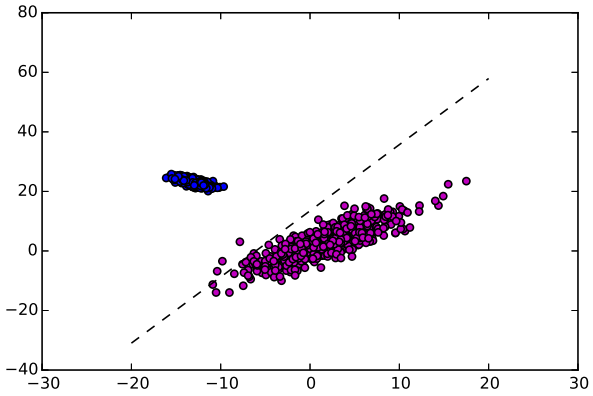

Figure 3: Generated dataset and its decision boundary, which was calculated using the Fisher's Linear Discriminant.

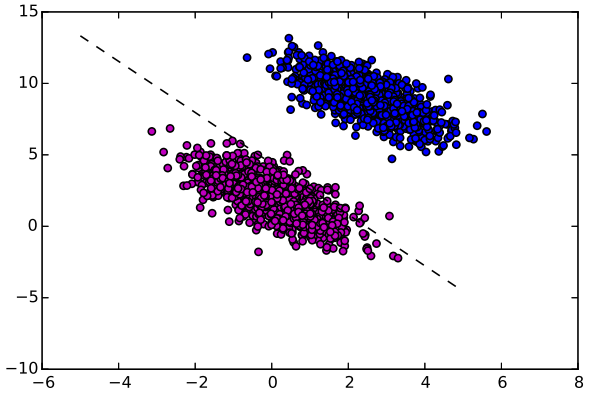

Figure 4: Generated dataset and its decision boundary, which was calculated using the Fisher's Linear Discriminant.

Fourth exercise

The train dataset is illustrated in Figure 5. The decision boundaries for the red, blue, and green classes are depicted in Figures 6, 7, and 8, respectively. These boundaries are plot together in Fig. 9. The classification rate in the test dataset was 99.83%.

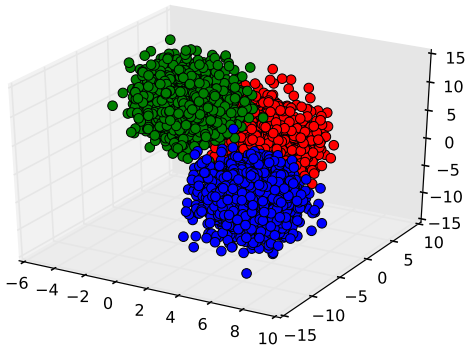

Figure 5: Train dataset

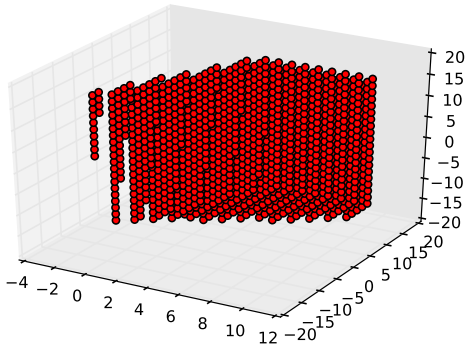

Figure 6: Decision boundary - red class

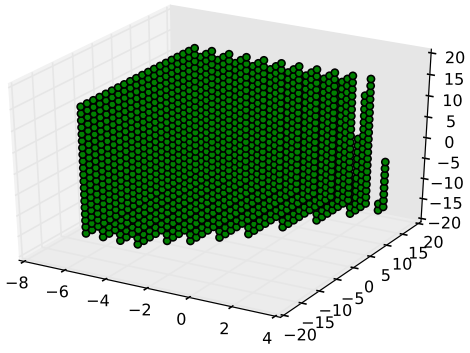

Figure 7: Decision boundary - green class

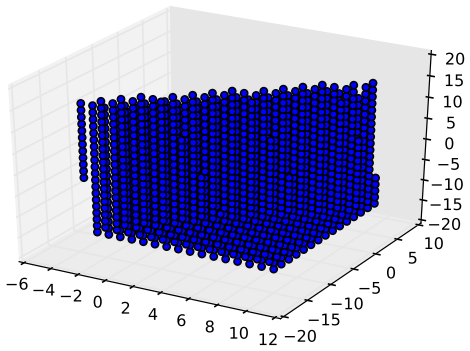

Figure 8: Decision boundary - blue class

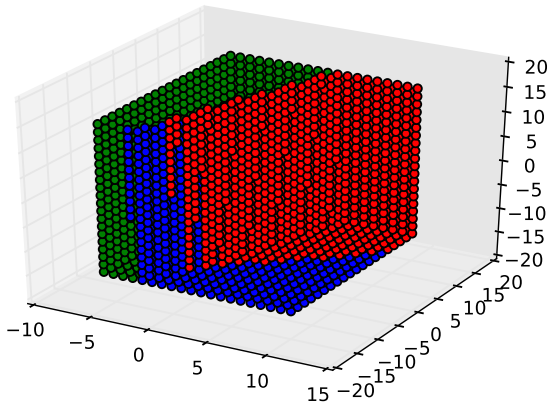

Figure 9: Decision boundaries

Fifth exercise

The classification rate provided by the multiclass logistic regression classifier in the test stage was 98.83%. This rate is inferior than that provided by the probabilistic generative model in the previous exercise.

Sixth exercise

The calculated statistics of the data attributes are shown below:

```
Attribute 0: max 7.9 min 4.3 mean 5.844 +- 0.824659929911
Attribute 1: max 4.4 min 2.0 mean 3.05733333333 +- 0.434410967735
Attribute 2: max 6.9 min 1.0 mean 3.76666666667 +- 1.76218298962
Attribute 3: max 2.5 min 0.1 mean 1.19933333333 +- 0.759692627902
```

PCA was employed to reduce the dimensionality of the IRIS dataset. After preserving only the two principal components of the original data, a dataset with two features was obtained. A linear classifier was used to obtain the decision boundaries. In Figure 10, the bidimensional dataset and its decision boundaries are depicted. The average classification rate was 92.67%.

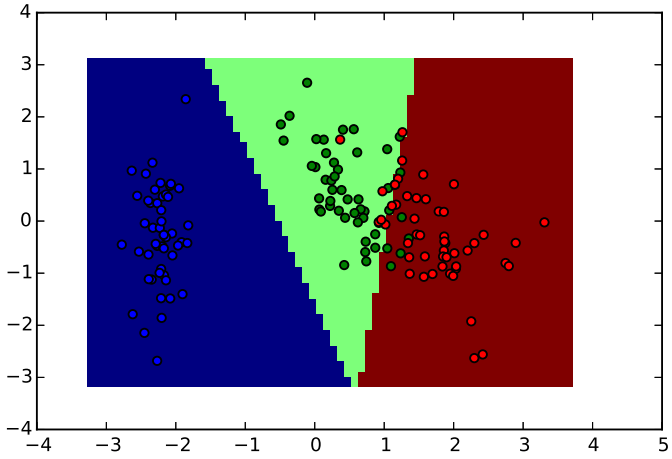

Figure 10: Decision boundaries considering the two principal components of the IRIS dataset.

4 Conclusion

Least-squares estimators, Fisher's Linear Discriminant, Multiclass Logistic Regression and other linear methods were employed to solve the exercises proposed in this assigment. Since most of the datasets were linearly separable, good classification rates were obtained.