Comp551 Mini Project 1 Report

Suofeiya (Sophia) Man, Minzhe Feng, Ye Yuan

Abstract

In this project, we investigated the performance of two classification models, namely k-nearest neighbours (abbreviated as KNN) and decision trees, to predict if the income of an adult exceeds $50K/yr and whether a person will accept the coupons in different scenarios. For **"Adult" dataset**, we found that KNN regression approach achieved lower accuracy than decision trees; yet for **"Vehicle" dataset**, KNN regression approach performed better . Moreover, the size of dataset has similar effects on the performance : the larger the size we have, the lower the mean squared error we achieved.

Introduction

For both of the datasets, similar features concerning different aspect of personal information e.g., age, sex & marital status, are provided, and the one consistent goal throughout the whole project is to make predictions on certain labels given the aforementioned information and different hyperparameters. To achieve this goal, we set the mean squared error (MSE) as the evaluation metric and employ KNN & Decision Trees, alongside various data cleaning techniques to analyze the datasets. Our most significant discoveries include that generally speaking, the larger the sample size (from the whole dataset) we have, the lower the MSE we will achieve.

Datasets

We preprocessed the data via the following procedures:

1. Handle the missing data

First discard the features with more than 10% values missing (across the whole respective datasets) completely, then delete rows that still contains missing values.

2. Data type conversion and normalization

Use the one hot encoder provided by sklearn and pandas to transform categorical features.

3. Feature investigation and engineering

We employed several techniques for feature engineering:

For "Adult" dataset:

(1) we removed the continuous attribute fnlwgt (final weight).

(2) We eliminated education-num because it is just a numeric representation of the attribute education.

(3) We mapped the following continuous features to categorical ones: age, education, hours_per_week, capital_gain and capital_loss.

For "Vehicle" dataset:

(1) we removed the discrete attribute 'car' due to the high percentage of missing data.

(2) we removed the continuous attribute 'temperature' because of limited information it provides.

(3) we removed the attribute 'direction_opp' because it is just the alternative of 'direction_same'.

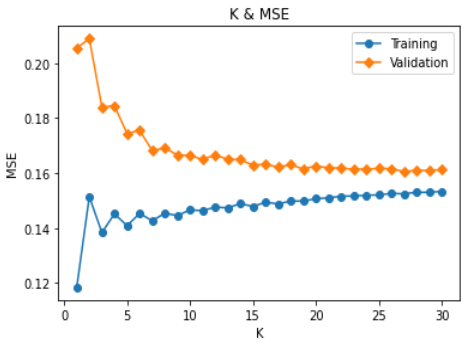
(4) we mapped the continuous attribute age to ordinal attribute as 'below 21', '21-29', '30-39', ' 40-49', '50plus'

Results

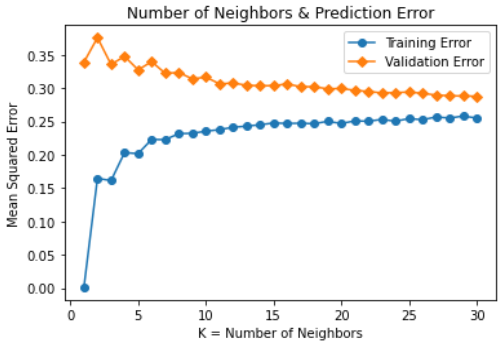
The key result is that the performances of models vary case by case. There no general rules.

Dataset	Model	Accuracy on the test set	MSE on the test set	Result
Adult	KNN	0.826	0.174	
	Decision tree with entropy	0.845	0.155	better
Vehicle	KNN	0.70 5	0.295	better
	Decision tree with entropy	0.655	0.345	

The only hyper-parameter we investigated for KNN model is the number of neighbours (hereafter abbreviated as k), which influences the performance of the KNN model evidently. As k increases, the MSE on the training set increases consistently. Whereas the MSE on the validation set decrease when k increases. Additionally, when k is small, our model will be overfitted to the training set, which result in worse performance on the validation set. In contrast, when k is too large, our model perform better on the validation set but worse on the training set, which means that the model underfits the training set. Consequently, we pick the number of neighbours at which the curve of training error achieves smoothness as the optimal hyper-parameter. In this case, k = 7 for dataset **"Adult"** and k = 20 for dataset **"Vehicle"**.



