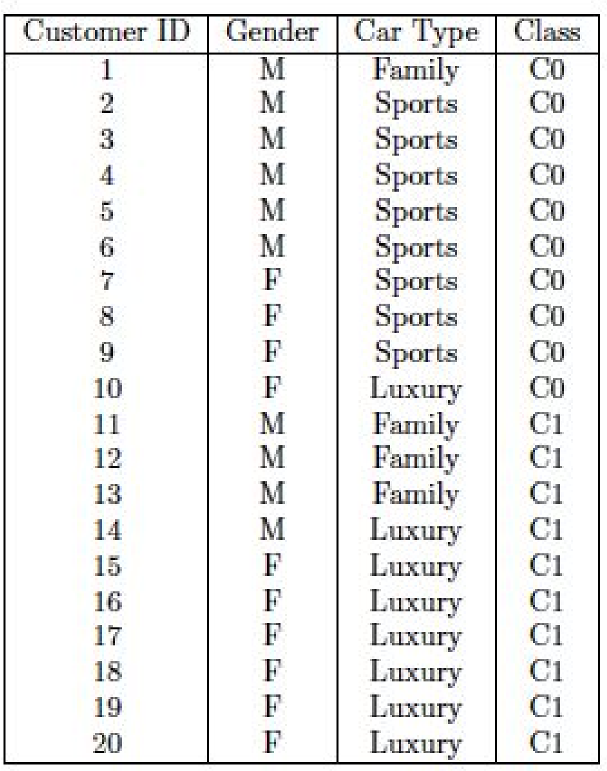
CSCI 5523 **Homework 2** **Due: Feb 18, 2019**

Please clearly mention your name, Student ID and Section(1 for morning, 2 for evening or UNITE) on your submission.

Question 1

Consider the training examples shown in Table 1 for a binary classification problem.

Table 1: Data set for Exercise 1.



Customer ID	Gender	Car Type	Class
1	M	Family	C0
2	M	Sports	C0
3	M	Sports	C0
4	M	Sports	C0
5	M	Sports	C0
6	M	Sports	C0
7	F	Sports	C0
8	F	Sports	C0
9	F	Sports	C0
10	F	Luxury	C0
11	M	Family	C1
12	M	Family	C1
13	M	Family	C1
14	M	Luxury	C1
15	F	Luxury	C1
16	F	Luxury	C1
17	F	Luxury	C1
18	F	Luxury	C1
19	F	Luxury	C1
20	F	Luxury	C1

(a) Compute the Gini index for the overall collection of training examples.

(b) Compute the Gini index for the Customer ID attribute.

(c) Compute the Gini index for the Gender attribute.

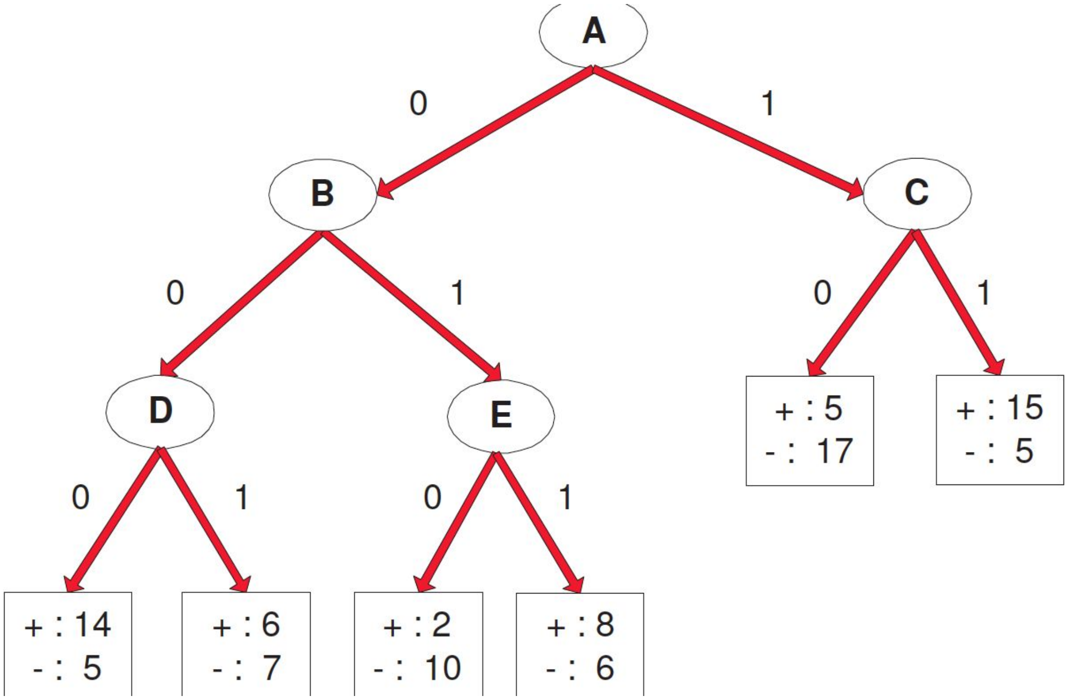
(d) Compute the Gini index for the Car Type attribute using multiway split.

