Report Machine Learning, Homework 3

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 k-Nearest Neighbors algorithm.

1 Introduction

In theory we would always like to predict qualitative responses using the Bayes classifier. But for real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible. Therefore, the Bayes classifier serves as an unattainable gold standard against which to compare other methods. Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability. One such method is the K-nearest neighbors (KNN) classifier. The k-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by N_0. It then estimates the conditional probability for class j as the fraction of points in N_0 whose response values equal j:

$$Pr\left(Y = j|X = x_0\right) = \frac{1}{K} \sum_{i \in N_0} I\left(y_i = j\right)$$

Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

2 k-Nearest Neighbors

As requested, after load and standardize Iris data set, PCA has been used to reduce to two dimensions. Successively, the dataset was split into the training and test set as 60%,40% by using the function train test split() and as requested was chosen K=$\{1, .., 10\}$ In figure 1 is shown the behaviour based on the change of value of K.

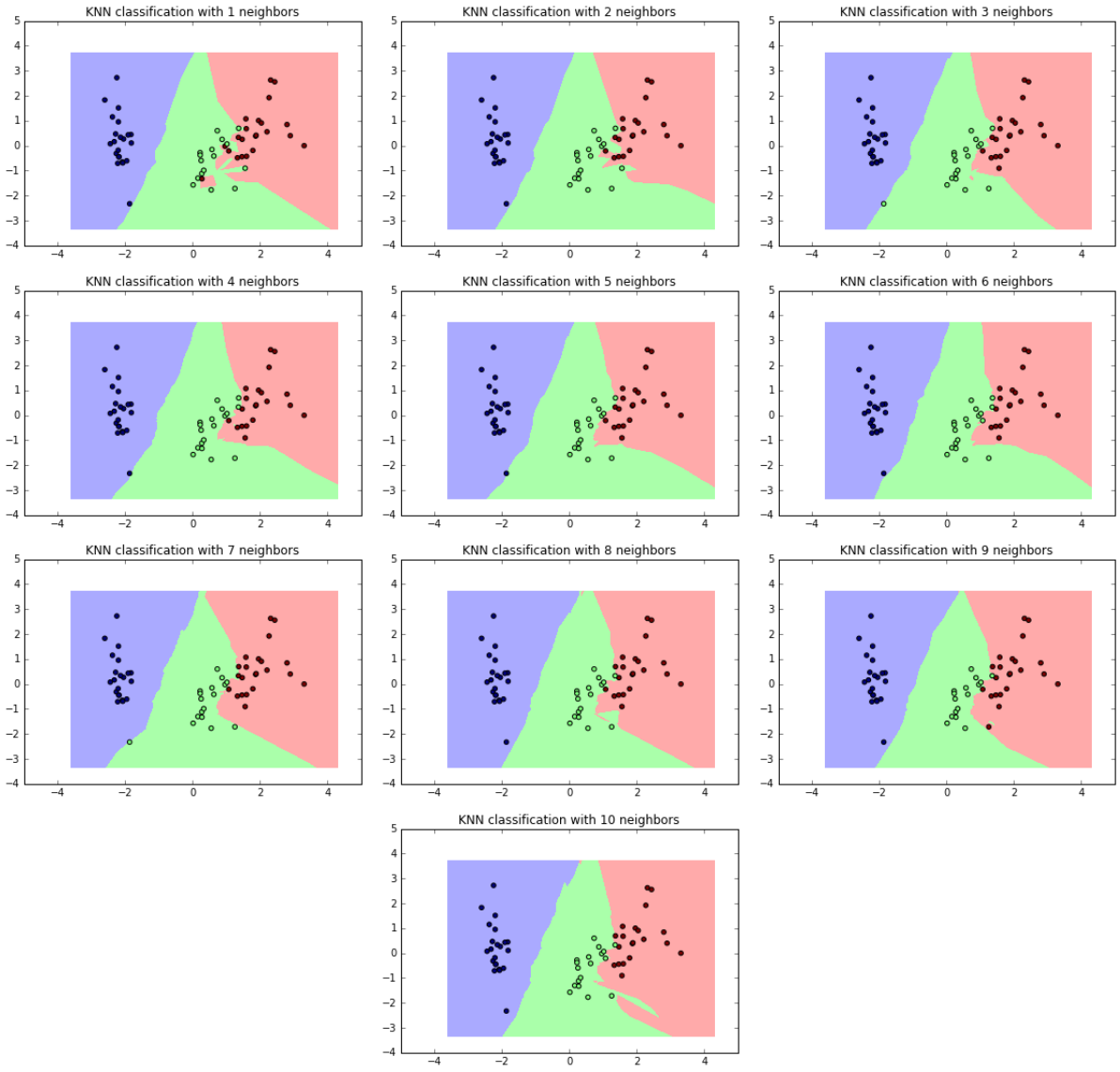

Figure 1

As we can see in Figure 1, when K = 1, the decision boundary is overly flexible and finds patterns in the data that dont correspond to the Bayes decision boundary. This corresponds to a classifier that has low bias but very high variance. As K grows, the method becomes less flexible and produces a decision boundary that is close to linear. This corresponds to a low-variance but high-bias classifier.

3 Weighted KNN

In this section, firstly, as requested, was chosen K=3 and shown how the decision boundary changes when using different weight functions like 'uniform' and 'distance'. This part is shown in Figure 2.

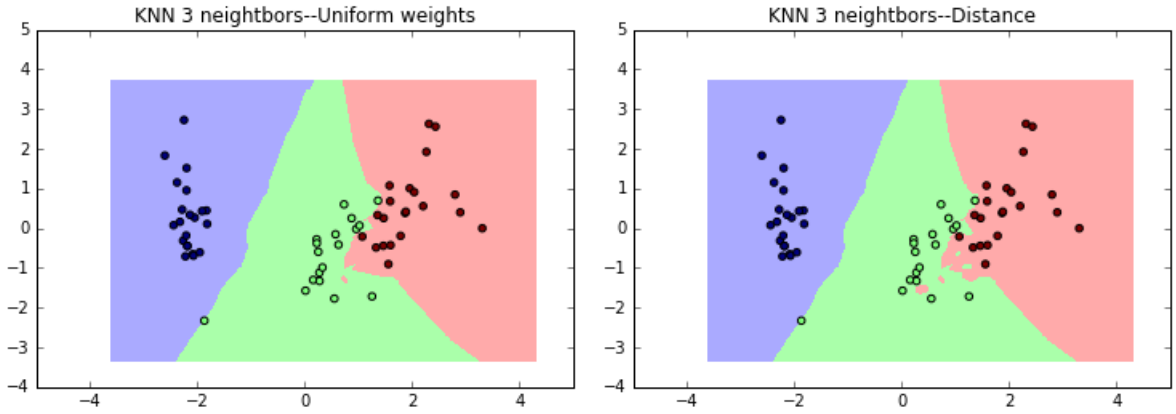

Figure 2

After, it was requested to write a weight function to compute the gaussian function of the square of the distance:

$$\omega = e^{-\alpha d^2}$$

and then has been plotted, as we can see in Figure 3, decision boundaries for $\alpha = [0.1, 10, 100, 1000]$ with K=3

