"K-NN Algorithm Implementation for Analysis of Breast Cancer in woman and Prostate Cancer in man"

ALY6020 Week 1 Assignment for Predictive Analytics

Submitted to:

Dr. Michael Prokle
College of Professional Studies
Northeastern University, Boston

Submitted by:

Anupam Maheshwari
Academic Term: Summer Quarter 2018
Northeastern University, MA

Abstract

The Paper talk in brief about the K-NN algorithm and how I have implemented a five step process to study two separate cases and database to predict breast Cancer in woman and paediatric cancer in man. The data was provided by the "Breast Cancer Wisconsin Diagnostic" dataset from the UCI Machine Learning Repository and Analytics Vidhya. The paper also talks about Testing alternative values of k, and reporting the findings in each distinct k value, the boundary values and condition of k in upper and lower scale after which it will turn ineffective to the result is also discussed. The paper also includes the usage, effectiveness, and limitation of K-NN.

With continuous evolving and development of computer brain, majorly called as discipline of artificial intelligence (AI), it has been producing some impressive technological achievements. Advances in image recognition, language understanding, and translation have led to the development of virtual assistants, smart home speakers, and gains in cybersecurity, and they are leading the charge toward autonomous driving. Now, companies have found a way to use those AI smarts to fight and diagnose cancer with nearly accurate precision.

The process involves the construction of artificial neural networks, using software and complex algorithms to recreate the capacity of the human brain to learn.

K-NN is one of the basic and prominent algorithm for classification and identification of dataset in supervised machine learning. Let's assume a sample dataset of items which are homogeneous in nature such that categorical classification exist perfectly for each item. Suppose we have an unlabelled or undiagnosed report which needs to be classified into one of the categorical group. This problem can be solved easily by using K-NN Algorithm by a majority vote of its k neighbours. The selection of K value will be discussed in later section.

Step one for any machine learning process is collecting data. As already stated I have used freely available Wisconsin Breast Cancer and Prostrate Cancer dataset in CSV format. Both of the data sets have features like radius, texture, area and 29 more of factors of similar kind. These are the common feature in the report of cancer diagnosis used by pathology laboratories, the dataset also have unique patient id and diagnosis result which is a classifier or a label. I decided to work on similar kind of data for part B problem so as to confirm and show the similar behaviour of K-NN on different dataset of same medical domain. The data collection these days are mostly structured in medical reporting.

In step two I have prepared the data by importing and reading both CSV files in the system global variable and dropping the patient id to make data set space efficient. This step further includes factoring and labelling diagnosis entry in datasheet more meaningfully, rounding up the values, normalization of all the numerical columns in order to maintain equal importance to each of the diagnosed feature.

Step three consist of dividing and creating test and train dataset for K-NN. In the implementation of part A of assignment there are 569 entries out of which 469 were taken for training and rest 100 are taken for testing. In the implementation of part B of assignment I have split the data in the ratio of 75:25 so as to evaluate how reducing data for training affect prediction results Here I have important point to note that the value of K is equivalent to the square root of trained rows. In the process I have implemented basic K-NN function to predict and diagnose the supplied test dataset.

Step four included evaluating the performance of the K-NN, for it I used GMODEL library cross table function which splits the observation in true positive, true negative, false positive, false negative cells in matrix for both of the dataset. The accuracy of K-NN and value of k is directly proportional to the percentage of true positive and true negative.

Step five is needed if want to perform standardization using z-transformation to see and upgrade the accuracy of the model. Though this never guarantee increased accuracy of model. On trying z-transformation on both of the data set the error percentage grew up by margin of even 30%.

It is failure of K-NN and K value if we get higher percentage of false negative. Below are the insights with different K values tried on each dataset and the resulting output and percentage accuracy of model.

Part A: Breast Cancer in woman with different value of K

Value of K	False Negative	False Positive	Error Percentage
5	**0**	**3**	**3**
11	**1**	**1**	**2**
15	**2**	**0**	**2**
21	**2**	**0**	**2**
27	**2**	**0**	**2**

Selected K value and its result for Breast Cancer diagnosis

```
K: 21
##                  | wbcd_test_pred
## wbcd_test_labels |    Benign | Malignant | Row Total |
## -----------------|-----------|-----------|-----------|
##           Benign |        77 |         0 |        77 |
## -----------------|-----------|-----------|-----------|
##        Malignant |         2 |        21 |        23 |
## -----------------|-----------|-----------|-----------|
##     Column Total |        79 |        21 |       100 |
##                  |     0.790 |     0.210 |           |
## -----------------|-----------|-----------|-----------|
```

Part B : Prostrate Cancer in man with different value of K

Value of K	False Negative	False Positive	Error Percentage
1	**2**	**9**	**44**
2	**2**	**8**	**40**
7	**0**	**6**	**24**
9	**0**	**6**	**24**
11	**0**	**6**	**24**

Selected K value and its result for Prostate Cancer diagnosis

```
K: 9
```

```
##                  | prc_test_pred
## prc_test_labels |  Benign   | Malignant | Row Total |
## ----------------|-----------|-----------|-----------|
##         Benign  |         7 |         6 |        13 |
## ----------------|-----------|-----------|-----------|
##       Malignant |         0 |        12 |        12 |
## ----------------|-----------|-----------|-----------|
##    Column Total |         7 |        18 |        25 |
##                 |     0.280 |     0.720 |           |
## ----------------|-----------|-----------|-----------|
```

- On applying the functions on the data, I can clearly show that the error percentage is proportional to accurately picked K nearest neighbour value. I am hereby getting different percentage error report than what is available in the book. I tried to tweak my function to Z- transformation but got no same result as mentioned in the text book.

- So on keeping k value as the square root of number of values given to training set I got close to diagnosing Breast Cancer accurately by 98%.

- The correctness of the recorded data also plays vital role in the diagnosis because it is only the existing and the labelled data which is predicting the test data. If the labelled data is wrong even if the algorithm accuracy shoots up but the real time results will be disaster.

- If the training data is huge we can take K value as the cube root of the training set. This is proved in the Part A, total trained set was 469, on square rooting we get ~21, while taking cube root ~8 (taking odd value of K to help in tie breaking while voting we take it as 9) which has similar error percentage

- We should definitely take odd value of K to make a tie breaking decision in dataset which have long list of labels and classifiers.

- While experimenting with data I got to see switching or error classification from False Positive to false negative, this kind of error should be checked because False Negative errors cause major chaos and losses.

- While experimenting I tried various combinational ratio of training and testing data set. The accuracy heavily depends on the number training data set provided. If the training set falls below 50% in the initial stage the error percentages are shooting up heavily. I can see this issue on a broader perspective as when we have continuous stream of data, the

bigger and the latest training set we have more will be the accuracy and efficiency of our prediction.

Technological limitations of K-NN :

- Classifying unknown records are relatively expensive
- Accuracy can be severely degraded by the presence of noisy or irrelevant features
- Soft value boundary testing and decision making can't be performed using K-NN

I am looking forward to investigate The MNIST database of handwritten digits using K-NN Euclidian function.

References

• Part- A dataset and analysis, Machine Learning With R: Brett Lantz, Retrieved 13th July, 2018
• Part- B dataset and analysis, https://discuss.analyticsvidhya.com/t/practice-dataset-for-knn-algorithm/3104, Retrieved 13th July, 2018
• Role of Google AI machine learning algorithm in diagnosing cancer, Retrieved 13th July, 2018