(e) Which of the three attributes has the lowest Gini index?

(f) Which of the three attributes will you use 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.

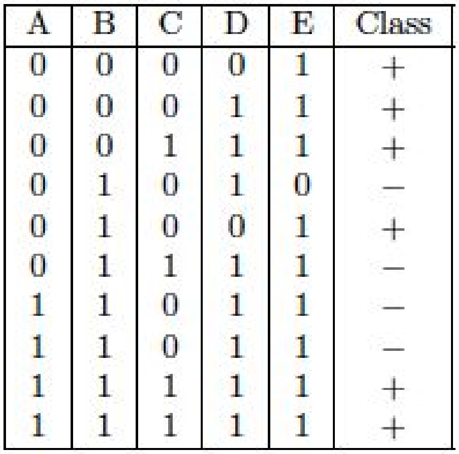


(a) Compute the training error rate for the tree.

(b) Estimate the generalization error for the tree using the pessimistic error rate approach (Assume the cost of each leaf is 2).

(c) Suppose the nodes labeled as D and E in the previous decision tree are replaced by their corresponding leaf nodes. Estimate the generalization error of the pruned tree using the pessimistic error rate approach. Use your answers for part (b) and (c) to determine whether the original tree should be pruned.

(d) Using the validation set shown below, determine whether the nodes D and E in the decision tree should be replaced by their corresponding leaf nodes.



A	B	C	D	E	Class
0	0	0	0	1	+
0	0	0	1	1	+
0	0	1	1	1	+
0	1	0	1	0	−
0	1	0	0	1	+
0	1	1	1	1	−
1	1	0	1	1	−
1	1	0	1	1	−
1	1	1	1	1	+
1	1	1	1	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	Hot	High	No
Overcast	Cool	Normal	No
Sunny	Mild	High	No
Overcast	Mild	High	No
Sunny	Hot	High	Yes
Rainy	Hot	High	Yes
Rainy	Mild	High	Yes
Rainy	Cool	Normal	Yes
Rainy	Cool	Normal	Yes
Sunny	Cool	Normal	Yes
Rainy	Mild	Normal	Yes
Sunny	Mild	Normal	Yes
Rainy	Mild	High	Yes
Rainy	Hot	Normal	Yes

Table 1. Running data for Question 3

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 = {Sunny, Mild, Normal}? Show your work.

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

d) In addition to knowing that temperature is mild that day, you also know that humidity is high. 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 mild and the humidity is high, assume you also know that the outlook is overcast. 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.

State whether the following statements are true or false, and provide brief explanations to support your answer.

a) The classification performance of decision trees will degrade in the presence of redundant attributes (attributes are duplicates of each other).

b) Decision Trees and RIPPER would perform equally well on a dataset in which there are many more instances of one class than the other.

c) The presence of noisy objects does not result in decision tree overfitting because decision trees are resistant to noise.

Question 5.

(a) If you had to choose between the naïve Bayes and k-nearest neighbor classifiers, which would you prefer for a classification problem where there are numerous missing values in the training and test data sets? Indicate your choice of classifier and briefly explain why the other one may not work so well?

(b) Consider the problem of predicting whether a person is a good credit risk given the following attributes: hair color, income, weight, time in current job, marital status, height, age, and birth month. If you had to choose between Ripper and a k-nearest neighbor classifier, which would you prefer? Indicate your choice of classifier and briefly explain why the other one may not work so well?

