TP1 KNN : K-nearest neighbors algorithm

Yan Chen & Dajing GU

September 2020

In this TP, a KNN algorithm is implemented in python and two datasets: Breast cancer wisconsin and Haberman are used to verify the code. Code associated can be found in the GitHub repository: TP1 KNN.

1 Principle and Steps of KNN

KNN means K nearest neighbors, which means that each sample can be represented by its K nearest neighbors. The category of one sample is decided by the categories of it's K(a number decided by us) nearest neighbors. This method is easy to understand and to implement, which doesn't need training.

The method is consisted by 4 steps:

1) Data preparation and and pre-processing

2) Distance calculation from the test sample point (the point to be classified) to each other sample point in the training set

3) Distance sorting and then selecting the K points with the smallest distance

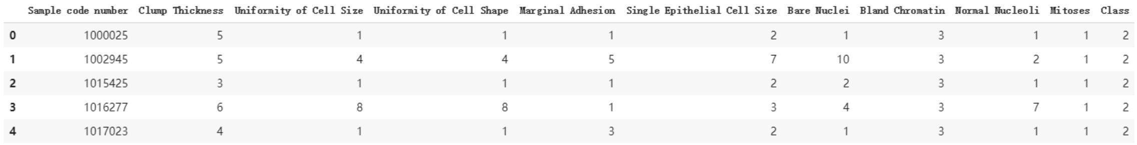
4) Comparing the categories to which the K points belong, and classifying the test sample points to the category with the most the points

2 Notion

Generally the value of K affects the accuracy of classification, the number of 3 is chosen in this TP for it has a relatively higher accuracy. The method of random is used in the split of the data set, with a random seed fixed to keep the result unchanged in each execution.

3 Breast cancer wisconsin

This breast cancer databases[1] was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. There are 11 types of attribute information in this data set whose names are respectively: Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and Class. The survival status is divided into two categories, one is 2 which means the benign cancer, the other is 4, which means the malignant cancer.



	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
0	1000025	5	1	1	1	2	1	3	1	1	2
1	1002945	5	4	4	5	7	10	3	2	1	2
2	1015425	3	1	1	1	2	2	3	1	1	2
3	1016277	6	8	8	1	3	4	3	7	1	2
4	1017023	4	1	1	3	2	1	3	1	1	2

Figure 3.1: The content of data set with attribute information

After data loading and data pre-procession (elimination of the whole line where it lacks data), the data set is splitted randomly into two parts: the training set and the test set. After procession, the total data set has 683 samples. The test set accounts for 20% of the total data set. The data in each feature have been plotted in figure 3.2.

