Machine Learning: Third Home Work

KN Neighbors

Edoardo Ghini

December 16, 2016

Dipartimento di Ingegneria dell'Università di Roma La Sapienza

Contents

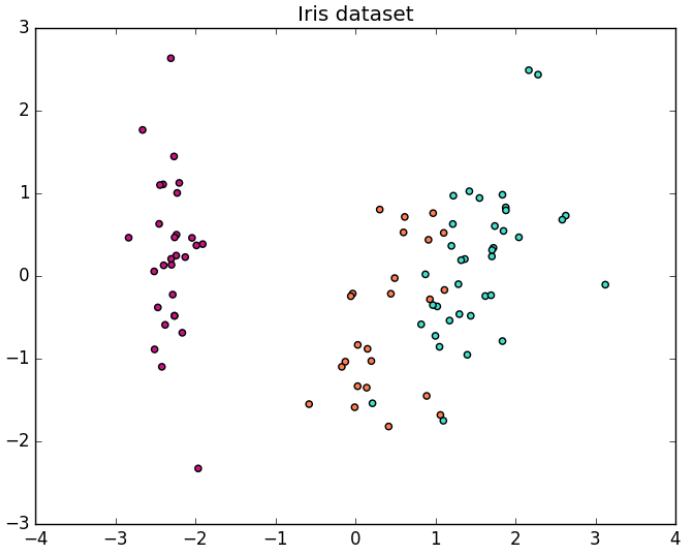


Figure 1

Part I

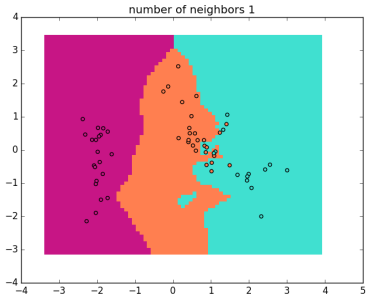
Introduction

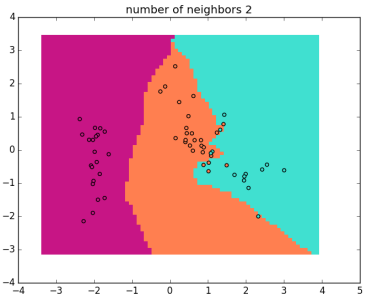
1 Scope

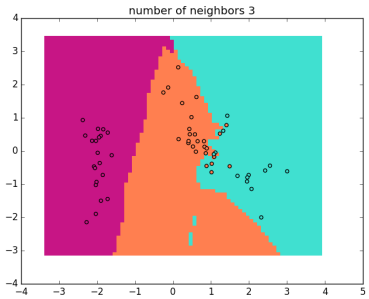
This experience was focused on the comprehension of the theoretic concepts around the classification through KNN algorithm.

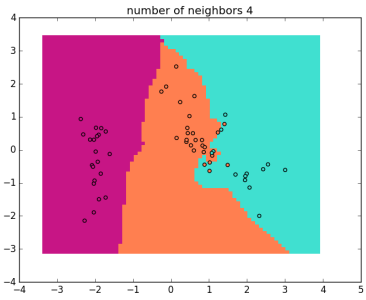
2 Objectives

The overall assignment was about performing a classification task on data in fig(1) experimenting various parameters of the KNN classifier like the number of neighbours considered and also experimenting variations about the metric used to compute the distances between points.









Part II

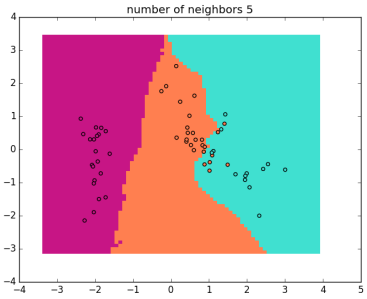
Development

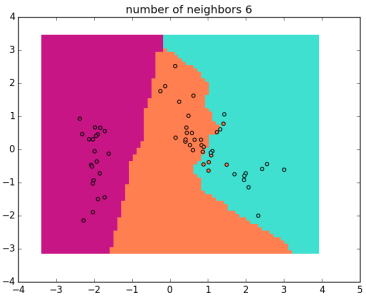
3 Data Manipulations

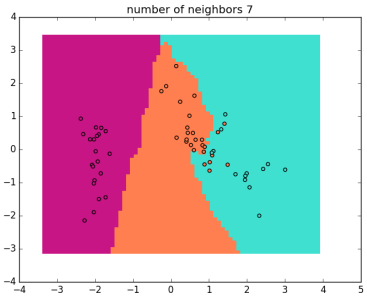
At first, I loaded the Iris dataset and I standardised the input data. After that, I applied a principal component analysis as common practice to maximise the variance carried from selected data. Finally I divided data in test and train subsets.

