Introduction to Machine Learning (CSCI-UA.473): Homework 4

Instructor: Lerrel Pinto

September 20, 2022

Submission Instructions

You must typeset the answers using LaTeX and compile them into a single PDF file. Name the pdf file as ⟨Your-NetID⟩_hw4.pdf and the notebook containing the coding portion as ⟨Your-NetID⟩_hw4.ipynb. The PDF file should contain solutions to both the theory portion and the coding portion. Submit the files through the following Google Form - https://forms.gle/g5t7GQ9i2hqwg3sL9 The due date is **November 3, 2022, 11:59 PM**. You may discuss the questions with each other but each student must provide their own answer to each question.

Part1

1.1 Methodology

1. Mention the libraries used in your solution.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
```

2. Explain the details of the methodology used to solve the homework - how did you read the data, scaler/normalizer used for the data, the SVM package used, data split, kernel functions, etc.

   I use Google Colab for this assignment. I first read the spambase.data file into a dataframe and create column names for each column. Then, I extract the features column into X and the output into y.
   I set 20 different random seeds using a for loop with a range 20 and fit these different random seeds to both `train_test_split` and my model. With each random seed, I randomly split 0.7 of the data as training set (3220 samples) and 0.3 of the data as testing set (1381 samples). I also normalize based on the training set on both the training and testing data using StandardScaler. Then, for different random seeds, I set my model's criterion as "gini" and "entropy" respectively and its random state using `DecisionTreeClassifier`.
   I set an max accuracy score as 0 and then calculate each accuracy score with different random seeds with criterion "gini" using the `accuracy_score` function. If the current accuracy score is larger than the max accuracy score, I replace the max accuracy score with the current accuracy score. And I do the same things to criterion "entropy". Then, I can find the best accuracy score for criterion "gini" and "entropy" among these 20 different seeds.

1.2 Results

According to the table, the best Test Accuracy using Gini Impurity is 0.9211 and the best Test Accuracy using Shannon I.G. is 0.9385. In this case, the tree classifier under Shannon I.G. has a higher test accuracy.

Criterion	**Best Test Accuracy**
Gini Impurity	0.9211
Shannon I.G.	0.9385

Part2

2.1 Methodology

1. Mention the libraries used in your solution.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
```

2. Explain the details of the methodology used to solve the homework - how did you read the data, scaler/normalizer used for the data, the SVM package used, data split, kernel functions, etc.

   I use Google Colab for this assignment. I first read the spambase.data file into a dataframe and create column names for each column. Then, I extract the features column into X and the output into y.
   I create a list of different number of estimators, which are 1,3,5,10,15,20,40,70. I also create 20 different random seeds using a for loop with a range 20 and fit these different random seeds to both `train_test_split` and my model.
   For each number of estimators, I randomly split 0.7 of the data as training set (3220 samples) and 0.3 of the data as testing set (1381 samples) through looping over 20 different random seeds. I also normalize based on the training set on both the training and testing data using StandardScaler. Then, for different random seeds, I set the criterion of my model `DecisionTreeClassifier` as "gini" and "entropy" respectively and its random state.
   I create two lists to record the best test accuracy of "gini" and "entropy" for each number of estimators. For each number of estimator, I set an max accuracy score as 0 and then calculate each accuracy score with different random seeds with criterion "gini" using the `accuracy_score` function. If the current accuracy score is larger than the max accuracy score, I replace the max accuracy score with the current accuracy score. At the end of the loop, I append the best test accuracy score for this specific random seed. And I do the same things to criterion "entropy". Then, I can find the best accuracy score for criterion "gini" and "entropy" among these 20 different seeds for different number of estimators.

2.2 Results

According to the table, the trees using Gini Impurity and Shannon I.G. splitting criterion both have an overall trend of increasing as the number of estimators increasing.
When the number of estimator are [1 , 5, 15, 20, 70], the best test accuracy of using Gini Impurity to split tree is higher than the best test accuracy

of using Shannon I.G. to split tree. When the number of estimators are [3, 10], the best test accuracy of using Shannon I.G. to split tree is higher. When the number of estimators is 40, these two criterion have the same best test accuracy. Overall, we could observe that the Gini Impurity has a higher best test accuracy.

Criterion	**# Estimators**	1	3	5	10	15	20	40	70
Gini Impurity	Test Accuracy	0.9167	0.9377	0.9486	0.9522	0.9594	0.9580	0.9587	0.9602
Shannon I.G.	Test Accuracy	0.9160	0.9442	0.9479	0.9551	0.9558	0.9573	0.9587	0.9594