Exploring Basic Supervised Machine Learning Techniques

XINYANG ZHANG, ZHENYU WANG, XUANZI WANG

1 ABSTRACT

In this project, we performed supervised learning, K-nearest neighbours(KNN) and decision trees, on two datasets. We made the prediction on the first dataset about whether a person makes over 50k a year from attributes such as age, education, and occupation. We also identified in the second dataset which letter the display belongs to from quantitative attributes of the display. By testing these two classification models on test data in the two datasets, we found that on unseen data, decision trees outperforms KNN in the first dataset, but KNN performs better in the second dataset. However, the KNN approach was significantly slower during training.

2 INTRODUCTION

The datasets Dua and Graff [2017] included in this experiment are Adult Dataset, which includes 48842 instances, and Letter recognition Dataset, which includes 20000 instances. Kohavi et al. [1996] has applied the Naive-Bayes classifier on the Adult Dataset and the Letter recognition Dataset. The accuracy of the Navie-Bayes method is close to the accuracy of our methods. After preprocessing the data, we split the data into training/validation part and test part, and uses the KNN and the decision trees classification models to test them.

KNN is a supervised learning technique that we could use to classify each instance through the K nearest neighbours near it. K is a hyper-parameter that we need to optimize for the model. KNN is an important method widely been used and studied. Zhang et al. [2017] explored on the idea of assigning different test data points with different k values.

On the other hand, decision trees is another classification technique that divides the inputs into different regions and assigns a prediction label to each region. In this case, we would like to find the optimal depth and min samples split for our models. Myles et al. [2004] provides descriptions and explanations of the traditional and current decision trees models, which gave us some inspiration.

To obtain the appropriate hyper-parameters and reduce the cost, we performed 5-fold cross-validation on the training and validation data. Besides, we also explored the effect of using different sizes of training data.

We have also searched for other papers comparing these two methods. Rajaguru and S R [2019] has compared these two methods on breast cancer and concluded that KNN outperforms decision trees in classifying the breast tumor. In our experiment, we found that the accuracy of KNN is better in the second dataset but worse in the first dataset on unseen data. But the training time of decision trees is much faster than KNN. As the size of the dataset increases, the accuracy achieved by the same model would increase.

3 DATASETS

The experiments are performed on two data sets, Adult Data Set and Letter Recognition Data Set, downloaded from UCI. For both of these data sets, we have split the data into training set and test set. The instances with any missing or invalid data entries are excluded. In both datasets, the discrete variables were converted into multiple variables using one-hot encoding. We have visualized the datasets for better understanding.

3.1 Adult Data Set

The adult dataset consists of 15 inputs, including continuous variables such as age, and discrete variables such as workclass. The goal is to predict whether a person makes over 50k a year based on the attributes of each person. We have plotted the total counts of individuals with income above 50k and below 50k with different education, workclass, marital status, race, and relationship backgrounds using the bar plots. We could briefly have an insight into the distribution of the income in each

Author's address: Xinyang Zhang, Zhenyu Wang, Xuanzi Wang.

2021.

category. For example, most of the people who have high school degrees make less than or equal to 50K each year. Please refer to appendix A.1 for the graphs.

3.2 Letter Recognition Data Set

The letter recognition dataset includes 16 attributes and we are trying to identify each rectangular pixel display as one of the 26 capital letters in the English alphabet. The instances we are working with are based on different fonts and are randomly distorted. The statistical attributes are acquired from measuring each stimulus and scaled to fit in a range of values. We trained on the first 16,000 instances and tested on the remaining 4,000 instances. We calculated the mean, median, and range of each letter for each attribute. It is noticeable that the attributes vary for different letters. Thus, it is reasonable to classify the display based on these attributes. Please refer to appendix A.2 for the tables.

4 RESULTS

We manually performed 5-fold cross-validation to find the best hyper-parameters and applied the generated models with these hyper-parameters to the test set. Besides, we have also explored the effect of sample size on accuracy, another way of handling missing data, and the most important features corresponding to the target.

