COMP551 Mini-Project 1 Write-Up

Member: Junjian Chen 260909101
Shichang Zhang 260890019
Xichong Ling 260888765
Date: September 26th, 2021

Abstract

In this project, we investigated the performance of two classification models, K-Nearest Neighbours (KNN) and Decision Tree (DT), on predicting whether the individual will earn $50k/year and whether a child should be sent to nursery. During the experiment, we found that DT had a higher accuracy of prediction on the testing data than the KNN when we chose the hyperparameter of best performance for them. We also discovered that DT was significantly faster than that of KNN. When we carried out the experiment on how training size affects accuracy, we found that the relationships between them are different in two datasets. Therefore, we concluded that accuracy does not have a clear relationship with the size of the training data, it depends on choice of dataset.

Introduction

In the project, we examined and modified the data , exploring the performance of learning frameworks, k-nearest-neighbor (noted as KNN in the rest of the report) and decision tree (noted as DT) on two datasets, the adult dataset displaying attributes of individuals to predict their income; the nursery dataset illustrating features of the child to predict whether the child should be sent to the nursery. During the experiment, we compared the performance of two models in the aspect of accuracy and running time and investigated the factors such as choice of hyperparameter, training data size and manipulating features, which may influence the performance of the models. We found that the DT model usually performed better in both running time and prediction accuracy. Besides, we realized decreasing the size of training data might not help to improve the performance of models. When decreasing the training data size, the prediction accuracy for the adult dataset increased while the accuracy for the nursery dataset decreased. So we regarded that more datasets shall be investigated to summarize the rule.

Datasets

When processing the Adult dataset, we briefly examined it at some significant features. In terms of gender structure, males outweigh females in both number and ratio of earning more than 50K$/year. We also counted the ratio of people with annual income greater than 50K with different education backgrounds and found that the ratio is generally proportional to education level, which inspires us to encoding education features with sequential integers instead of one hot coding. Moreover, we examine the countries by region and do visualization, finding that people in countries from the same regions tend to have similar income structures.

Second dataset includes 8 attributes: parents' occupation, has_nurs, form of the family, children number, housing condition, financial standing, social conditions and health conditions. The class values are not recommend, recommend, very recommend, priority and spec_prior. The dataset has 12960 instances in total.

To handle missing and duplicate data, we delete the instances with missing features and duplication. We dropped the "fnlwgt" and "education" fields to avoid noises and meanwhile attempted to categorize country fields into different regions to prevent overfitting.

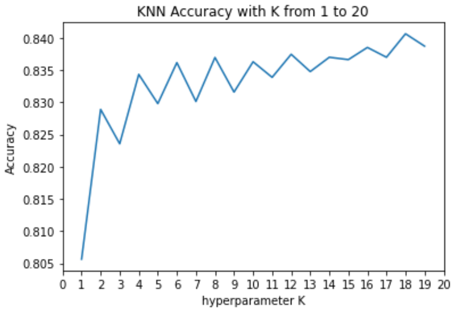
Results

To investigate how hyperparameters influence the performance of models, we calculated accuracies by the means of 5-fold cross validation in different hyperparameters from 1 to 20.

For the adult dataset, as shown in Figure 1.0, for the KNN model, increasing hyperparameter K generally causes a rising accuracy and a falling accuracy fluctuation so that the accuracy remains nearly constant at higher K. Nevertheless, Figure 1.1 shows that the plot of accuracy of DT is like a upwarding parabola with the maximum at D=9. Overall, in cross validation tests, DT generates a higher maximum accuracy and better average accuracy than that of KNN. For the nursery dataset, the plots of two models are similar. For the KNN model, as provided in Figure 1.2 the plot is also like a parabola and accuracy decreases slightly at higher K. When it comes to the DT model, shown in Figure 1.3 the plot grows and approximately reaches a constant value at higher D. Overall, DT also performs better in prediction for the nursery dataset.

We also conducted an experiment on how the size of the training data will affect the performance of models. We cut the training data size with $\frac{3}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$.

