iversité Libanaise
iculté d'Information
Br I

الجامعة اللبنانية
كلية الاعلام
الفرع الاول

الامتحانات النهائية الفصل الخامس

المادة: Data Mining and warehousing	المرحلة: اجازة
المدة: ساعة ونصف	السنة المنهجية الثالثة
الأستاذ: د. عباس رمال	الاختصاص: علم البيانات

Short Questions (30 pts)

1. You observe that the training accuracy of your Decision Tree classifier is much higher than the test accuracy. What is this phenomenon called? What would you change in your decision tree to combat this behavior? (10pts)

2. Describe briefly the k-fold cross validation algorithm. Explain why this method is used (5pts)

3. A classifier was built to predict a dependent variable categorized as "Yes", "No". 80% of the data set were used to train the classification model and the remaining 20% was used to test the resulting model. The prediction accuracy was evaluated using the test set. The confusion matrix is below.

		Actual	
		Yes	No
Predicted	Yes	100	30
	No	10	37

   a. How many observations were used to train the model? How many observations were used to test the model? (7pts)
   b. What is the accuracy, false-positive rate, and false-negative rate? Assume that "Yes" is positive and "No" is negative. (8 pts)

Association Rule and Decision Tree (30 pts)

Consider a database as shown below. Suppose each transaction is considered as a set of items.

ID	Items
T-1	A, B, C, D
T-2	A, B, C, F
T-3	A, C, F
T-4	B, C, D

a. Let minimum support be 3. Derive the frequent itemsets using either the Apriori algorithm, or by drawing a lattice. (8pts)

b. List the set of all **max** frequent patterns. (7pts)

c. Let F and D be two class labels; that is, our target variable can have one of the values F or D. A, B and C are considered as binary attributes. Transform the transactions database above into a **labeled dataset.** (8pts)

d. Construct the corresponding decision tree using Entropy as split criterion. (7pts)

Decision Tree (40 pts)

Consider the training examples shown in 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	ExtraLarge	C0
6	M	Sports	ExtraLarge	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	ExtraLarge	C1
13	M	Family	Medium	C1
14	M	Luxury	ExtraLarge	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

a) Compute the Gini index for the overall collection of training examples (5pts).

b) Compute the Gini index for the Customer ID attribute (5pts).

c) Compute the Gini index for the Gender attribute (5pts).

d) Compute the Gini index for the Car Type attribute using multiway split (7pts).

e) Compute the Gini index for the Shirt Size attribute using multiway split (8pts).

f) Which attribute is better, Gender, Car Type, or Shirt Size? (5pts)

g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini (5pts).

Good luck!

Exercise2: Association Rule & Decision Tree

ID	Items
T1	A, B, C, D
T2	A, B, C, F
T3	A, C, F
T4	B, C, D

minSup = 3

Sup(A) = 3 ✓
Sup(B) = 3 ✓
Sup(C) = 4 ✓
Sup(D) = 2
Sup(F) = 2

Sup(A, B) = 2
Sup(A, C) = 3 ✓
Sup(B, C) = 3 ✓

a) Freq. Itemsets:
A, B, C, AB, AC, BC

b) {A, C} {B, C}

c)

A	B	C	Class
1	1	1	D
1	1	1	F
1	0	1	F
0	1	1	D

d) Before Splitting

~~E(B)~~ =

$E(\text{Before}) = -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4})$

$= 1$

→ Entropy for each attribute

1) A : $1: \{F^2, D^1\}$ $0: \{F^0, D^1\}$

$E(1) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.918$

$E(0) = 0$

$E(A) = 3/4 \ 0.918 = 0.6885$

$Gain(A) = 1 - 0.6885 = 0.3115$

2) B: $1: \{F^1, D^2\}$ $0: \{F^1, D^0\}$

$E(1) = 0.918$ $E(0) = 0$

$E(B) = 0.6885$

$Gain(B) = 0.3115$

3) C: $1\{F^2, D^2\}$ $0: \{F^0, D^0\}$

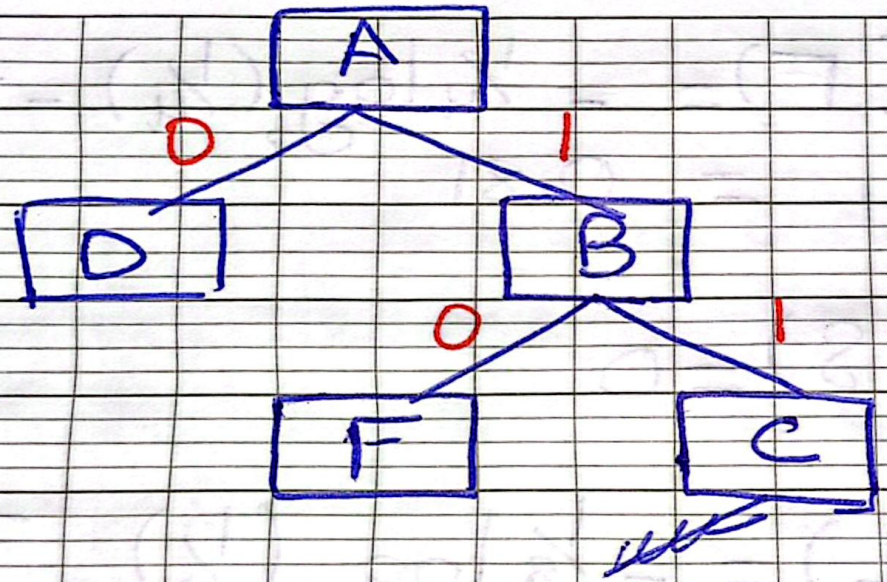
$E(1) = -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4})$

$= 1$

$E(C) = 1$

$Gain(C) = 1 - 1 = 0$

A and B have the same Info. Gain we can choose on of them as the root


A
0
1
D
B
0
1
F
C

Exercise3: Decision Tree

2. Consider the training examples shown in Table 4.1 for a binary classification problem.

Table 4.1. Data set for Exercise 2.

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	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	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

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

Answer:

Gini $= 1 - 2 \times 0.5^2 = 0.5$.

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

Answer:

The gini for each `Customer ID` value is 0. Therefore, the overall gini for `Customer ID` is 0.

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

Answer:

The gini for `Male` is $1 - 2 \times 0.5^2 = 0.5$. The gini for `Female` is also 0.5. Therefore, the overall gini for `Gender` is $0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$.

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

Answer:

The gini for `Family` car is 0.375, `Sports` car is 0, and `Luxury` car is 0.2188. The overall gini is 0.1625.

(e) Compute the Gini index for the `Shirt Size` attribute using multiway split.

Answer:

The gini for `Small` shirt size is 0.48, `Medium` shirt size is 0.4898, `Large` shirt size is 0.5, and `Extra Large` shirt size is 0.5. The overall gini for `Shirt Size` attribute is 0.4914.

(f) Which attribute is better, `Gender`, `Car Type`, or `Shirt Size`?

Answer:

`Car Type` because it has the lowest gini among the three attributes.

(g) Explain why `Customer ID` should not be used as the attribute test condition even though it has the lowest Gini.

Answer:

The attribute has no predictive power since new customers are assigned to new `Customer IDs`.