4.1 KNN

We tried on different numbers of neighbors (K) ranging from 1 to 30. For each K, we took the means of the accuracy score from each training and validation datasets, and we plot them against the hyper-parameter. We would like to choose the K at which the validation set reaches its best accuracy. For the Adult dataset, we chose $K = 9$ since the accuracy is approximately the highest at $K = 5$ and the slope of the curve becomes flat for $K > 9$. For the second dataset, we chose $K = 3$ since $K = 3$ would produce the second highest accuracy. The validation accuracy achieves the highest point when $K = 1$, but $K = 1$ clearly would have the risk of overfitting.

4.2 Decision Trees

To acquire the best hyper-parameter, we performed 5-fold cross-validation with different combinations of two hyper-parameters. Hyper-parameters being tested

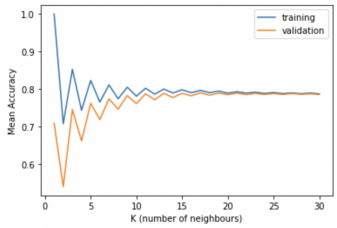

Fig. 1. KNN accuracy plot for Adult dataset

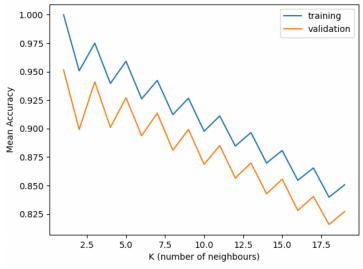

Fig. 2. KNN accuracy plot for Letter Recognition dataset

are the max depth of the decision tree, which varies from 1 to 30, and the minimum number of samples required to split a node, which varies from 2 to 9. We have also attached in appendix B.1 the accuracy plot generated from cross-validation containing only one hyper-parameter. The method taking multiple hyper-parameters into account considers not only each individual hyper-parameter, but also the interaction between them. As the accuracy obtained by varying these two hyper-parameters would generate a 3-D plot, we plotted 2-D plots against each hyper-parameter for a better understanding. For Adult Dataset, we chose max depth as 9 and min samples split as 2 since the curve reached the highest level at these points. For the second dataset, we acquired the highest accuracy when the max depth is 20 and the minimum samples split is 2. Please refer the graphs for the second dataset in appendix B.2.

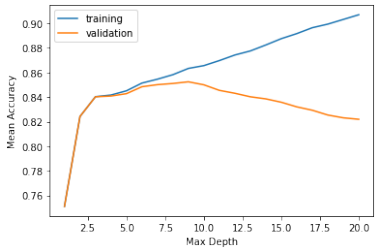

Fig. 3. Adult: Decision tree accuracy plot varying max depth

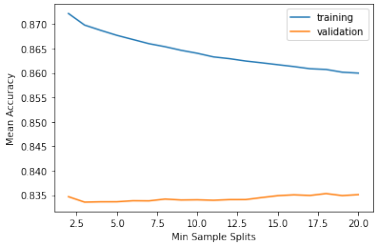

Fig. 4. Adult: Decision tree accuracy plot varying min samples split

4.3 Sample Growing

We have applied both KNN and decision trees models on the two datasets while reducing the training/validation data size to 75%, 50%, and 25%, respectively. Again, using the method above to plot the accuracy corresponding to the hyper-parameter, we found that for a fixed hyper-parameter, the corresponding accuracy decreases as the sample size decreases. Thus, it would be better to have more data for training/validation. The graphs are added to appendix B.3.

4.4 Test Set

After determining the hyper-parameters, we generated a model corresponding to the dataset. By applying the model to the test set, we found that for the first dataset, we have achieved an accuracy score of 0.78 using KNN and 0.85 using decision trees. And for the second dataset, the KNN score is 0.93, and the decision trees score is 0.86. The performance of KNN and decision trees are related to the features of the datasets. If there are many random occurrences in the dataset, the KNN would outperform the decision trees.

4.5 More experiments

As there are many missing values in the Adult dataset, instead of simply removing the whole row, we replaced them with the value that appears the most in their columns for categorical data and the mean of other values in their columns for numerical data. After re-running the models on the new dataset, we found that the accuracy did not change greatly. This is because the size of the data we previously used is already large enough to make the result general. We have attached the graphs generated by the new data in appendix B.4.

