Report Machine Learning, Homework 1

Domenico Alfano

The report describes the results obtained during the development of Homework 1. The main goal of the assignment has been to evaluate the capabilities and functionalities of Principal components analysis and the Naive Bayes Classifier.

1 Data Preparation

There comes a set of approximately 7000 images classifiable in 100 differents classes (categories). Each image is represented with 128x128 pixels. As requested, 4 classes have been chosen among the 100 availables and it has been created the X matrix. It has a number of columns equal to 49152 (128x128x3, the number of pixels of each image has been multiplied by three becouse every pixel is represented with a 3-d array) and a number of rows equal to 288 (72x4, where 72 is the number of images for category). X, then, is a 288x49152 matrix where the rows are the examples and the columns are the features. In particular the classes ”obj10”, ”obj30”, ”obj60”, ”obj90” have been chosen. Subsequently, vector y, holding ordinal labels of the images, has been created. This vector has 288 positions (one for each image). With this last operation the section of the data preparation has been completed.

2 PCA

The PCA method is used to extract relevant information from a data set. After standardize X, PCA has been used to obtain the first eleven principal components from the X matrix. In the figures 1-3 the scatter plot of different components are shown . Figure 1 shows that the first two components contain the most significant information of the images features, making easy to distinguish the different class of images. However as higher components of PCA are used, the distinction between classes becomes hard. This behavior has shown in figures 2 and 3. First case uses the third and the fourth component, where some of the features of each class are mixed between them. In second case the situation, using the tenth and the eleventh component, is more complicated because most of the features of each class are in the same area, making hard the separation of the dataset. From a theoretical point of view the results have sense. Each component of PCA represents a linear combination of the all features where the variance of the dataset is maximized. The first component is the subset of data with the biggest variance, the second component correspond to the subset with the second biggest variance an so on. As regard the second question of this section, I would use the graphic of percentage of variance explained respect to the number of components to determine the components needed to preserve data without much distortion. The plot in figure 4 is obtained with 20 components of PCA. From the plot I would choose the components which has the greatest percentage of variance explained, in this case the first three. Mainly because these components collects the biggest part of the variance, making easy the distinction among the classes of the dataset.

