Machine Learning: First Home Work

Principal Component Analysis and Na??ve Bayes Classification

Edoardo Ghini

December 14, 2016

Dipartimento di Ingegneria dell'Università di Roma La Sapienza

Contents

Part I

Introduction

1 Scope

In the first place, during this experience I tried to analyze a dataset made of a sub set of pictures in which it has been necessary to perform a feature extraction.

2 Objectives

Some important concepts to underline will be the impact on the analysis brought from the choice of the feature reduction and also the magnitude of the splitting factor between train and test data.

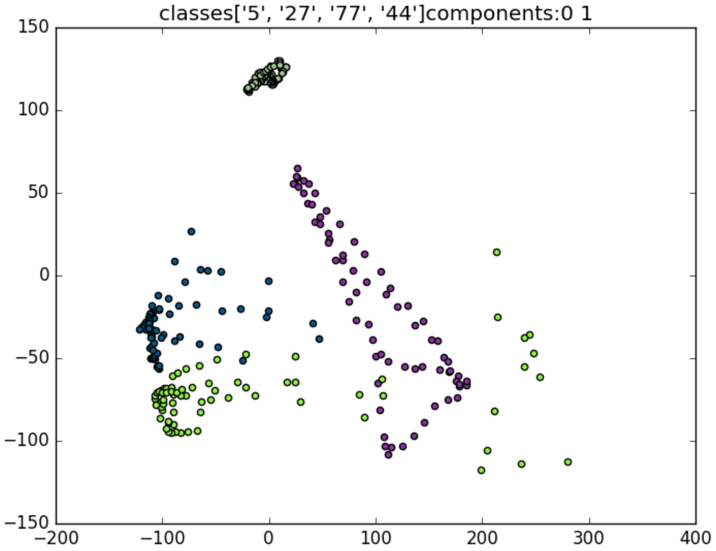

Figure 1

Part II

Development

3 Data loading and feature manipulation

At first, I loaded in suitable data structures the features coming from the pixels of the images of the chosen objects. After that I standardised and reshaped the matrix in order to obtain an optimal approximation of the original dataset. Then , the final dataset manipulation was to apply a principal component analysis to take in consideration the most significative vectors in terms of the variance representation. The plots shown above represent different choices of PCA vectors according to the expressive capabilities that are embodied in the components near the first.

As shown in fig.(1) , fig(2), and fig(3) there are significative discrepancies between data point concentrations: in fact, as expected vector choices that are far away from the first principal components gave more vague results.

