For the adult set, I am importing 32561\*15 table a input, using one-hot encoding for the categorical terms in the first 14 terms, I can extend this to 32561\*74 matrix. Then write the last column of K<= 50 and K> 50 into 1,0 as the output. For this set of data, there are some missing data in the showed by ‘?’, in order to avoid bias, in this case, I am erasing lines where ‘?’ exists. The remaining intact cases in total are 32321\*74. By using 5-fold cross-validation on the 32321cases, which 6465 terms implemented as validation and the rest 25856 used as training cases. The optimal K is 26 which provides the accuracy of 0.7993. For this KNN, I am trying 20 values of K range from 1-101 to find the optimal K since I find the accuracy for validation set stops growing up before K reaches 100. Once getting the optimal K, I am using the best model I get from predicting the validation set to predict the test set with its size 16281\*15, cutting the incomplete data, there are 16171 left, using one-hot encoding again, we are predicting the label of 16171\*74 test input. Comparing that with the output, it has an accuracy of 0.8025. As I am using a larger dataset than in the paper, the accuracy is about 5% higher. However, it seems that KNN is not a good classifier.

79.00% knn 1

85.44% for the random forest of adult Training 85.64%

85.54%

|  |  |  |  |
| --- | --- | --- | --- |
|  | Knn | Random Forest | Decision Tree |
| Adult | 79% | 85.44% | 85.54% |
| Letter | 94.96% | 94.43% | 90.74% |
| Bank | 87.21% | 87.27% | 87.44% |
| Adult(With Missing) |  | 85.49% |  |

Treeoptimal 6

Optimal Feature Set Size 2

89.73% Trainmax 89.88%

Letter Knn

Random Forest 91.97%

Feature optimal is 1

If not setting into 2 categories 60.79%

Letter to 2 categories

Making the decision tree boosting 90.74% setting data into 2 categories

Random Forest 94.43%

Knn 94.96%

Bank Investment

Issue is that there is one component plays most important role in doing the classification

There are only two types of categories in terms of this

Knn prediction 97.04% accuracy 87.21%

Accuracy 87.44% 96.66%

Ramdom Forest 87.27% 97% accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Training set | Testing Set | Features | Classification Type |
| Adult | 5000 | 27321 | 15 | 2 |
| Letter | 5000 | 15000 | 171 | 2 |
| Letter(not binary) | 5000 | 15000 | 17 | 26 |
| Bank | 5000 | 40211 | 15 | 2 |
| Adult with missing | 5000 | 27561 | 15 | 2 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Grading

The merit and grading of your project can be judged from aspects described below that are common

when reviewing a paper:

1. How challenging and large are the datasets you are studying? (10 points)

2. Any aspects that are new in terms of algorithm development, uniqueness of the data, or new

applications? (10 points)

3. Is your experimental design comprehensive? Have you done thoroughly experiments in tuning

highper parameters and performing cross validation? (50 points)

4. Is your report written in a professional way with sections including abstract, introduction, data

and problem description, method description, experiments, conclusion, and references? (30

points)

5. Bonus points will be assigned to projects in which new ideas have been developed and implemented, or thorough experiments where extensive empirical studies have been carried out (e.g. evaluated on many additional classifiers).