FTML practical session 6: 2023/04/21

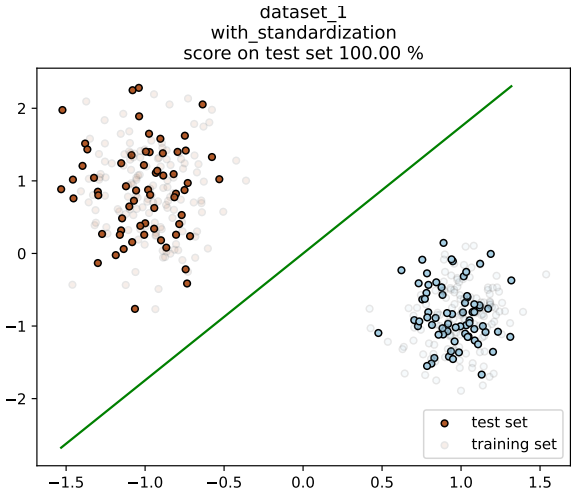
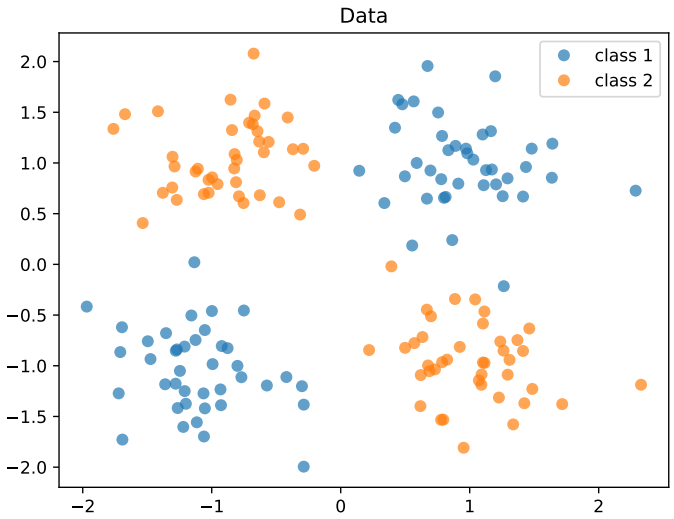


TABLE DES MATIÈRES

INTRODUCTION

1 KERNELS FOR SUPPORT VECTOR MACHINES

We would like to classify these data with a SVC.



Obviously, a linear separator will not work to separate these two classes. Hence, we would like to explore kernels and feature maps in order to classify them.

Explore the documentation from scikit-learn and read about the definition of the most commonly used kernels. Try to guess which kernel(s) might work or not in order to classify these data. Only after this stage, test your assumptions by using the library to train a Support vector classifier using different types of kernels and by monitoring the decision function values and a test error.

The data are stored in **svm_kernels/data/**

2 INFLUENCE ON DATA SCALING ON THE CONVERGENCE OF SGD

2.1 Introduction

We would like to perform a binary clasification on the following 3 datasets (figures 1, 2, 3). Each one has a different difficulty for a linear classifier, such as a SVM. We also note that the scales are different for some axes.