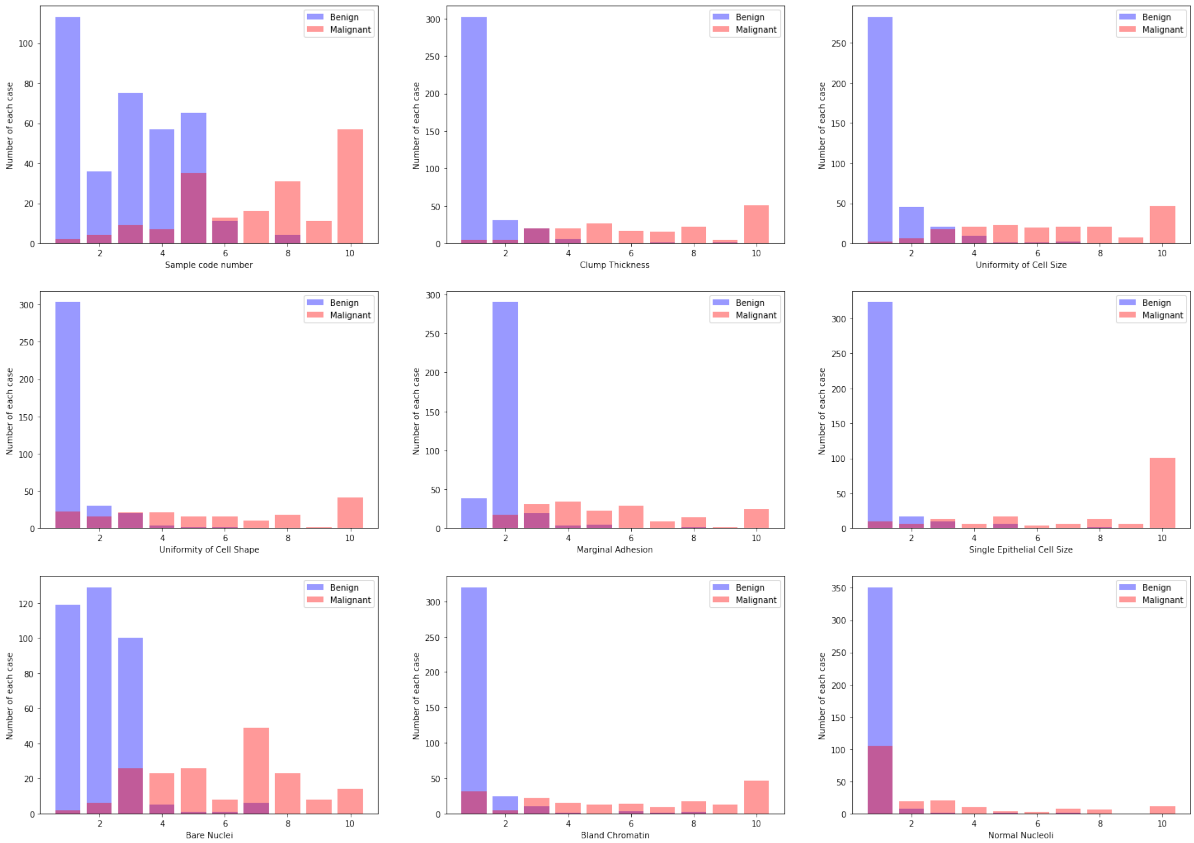


Figure 3.2: Illustration of data in each feature of data set Breast- cancer-wisconsin; *Blue* - [Benign]; *Red* - [Malignant]

It can be seen that the benign cancer has small values in all the features while the values for malignant cancer are more various and often bigger. In this case, the euclidean distance is chosen to be the distance. With the python program, we have calculated the confusion matrix and plotted it in figure 3.3.

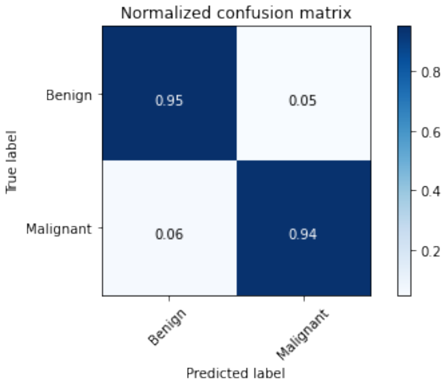


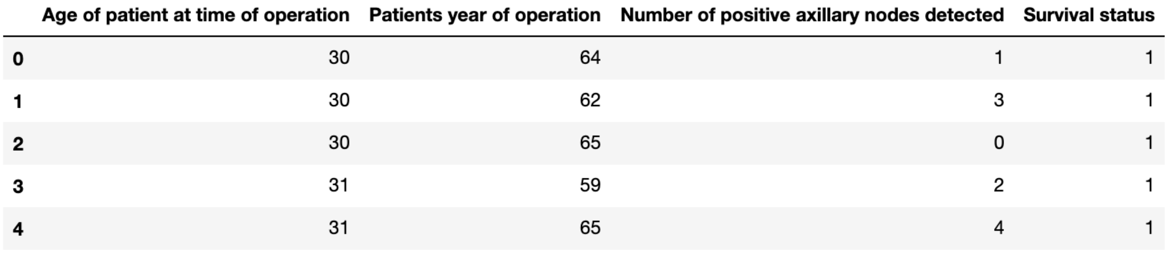
Figure 3.3: Normalized matrix of data set Breast-cancer-wisconsin

It shows that the KNN has a good effect in classifying the test samples, with an accuracy of over 95%.

4 Haberman

This data set contains cases from a study conducted at the University of Chicago Billings Hospital from 1958 to 1970, which involved patients for undergoing surgery for breast cancer[1].

The Attribute information of the data set is respectively: Age of patient at time of operation Patient's year of operation Number of positive axillary nodes detected Survival status (class attribute). The survival status is divided into two categories. One is 1=the patient survived 5 years or longer, the other is 2=the patient died within 5 year. The content of data set with attribute information are shown in the figure 4.1.



	Age of patient at time of operation	Patients year of operation	Number of positive axillary nodes detected	Survival status
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1

Figure 4.1: The content of data set with attribute information

The data set Haberman and the data set breast cancer wisconsin have been performed the same pre-procession. After procession, the total data set has 306 samples. The test set accounts for 20% of the total data set. These diagrams shown in the figure 4.2 present intuitively the relation between each attribution information and survival status and it has been plotted by the same way as the previous data set. Note : the data diagram shown below is that of train set, not whole data set.

There are fewer patients who can survive for more than 5 yeas than those who can't. Most patients who are ill and undergo surgery are always middle-aged or elderly. We have similar distribution about patient's year of operation. Both survivors beyond 5 years and deaths within 5 years had mostly low level of number of positive auxiliary nodes detected.

