551 MiniProject 1: Machine Learning 101

Moyang Chen [260787220], Nianzhen Gu [261044523], Yinan Zhang [260764274]

September 2021

Abstract

This project experiments with two machine learning algorithm, K-Nearest Neighbours (KNN) and Decision Tree, by testing on different hyperparameters. It was conducted on two data sets, Adult Income Dataset[1] and Letter Recognition Dataset[2]. The Adult Income Dataset consists of a mix of characters and discrete features, while the Letter Recongnition Dataset contains only features with real numbers. Through building a 5-fold cross-validation method on each model, both algorithms yield an accuracy of 81% for the Adult Income Dataset but KNN model has a higher accuracy, 96%, than Decision Tree, 86% for the Letter Recognition Dataset. In general, KNN has a better performance than Decision Tree on these two different datasets.

1 Introduction

The application of machine learning in daily life is becoming more and more extensive, but many people are not very thorough in the use and understanding of machine learning models. They tend to focus on the direction of building extremely complex models or deep learning, and it is easy to overlook the advantages of traditional machine learning.

In our project, two small sample datasets, Adult Income Dataset and Letter Recognition Dataset, are selected as the analysis objects. Before building the machine learning model, we first cleaned up the two data sets. Among them, we mainly target the missing and wrong values in the data. We use the simplest method to eliminate all of them to ensure the purity of the data. On the cleaned data set, we respectively constructed a 5-fold cross-validated decision tree and KNN model.

Since the two machine learning models have different hyperparameters, we use the accuracy on predicting unseen data (ACC) during cross-validation as the evaluation parameter to find the optimal hyperparameters of the two models within the specified range. The results show that using overall data, models with optimal hyperparameters tend to have better predictive effects. This shows that model tuning is a very important part of machine learning.

2 Dataset

The Adult Income Dataset has the background information of a person, such as age, education and work class, and whether he or she makes more than 50k dollars per year. We first preprocessed the dataset by deleting the rows that contains missing data, which is represented by "?". Then we used one-hot encoding method to turn the discrete variables into multiple columns with all the answers appeared in the dataset, and use 0 and 1 to represent the answer. The target variable is set to income ('Income≥50K' =1 0, 'Income<50K' =0 1). Secondly, by observing the relationship between attributes and income in the Adult Income Dataset, irrelevant features, such as 'fnlwgt', 'education', 'capital-gain', and 'capital-loss', are deleted, and effective features are finally obtained.

The Letter Recognition Dataset contains letters in rectangular pixel display from A-Z in 20 different fonts, the dataset provides the information on the pixel image of each letter. For this dataset, the data is cleaned in the similar fashion as above, the 26 letters of A-Z are sequentially encoded by one-hot encoding. In addition, the first 16,000 of the letter-recognition dataset (80% of the whole dataset) are used as the training and validation set, and the rest are used as the test set.

3 Result

We established a decision tree model and a KNN model for the two data sets, and performed 5-fold cross-validation. The specific results are as follows:

3.1 Comparing performances between KNN and Decision Trees

3.1.1 Ault Income Dataset

For the Adult dataset, we first established a 5-fold cross-validation Decision Tree, which was trained with 1/3, 2/3 and all training data respectively. The hyperparameters includes the maximum depth of the Decision Tree, as well as the function to measure the quality of a split selected from ‘gini’ or ‘entropy’. Select and set the merit function to ACC. After comparison and optimization, the optimal decision tree model was finally obtained. The specific iterative result visualization is shown in Figure 1.

