Homework 1 Machine Learning

Marco Treglia

1 K nearest neighbors

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.

1.1 Metrics

Distance fuction :

Euclidean

$$d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Manhattan

$$d = \sum_{i=1}^{k} |x_i - y_i|$$

Minkowski[1]

$$d = \left(\sum_{i=1}^{k} (|x_i - y_i|)^p \right)^{\frac{1}{p}}$$

1.2 Weights

One of the straight forward extension is not to give 1 vote to all the neighbors. A very common thing to do is weighted kNN where each point has a weight which is typically calculated using its distance. For eg under inverse distance weighting, each point has a weight equal to the inverse of its distance to the point to be classified. This means that neighboring points have a higher vote than the farther points. Weights :

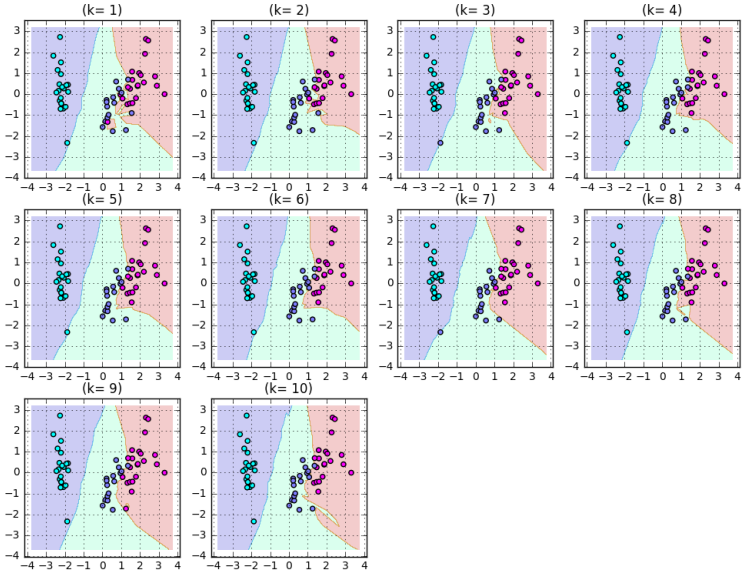
Uniform
All points in each neighborhood are weighted equally.

Distance
Weight points by the inverse of their distance.

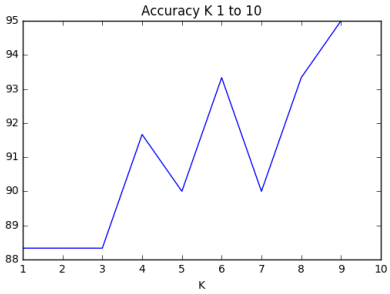
[1] With $p = 1$ and $p = 2$ rapprensent respectively Manhattan and Euclidean distance

2 K nearest neighbors Visualizzation

2.1 Number of neighbors one to nine

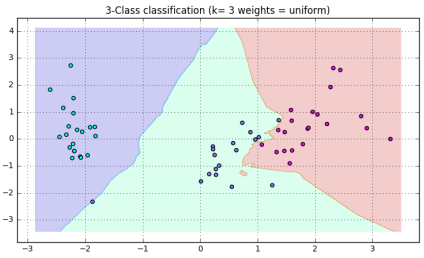


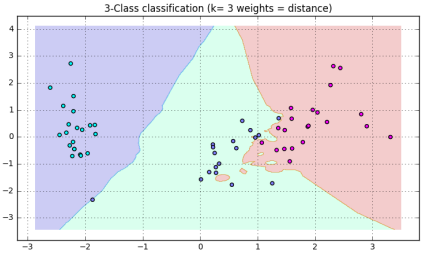
We can notice that the boundaries change for the result of considering respectly 1 to 10 neighbors for determinig at wich class the feature belong to. We can observe that if considering only one neighbor then we can be effected of overestimate the class with the noise of the data and considerign more neightbors the accuracy "might" increase but the computation cost also increases.



Accurancy according the number of neighbors.