In both datasets, it is noticeable that there are multiple attributes. Take the second dataset as an example, after applying PCA, we found that in order to explain 70% of the variance, we have to use at least 13 principal components, which did not make further analysis easier. Thus, we found the most important features in the dataset in order to obtain much insight into the dataset. And it is possible to scale up important features to obtain a better performance of the model. Thus, we have generated a correlation matrix. Simply looking at the correlation between the features and the target, the feature that possesses the highest correlation is the most important feature related to the result. For the Adult dataset, marital status and relationship are the two most important features related to the target. In the second dataset, x2ybr and y-bar are the two most useful features. The correlation matrices generated are attached in appendix B.5.

5 DISCUSSION AND CONCLUSION

5.1 Takeaways

For different datasets, the suitable model for classification would be different. Even for a certain technique, the choice of hyper-parameter would greatly affect the model accuracy. More training and validation data would have a positive impact on the model generated. If the size of training and validation data is large enough, there is nearly no difference in how to handle the missing data.

5.2 Future direction

Due to the lack of time, we have not considered all combinations of the hyper-parameters in the two models. In future studies, it would be useful to conduct an experiment to explore more hyper-parameter combinations for better model performance. Moreover, it is also significant to scale up important features to make the classification more precise.

6 STATEMENT OF CONTRIBUTION

Xinyang Zhang searched related papers and wrote most of the report. Zhenyu Wang preprocessed the data and implemented the experiments for the Adult dataset. Xuanzi Wang implemented the experiments for the Letter Recognition dataset and wrote part of the report.

APPENDIX A DATA VISUALIZATION

A.1 Adult Data Set

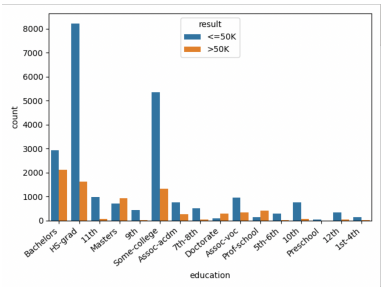

Fig. 5. Counts for each education category

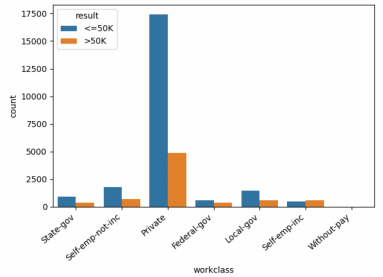

Fig. 6. Counts for each workclass category

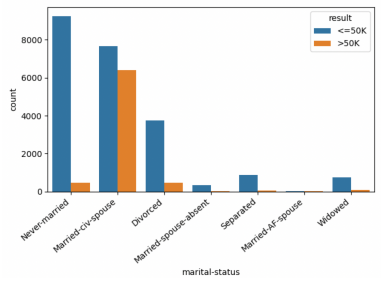

Fig. 7. Counts for each marital status category

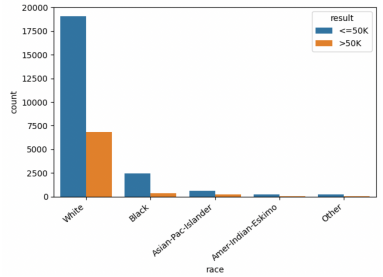

Fig. 8. Counts for each race category

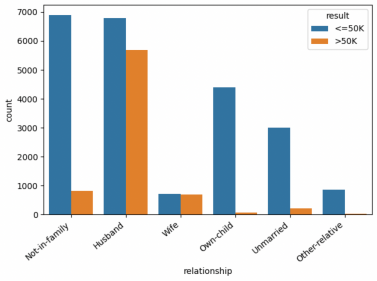

Fig. 9. Counts for each relationship category

A.2 Letter Recognition Data Set

letter	x-box horizontal position of boxÊ	...	yegvx correlation of y-ege with x
		...	
A	3.337136	...	7.468948
B	3.985640	...	9.100522
C	4.031250	...	8.555707
D	4.023602	...	7.628571
E	3.727865	...	8.506510