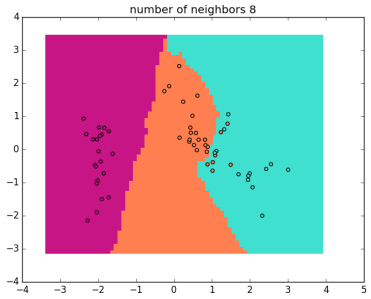
4 Classification

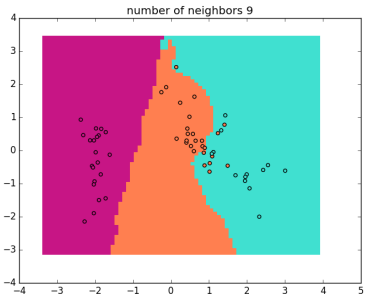
As shown in these plots, I performed the classification of a KNN model with parameters which define the number of neighbours that it would take into consideration assigning the class to new data. This process has been iterated for ten times.

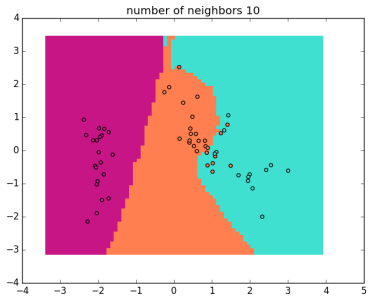

number of neighbors 5


number of neighbors 6


number of neighbors 7


number of neighbors 8


number of neighbors 9


number of neighbors 10

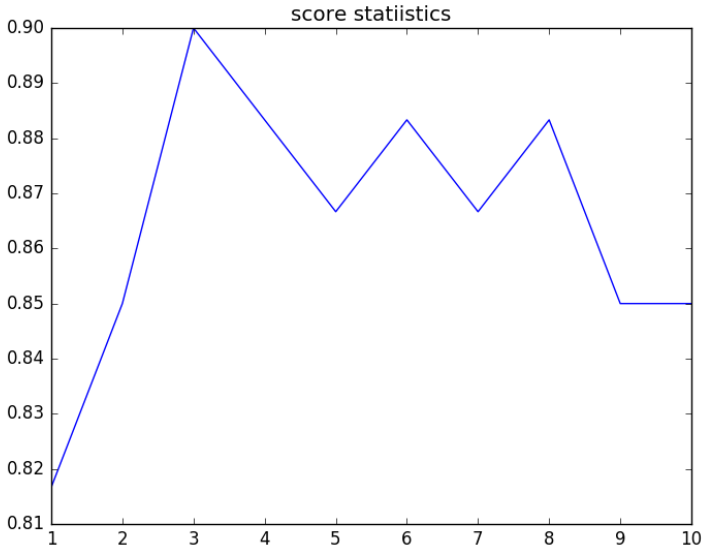


Figure 7

5 Score Analysis

In this particular case, with this particular train and test dataset, the best classifier is a KNN that take into consideration three neighbours (fig(7)).

6 Weights of the Metric

By the time, when the best K parameter was found, I tested the classification with two different weight types for distance calculation: in fig(8) is shown the behaviour of weights that consider the euclidean distance between points, instead in fig(9) there is the result due to the adoption of uniform weights so the classification is made on the absolute number of neighbours regardless their distances from the data point that is being classified.

