Comp 551 MiniProject 1 Report

Group-88

Barry Li (260912069)

Haochen Liu (260917834)

Yiran Fu (260904135)

Abstract

This project aims to explore the performance of two machine learning models, KNN and Decision tree by using two classification datasets, Adult dataset and Occupancy Detection dataset. For Adult dataset, we want to predict whether the income of an adult exceeds 50k/year from fourteen features while in Occupancy Detection dataset, we wish to predict whether a room is occupied or not from five features. We want to investigate how the choice of different models, hyperparameters, sample size, normalization techniques, and features manipulation impacts our results. We found that decision tree model had a better performance in Adult dataset while KNN model achieved better accuracy in Occupancy dataset. Since each dataset has its unique characteristics, the best way to find the fitted model with the best combination of hyperparameters is just a case-by-case analysis. In terms of time, in general decision tree will be a little faster than KNN from the hyperparameter set we are using.

Introduction

Both KNN and decision tree are useful machine learning models to deal with classification types. KNN predicts the label by finding k similar examples in the training set while decision tree divides the inputs into different regions using a binary tree structure and assigns a predictable label to each region. In this project, we implemented the 5-fold cross validation step, used KNN and decision tree classifier from sklearn library to design tests for both models and investigated how changing hyperparameters, size of training data, normalization techniques and manipulating features impacts performance. To perform such tasks, we decided to use the Adult dataset and the Occupancy dataset and found that one had a better accuracy with KNN model while the other fitted decision tree model better. To help us better formulate our test, we decided to do some research both on the models and the datasets. The first article we found useful is on the various hyperparameters we can use to tune our decision tree. While the second performs a whole analysis on the Adult dataset which guides us when we manipulate features.

Datasets

Both of our datasets are classification types. The Adult dataset contains 48842 instances with 14 attributes. To handle the missing values, we first checked the number of instances containing missing values and decided to delete them after finding that they were only a small portion of the entire dataset. Since the Adult dataset includes both categorical and continuous features, we used one hot encoding provided by sklearn to convert discrete variables into multiple binary variables. While processing the Adult dataset, we found the training and testing data were in different dimensions since one has the Netherlands as one of its inputs while the other doesn't. To fix it we merged the two datasets, encoded them together and split them afterwards. Then we normalized the data by using sklearn normalizer to finish the cleaning and preprocessing steps. We also do feature manipulation by deleting different features and explore how they influence the accuracy (Details in result section). The Occupancy detection dataset includes 20560 instances with 7 continuous attributes and no missing values. In data cleaning, we deleted the useless features "Date" and "line number" and then normalized this dataset using sklearn as well.

Results

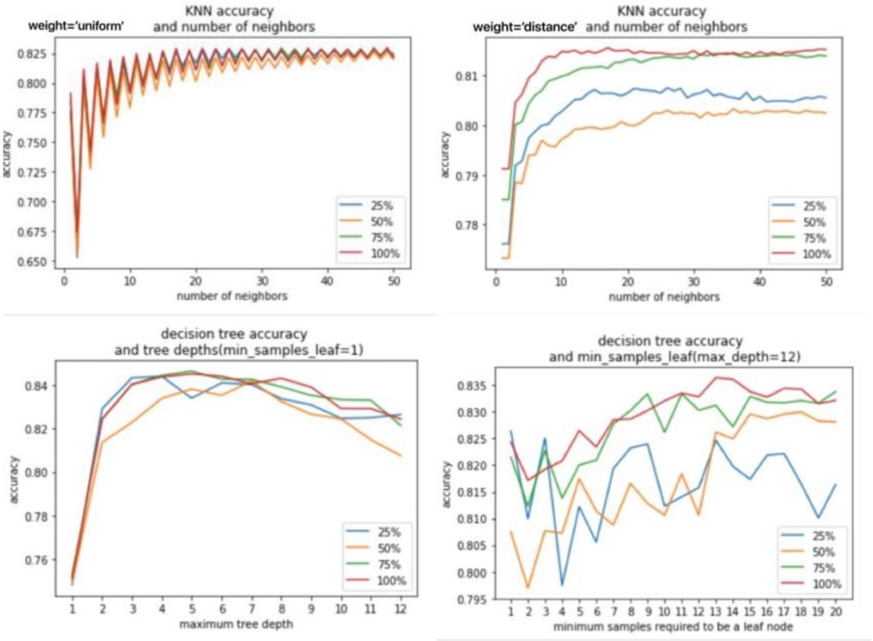
According to the given procedure, we first performed 5-fold cross validation on the training/validation set for both datasets to find the best combination of hyperparameters. The hyperparameters we tuned in KNN are k and weight = "uniform" / "distance" in order to explore how different k and weight influence the final results. We tested k from 1 to 50 in the Adult dataset with more instances and features. While for the smaller Occupancy dataset, we tried less k from 1 to 25. For the decision tree model, we tuned 2 hyperparameters, max_depth and min_samples_leaf that not only control the overall structure of the tree but also the little detail such as size of each leaf node. We tested max_depth from 1-12 and min_samples_leaf from 1-20 for both datasets. The best validation/test accuracy with size 100% of training data for both models are shown in the table below.

