Homework Assignment 3

IT 531, Summer 2018

**Question 2:** Classification Analysis

Perform classification using Decision Tree and Naïve Bayes using R on following dataset:

Spambase Dataset - <http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/>

You can find more details for the dataset on the UCI repository page. For each algorithm and each dataset, do 10-fold cross validation.

Deliverables:

1. Briefly describe the dataset.
2. Generate confusion matrix on the test data. Calculate TP Rate, FP Rate, Precision and Recall with hand. Describe how you calculate these numbers.
3. Submit a zip file containing the data and R script file(s) and the document describing dataset, output results and TP, FP, precision and recall.

Answer:

1. This dataset include 68 attributes. Among this dataset, 48 attributes are numeric attributes, which represent the occurrence ratio of a some specific word.There are also 6 attributes to present the percentage for some characters. And ther are also attributes to present the features of capital letters. The classifier of the dataset is spam, when it is 1,it means the letter is considered as spam letter.
2. In decision tree

|  |  |  |
| --- | --- | --- |
| Test Pred | email | spam |
| email | 104 | 2 |
| spam | 740 | 514 |

TP= true positive = 104

FP= false positive =740

Precision=TP/TP+FP = a/a+c = 104/(104+740) =~0.123

Recall=TP/TP+FN = = a/a+b =104/(104+2) = 0.98

I used party library also used the summary to verify the number, and also used rPart to generate again, but I don’t know why the precision is so low.

In Naïve Bayes

|  |  |  |
| --- | --- | --- |
| Test Pred | email | spam |
| email | 1564 | 1224 |
| spam | 94 | 1719 |

TP= true positive = 1564

FP= false positive =94

Precision=TP/TP+FP = a/a+c = 1564/(1564+94)= 0.943

Recall= sensitivity = a/a+b = 1564/(1564+1224) = 0.561