2.2 Number of neighbors 3 with different weights fuction

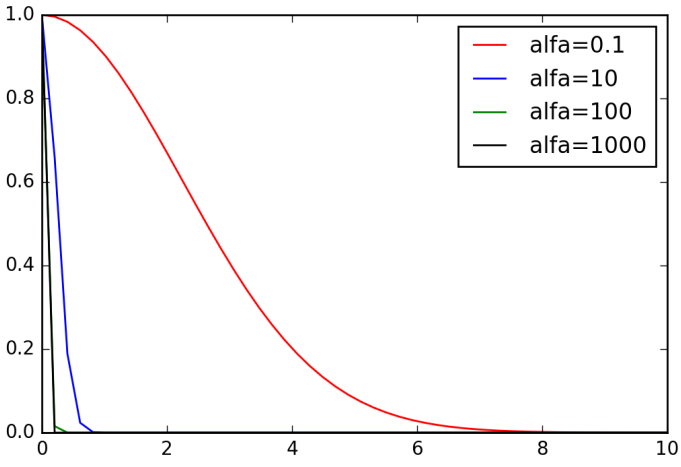




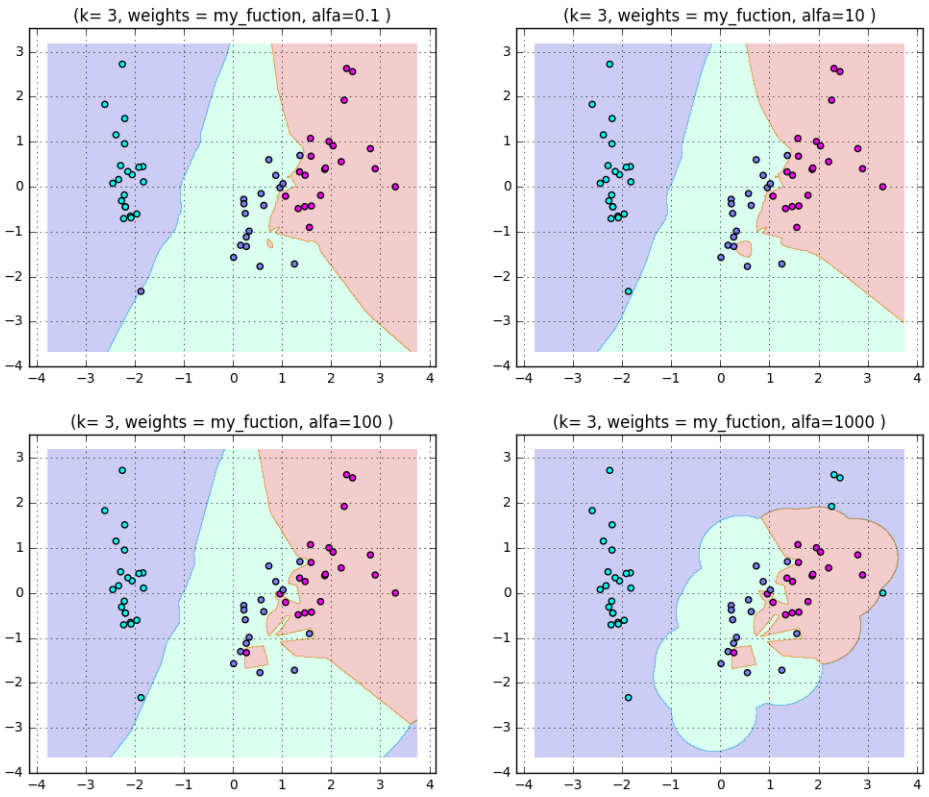
Weight	Accuracy
Uniform	88.33 %
Distance	88.33 %

2.3 Gaussian function as weight function

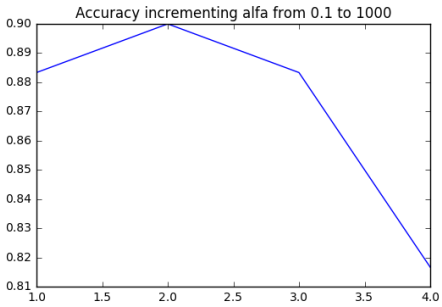
$$w = e^{-\alpha d^2}$$



Distance is positive so the fuction is only decremental.

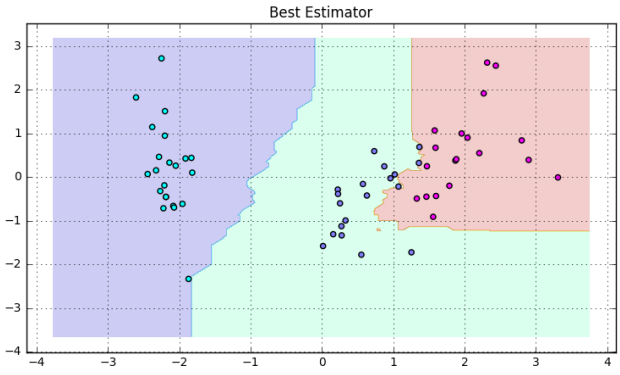


Accuracy incrementing alfa . K = 3



In this case we are using the gauss fuction as weight for estimante the k near neighbors , we can see that in the case of a big valuea of α the fuction tend to zero faster then the case of the small one. This means that when we are weighting the distance we can get some probability of commit an erroneous valutation. Infact, as shown at the end, the gauss fuction with a smaller value of α can predict with a nice accuracy.

2.4 Best solution finded with GridSearchCV



$$\text{Accuracy} = 96.6\ \%$$

After a grid search, the best parameter wich gives the higest accuracy are nine number of neightbors with a manhattan metric and the gaussian fuction as weight value.

Number Neighbors	Weight	Metric
9	Gaussian Func.	Manhattan distance