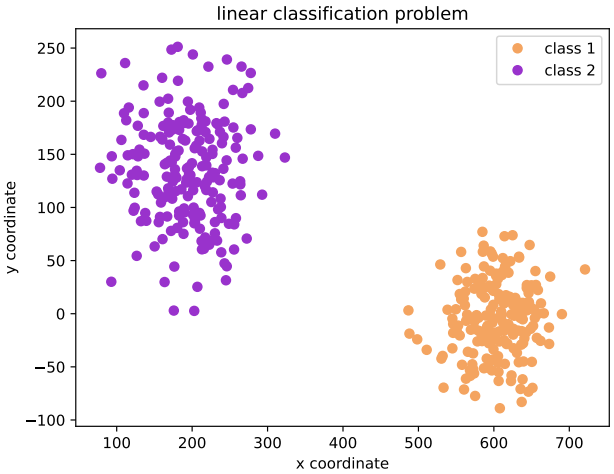


FIGURE 1 – Dataset 1

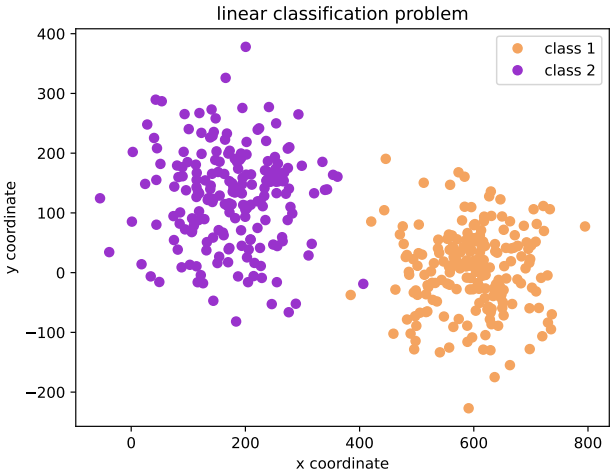


FIGURE 2 – Dataset 2

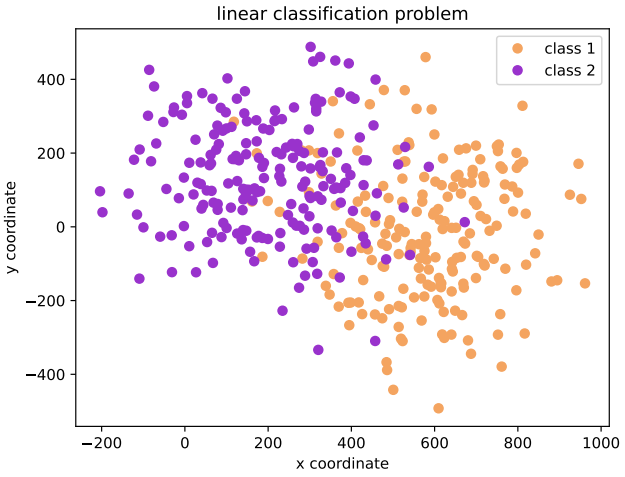


FIGURE 3 – Dataset 3

The data are located in **svm_sgd/data/**.

Data standardization consists in transforming the data so that each feature (each column) is centered (zero mean) and has a variance equal to 1.

https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-scaler.

It is experimentally often noticed that algorithms trained by SGD give a better performance (generalization error) when the data are standardized.

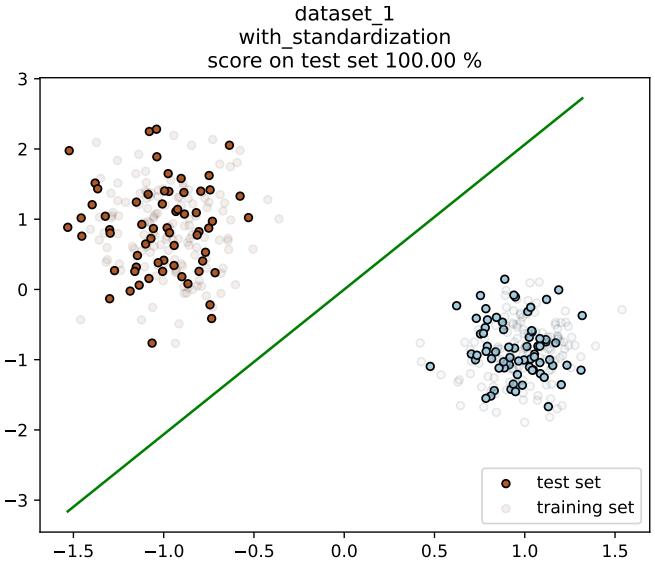
Perform a linear classification on these data, optimized by SGD, and compare quality of the results with and without preprocessing the data by standardizing them. You may use scikit-lean to do it, in particular :

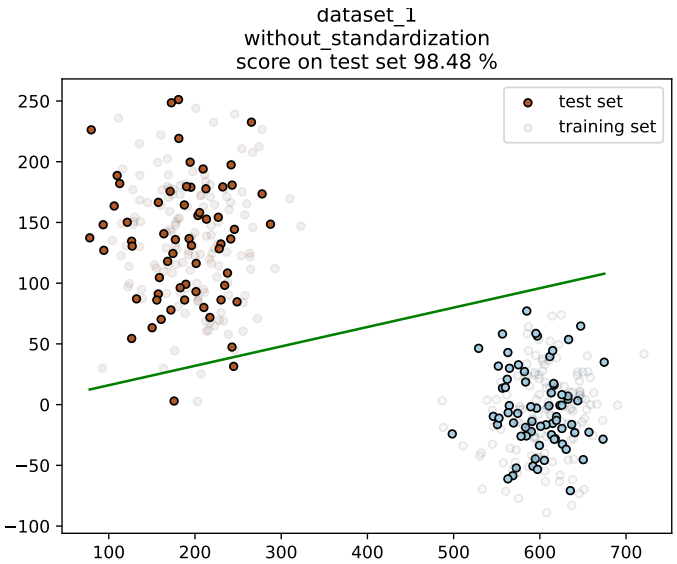
— https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html. By default, this class trains a liner SVM (no kernel).

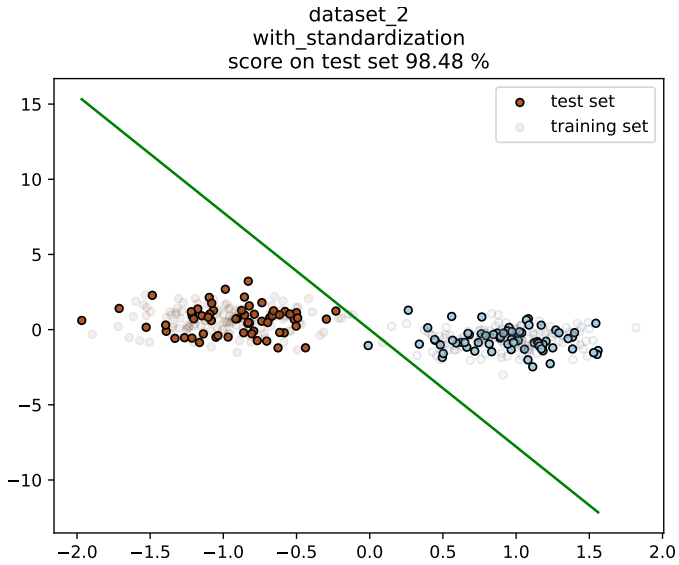
— https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