	KNN	Decision tree
Adult	Validation:82.98%; Test: 83.40% (k=49, weight= 'uniform')	Validation:84.60%; Test: 85.06% (max_depth=6, min_samples_leaf=13)
Occupancy	Validation:98.98%; Test: 96.09% (k=8, weight= 'distance')	Validation:98.73%; Test: 96.39% ((max_depth=10, min_samples_leaf=5)

Overall, the decision tree model fits better for the Adult dataset while KNN has a higher accuracy in Occupancy dataset. As for speed, decision tree has a clear edge for both datasets. This could be due to a large k we are choosing. The overall accuracy for occupancy dataset is higher than adult dataset, which may be a cause of its relatively simple structure and fewer noisy features.

To better understand how hyperparameter tuning influences our accuracy, we made plots for each hyperparameter vs. accuracy. Here for the decision tree model, we plotted the two hyperparameters separately by changing one hyperparameter while keeping the other as its default value to visualize each of their trends. The default value for max_depth is none but we chose 12(largest value in our tuning) since using none will easily cause overfitting and also be a huge load for the machine to process.

For training accuracy in both models and datasets (plots in Appendix), as we changed hyperparameters away from their defaulted values, the accuracy decreased. KNN model with weight= 'distance' is an exception that no matter how we changed hyperparameters, the accuracy was always equal or very close to 1 due to its characteristic of having more weight for closer elements.



(Due to limited space, only some plots for Adult dataset are shown here, others are in Appendix)

For validation sets in KNN model, as we increased the value of k, the accuracy changed substantially when k was relatively small and became more stable as k became larger. Interestingly, odd ks often have better validation accuracy than even ks for uniform weight and the plot is in zigzag shape.

Changing weight to ‘distance’ smooths the shape of the line but its influences the validation accuracy is uncertain: in Adult dataset uniform has a better accuracy but for Occupancy dataset its vice versa. In the decision tree model, too big or too small max_depth won’t lead to a good accuracy and increasing min_samples_leaf in a large dataset may be a good way to avoid overfitting and enhance accuracy. Although we analyzed some trends within the two datasets, we think it is very hard to summarize rules that fit all datasets when doing hyperparameter tuning. We do believe a case-by-case analysis is essential to find the best combinations of hyperparameters.

We also investigated how the size of training datasets impacts our accuracy by using 25%, 50%, 75% and 100% of the original dataset to test. From the plots we can find, a relatively large data size often has a higher validation accuracy than the smaller size. There might be some outliers, for example, 50% in Adult dataset and 75% in Occupancy dataset, but since our data are randomly selected, it is common to have more noisy features in some sizes and less in others which causes the larger size to have a lower validation accuracy than the smaller size. The results for test sets are similar to the validation set, having fluctuations but will not be much lower than validation sets since we used 5-fold cross validation.

We tried different sklearn normalization methods(‘l1’,’l2’,’max’) and found that they perform differently in different datasets (‘l2’ has the highest accuracy in Occupancy dataset but the lowest in Adult.) However, the difference in accuracy is small so we think it is not as important a factor as others.

To explore the importance of features, in Adult dataset, we deleted sex, education and other features that the report we found online analyzed to be important but found that there was no big change in accuracy. Then we removed three highly skewed features which should be unuseful but still got similar accuracy. However, for Occupancy dataset, we only deleted “CO2” but got an obvious decrease in accuracy. We think this is due to the Adult dataset has a lot of features and obviously they are correlated with each other so deleting one or a few of them will not have that big of an influence on the accuracy. But in the Occupancy dataset, there are only 5 features so deleting one important feature can lead to a big drop in accuracy.

Discussion and conclusion

From our results, we found that in Adult dataset with large data size and lots of features, decision tree model tended to have better performance. But for a less complex dataset with fewer features like in Occupancy dataset, KNN may have a better accuracy. For KNN model, a relatively large dataset with more features often results in a larger k being its best hyperparameter and for decision tree model, limiting max_depth and increasing min_samples_leaf properly can avoid overfitting and get a higher accuracy. When we choose how much data for training and validation, generally, larger size often has better accuracy and are less likely to overfit. However, these are just very general trends. In order to find the best hyperparameter/size, a case-by-case study is required. From the no free lunch theorem, we also know there isn’t a best model for every case. To choose the best model for an individual case, experiments and tests need to be conducted. In this project we focused on the classification use of both models so in the future we want to explore their performance for regression datasets as well.