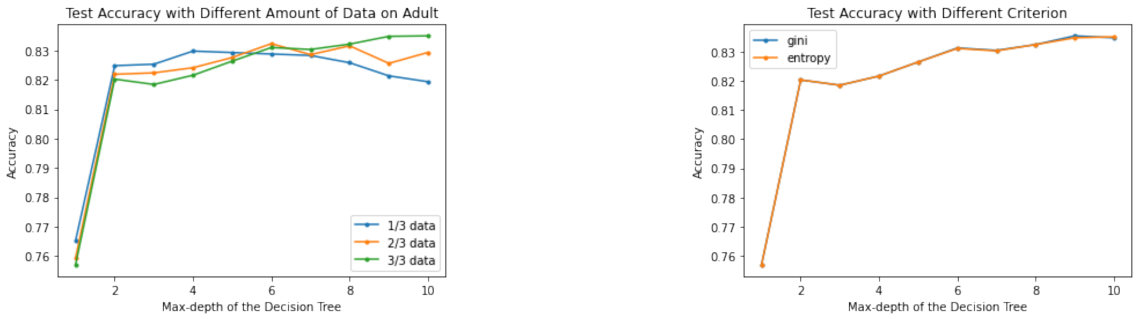

Figure 1: the Decision Tree interative result visualization

The results show that the optimal model of the decision tree would be the one that is trained using all training data with gini evaluation method, and the Max-depth of decision tree stetted to 9. This optimal model has a prediction on the test set ACC = 0.8153 .

Secondly, we use the 5-fold cross-validated KNN model, which was trained with 1/3, 2/3 and all training data respectively, with the hyperparameter of the value ‘K’, as in the number of nearest neighbours would be used to evaluate, tested within the range of 1-20. The evaluation criteria of the KNN model selects the knn.score() function in the sklearn package that comes with Python. From Figure 2, we can find that as K increasing, the training accuracy is decreasing, but validation accuracy is increasing. Also, more data will give more accurate result in both training and validation.

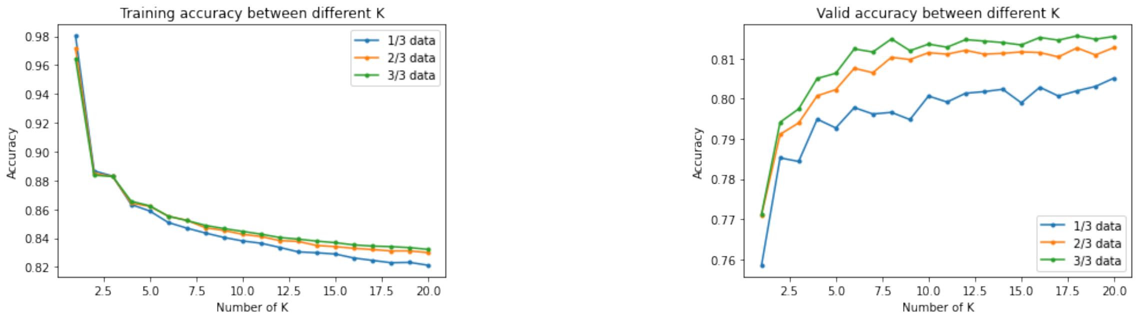

Figure 2: the KNN interative result visualization

After iterations, the optimal K and the data ratio required for model training are selected. The results showed that the K of the optimal KNN model should be set to 18. Then, all the training data is used to build the optimal model, and finally get ACC = 0.819 on the test set (shown in Table 1 in Appendix).

3.1.2 Letter Recognition Dataset

For the Letter Recognition Dataset, we implemented the same cross-validation method and hyperparameters as Adult Income Dataset, and the results are obtained by running in the same environment. The result of decision tree is shown in Figure 4:

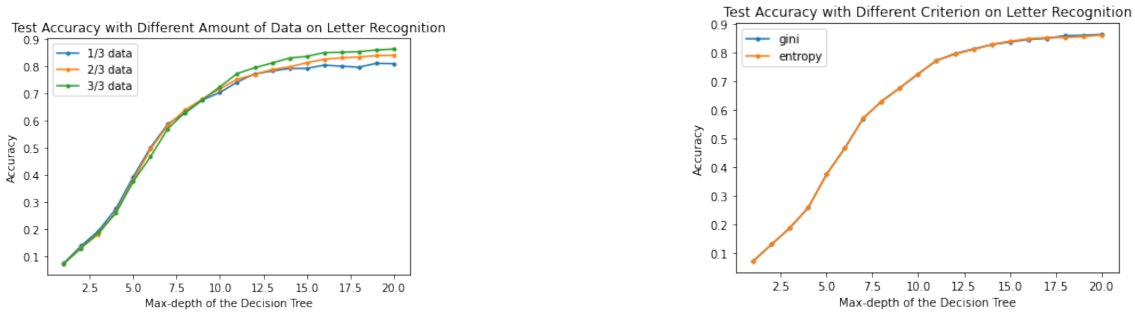

Figure 3: the Decision Tree interative result visualization

The results show that the optimal model of the decision tree would be the one that is trained using all training data, with gini evaluation method, and the Max-depth of decision tree stetted to 9. This optimal model has a prediction on the test set ACC = 0.8635.

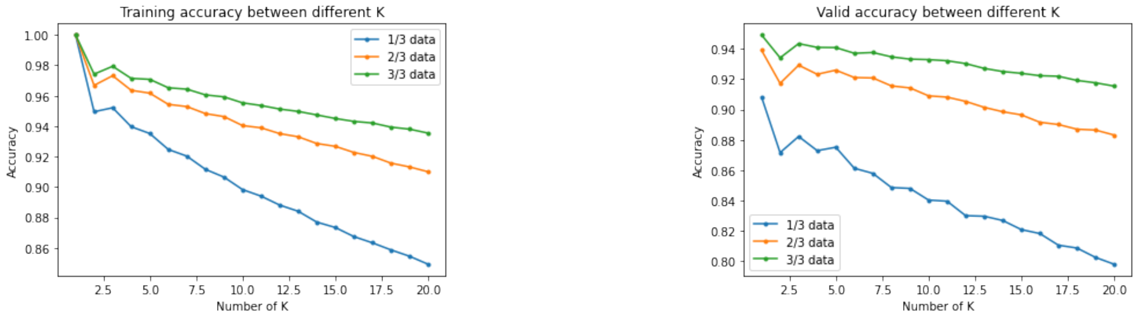

Figure 4: the KNN interative result visualization

After iteration, the optimal K and the data ratio required for model training are selected. The results showed that the K of the optimal KNN model should be set to 1. Then, all the data is selected for training, and finally get ACC = 0.958 on the test set. Same as the previous evaluation, the accuracy increases as the number of data become larger. However, in this dataset, both training and validation accuracy will decrease if we set K larger and larger.

3.1.3 Comparison of Two Models to a Conclusion

- When machine learning models are used for prediction, different models have different performances on different data sets.
- Some hyperparameters can heavily affect the performance of a model.
- Generally, the larger and more comprehensive the data used in model training, the better the training results will be on the test set.

4 Discussion and Conclusion

The most critical part of this project is the implementation of 5-fold cross-validation and the selection of the optimal hyperparameters for the model. We implemented these two parts on our own with experiments on different ranges. In addition, in the part of the Decision Tree model, we also compared and verified the effects of the two evaluation methods. In the project, 'gini' and ‘entropy' methods are used as hyperparameters, but the differences between the two is insignificant. We believe that there are still some areas deserving to be improved in this project. For example, the data cleaning part is still relatively rough. We directly eliminated the row when missing data appeared. In the future, we will consider using hot platform interpolation and other possible methods to handle the missing data better.

5 Statement of Contributions

In this project, Moyang developed the Decision Trees model for both the datasets, and NianZhen was responsible for development of the KNN model. Yinan applied the model to test set and helped with the write-up.