Question 6.

Consider the data shown in the table below, where X and Y are the two attributes, and the class label is either positive (+) or negative (-):

X	*Y*	*# of (+) records*	*# of (-) records*
x_0	y_0	80	10
x_0	y_1	35	26
x_1	y_1	27	40
x_1	y_0	48	4
x_2	y_1	10	20

(a) Use Bayes theorem to answer the following: (Show all steps. Assume conditional independence.)

What would be the class label for an instance for which we do not have information about the Attributes?

(b) What is the class label of an instance if it is known that X = x_2 ?

(c) What is the class label of an instance if it is known that X = x_2 and Y = y_0 ?

PRACTICE QUESTIONS

Question 1

Consider the dataset shown in Table 1 for a binary classification problem.

Movie ID	Format	Movie Category	Class
1	DVD	Entertainment	C0
2	DVD	Comedy	C0
3	DVD	Documentaries	C0
4	DVD	Comedy	C0
5	DVD	Comedy	C0
6	DVD	Comedy	C0
7	Online	Comedy	C0
8	Online	Comedy	C0
9	Online	Comedy	C0
10	Online	Documentaries	C0
11	DVD	Comedy	C1
12	DVD	Entertainment	C1
13	Online	Entertainment	C1
14	Online	Documentaries	C1
15	Online	Documentaries	C1
16	Online	Documentaries	C1
17	Online	Documentaries	C1

18	Online	Entertainment	C1
19	Online	Documentaries	C1
20	Online	Documentaries	C1

Table 1

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: Movie ID, Format, and Movie Category, using multiway splits for Movie ID and Movie Category, and binary split for Format.

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) For splitting at the root node, would you choose the attribute that provides the maximum IG, or the attribute that provides maximum Gain Ratio? Briefly explain your choice.

Question 2

Consider the decision tree shown in Figure 1, and the corresponding training and test sets in Tables 2 and 3 respectively.

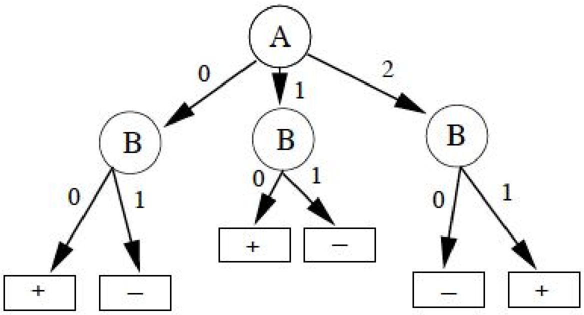
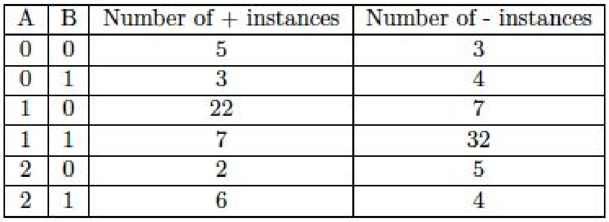
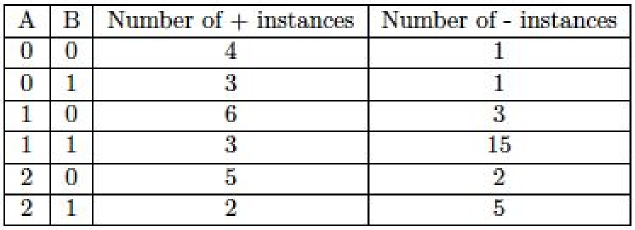


Figure 1: Decision tree for Question 6.



A	B	Number of + instances	Number of - instances
0	0	5	3
0	1	3	4
1	0	22	7
1	1	7	32
2	0	2	5
2	1	6	4

Table 2: Training set for Question 6



A	B	Number of + instances	Number of - instances
0	0	4	1
0	1	3	1
1	0	6	3
1	1	3	15
2	0	5	2
2	1	2	5

Table 3: Test set for Question 6

(a) Estimate the generalization error rate of the tree using both the optimistic approach and the pessimistic approach. While computing the error with the pessimistic approach, to account for model complexity, use a penalty value of 2 for each leaf node.

