R Report

2.KNN

2.1 Business Understanding

Using KNN way to analysis this data fertility_diagnosis.

The data comes from UCI: http://archive.ics.uci.edu/ml/datasets/Fertility. It collected 10 features about the fertility. From the season of analysis, the age at the time of analysis, whether there was any disease in childhood, whether it was accidental or serious trauma, whether it was an operation, whether it had a high fever last year, the number of drinking alcohol times, whether smoking, the number of hours of sitting per day-16 to explore fertility. Whether the output is normal.

2.2 Data Understand & Preparation

1. season:1) winter, 2) spring, 3) Summer, 4) fall. (-1, -0.33, 0.33, 1)
2. age: 18-36 (0, 1)
3. children diseases: (ie , chicken pox, measles, mumps, polio) 1) yes, 2) no. (0, 1)
4. accident/trauma: 1) yes, 2) no. (0, 1)
5. surgical: 1) yes, 2) no. (0, 1)
6. fever: 1) less than three months ago, 2) more than three months ago, 3) no. (-1, 0, 1)
7. alcohol: 1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never (0, 1)
8. smoke: 1) never, 2) occasional 3) daily. (-1, 0, 1)
9. sitting:number of hours spent sitting per day ene-16 (0, 1)
10. diagnosis: normal (N), altered (O)

```
> str(datafertility)
'data.frame':   100 obs. of  10 variables:
 $ season           : num  -0.33 -0.33 -0.33 -0.33 -0.33 -0.33 -0.33 -0.33 1 1 ...
 $ age              : num  0.69 0.94 0.5 0.75 0.67 0.67 0.67 1 0.64 0.61 ...
 $ children_diseases: int  0 1 1 0 1 1 0 1 0 1 ...
 $ accident/trauma  : int  1 0 0 1 1 0 0 1 0 0 ...
 $ surgical         : int  1 1 0 1 0 1 0 1 1 0 ...
 $ fever            : int  0 0 0 0 0 0 -1 0 0 0 ...
 $ alcohol          : num  0.8 0.8 1 1 0.8 0.8 0.8 0.6 0.8 1 ...
 $ smoke            : int  0 1 -1 -1 -1 0 -1 -1 -1 -1 ...
 $ sitting          : num  0.88 0.31 0.5 0.38 0.5 0.5 0.44 0.38 0.25 0.25 ...
 $ diagnosis        : chr  "N" "O" "N" "N" ...
> summary(datafertility)
     season             age        children_diseases accident/trauma    surgical
 Min.   :-1.0000   Min.   :0.500   Min.   :0.00      Min.   :0.00    Min.   :0.00
 1st Qu.:-1.0000   1st Qu.:0.560   1st Qu.:1.00      1st Qu.:0.00    1st Qu.:0.00
 Median :-0.3300   Median :0.670   Median :1.00      Median :0.00    Median :1.00
 Mean   :-0.0789   Mean   :0.669   Mean   :0.87      Mean   :0.44    Mean   :0.51
 3rd Qu.: 1.0000   3rd Qu.:0.750   3rd Qu.:1.00      3rd Qu.:1.00    3rd Qu.:1.00
 Max.   : 1.0000   Max.   :1.000   Max.   :1.00      Max.   :1.00    Max.   :1.00
     fever          alcohol          smoke          sitting        diagnosis
 Min.   :-1.00   Min.   :0.200   Min.   :-1.00   Min.   :0.0600   Length:100
 1st Qu.: 0.00   1st Qu.:0.800   1st Qu.:-1.00   1st Qu.:0.2500   Class :character
 Median : 0.00   Median :0.800   Median :-1.00   Median :0.3800   Mode  :character
 Mean   : 0.19   Mean   :0.832   Mean   :-0.35   Mean   :0.4068
 3rd Qu.: 1.00   3rd Qu.:1.000   3rd Qu.: 0.00   3rd Qu.:0.5000
 Max.   : 1.00   Max.   :1.000   Max.   : 1.00   Max.   :1.0000
> #labels
> table(datafertility$diagnosis)

 N  O
88 12
```

2-1

When finish set the name, can check the data type. The output data(diagnosis) is character type.others have numeric and integer. In total, it has 100 data. The normal diagnosis has 88, the altered has 12. So the training data can choose 70, others belong to testing.

2.3 Modeling

```
> #predict1
> predictions1=knn(train=datafertility_train,test=datafertility_test,
+                  cl=datafertility_train_labels,k=1)
> #predict2
> predictions2=knn(train = datafertility_train, test = datafertility_test,
+                    cl = datafertility_train_labels, k=2)
> #predict3 27/30 is good
> predictions3=knn(train = datafertility_train, test = datafertility_test,
+                  cl = datafertility_train_labels, k=3)
> #predict4
> predictions4=knn(train = datafertility_train, test = datafertility_test,
+                  cl = datafertility_train_labels, k=4)
```

```
> cm=table(predictions1,datafertility_test_labels)
> sum(diag(cm))/sum(cm)
[1] 0.8333333
> cm=table(predictions2,datafertility_test_labels)
> sum(diag(cm))/sum(cm)
[1] 0.8
> cm=table(predictions3,datafertility_test_labels)
> sum(diag(cm))/sum(cm)
[1] 0.9
> cm=table(predictions4,datafertility_test_labels)
> sum(diag(cm))/sum(cm)
[1] 0.8666667
```

2-2

Dividing the data into 70 and 30 parts. The KNN does not need build model. Setting datafertility_n has nine features, then distributing 70 data to datafertility_train and 30 data to datafertility_test. The datafertiity has ten features, then distributing 70 data to datafertility_train_labels and 30 data to datafertility_test_labels. To finish the training data and testing data, setting k=1, the accuracy rate is 0.8333. Setting k=2, the accuracy rate is 0.8. Setting k=3, the accuracy rate is 0.9. It is good.

```
> CrossTable(predictions3, datafertility_test_labels,
+            prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE)


   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  30


             | datafertility_test_labels
predictions3 |    Normal |   Altered | Row Total |
-------------|-----------|-----------|-----------|
      Normal |        26 |         2 |        28 |
             |     0.867 |     0.067 |           |
-------------|-----------|-----------|-----------|
     Altered |         1 |         1 |         2 |
             |     0.033 |     0.033 |           |
-------------|-----------|-----------|-----------|
Column Total |        27 |         3 |        30 |
-------------|-----------|-----------|-----------|
```

2-3

In total has 30 data. The result is very close to actual. Comparing with predictions and actual, the predict has two errors in normal and one error in altered. In actual, normal has 27, altered has 3. In predict, normal has 28, altered has 2.

2.4 Evaluation

In conclude, KNN is very powerful. In the absence of a training model, it selects the appropriate range to predict the results. When k is not set properly, it will result in overfitting. But when the range is too wide, unnecessary data will be included in the reference. So the value of k is appropriate, not too big or too small. The results are not accurate. In this analysis, when k=2 and k=4, their predictions are the same. But at k=3, the prediction is the best. Via trying to set the k value, can explore the suitable value.