Fig. 10. Mean of each attribute in Letter Recognition Dataset

letter	x-box horizontal position of boxÊ	...	yegvx correlation of y-ege with x
		...	
A	3.0	...	8.0
B	4.0	...	9.0
C	4.0	...	9.0
D	4.0	...	8.0
E	4.0	...	8.5

Fig. 11. Median of each attribute in Letter Recognition Dataset

letter	x-box horizontal position of boxÊ	...	yegvx correlation of y-ege with x
		...	
A	9	...	10
B	11	...	9
C	10	...	9
D	9	...	9
E	10	...	10

Fig. 12. Range of each attribute in Letter Recognition Dataset

APPENDIX B RESULT PLOT

B.1 Accuracy Plots for Decision Trees

This section only includes the accuracy plots generated for decision trees while taking max-depth as the only hyper-parameter.

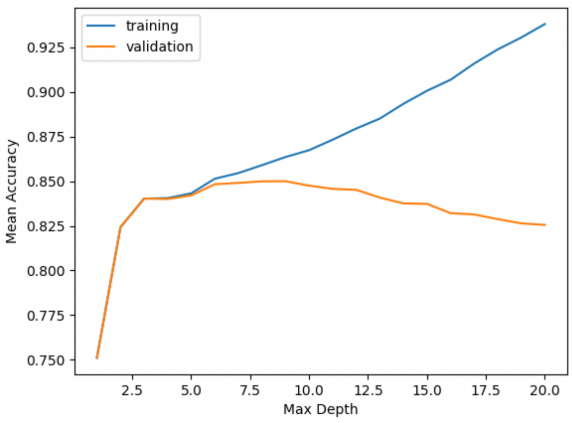

Fig. 13. Decison Trees accuracy plot for Adult dataset

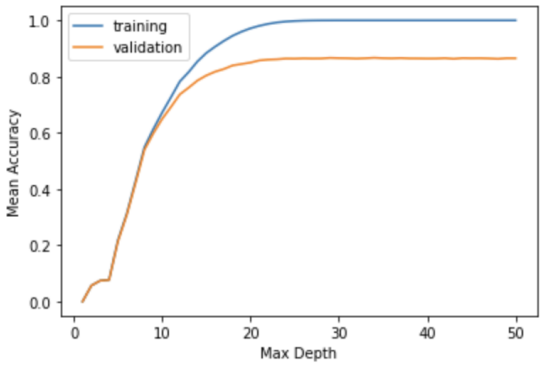

Fig. 14. Decison Trees accuracy plot for Letter Recognition dataset

B.2 The other set of graphs tuning two hyper-parameters together

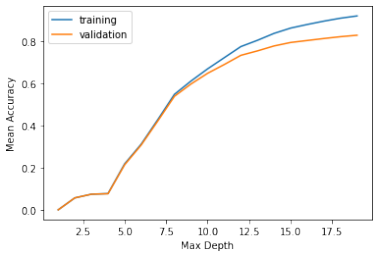

Fig. 15. Letter: Decision tree accuracy plot varying max depth

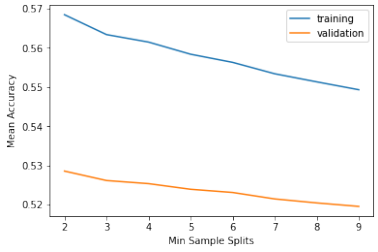

Fig. 16. Letter: Decision tree accuracy plot varying min samples split

B.3 Plots for different sample sizes

This section includes the graphs for section 4.3.