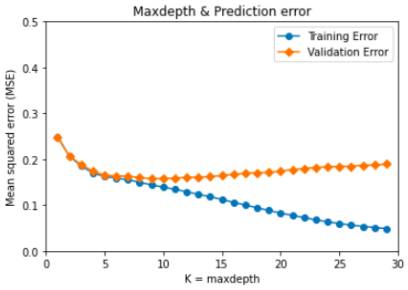
(For **"Adult"**)



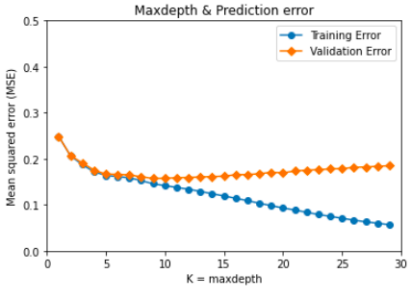
(For **"Vehicle"**)

The two hyper-parameters we investigated are maxdepth and criterion. For the criterion, the differences of the performance between 'gini' and 'entropy' is subtle, and the trends are similar. We choose the criterion whose average MSE is lower. In both of the cases, it is entropy by accident.

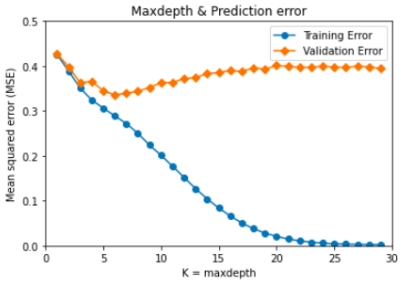
The effect of hyper-parameter maxdepth is obvious: as maxdepth increases, the MSE will decrease to a local minimum and then increases. Thus, we pick the maxdepth that give the smallest MSE as the optimal hyper-parameter.



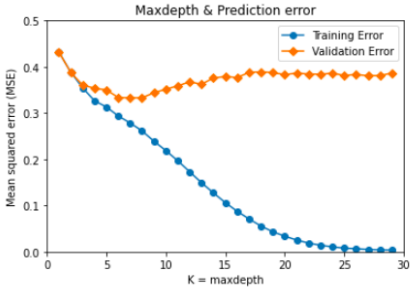
(When using Gini for **"Adult"**)



(When using Entropy for **"Adult"**)



(When using Gini for **"Vehicle"**)



(When using Entropy for **"Vehicle"**)

Refer to the figures attached in appendix of this report, for the k-nearest neighbours model, it is obvious that as the size of data increases, the MSE on both training set and validation set decrease. In order to research the effect of size to the result, we use fixed hyper-parameters for this part. We set the number of neighbours k = 7 **for "Adult"** and k = 20 **for "Vehicle"**.

Meanwhile, for the decision tree model, we have almost same conclusion. With larger sample sizes, we achieved lower MSE on the validation set. Even though the training error will increase as the size of dataset increases, we still have better performance on the validation set in general.

Conclusion and Discussion

The performances of KNN model and decision tree are pretty similar. It is hard to say which model will always have higher accuracy score on the test set. Moreover, the size of dataset correlates with the performance. Generally, the larger size of dataset we have, the lower the mean squared error . However, The size cannot determine the performance. It is possible to achieve fairly good performances with a small dataset when the data itself is unbiased and has small variance. Conversely, we cannot perform well when the dataset is biased, or the variance is really large even if we have massive amount of data. In the future, we could explore the relationship between the distribution of data and the performance of models.

Statement of Contributions

Ye Yuan: Independently completed the KNN part for the second "Vehicle" dataset & partially implemented KNN for the first "adult" dataset.

Minzhe Feng: Completed the KNN part for the first "Vehicle" dataset (on top of Ye Yuan's work) & performed some code optimizations (by replacing Python native methods with Numpy equivalents to achieve better speed).

