# CAP 5610: HW3: Decision Tree and Ensemble Learning

# **Question 1. (10 points) Understanding Training Error and Testing**

Consider the decision tree shown in the diagram below. The counts shown in the leaf nodes correspond to the number of training records associated with the nodes.



1. What is the training error rate for the tree? Explain how you get the answer?

**With an optimistic error rate, we find: 29/100 = 29%**

**With a pessimistic error rate, we find: 29/100 + 1\*6/100 = 35/100 = 35%**

1. Given a test instance T={A=0, B=1, C=1, D=1, E=0}, what class would the decision tree above assign to T? Explain how you get the answer?

Is this a random forest classifier question?

Since there are three 1’s and 2 0’s, the decision tree will assign a 1 to the test instance.

# **Question 2 (16 points) Understand Splitting Process**

Consider the following data set for a binary class problem.

A close up of a logo

Description automatically generated

**Number of True: 4**

**Number of False: 6**

**True rate: 40%**

**False rate: 60%**

Q1: What is the overall gini before splitting?

**Original Gini = 1 – (0.4)^2 – (0.6)^2 = 0.48**

**This means that 48% of a randomly chosen variables will be wrongly classified**

Q2: What is the gain in gini after splitting on A?

**Number of True in A: 7**

**Number of False in A: 3**

**Gini(A)= T = 1 – (4/7)^2 – (3/7)^2 = 0.49**

**Gini(B = F = 1 – (3/3)^2 – (0/3)^2 = 0**

**Gain = GOrig(0.48) – (7/10\* GTrue(0.49)) – (3/10 \* GFalse(0)) = 0.48 – 0.343 – 0 = 0.137**

Q3: What is the gain in gini after splitting on B:

**Number of True in B: 4**

**Number of False in B: 6**

**Gini(B) = T = 1 - (3/4)^2 - (1/4)^2 = 0.375**

**Gini(B) = F = 1 – (1/6)^2 – (5/6)^2 = 0.2778**

**Gain = GOrig – (4/10 \* GTrue(0.375)) – (6/10 \* GFalse(0.2778)) = 0.48 – 0.15 – 0.16668 = 0.16332**

Q4: Which attribute would the decision tree choose?

**Since splitting B yields a higher gain in Gini, we would choose B to split the node**

# **Question 3: (15 points) Please answer and explain.**

Q1: Are decision trees a linear classifier? Why?

* **Decision Trees are non-linear because there are no linear relationships between the independent and the dependent variables within the tree.**

Q2: What are the weaknesses of decision trees? Why?

* **Disadvantages to a decision tree can be :**
  + **The interacting attributes may be outweighing other attributes that are less discriminating. This is bad because one attribute can essentially decide the class of something without taking anything else into account.**
  + **The decisions can only take in a single attribute. This is bad because everything has to be either a yes or no answer to something and not a continuous gradient of answers.**

Q3: Is Misclassification error better than Gini index as the splitting criteria for decision trees? Why?

* **The Gini index is preferred over Misclassification because it is more sensitive to the data. It will allow for further splits than misclassification which means the tree will become more accurate.**

# **Question-4 (35 points) Build decision tree and random forest using Scikit Learn （**[**https://scikit-learn.org/stable/**](https://scikit-learn.org/stable/)**）**

For the Titanic challenge (https://www.kaggle.com/c/titanic), we need to guess whether the individuals from the test dataset had survived or not. Please:

1. Preprocess your Titanic training data; Please briefly describe what preprocess you have done.

**When looking at the data for preprocessing, I took a look at the following:**

**Null Values**

* + **I found there were 177 Null Values in the Age category and 687 Null values in the Cabin category. I left the Age category alone but changed the Cabin category to a numeric representation with 0 representing Null and 7-1 representing cabins A-G. I set cabin A as a higher value than cabin G due to the nature of cabin A generally having a higher priority than lower cabins.**
  + **I also found there were 2 Null Values in the Embarked categories, since there was such a miniscule amount in the embarked category, I decided to just drop those two rows from the training set all together.**

**Categorical Data:**

* + **I changed ‘Sex’ from ‘Male’ and ‘Female’ to 1 and 2 respectively**
  + **I changed the ‘Cabin’ category to numeric, A=1, B=2 and so on through G**

**Names**

* + **I have decided to drop the names column entirely. There was a correlation between names with ‘Dr.’ as they had a higher chance of survival. But there were only 7 ‘Dr.’ total in the whole training set so I just went and removed them.**

1. Select a set of important features. Please show your selected features and explain how you perform feature selection.

**For the important features I used the selectKBest function from the sklearn library to return the top features.**

**The top 5 features according to the chi2 score are:**

1. **Fare**
2. **Cabin**
3. **Sex**
4. **Age**
5. **PClass**

**So, I will be using just these 5 features in the training.**

1. Learn and fine-tune a decision tree model with the Titanic training data, plot your decision tree;
2. Apply the five-fold cross validation of your fine-tuned decision tree learning model to the Titanic training data to extract average classification accuracy;
3. Apply the five-fold cross validation of your fine-tuned random forest learning model to the Titanic training data to extract average classification accuracy;
4. Which algorithm is better, Decision Tree or Random Forest?
5. What are your observations and conclusions from the algorithm comparison and analysis?

# **Question-5 (20 points) Build a bagging classifier using Scikit Learn for the above Titanic challenge.**

# **Question-6 (20 points) Build an Adaboost classifier using Scikit Learn for the above Titanic challenge.**

**Please write down your answers and submit a PDF report. In your report, please answer each question with your explanations, plots, results in brief. DO NOT paste your code or snapshot into the PDF. At the end of your PDF, please include a website address (e.g. Google colab) that can allow the TA to read your code (for Questions 4-6). Or you can submit your Python or Jupyter Notebook code together with the PDF report.**