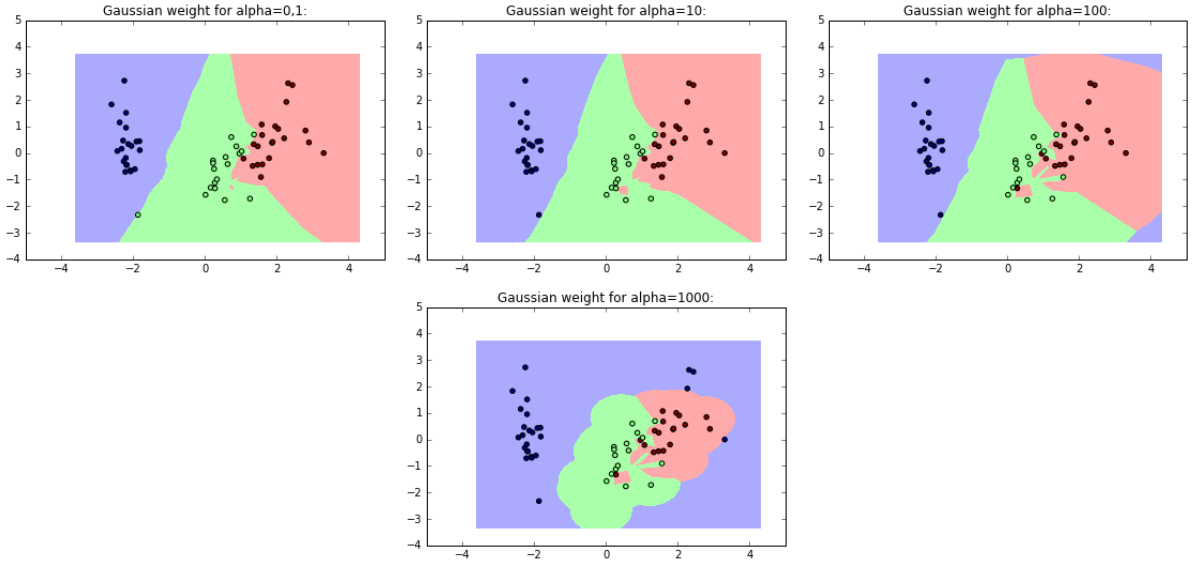

Figure 4

In this case, when α assumes highest value, the function tends faster to 0. For this reason the boundaries depends from α value.

4 Best Weight

The last part of this Homework requested to find the optimal number of neighbours and the best weight function for this data trying uniform, distance and my gaussian function. Figure 5 shows as follow all the accurancies.

N-Neighbors	1	2	3	4	5	6	7	8	9	10
KNN_unweighted	0.883	0.883	0.883	0.916	0.9	0.933	0.9	0.933	0.95	0.95
KNN_distance	0.883	0.883	0.883	0.9	0.9	0.9	0.883	0.916	0.916	0.916
KNN_Gaussian_α=0.1	0.883	0.883	0.883	0.9	0.9	0.9	0.9	0.916	0.95	0.916
KNN_Gaussian_α=10	0.883	0.883	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
KNN_Gaussian_α=100	0.883	0.883	0.883	0.883	0.883	0.883	0.883	0.883	0.883	0.883
KNN_Gaussian_α=1000	0.816	0.816	0.816	0.816	0.816	0.816	0.816	0.816	0.816	0.816

Figure 5

As I have said previously, the α must be equal to a lower value, and the K value when K grows, the method becomes less flexible and produces a decision boundary that is close to linear. Fot these reasons, it has been chosen the gaussian weight function with $\alpha = 0,1$ and K=9. Figure 6 shows decision boundary of best solution with an accuracy equal to 0,95.

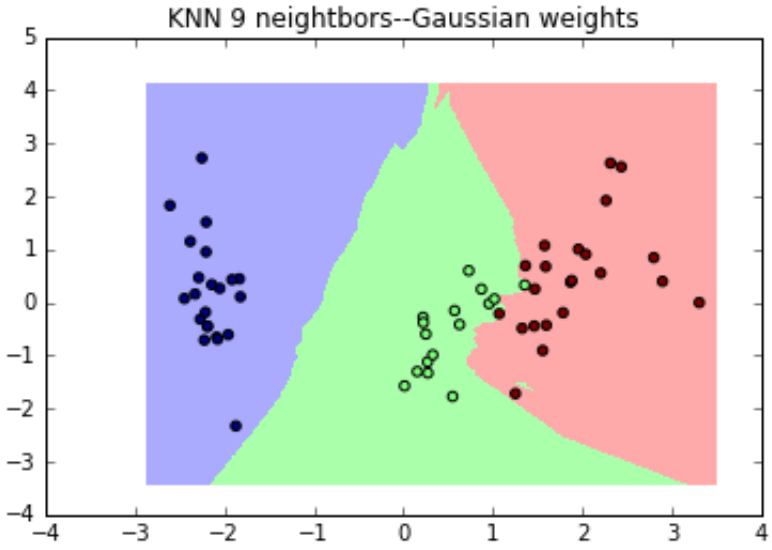

Figure 6