ROB311 - Apprentissage pour la robotique

TP1 - KNN

Mateus Lopes Ricci
Matheus Melo Monteverde

20 Septembre 2020

1 Introduction

The Nearest Neighbor (KNN) based classification algorithm is a widely used technique to recognize patterns. The core of its operation is to find the nearest k neighbours of a given element and determine the most recurrent class among those neighbours. In the following we propose to analyse two case studies and verify the influence of some parameters on the results obtained.

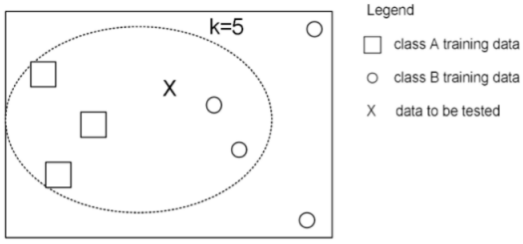

FIGURE 1 – Knn for k = 5

1.1 Breast Cancer Wisconsin (Diagnostic)

The first data set is the "Breast Cancer Wisconsin (Diagnostic) Data Set". The data set contains 2 classes : benign and malignant and the features are obtained by computing a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The attribute information are described below :

1. Sample code number ID number
2. Clump Thickness 1 - 10
3. Uniformity of Cell Size 1 - 10
4. Uniformity of Cell Shape 1 - 10
5. Marginal Adhesion 1 - 10
6. Single Epithelial Cell Size 1 - 10
7. Bare Nuclei 1 - 10
8. Bland Chromatin 1 - 10
9. Normal Nucleoli 1 - 10
10. Mitoses 1 - 10
11. Class : 2 for benign, 4 for malignant

1.2 Haberman's Survival Data Set

The second data set is the "Haberman's Survival Data Set". The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. The attribute information are described below :

1. Age of patient at time of operation numerical
2. Patient's year of operation year - 1900, numerical
3. Number of positive axillary nodes detected numerical
4. Class (Survival status) : 1 = the patient survived 5 years or longer, 2 = the patient died within 5 year

2 Breast Cancer Wisconsin (Diagnostic)

For this study case, we programmed an algorithm based on KNN method. There was 16 elements with missing parameters and as it was not desirable to include these data in the analysis, it was necessary to replace these parameters values for high ones so that they would not influence on the tests. Moreover, for a better result analysis it was decided to shuffle the data before the training and run the algorithm multiple times (100 times for this case). At the end it was obtained the accuracy ; the rate of true positive, true negative, false positive and false negative ; the standard deviation, and the confusion matrix. For this case we considered 3 cases with the number of nearest neighbors, k = 3, k = 5 and k = 7.

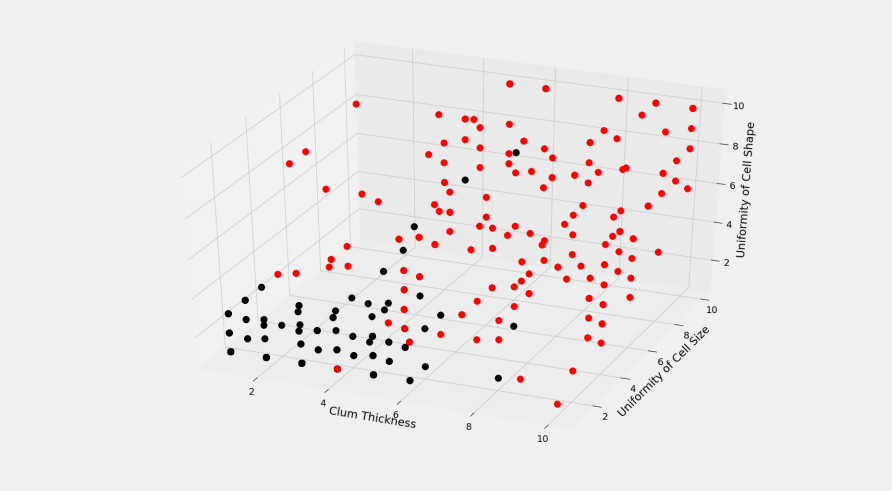

FIGURE 2 – Plot of the Breast Cancer Wisconsin's data set. The red dots represent the malign cases and the black ones represent the benign cases

Mean Accuracy	Standard Deviation
0.96453	0.01457

(a) Mean Accuracy and Standard Deviation

	Actual Positive	Actual Negative
Predicted Positive	0.63611	0.01582
Predicted Negative	0.01964	0.32841

(b) Confusion Matrix

TABLE 1 – Results for k = 3

Mean Accuracy	Standard Deviation
0.97071	0.01291

(a) Mean Accuracy and Standard Deviation

	Actual Positive	Actual Negative
Predicted Positive	0.64215	0.01359
Predicted Negative	0.01568	0.32856

