CS653: Data Mining

Mid-term Exam

Fall 2025

Name: _____ REDID: _____

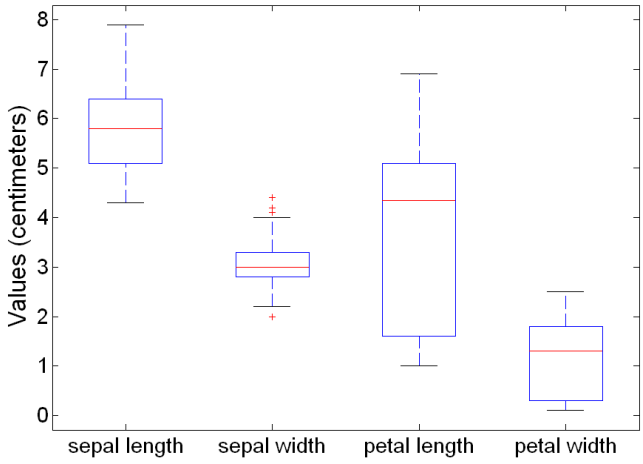
Note:

1. Write your answers on the provided answer sheet.
2. Submit both the answer sheet and the question sheet at the end of the exam.
3. The duration of the exam is 75 minutes.

Question Set I

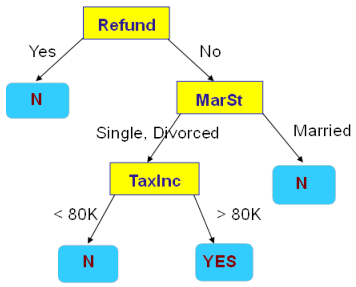
Question 1. Please read the following statements each describing a computer system, and identify which system(s) is (are) not addressed by data mining techniques. Put 'Yes' or 'No' after each activity's number in your answers (e.g., a: No).
(a) A system that can group the customers of a company according to their genders
(b) A system that can intelligently recommend a new product to existing customers who are likely to buy it.
(c) A system that can sort a student database based on student identification numbers.
(d) A system that can calculate the total sales of a company given the sale records.
(e) A system that can extract the frequencies of a sound wave.
(f) A system that can predict the future stock price of a company using its historical records.
(g) A system that can monitor the heart rate of a patient for abnormalities.
(h) A system that can predict seismic waves for earthquake activities.
(i) A system that can identify fraud transactions which are substantially different from other transactions

Question2 The following figure shows the box-plot the Iris flower dataset, where each flower includes 4 attributes: sepal length, sepal width, petal length, petal width. Describe how a box plot can give information about each feature.



Question Set II

Suppose you have trained the following decision tree to predict if a user has cheated on tax claim.



Question 3 Please apply the tree model to the following test samples and fill in the predictions (Y or N) in the last column.

Refund	Married Status(MarSt)	Tax Income (TaxInc, $K)	Ground-truth	Predictions
Yes	Divorced	10	Y	
No	Married	82	N	
No	Single	60	N	
No	Divorced	100	N	
No	Divorced	70	Y	
No	Single	90	N	
No	Married	89	Y	

Question 4. Based on your predictions in the last question, please calculate the confusion matrix, and per-class precision and recall.

Question Set III

Consider the training samples shown in the following table for a binary classification problem.

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

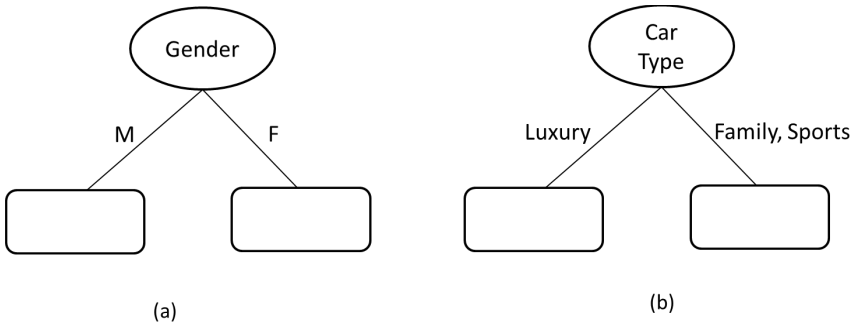
Question 5. Compute the Gini Index (i.e., GINI) for the overall collection of training examples.

Question 6. When a set of samples is split into k partitions (children), the quality of this split is computed as:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

where n_i is the number of samples at children i, n is the total number of samples.

Please calculate the qualities of the following two ways of splitting, i.e., (a) and (b), respectively.



(a) (b)

Question Set IV

Question 7. Consider the one-dimensional data set shown in the following table. Classify the data point $x = 5.0$ according to its 1-, 3-, and 5- nearest neighbors (using majority vote). Please briefly discuss the consequences of using small or large K for K-NN (K-nearest neighbor method).

x	0.5	3.0	4.2	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

Question 8. Consider the data set shown in the following table. We will apply the Naïve Bayes method to predict the class of an unseen record. Please accomplish the following two steps.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

(a) Estimate the conditional probabilities: P(A|+), P(B|+), P(C|+), P(A|-), P(B|-), and P(C|-).
(b) Use the estimate of conditional probabilities to predict the class label for a test sample (A=0, B=1, C=0) using the Naïve Bayes method.