Statement of Contributions

Haochen Liu, designed the test functions, graphs, collected data and maintained and debugged part of the code. Barry Li, did the majority of the coding and debugging, cleaned and preprocessed the data and implemented the 5-fold cross validation for both models also proofread the report. Yiran Fu, did project write-up and gave advice when debugging.

Citation

How to tune a decision tree?
Mukesh Mithrakumar, Nov 11, 2019
https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680

Predicting Earning Potential using the Adult Dataset
Haojun Zhu, December 5, 2016
https://rpubs.com/H_Zhu/235617

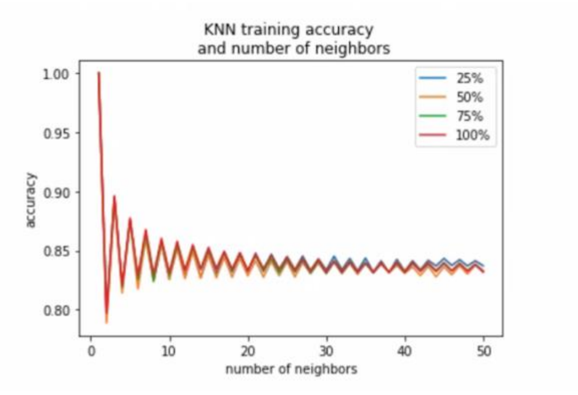
Our second dataset

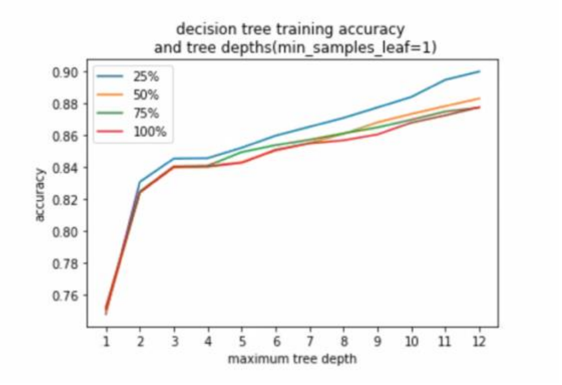
Occupancy Detection Dataset
https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

Note: We rearranged the original training and testing set by combining files "datatest.txt", "datatraining.txt", and "datatest2.txt". During this process, we set the data "datatest.txt", "datatraining.txt", and the top 3582 rows of data (without header) in "datatest2.txt" as the training set, and the rest of "datatest2.txt" as the testing set. This makes a 70%-to-30%-training-testing split.

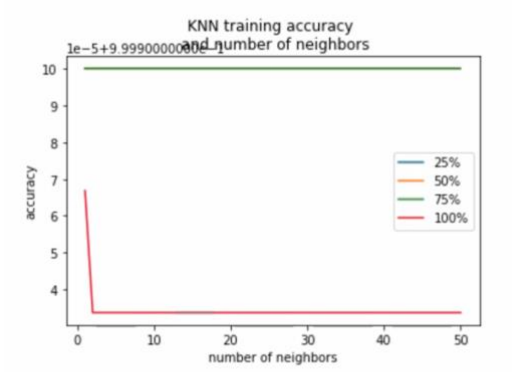
Appendix

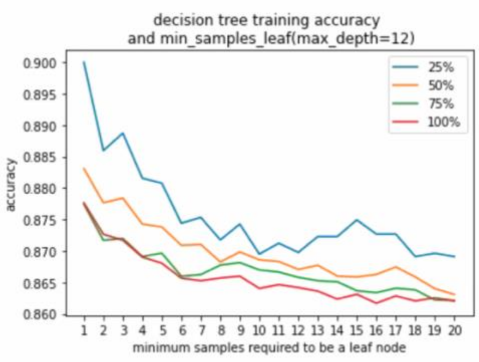
Adult Training accuracy





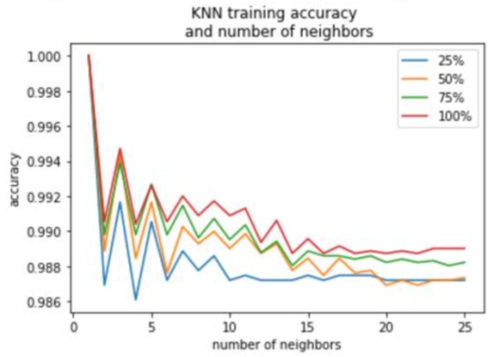
(weight = 'uniform')

