ELG 5255 Applied Machine Learning Fall 2021

Assignment 4 (Decision Tree and Ensemble Learning)

Submission

You must submit your assignment on-line with Bright Space. This is the only method by which we accept assignment submissions. We do not accept assignments sent via email, and we are not able to enter a mark if the assignment is not submitted on Bright Space! The deadline date is firm since you cannot submit a assignment passed the deadline. It is a student's responsibility to ensure that the assignment has been submitted properly. A mark of 0 will be assigned to any missing assignment.

Goal

This assignment aims to understand how to select criteria for decision tree, how to use tree related machine learning algorithms to reduce dataset dimension, and how to tune hyperparameters for ensemble methods.

Part 1: Numerical Questions:

Please note that Part 1 is not programming questions and should be solved manually, and you should answer the following questions and show your calculation process and the formulas you use. You won't receive marks if you only provide answers.

Let's assume that TAs would go hiking every weekend, and we would make final decisions (i.e., Yes/No) according to weather, temperature, humidity, and wind. Please create a decision tree to predict our decisions.

Weather (F1)	Temperature (F2)	Humidity (F3)	Wind (F4)	Hiking (Label)
Cloudy	Hot	High	Weak	No
Sunny	Hot	High	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Cloudy	Mild	High	Strong	Yes
Rainy	Mild	High	Strong	No
Rainy	Cool	Normal	Strong	No
Rainy	Mild	High	Weak	Yes
Sunny	Hot	High	Strong	No
Cloudy	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Please show the whole process. You will not receive any marks if you only show the final results.

If two features have same Gini or IG, select the feature that is closer to the left; for instant, if F1 and F2 have the same Gini or IG, select F1 because 1 < 2

Q1 (15 Marks): Please build a decision tree by using Gini Index (i.e., $Gini = 1 - \sum_{i=1}^{N_C}(p_i)^2$, where N_C is the number classes)

Q2 (15 Marks): Please build a decision tree by using Information Gain (i.e., $IG(T, a) = Entropy(T) - Entropy(T|a)$). More information about IG

Q3 (5 Marks): Please compare the advantages and disadvantages between Gini Index and Information Gain

Part 2: Programming Questions

Datasets

1. Circle Dataset

   To plot figures and understand how ensemble algorithms work, we will use a 2D dataset generated by make_circles.

```
from sklearn.datasets import make_circles
from sklearn.model_selection import train_test_split

rs = 0
X, y = make_circles(300, noise=0.1, random_state=rs)
trX, teX, trY, teY = train_test_split(X, y, test_size=0.2,
random_state=rs)
```

2. Classification Dataset

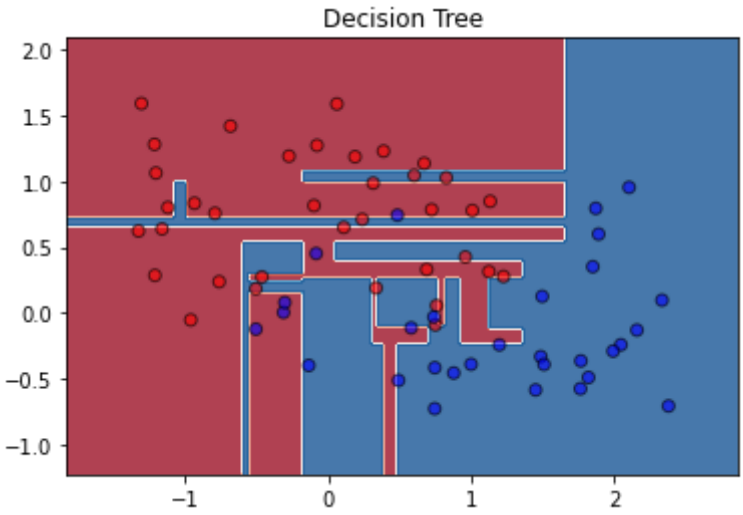
```python
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split

rs = 0
X, y = make_classification(300, random_state=rs)
trX, teX, trY, teY = train_test_split(X, y, test_size=0.2,
random_state=rs)
```

Decision Tree

Q4 (5 Marks): Use Circle Dataset. Apply decision tree on the Circle Dataset, set criterion as gini and entropy, get the accuracy of the testing results, plot the decision boundaries as we showed in the following figure, and explain the difference between these criterion

Decision boundary example:



Q5 (10 Marks): Use Classification Dataset. Use training set to obtain the importance of features. Plot `Validation Accuracy (y-axis)` vs `Top K Important Feature (x-axis)` curve; where 4-fold cross validation should be used, and also plot `Test Accuracy` vs `Top K Important Feature` curve

Bagging

Bagging is to generate a set of bootstrap datasets, create estimators for each bootstrap dataset, and finally utilize majority voting (soft or hard) to get the final decision.

Q6 (5 Marks): Use Circle Dataset. Set the number of estimators as 2, 5, 15, 20 respectively, and generate the results accordingly (i.e., accuracy and decision boundary)

Q7 (5 Marks): Explain why bagging can reduce the variance and mitigate the overfitting problem

Random Forest

Q8 (5 Marks): Use Circle Dataset. Set the number of estimators as 2, 5, 15, 20 respectively, and generate the results accordingly (i.e., accuracy and decision boundary)

Q9 (5 Marks): Compare with bagging results and explain the difference between Bagging and Random Forest

Boosting

Q10 (10 Marks): Use Circle Dataset. There are 2 important hyperparameters in AdaBoost, i.e., the number of estimators (ne), and learning rate (lr). Please plot 12 subfigures as the following table's setup. Each figure should plot the decision boundary and each of their title should be the same format as `{n_estimaotrs}, {learning_rate}, {accuracy}`

ne=10; lr=0.1	ne=50; lr=0.1	ne=100; lr=0.1	ne=200; lr=0.1
ne=10; lr=1	ne=50; lr=1	ne=100; lr=1	ne=200; lr=1
ne=10; lr=2	ne=50; lr=2	ne=100; lr=2	ne=200; lr=2

Stacking

Q11 (10 Marks): We have tuned the Decision Tree, Bagging, Random Forest, and AdaBoost in the previous section. Use these fine tuned model as base estimators and use Naive Bayes, Logistic Regression, and Decision Tree as aggregators to generate the results accordingly (i.e., accuracy and decision boundary)

Conclusion

Q12 (10 Marks): Provide a conclusion section on your report. Include overview of what you have done and learnt during the assignment. Aim no less than one third of a page and no more than half page.