CS 2756, Spring 2022 Homework 2

Due: Feb 25, 2022, 11:59 PM, Gradescope

Instructions:

- The main submission of HW 2 must be uploaded on **Gradescope**. Your submission will be graded via Gradescope platform.
- Your PDF file should be named as "lastname_firstname_hw2". For example, if your name is Jane Doe, name your file as "doe_jane_hw2".
- Please write your name (First Name, Last Name) legibly, as it appears on Canvas. Please also include your PITT email.
- Type out your solutions on a separate blank file (using Word, Latex, etc.), and don't forget to convert the file to PDF extension before submitting. You should not type out on the original pdf file.
- Only a subset of the questions will be graded. However, such questions are not determined a priori, therefore please do your best to answer all the questions correctly.

Question 1.

Consider the training examples shown in the table below for the binary classification problem for a pet store's loyalty program members.

Customer ID	Shops Weekly	Pet	Class
1	yes	cat	C0
2	yes	dog	C0
3	yes	both	C0
4	yes	cat	C0
5	yes	cat	C0
6	yes	cat	C0
7	no	dog	C0
8	no	cat	C0
9	no	cat	C0
10	no	cat	C0
11	yes	cat	C1
12	yes	both	C1
13	no	both	C1
14	no	dog	C1
15	no	both	C1
16	no	dog	C1
17	no	dog	C1
18	no	dog	C1
19	no	dog	C1
20	no	dog	C1

a) Compute the Gini, Misclassification Error, and the Entropy for the overall collection of training examples.

b) Compute the Gini, Misclassification Error, and Entropy for all the three attributes: Customer ID, Shops Weekly, and Pet, using multi-way splits for Customer ID and Pet, and binary split for Shops Weekly.

c) Compute the Information Gain for all the three attributes. Which attribute provides the highest Information Gain?

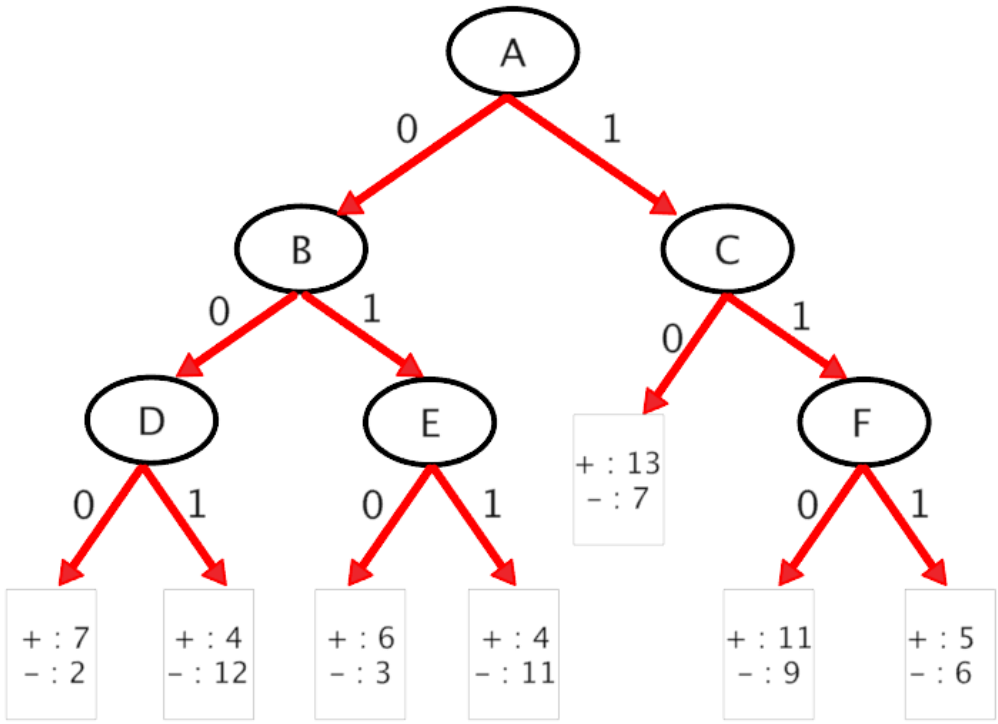
d) Compute the Gain Ratio for all the three attributes. Which attribute provides the highest Gain Ratio?

e) Is there a difference between the attribute that provides highest Information Gain and the attribute that provides highest Gain Ratio? Which attribute would you finally consider for

splitting at the root node? Briefly explain your choice.

Question 2.

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.



a) If each leaf node is labeled according to the majority class of the training instances that reach the node, compute the training error for the tree.

b) Estimate the generalization error for the tree using the pessimistic error rate approach, assuming the cost of each leaf is $\Omega = 0.5$.

c) Suppose the nodes labeled as E and F in the tree are replaced by their corresponding leaf nodes. Estimate the generalization error of the pruned tree using the pessimistic error rate approach. Compare with your answer from part (b) to determine whether the original tree should be pruned.

d) Using the validation set below, determine whether nodes E and F should be pruned by considering the validation error rate of the two trees.