A Appendix

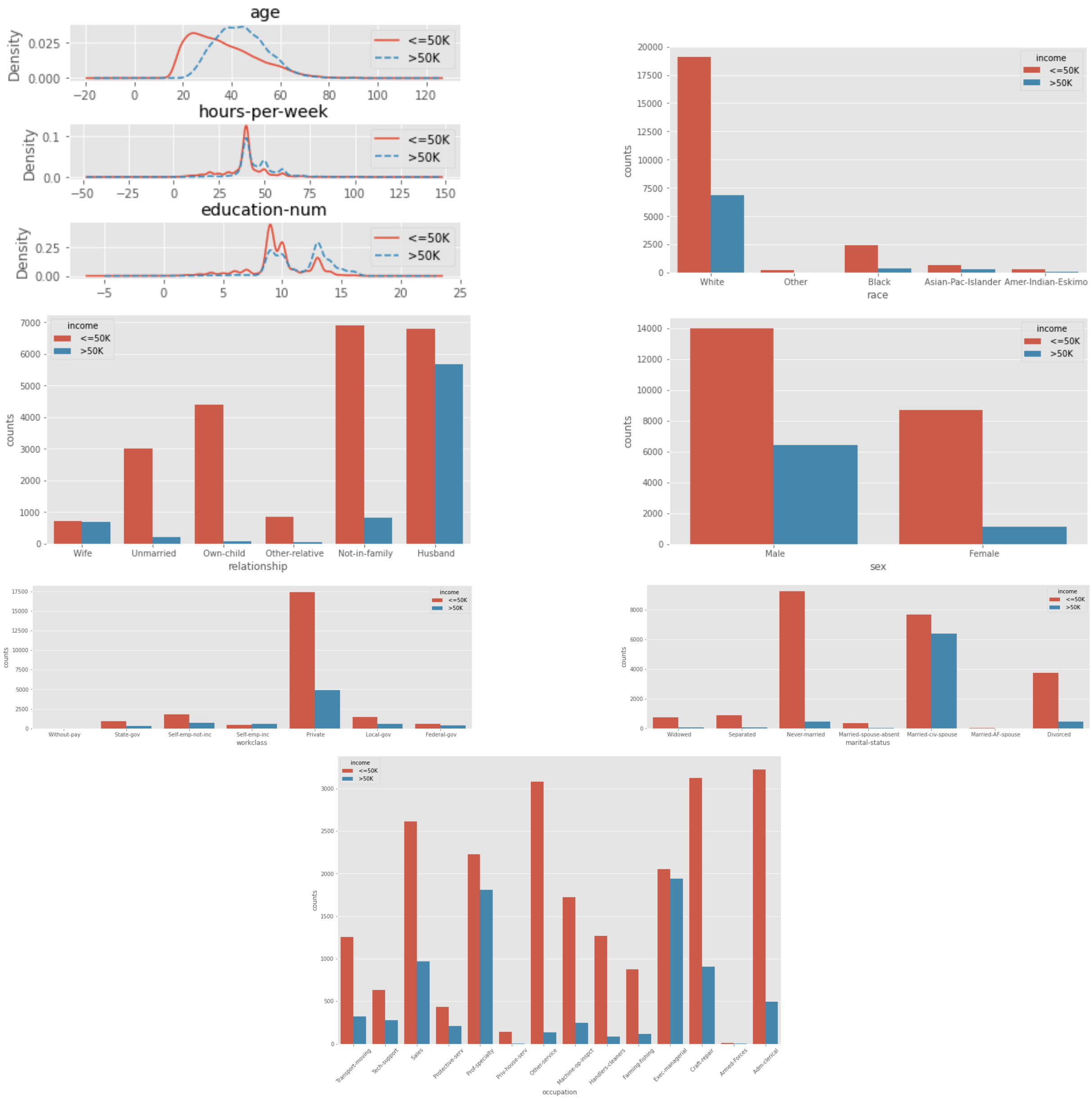

Figure 5: Visualized Results of Income at Different Levels of Some Indicators in the Adult Income Dataset

Decision Tree	KNN
ACC = 0.8153	ACC = 0.819

Table 1: Comparison Results of the Predictive Effects of the Two Models on Adult Income Dataset

Decision Tree	KNN
ACC = 0.8635	ACC = 0.958

Table 2: Comparison Results of the Predictive Effects of the Two Models on Letter Recognition Dataset

References

[1] UCI Machine Learning Repository, Adult Income Dataset, https://archive.ics.uci.edu/ml/datasets/Adult, accessed: 09.2021.

[2] UCI Machine Learning Repository, Letter Recognition Dataset, https://archive.ics.uci.edu/ml/datasets/Letter+Recognition, accessed: 09.2021.