(b) Confusion Matrix

TABLE 2 – Results for k = 5

Mean Accuracy	Standard Deviation
0.96863	0.01440

(a) Mean Accuracy and Standard Deviation

	Actual Positive	Actual Negative
Predicted Positive	0.63417	0.01345
Predicted Negative	0.01791	0.33446

(b) Confusion Matrix

TABLE 3 – Results for k = 7

3 Haberman's Survival Data Set

For this study case, we programmed an algorithm based on KNN method. For a better result analysis it was decided to shuffle the data before the training and run the algorithm multiple times (100 times for this case). At the end it was obtained the accuracy ; the rate of true positive, true negative, false positive and false negative ; the standar deviation, and the confusion matrix. We considered the number of nearest neighbors k = 5.

For this case, we analyzed the influence of the different features in the results of the prediction.

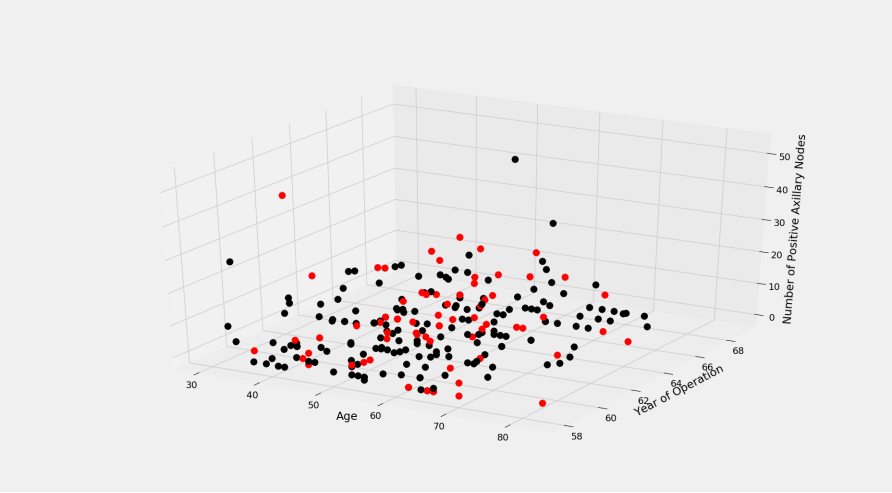

FIGURE 3 – Plot of the Haberman's data set. The red dots represent if the patient died within 5 year and the black ones represent if the patient survived 5 years or longer

Mean Accuracy	Standard Deviation
0.72393	0.04570

(a) Mean Accuracy and Standard Deviation

	Actual Positive	Actual Negative
Predicted Positive	0.65377	0.08229
Predicted Negative	0.19377	0.07016

(b) Confusion Matrix

TABLE 4 – Results with all features

Mean Accuracy	Standard Deviation
0.70327	0.05922

(a) Mean Accuracy and Standard Deviation

	Actual Positive	Actual Negative
Predicted Positive	0.65508	0.07459
Predicted Negative	0.22213	0.04819

(b) Confusion Matrix

TABLE 5 – Results without "Age" feature

Mean Accuracy	Standard Deviation
0.75295	0.05343

(a) Mean Accuracy and Standard Deviation

	Actual Positive	Actual Negative
Predicted Positive	0.66557	0.06295
Predicted Negative	0.18409	0.08737

(b) Confusion Matrix

TABLE 6 – Results without "Year of Operation" feature

4 Conclusion

We find that the features utilized for the training is very important and the results may vary in function of the features chosen. As an example, we may cite the features of the study case 1 (section 2) in which we had to drop the feature "ID" as it didn't have any correlation with the diagnosis. We have also analyzed the influence of the number of nearest neighbors to be considered for the determination of the prediction for the study case in section (2). In the case analyzed we haven't find a big difference in the results, obtaining the best result for k = 5, with accuracy of 97.071% and standard deviation of 1.291%. As for the second study case we analyzed the influence of the features "Age" and "Year of Operation" (section 3) and we obtained the best mean accuracy for the case in which we don't consider the "Year of Operation" with an accuracy of 75.295% and standard deviation of 5.343%.

After these analysis, we conclude that this method can be very efficient to predict the class of a given element if we have a data set with the class already determined. Moreover, it can be concluded that the choice of the features is very important for the final results.

Another important aspect, is the distribution of the elements in the data set. For the first case we can see that the elements have a clear separation concerning the three features showed in figure 2 (Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape). As for the second case 3, concerning the three features showed (Age, Year of Operation, Number of Positive Axillary Nodes), the elements do not present a clear separation. The influence of the distribution of the data set may be seen in the final results, as the rate of accuracy of the second case is lower than the first one.