VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Essentials of Data Analytics - (CSE3506)

Faculty – Lakshmi Pathi Jakkamputi **Sir**

Lab-6

Harshanth k Prakash

19BCE1293

L21-22 Slot

Tasks for Week-6: K- NN algorithm

Understand the following operations/functions on to perform K- NN algorithm and perform similar operations on 'wdbc' dataset based on given instructions.

AIM

To understand operations/functions on to perform K- NN algorithm and perform similar operations on 'wdbc' dataset based on given instructions.

Algorithm

1. Start
2. Import the dataset "wdbc" using function file.choose().
3. mynorm function is created to find the normalize the values separating each and every value with its min value and dividing it the difference of max and min value in the column.
4. Create a new dataframe named mydata and store all the normalized value of every in that new dataframe except the first column as it is an categorical data.
5. For comparing the original dataset and normalized dataset take 2 to 5 columns of both data set and apply summary () function to find summary.
6. Divide the first 400 values as the train dataset and remaining 169 values to the test dataset from mydata(normalized dataset).
7. Do the knn algorithm and store all the predicted values in pred variable.
8. Form the confusion matrix form using predicted data from pred variable and from 401 to 569 rows in first dataset.
9. Find the accuracy of the data by adding the [1,1] element and [2,2] element and dividing its summation with the whole sum.
10. Stop

Result

```
> head(wdbc)
        V1 V2    V3    V4     V5     V6      V7      V8     V9     V10    V11     V12    V13    V14   V15    V16      V17     V18      V19      V20      V21
1   842302  M 17.99 10.38 122.80 1001.0 0.11840 0.27760 0.3001 0.14710 0.2419 0.07871 1.0950 0.9053 8.589 153.40 0.006399 0.04904 0.05373 0.01587 0.03003
2   842517  M 20.57 17.77 132.90 1326.0 0.08474 0.07864 0.0869 0.07017 0.1812 0.05667 0.5435 0.7339 3.398  74.08 0.005225 0.01308 0.01860 0.01340 0.01389
3 84300903  M 19.69 21.25 130.00 1203.0 0.10960 0.15990 0.1974 0.12790 0.2069 0.05999 0.7456 0.7869 4.585  94.03 0.006150 0.04006 0.03832 0.02058 0.02250
4 84348301  M 11.42 20.38  77.58  386.1 0.14250 0.28390 0.2414 0.10520 0.2597 0.09744 0.4956 1.1560 3.445  27.23 0.009110 0.07458 0.05661 0.01867 0.05963
5 84358402  M 20.29 14.34 135.10 1297.0 0.10030 0.13280 0.1980 0.10430 0.1809 0.05883 0.7572 0.7813 5.438  94.44 0.011490 0.02461 0.05688 0.01885 0.01756
6   843786  M 12.45 15.70  82.57  477.1 0.12780 0.17000 0.1578 0.08089 0.2087 0.07613 0.3345 0.8902 2.217  27.19 0.007510 0.03345 0.03672 0.01137 0.02165
       V22   V23   V24    V25    V26    V27    V28    V29    V30    V31     V32
1 0.006193 25.38 17.33 184.60 2019.0 0.1622 0.6656 0.7119 0.2654 0.4601 0.11890
2 0.003532 24.99 23.41 158.80 1956.0 0.1238 0.1866 0.2416 0.1860 0.2750 0.08902
3 0.004571 23.57 25.53 152.50 1709.0 0.1444 0.4245 0.4504 0.2430 0.3613 0.08758
4 0.009208 14.91 26.50  98.87  567.7 0.2098 0.8663 0.6869 0.2575 0.6638 0.17300
5 0.005115 22.54 16.67 152.20 1575.0 0.1374 0.2050 0.4000 0.1625 0.2364 0.07678
6 0.005082 15.47 23.75 103.40  741.6 0.1791 0.5249 0.5355 0.1741 0.3985 0.12440
>
```

```
> summary(wdbc[,2:5])
       V3              V4              V5               V6
 Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.   : 143.5
 1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu.: 420.3
 Median :13.370   Median :18.84   Median : 86.24   Median : 551.1
 Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean   : 654.9
 3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu.: 782.7
 Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.   :2501.0
> summary(mydata[,1:4])
       V3               V4               V5               V6
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.2233   1st Qu.:0.2185   1st Qu.:0.2168   1st Qu.:0.1174
 Median :0.3024   Median :0.3088   Median :0.2933   Median :0.1729
 Mean   :0.3382   Mean   :0.3240   Mean   :0.3329   Mean   :0.2169
 3rd Qu.:0.4164   3rd Qu.:0.4089   3rd Qu.:0.4168   3rd Qu.:0.2711
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

Confusion matrix:

```
> cf

pred   B   M
   B 128   3
   M   2  36
```

Accuracy of the model:

```
> acc
[1] 0.9704142
>
```

Inference

The accuracy of the KNN model is 97%, so as per the accuracy its clear that the model is best fit model and will predict 97% chance of correct result.

Program

```
rm(list=ls())
wdbc<-read.table(file.choose(),sep=',')
view(wdbc)
head(wdbc)
wdbc<-wdbc[,-1]
mynorm<-function(x){((x-min(x))/(max(x)-min(x)))}
mydata<-as.data.frame(lapply(wdbc[,-1], mynorm))
summary(wdbc[,2:5])
summary(mydata[,1:4])
train<-mydata[1:400,]
test<-mydata[401:569,]
library(class)
pred<-knn(train,test,wdbc[1:400,1],k=21)
cf<-table(pred,wdbc[401:569,1])
cf
acc=(cf[[1,1]]+cf[[2,2]])/sum(cf)
acc
```