(b) Compute the error rate of the tree on the test set shown in Table 3.

(c) Figure 2 shows a pruned version of the original decision tree. Estimate the generalization error rate of this tree using both the optimistic approach and the pessimistic approach, as in Part (a). Also compute the error rate of this tree on the test set shown in Table 3.

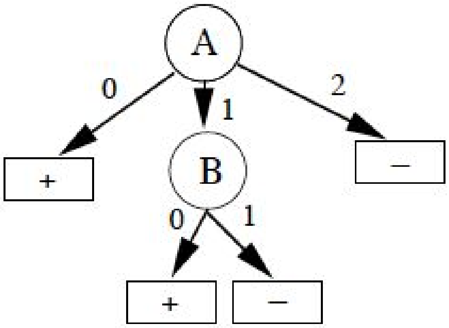


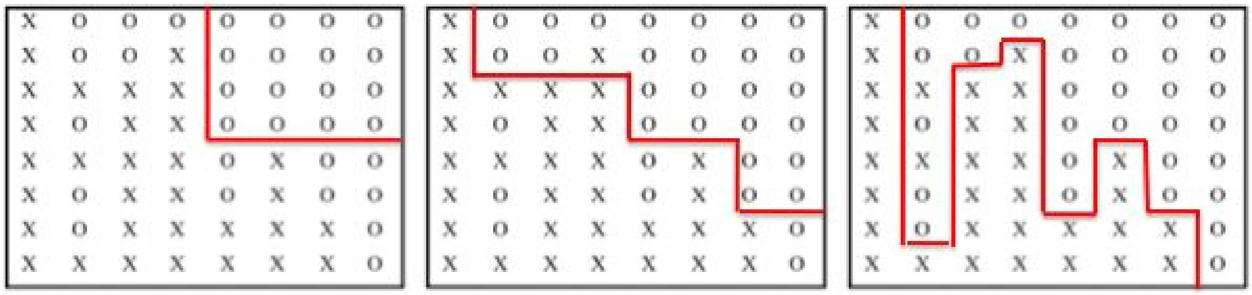
Figure 2: Pruned version of tree in Figure 1

(d) Use your answers of the pessimistic error rate approach from part (b) and (c) to determine whether the original tree should be pruned, and briefly explain.

(e) Comment on the utility of incorporating model complexity in building a predictive model.

Question 3

Figure 3 shows three different decision tree classifiers built on a training sample of the same data set. The dataset consists of instances of class ‘X’ and instances of class ‘O’. The decision boundary of the classifiers is indicated by the boundary line inside the rectangle and the classification decision is made as follows: everything above the boundary is classified as ‘O’ and everything below the boundary is classified as ‘X’. Assume the future instances of the data are similar in terms of distribution and class composition.



(a) (b) (c)

Figure 3: Three decision tree classifiers, a, b, and c. Border of the classifier is indicated by the boundary line inside the rectangle.

a) Which decision tree is the best fit to the training sample and why?

b) Which model will best predict future data? Explain.

c) What are the phenomena in Figure 1a) and 1c) called?

Question 4

Imagine you are given the task to predict the educational qualification of each person using their demographic data with the following attributes: (1) Annual Income (real-valued), (2) Income Tax filed (real-valued), (3) Age (integer), (4) State of residence in the US (categorical), (5) Gender (categorical), (6) House Owner or not (Boolean) and (7) Height (in inches). Assume the target classes are (a) college degree and (b) without a college degree. Also, assume that the fraction of the population that has a college degree is roughly equal to the fraction that does not have a college degree. State one strength and one weakness of decision trees for this task?

Question 5

Both Minimum Description Length (MDL) and the pessimistic error estimate are techniques used for incorporating model complexity. State one similarity and one difference between them in the context of decision trees.

Question 6

Consider a data set with four binary attributes X1, X2, X3 and X4. The attribute X4 takes exactly the same value as X3 for each record, i.e. X4 is equal to X3. In the following scenario, find whether the decision boundary learnt by the two models would be similar; otherwise find which of the two models would perform better. Provide a brief justification.

We build two decision trees:

a. T1, which is learnt using all the four attributes

b. T2, which is learnt using only three attributes X1, X2, and X3.