CSC-691 — Data Mining
Assignment 1

Esteban Murillo

Saturday 7th September, 2019

Analysis

As it can be seen down below in Figure 1, the accuracy rate depends on the value we choose for our ***k***. For this particular example the ten different examples provided with 'phoneme-10-fold' were used. The main reason for this to happen is because as the ***k*** grows, so does the neighborhood around our point. When this happens, new values might interfere with the actual prediction as can be appreciated in Figure 2.

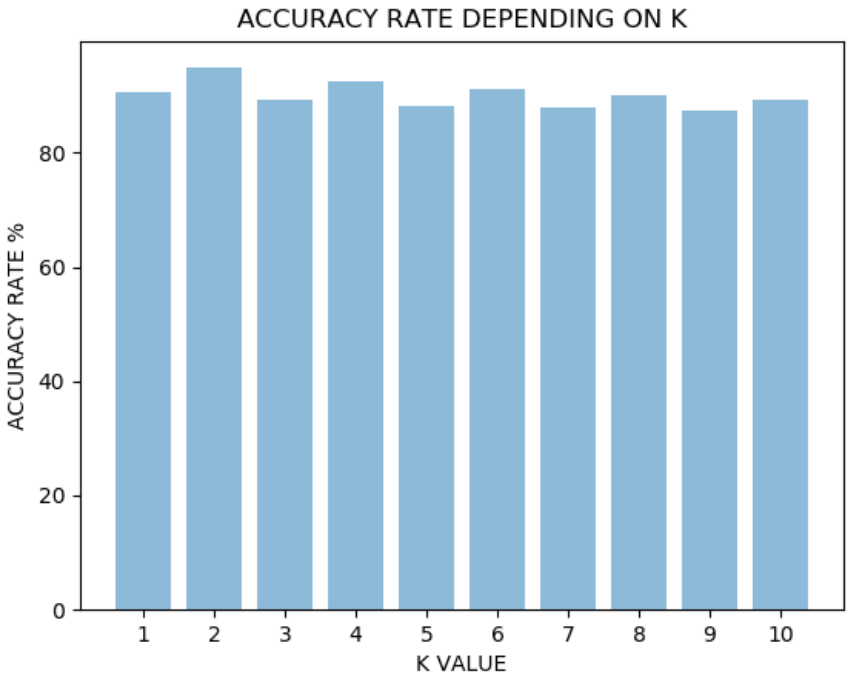

Figure 1: Yielded accuracy rates for the dataset 'phoneme'

As seen above, the final values for the chosen dataset ranged between $\sim$ 87% and $\sim$ 95%. From all the testing done, we saw that accuracy rates got really high whenever the training dataset was very large. Finally it is worth noting that when this condition was met, the higher rates were gotten whenever we tried with a relatively small $\boldsymbol{k}$. The final yielded values were: 90.61%, 94.85%, 89.11%, 92.35%, 88.11%, 91.06%, 87.87%, 90.06%, 87.45%, 89.24% for $\boldsymbol{k}$ = 1, 10.

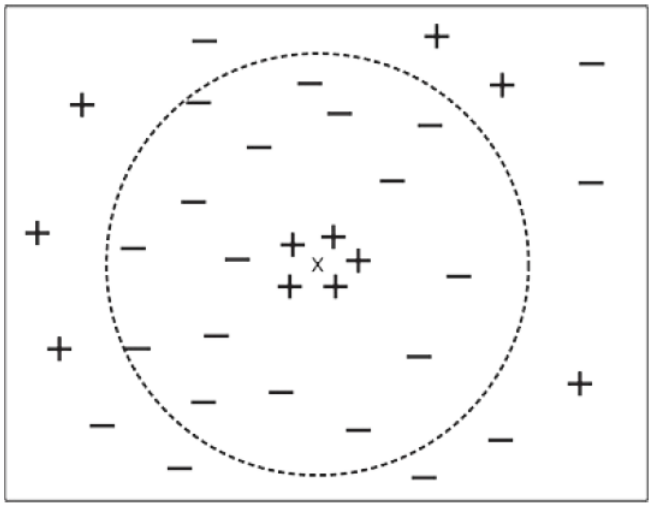

Figure 2: Possible problems for large $\boldsymbol{k}$ values. Image taken from [1]

From Figure 2 we can see that the '-' symbol made some noise to the actual prediction and might end up compromising the validity of our guess.

Notes

- When entering the name of the dataset that you want to analyze, it should be done in the following way: "path_to_folder/phoneme-10-fold/phoneme-10"
- An extra library is needed for plotting. The name of the library is **matplotlib**
- Just run the program and a menu is going to guide you through
- A plot is going to be shown after completing all values $\boldsymbol{k}$**'s**
- You might want to go for a walk if you decide to go for option 3

References

[1] *Introduction to Data Mining.* Accesed on September 6th.