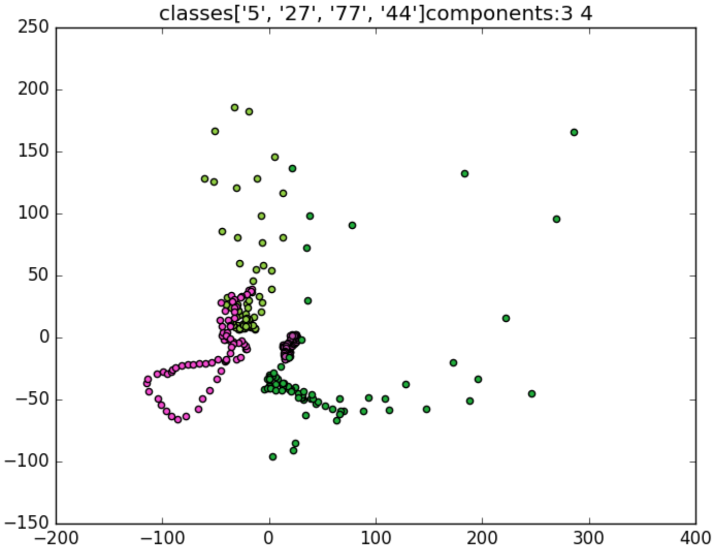

Figure 2

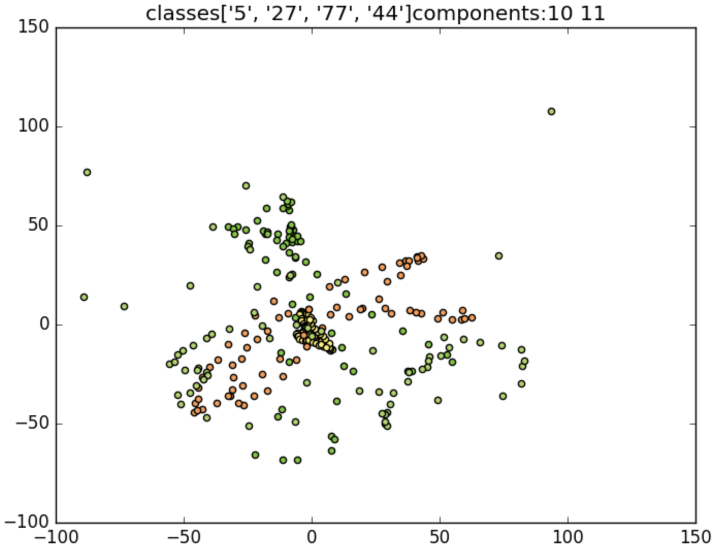

Figure 3

4 Training and Testing

4.1 Data spitting

In order to have a feed back on the effectiveness of the model in use, I partitioned my dataset in two subsets. The former will be used to train the model and the latter will be employed to certificate the accuracy of this model.

4.2 Naive Bayes Classification

At this point, with a coherent dataset available, I trained a classifier with the subset data chosen to be the prior knowledge for the model. I used a ”naive” classifier, a model that works by approximation and in particular it applies a sort of constraint relaxation assuming that all the data vectors where conditionally independent from each others.

4.3 Checking model effectiveness

Finally I requested a prediction to the trained classifier and I computed a score according to the miss-classification errors encountered. Obviously I managed to check the rightness of the predictions thanks to the subset of labels that I have preserved for the test. There are some examples below in fig(4), fig(5) and fig(6) that prove that the level of rightness of predictions depends strongly by the rateo chosen to split the subsets of training and testing.

