K-Nearest Neighbors

Author: Rodriguez Noh Santiago Miguel
Professor: Dra. Anabel Martin Gonzalez
Link to code: https://github.com/Santiagomrn/K-NN

I. Development of practice

A. Implement the kNN function inside run knn, you can use the utility function from l2 distance to compute distance between different data points.

Write a separate script that runs kNN for different values of k and plot the classification rate on the validation set (number of correctly predicted cases, divided by total number of data points) as a function of k.

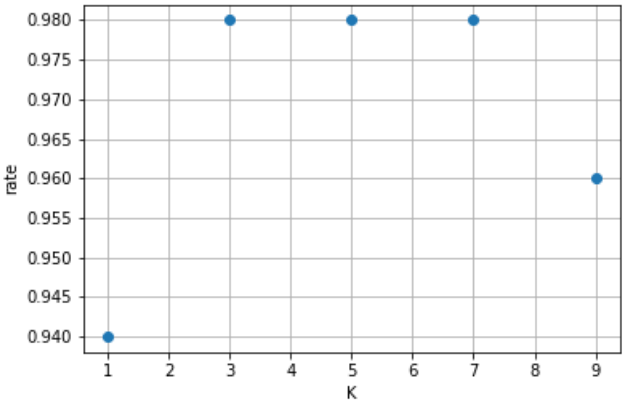

Fig. 1. Classification Rate versus K in validation set

1) Comment on the performance of the classifier and argue which value of k you would choose.

   The performance is good.
   Based on the graph k = 5 it looks like the best option to test with the test set.

   What is the classification rate on test set of your chosen value k*? Also compute the rate for k*+2 and k*-2 on test set.

   I choose K=5

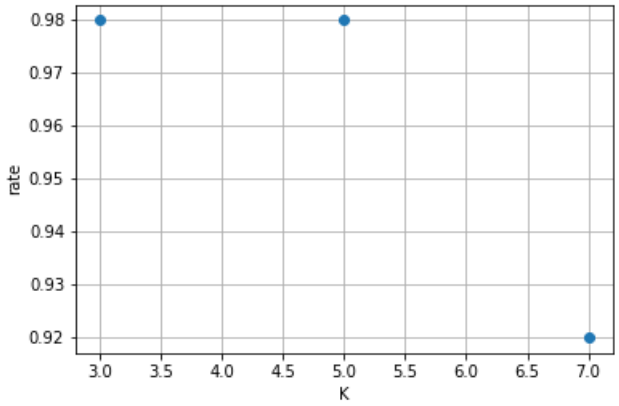

Fig. 2. Classification Rate versus K in test set

2) Does the test performance for these values of k correspond to the validation performance? Why or why not?

   The classification rate at k = 3 and k = 5 is the same, but not at 7. This result is because the data is not the same and the model is not perfect, it is also likely that more training data is needed.