For the adult dataset, from Figure 1.4 and 1.5, both models witnessed an increase in training/validation process.



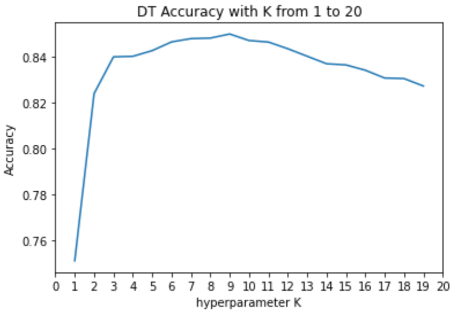
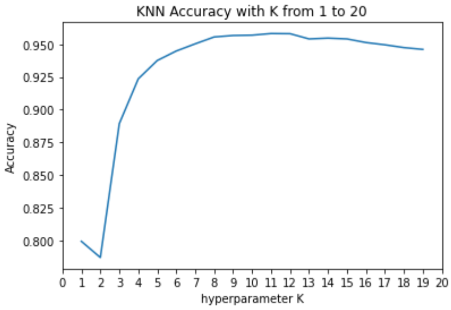


Figure 1.0: Cross Validation Result on KNN, adult dataset Figure 1.1: Cross Validation Result on DT, adult dataset



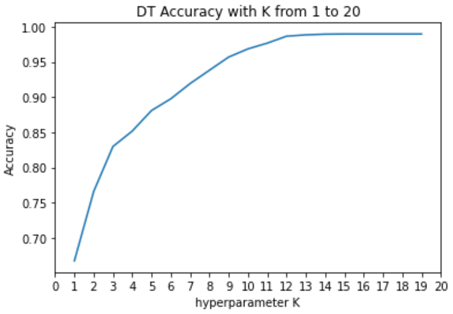
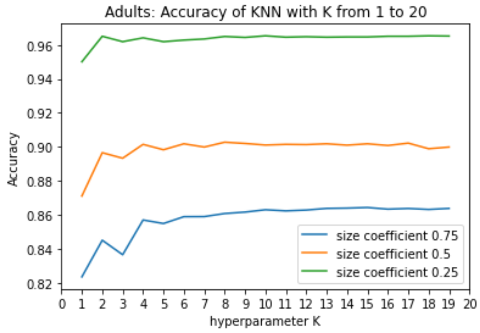


Figure 1.2: Cross Validation Result on KNN, nursery dataset Figure 1.3: Cross Validation Result on DT, nursery dataset



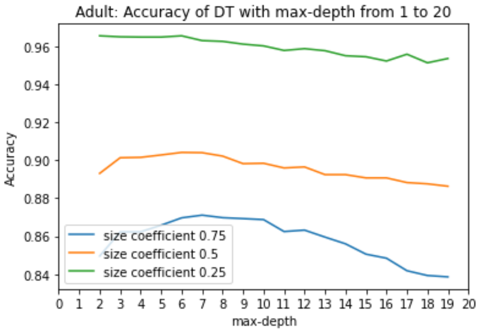
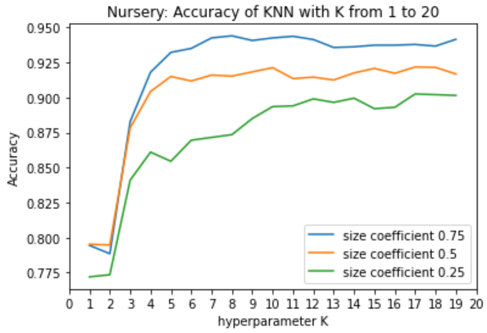


Figure 1.4: Variation of Training Size on KNN, adult dataset Figure 1.5: Variation of Training Size on DT, adult dataset

For the nursery dataset, the result concluded from Figure 1.6 and 1.7 is contrary to the adult dataset. With the size decreasing, the accuracy of the KNN plot falls and tends to fluctuate significantly. The DT model also shows a decreasing accuracy as the training size declines.