A	B	C	D	E	F	Class
0	1	1	1	1	0	-
0	1	1	1	1	1	-
0	0	0	1	0	1	-
1	0	1	0	1	1	-
0	1	1	1	0	0	-
1	0	1	1	1	1	-
0	1	0	1	0	0	+
1	0	1	1	0	1	+
0	1	1	0	0	0	-
0	1	0	1	0	1	-

Question 3.

Table 1 shows data collected on a runner's decision to go for a run or not go for a run depending on the weather conditions that day. We will use Naïve Bayes (NB) classifier to answer several questions related to this dataset.

Outlook	**Temperature**	**Humidity**	**Run**
Sunny	Mild	Normal	Yes
Rainy	Hot	High	No
Overcast	Mild	Normal	Yes
Overcast	Hot	Normal	Yes
Rainy	Mild	High	Yes
Sunny	Hot	High	No
Rainy	Cool	Normal	Yes
Rainy	Cool	High	Yes
Rainy	Mild	Normal	Yes
Sunny	Hot	High	Yes
Tornado	Cool	Normal	No
Tornado	Mild	Normal	No
Rainy	Hot	High	Yes
Rainy	Cool	Normal	Yes
Sunny	Mild	Normal	Yes

a) Given the data in Table 1 is a person more likely to go for a run or not? Justify your answer.

b) How would Naïve Bayes classify an unseen data point X = {Rainy, Mild, High}? Show your work.

c) Assume that the only information you have about the weather outside is that the temperature is hot. What is NB's prediction whether a person will run or not? Show your work.

d) In addition to knowing that temperature is hot that day, you also know that the humidity is normal. What is NB's prediction whether a person will go for a run or not?

e) Given results in c) and d) comment on the behavior of Naïve Bayes when handling missing data.

f) Now let us go back and compute prediction for a complete data point. In addition to knowing that the temperature is hot and the humidity is normal, assume you also know that the outlook includes a tornado. Is a person more likely to go for a run or not?

g) What went wrong in (f)? What approach would you use to fix it? Explain your answer.

Question 4.

Amazon announces a hiring challenge for building the best classification model that can predict an item's profitability (high or low) based on its various characteristics. Amazon provides everyone with a labeled dataset, A1, that can be used to build a classification model (e.g., using any of the techniques discussed in the class). In addition, Amazon allows participants in this competition to check the accuracy of their models on another data set, A2, for which labels are hidden from the participants. Participants can submit their labels for A2 to Amazon, and it returns the accuracy of these labels on A2 back to them. Thus, this evaluation on A2 can be used by the participants to evaluate various models they are developing before they choose one of them to submit as their final solution to the challenge. The A2 data set is much smaller than A1.

You choose to participate in the challenge, so you build a classifier N1 using A1. For evaluation, you run it on both A1 and A2 and note the accuracy. You are also able to come up with 2 other classifiers N2 and N3. You repeat the above training and evaluation process with them. This gives you a table of accuracies for these 3 classifiers as shown in the following table [Accuracy% = 100% - Error Rate]

Classifier	Accuracy on A1	Accuracy on A2
N1	77%	72%
N2	84%	68%
N3	83%	65%

a) Which one of these classifiers - N1, N2, N3 would you submit as your entry into the challenge and why?

b) You've recently been looking at some Deep Learning courses online and decide to develop a number of highly complex deep neural networks and choose one of them that has the highest accuracy on A2. For this model, N4, the accuracy on A1 and A2 is given in the following table:

N4	90%	72%

Should you submit this classifier instead of the earlier three? Why or why not?

c) Which dataset(s) should Amazon use to evaluate and rank the final submission by each participant?

i) A1, the larger dataset
ii) A2, the smaller but hidden dataset
iii) Both A1 and A2
iv) Neither A1 nor A2

Please state the reasons behind your choice.

Question 5.
In each of the classification scenarios listed below, you are given a set of classifiers and a description of the classification scenario. For each scenario, state the choice of one classifier that is best suited and another one that is worst suited. Give a brief explanation to support your answer.

a) **Scenario:** The data has mixed attributes (some are continuous, some are categorical) and some attributes are redundant (duplicate copies of other attributes).
Classifiers for comparison: kNN, Decision Trees, Naïve Bayes

b) **Scenario:** Data contains missing values, both during training and testing.
Classifiers for comparison: Naïve Bayes, decision trees, kNN

c) **Scenario:** Computation time during training has to be minimized.
Classifiers for comparison: Naïve Bayes, Neural Networks, kNN

Question 6.

For the following scenario, State one major strength and one major weakness of Decision Trees

Scenario: Some attributes in the data are not discriminative by themselves (when considered individually) but are discriminative when used in combinations. Data also contains some redundant (duplicate) and irrelevant attributes.

Question 7. True/False with short explanations
State whether the following statements are true or false and provide brief explanations to support your answer.

a) The presence of noise does not result in decision tree over-fitting because decision trees are resistant to noise.

b) Consider a data set with four binary attributes X1, X2, X3 and X4. The attribute X4 takes the same value as X3 for each record, i.e., X4 is equal to X3.
We build two decision trees:
T1, which is learnt using all the four attributes
T2, which is learnt using only three attributes X1, X2, and X3.
In this scenario, T1 will perform better than T2.

c) Decision tree and rule-based classification automatically perform variable selection.