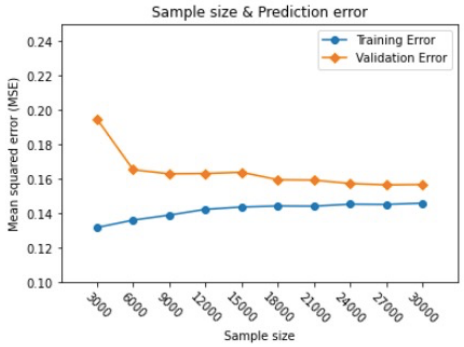
Suofeiya (Sophia) Man: Independently completed the decision tree part for both datasets.

Citations

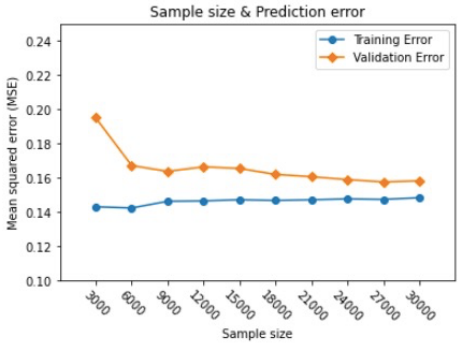
Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. 'A bayesian framework for learning rule sets for interpretable classification.' The Journal of Machine Learning Research 18, no. 1 (2017): 2357-2393.

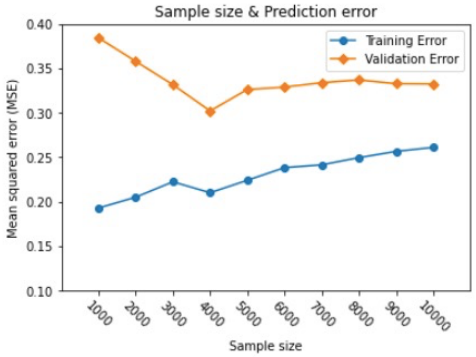
Appendix



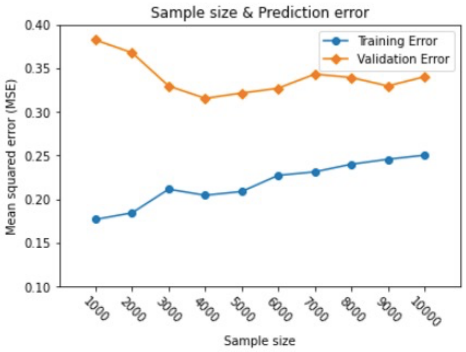
(When using Gini for **"Adult"**)



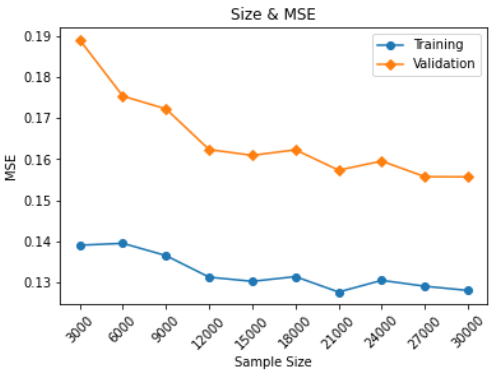
(When using Entropy for **"Adult"**)



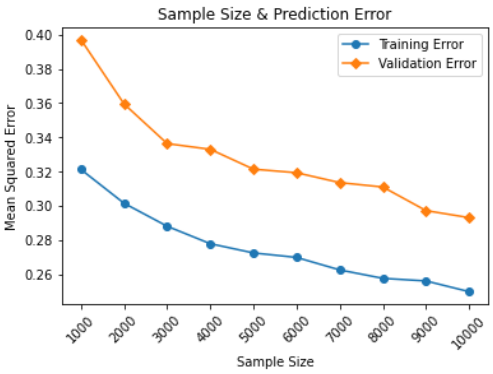
(When using Gini for **"Vehicle"**)



(When using Entropy for **"Vehicle"**)



(For **"Adult"**)



(For **"Vehicle"**)