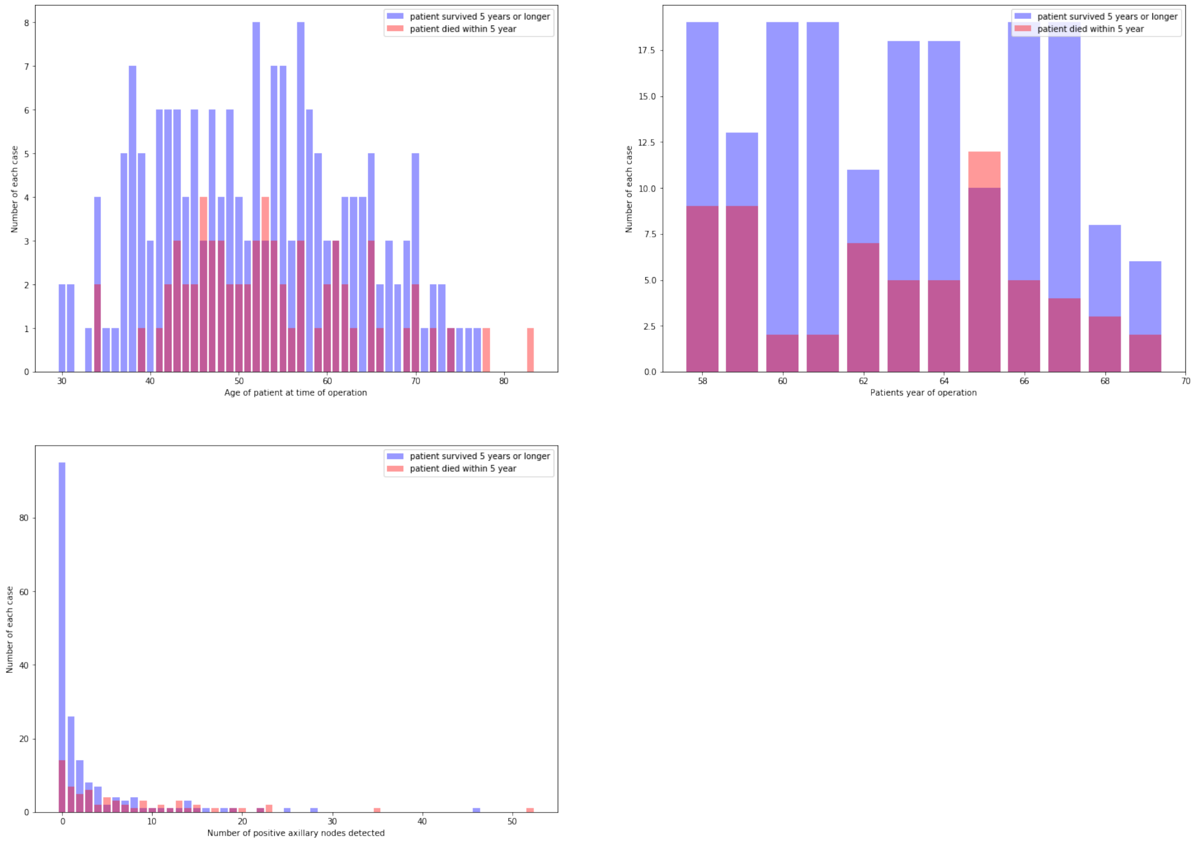


Figure 4.2: Illustration of data in each feature of data set Haberman; *Blue* - [Patient survived 5 years or longer]; *Red* - [Patient died within 5 years]

In order to check intuitively the accuracy, we have plotted the relative confusion matrix shown in the figure 4.3.

According to the confusion matrix, more than 55% of patient died within 5 years samples were misjudged as patient survived 5 years or longer, which is not a great performance. But 93% patient survived 5 years or longer have been judged accurately. The total accuracy is around 77%.

The low accuracy can be explained by the lack of obvious difference between the two classed in the three existent features. More features should be measured in order to improve the accuracy.

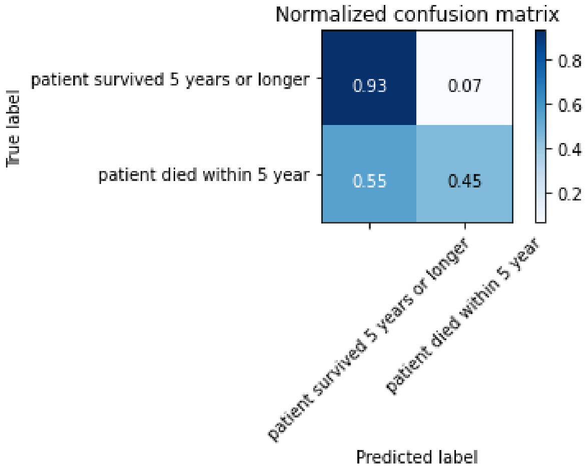


Figure 4.3: Normalized Confusion matrix of data set Haberman

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.