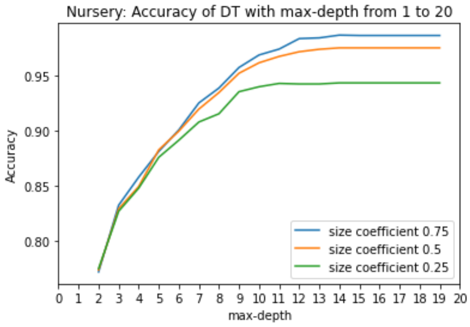
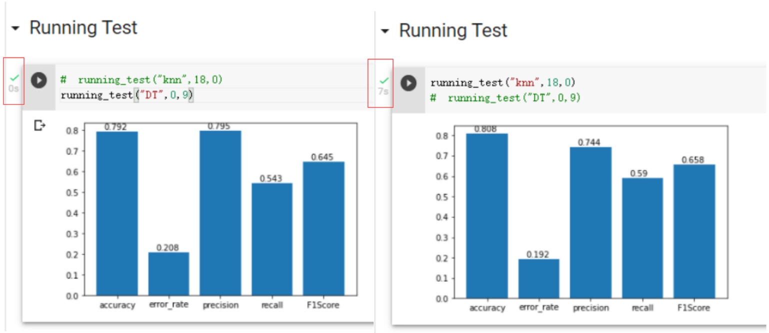


Figure 1.6: Variation of Training Size on KNN, nursery dataset Figure 1.7: Variation of Training Size on DT, nursery dataset



Running Test

```
#  running_test("knn",18,0)
running_test("DT",0,9)
```

Running Test

```
running_test("knn",18,0)
#  running_test("DT",0,9)
```

Figure 1.8: Final Tests of Adult dataset

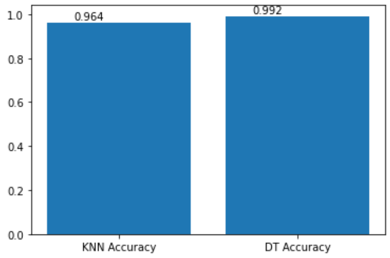


Figure 1.9: Final Tests of Nursery

During the final test, we used the best hyperparameters provided from cross validation to build the model and evaluated the model performance by the accuracy of prediction.

For the adult dataset, from Figure 1.8, we found that the accuracy of the KNN model reaches maximum when K is 18 and the accuracy of the DT model has a maximum when D is 9. In the final test, the accuracy of KNN for this dataset is marginally higher than that of DT. However, the time consumed by KNN to perform prediction is significantly higher than DT. For example, as highlighted from Figure 1.8, KNN took 7 seconds while DT spent less than 1 seconds. For the nursery dataset, as is shown in Figure 1.2 and 1.3, the previous cross validation step showed that the best performing hyperparameters are K=11 and D=13. Figure 1.9 shows the performance of models in the final test. DT performed almost perfectly to have a 0.992 prediction accuracy, 0.028 higher than that of the KNN model.

Other interesting findings:
Zero real positive: When we set the maximum depth of a decision tree to 1 and performed a test, we always found that the real positive is 0. The possible reason is that, in our training data, the proportion of instances whose income is lower than 50k$/year (negative results) is higher than that of those who earns higher than 50k/year(positive results). If the depth of the tree is 1, the model can basically predict all outputs to negative to gain a higher accuracy.

Discussion and Conclusion

We got equipped with numpy operations and got more experienced in data handling. Before applying machine learning models, we should carefully clean the input data and encode features in favored manners. It also helps to take a brief data overview to understand the distribution of each feature, thus picking up the most significant ones.

For future investigation, firstly, one remaining question is why a shrinking data sample has distinct impacts on different datasets. We assume it is related to the size of datasets, while remaining further experiments to verify.

Statement of Contribution

Shichang Zhang: Cross validation implementation, testing, write-up
Xichong Ling: Data overview, data visualization, testing, Jupyter Notebook documentation
Junjian Chen: Data preprocessing, testing, write-up