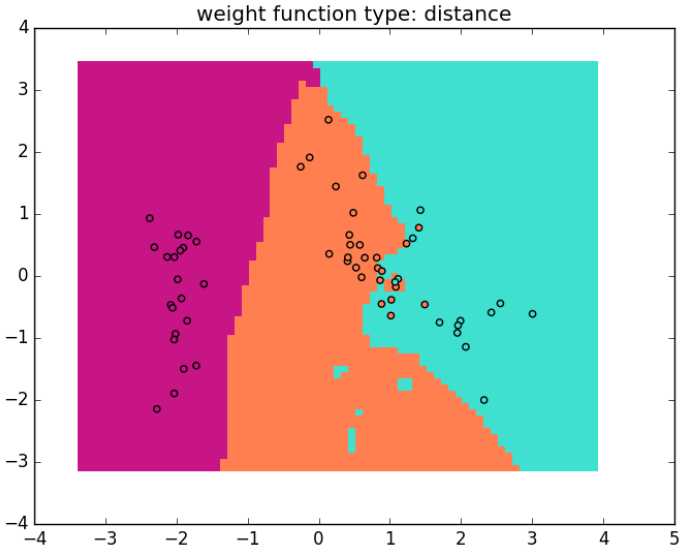


Figure 8

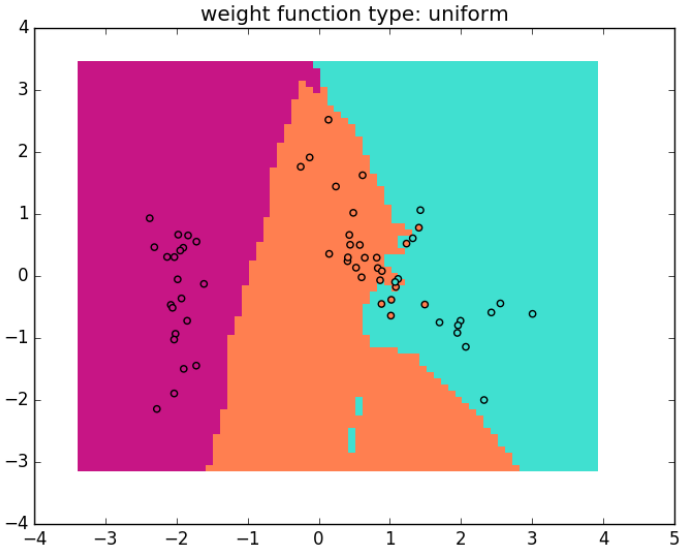
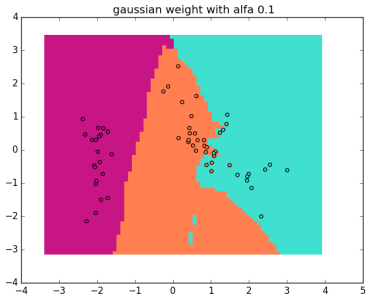
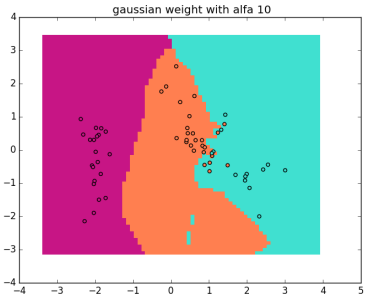
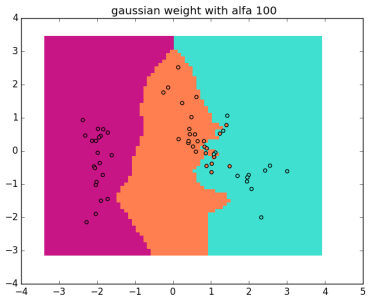
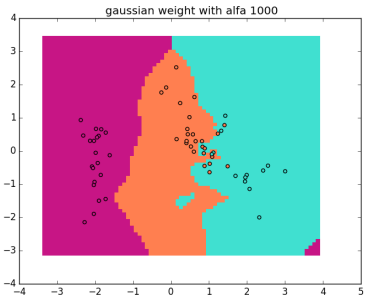


Figure 9









7 Gaussian Distance Metric

Finally I implemented a function on my own in order to compute the distance in a non linear way. These shown below are the results of the classification with the respect of this new metric with different alfa coefficients that affect the steepness of the gaussian function, as it can be seen, in particular, in the last plot with a huge value for alfa

Part III

Conclusions

At the end, collecting the scores of the various classifications, It turned out that for this dataset, the best score around ninety per cent has been achieved with small values of the alfa coefficient.