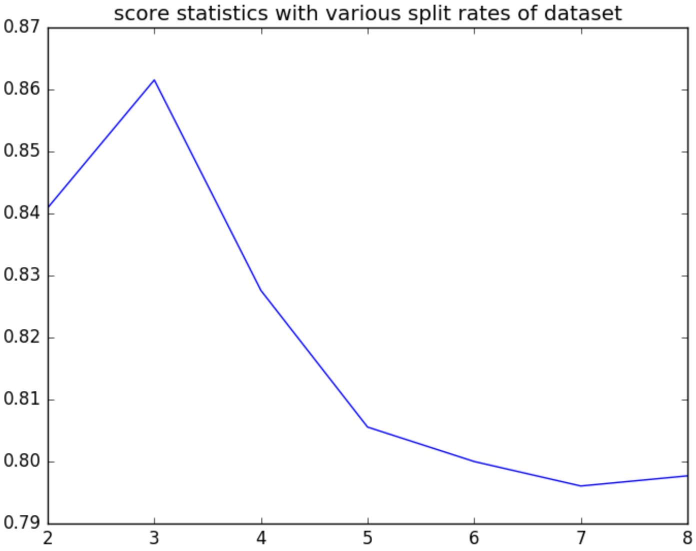

Figure 4

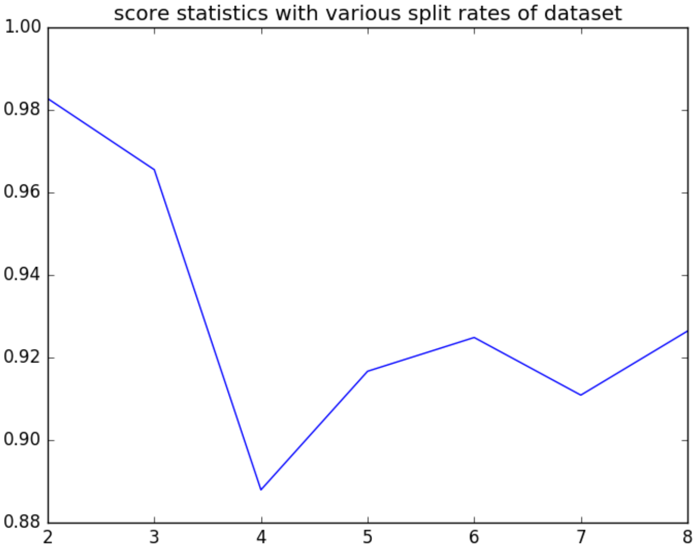

Figure 5

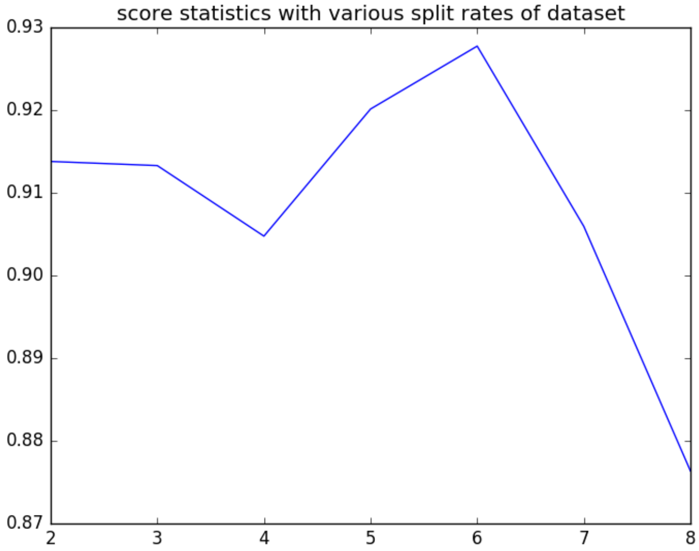

Figure 6

Part III

Conclusions

In conclusion, the difficulties which I would meet with trying to visualize a multidimensional dataset can be mitigated through a feature extraction with a principal component analysis approach. However the difficulties in the data representation will grown linearly with the increment of the different classes taken in consideration from the model. For instance in fig(7) there is the case in the dataset representation of much more classes than the aforementioned cases.

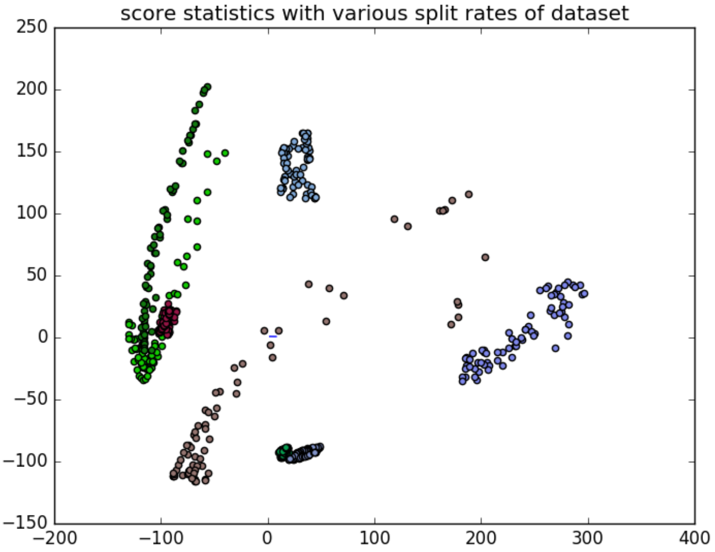
score statistics with various split rates of dataset
250
200
150
100
50
0
−50
−100
−150
−200
−100
0
100
200
300
400

Figure 7