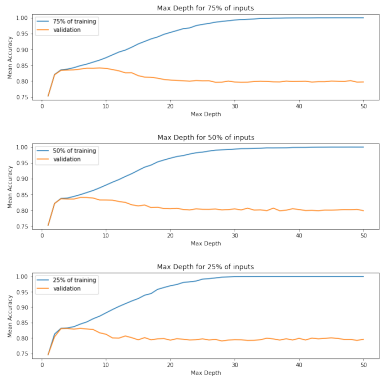

Fig. 17. Adult: Decision tree accuracy plot for different sample sizes

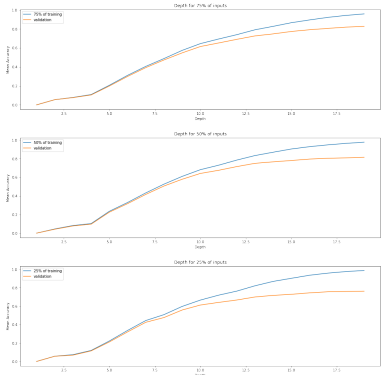

Fig. 18. Letter: Decision tree accuracy plot for different sample sizes

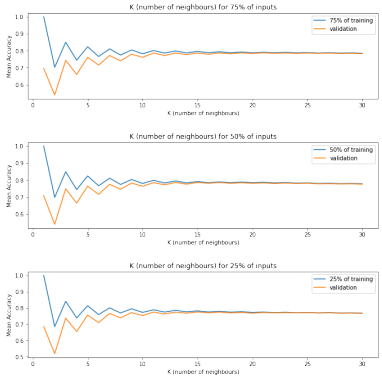

Fig. 19. Adult: KNN accuracy plot for different sample sizes

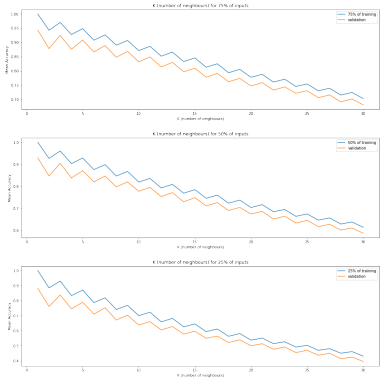

Fig. 20. Letter: KNN accuracy plot for different sample sizes

B.4 Graphs generated after data imputation

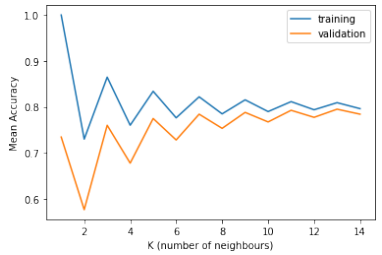

Fig. 21. Letter: KNN accuracy plot after data imputation

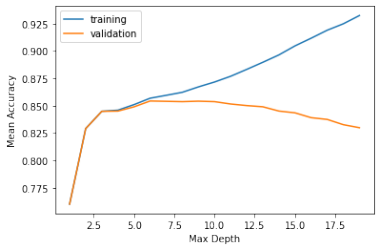

Fig. 22. Letter: Decision tree accuracy plot after data imputation

B.5 Correlation matrices

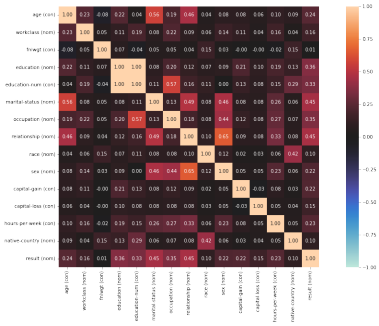

Fig. 23. Correlation matrix for Adult dataset

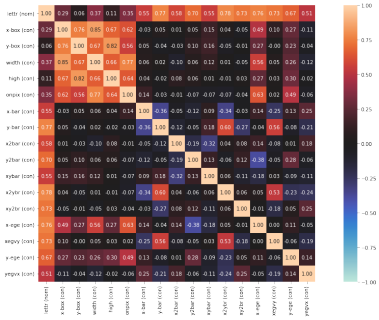

Fig. 24. Correlation matrix for Letter Recognition dataset

REFERENCES

D. Dua and C. Graff. 2017. UCI Machine Learning Repository. (2017). http://archive.ics.uci.edu/ml

R. Kohavi et al. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.. In *Kdd*, Vol. 96. 202–207.

A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown. 2004. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society* 18, 6 (2004), 275–285.

H. Rajaguru and S. C. S R. 2019. Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer. *Asian Pacific Journal of Cancer Prevention* 20, 12 (2019), 3777–3781. https://doi.org/10.31557/APJCP.2019.20.12.3777

S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng. 2017. Learning k for KNN Classification. *ACM Trans. Intell. Syst. Technol.* 8, 3, Article 43 (Jan. 2017), 19 pages. https://doi.org/10.1145/2990508