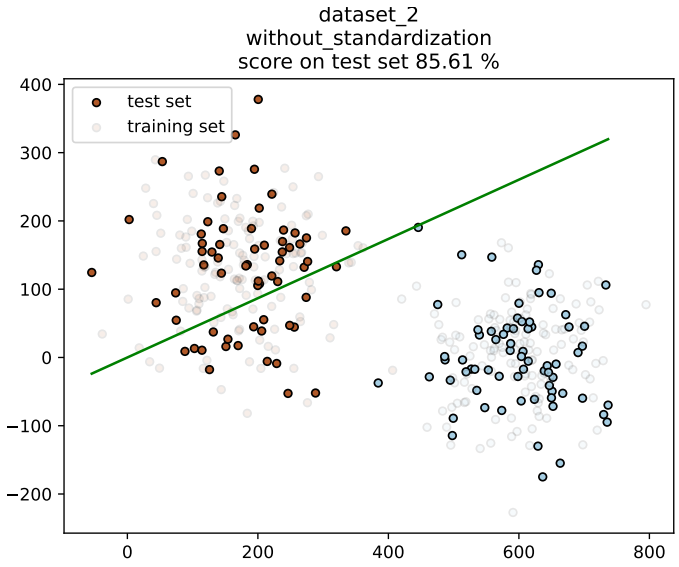
— https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

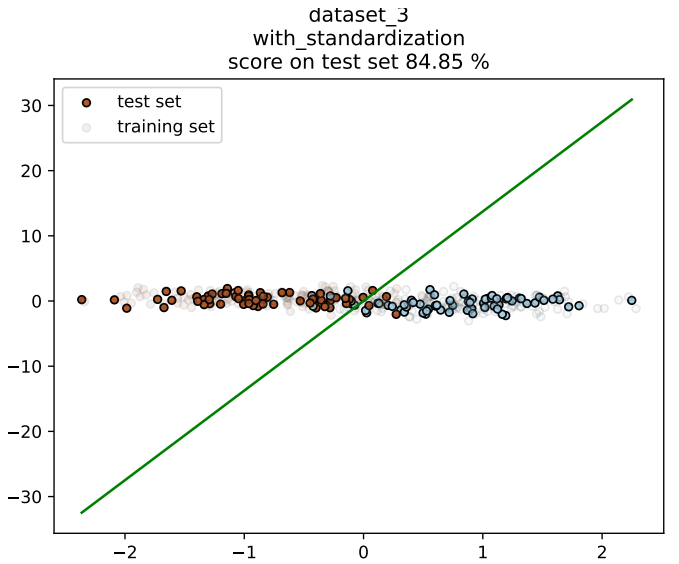
You should obtain results like the following figures.

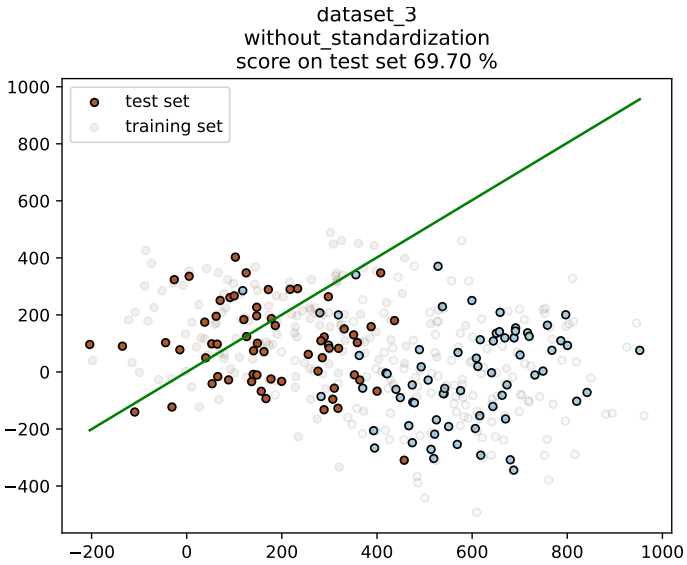





dataset_2
with_standardization
score on test set 98.48 %
test set
training set
15
10
5
0
−5
−10
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


dataset_2
without_standardization
score on test set 85.61 %
test set
training set
400
300
200
100
0
−100
−200
0
200
400
600
800


dataset_3
with_standardization
score on test set 84.85 %
test set
training set
30
20
10
0
−10
−20
−30
−2
−1
0
1
2



2.2 Conclusion

In this example, we saw that data standardization may give an improved performance, on a linear SVM trained by SGD. You can perform the same study on different datasets such as the toy datasets from scikit or other datasets.