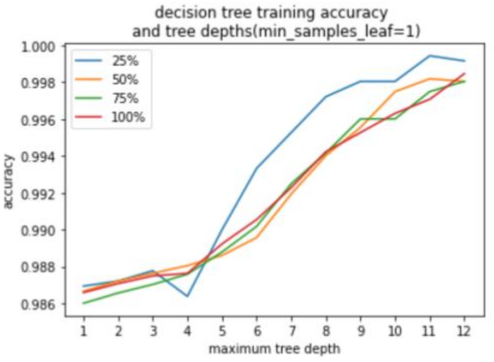




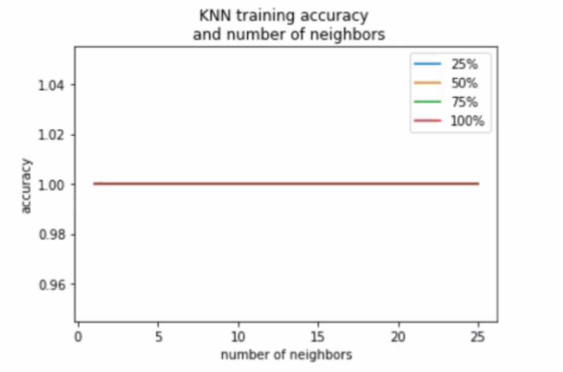
(weight='distance')

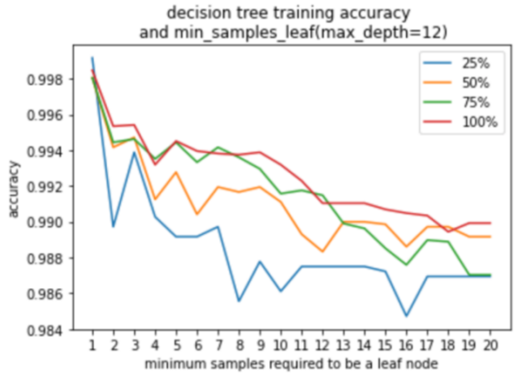
Occupancy Training accuracy





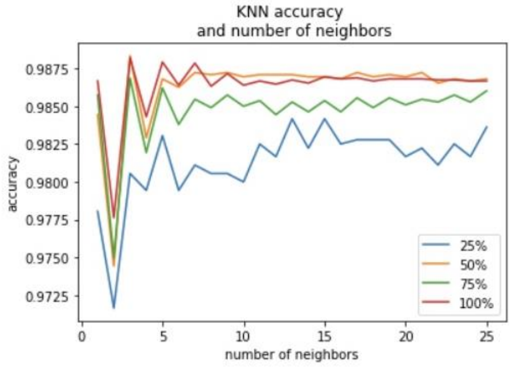
(weight= 'uniform')

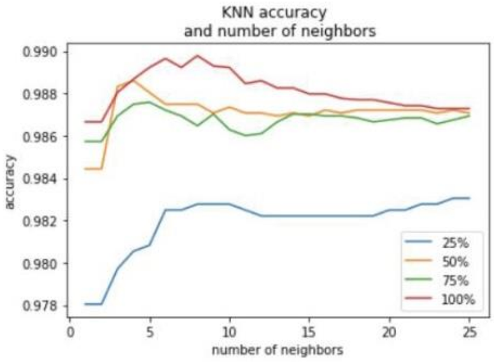




(weight= 'distance')

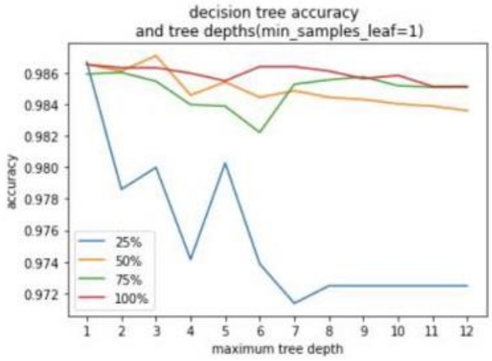
Occupancy Validation accuracy

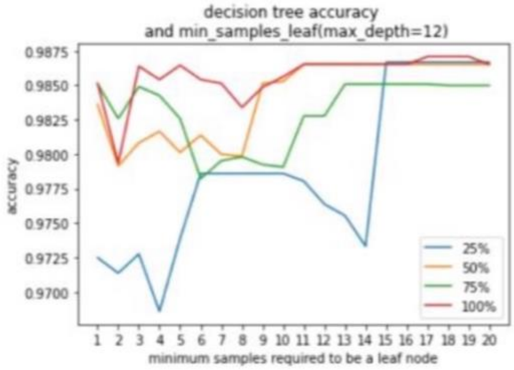




(weight= 'uniform')

(weight= 'distance')


decision tree accuracy
and tree depths(min_samples_leaf=1)
accuracy
maximum tree depth


decision tree accuracy
and min_samples_leaf(max_depth=12)
accuracy
minimum samples required to be a leaf node