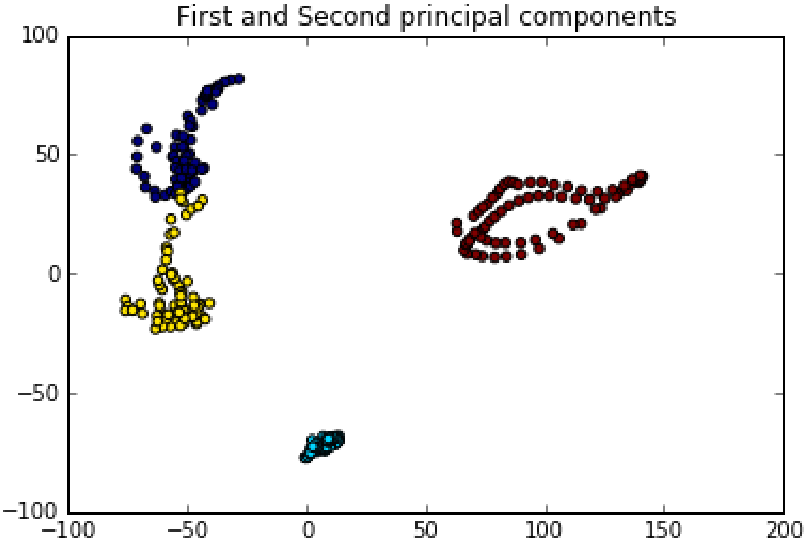

Figure 1

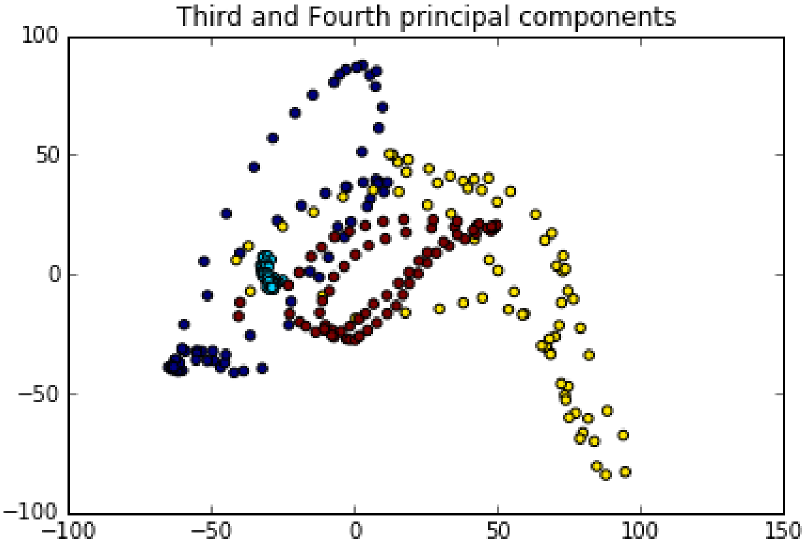

Figure 2

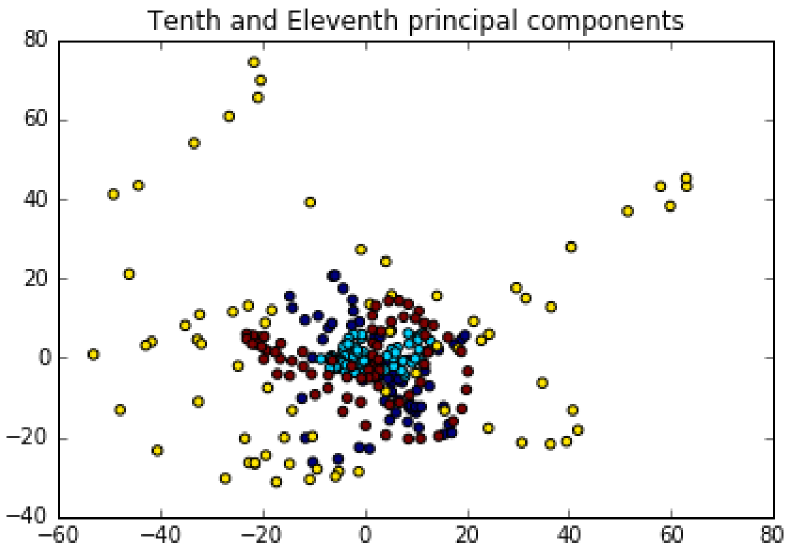

Figure 3

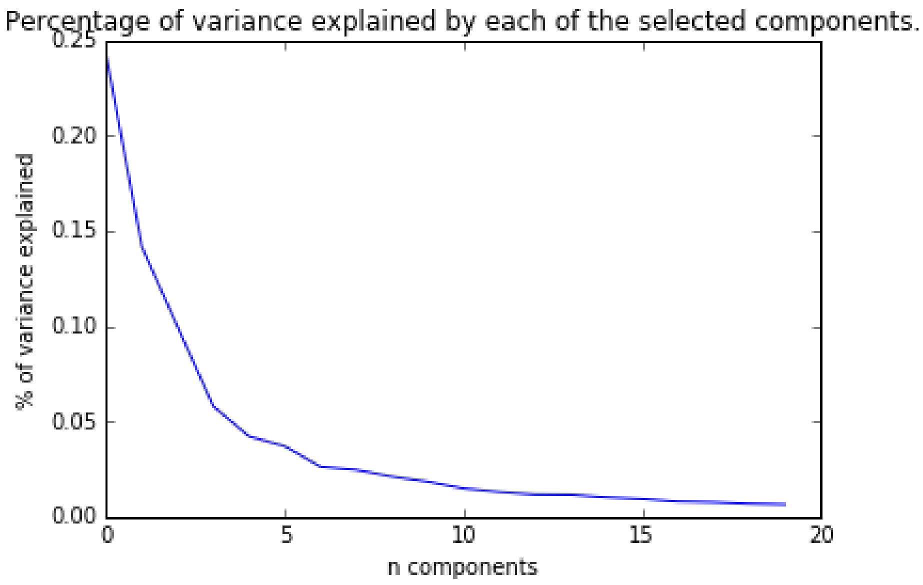

Figure 4

3 Classification

In Machine Learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. We start from the following formula:

$$\hat{y} = \operatorname*{argmax}_{y \in 1,..,k} p(y|x_1, .., x_d)$$

where y is a predicted label, k is the number of classes, $\{x_i\}_{i=1}^{d}$ are examples, $p(x|y)$ is a Gaussian, and distribution of labels is unifom. In this case k is equal to 4 (becouse we are considering four classes) and d is equal to 49152 becouse the features correspond to the columns of the X matrix (49152). In order to apply the Naive Bayes classificator, the X matrix of data and the vector y of the labels have been splitted into two sets, one for training and one for testing. The set of training data is used to model the Gaussian probability distribution, the set of testing data is used to compute the accuracy of the model created by the training set. The split has been done before with the original data matrix X, then using the matrix made with the projection of data on the first two principal components (Figure 5) and, at the end, using the matrix made with the projection of data on the third and fourth principal components (Figure 6). Accuracies resulting were the following:

Accuracy of Classification by the first two principal component = 96%

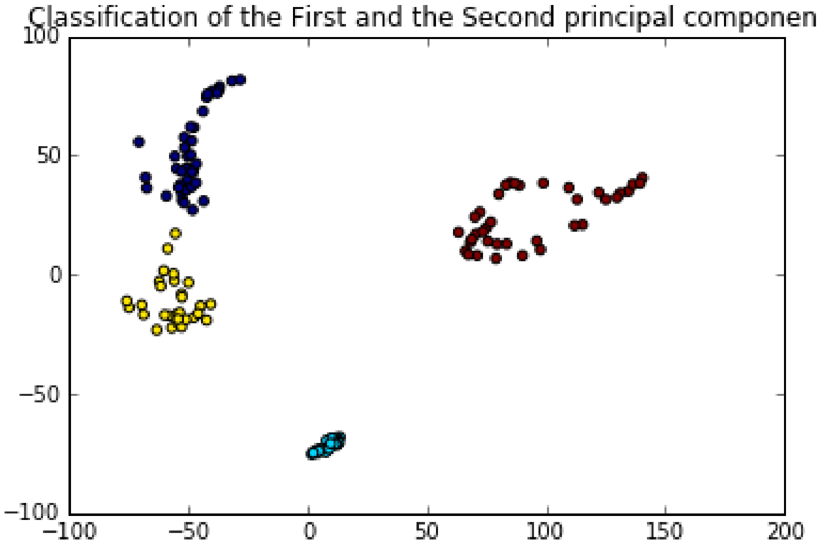

Figure 5

Accuracy of Classification by the third and the fourth principal component= 79%

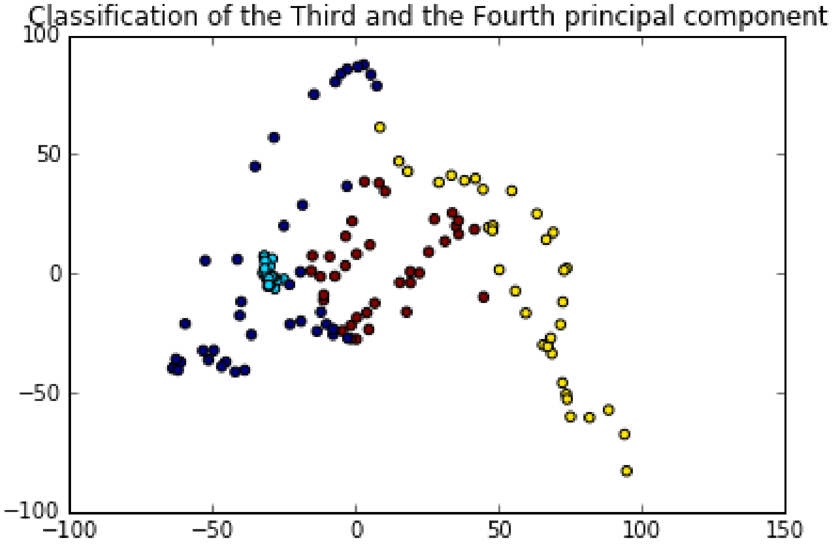

Figure 6

As it is expected the best results are obtained using the dataset of the two first components. In this case the performance of the classifier is excellent by classifying the 96% of all the training dataset. Instead, using the third and the fourth principal component can be seen an accuracy of 79%. From the Figure 2 of the second part of the report, in fact, can be notices that data classifcation is less clear than in the Figure 1. Finally, a plot with the decision boundaries of the classifier from the two first components of Matrix X is presented in Figure 7.

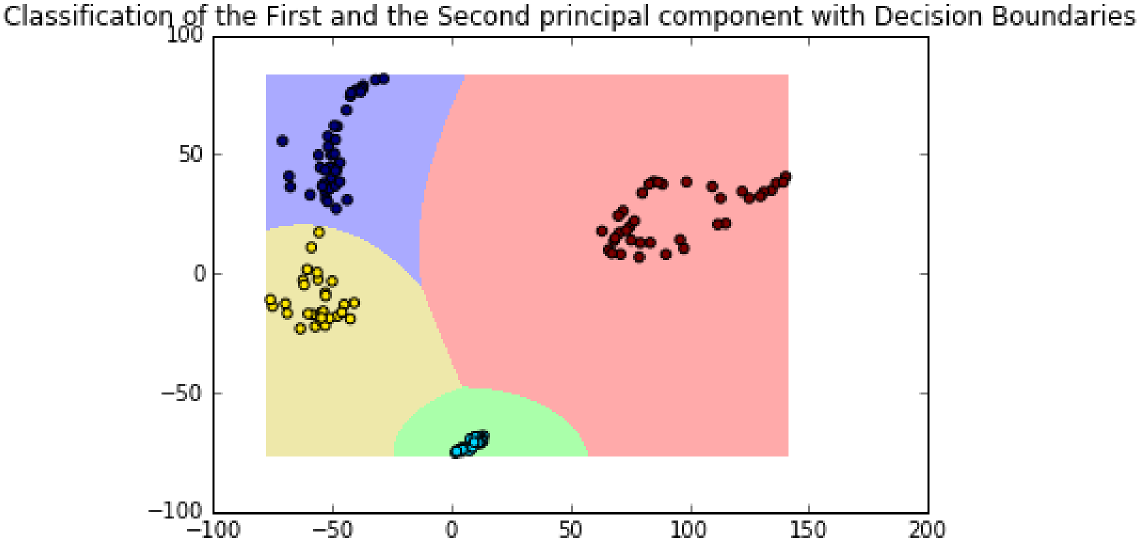

Figure 7