LEBANESE UNIVERSITY

Faculty of Sciences – Section I

الجامعة اللبنانيّة

كلية العلوم - الفرع الأول

FINAL Exam 2023: INFO 437 -E -- Duration: 1h:30m

Exercise I:

1. Explain the principle of supervised classification and the difference with unsupervised classification.
2. Among the algorithms presented in class, specify which ones belong to supervised methods and which ones belong to unsupervised methods.
3. What are the qualities of a good clustering?

Exercise II:

A) Consider the following data set for a binary class problem (table 1).

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Table 1

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	+

Table2

a) Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

b) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

B) Consider the data set shown in Table2.

a) Estimate the conditional probabilities for P(A|+), P(B|+), P(C|+), P(A|−), P(B|−), and P(C|−).

b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A = 0,B = 1, C = 0) using the naive Bayes approach.

Exercise III: consider the following table.

Tableau des données

N°	Attribut 1	Attribut 2	Classe
1	1	2	C1
2	2	6	C1
3	2	5	C2
4	2	1	C3
5	4	2	C5
6	5	6	C4
7	6	5	C3
8	6	1	C6

a) We want to classify **U(1, 4)** using KNN. What will be the class of U if we choose k=3. Justify.

b) We now use the variant of KNN which uses the distance **1/d2** (inverse of the distance squared) to calculate the neighbors. What will be the class of U with k=3? Justify.

LEBANESE UNIVERSITY

Faculty of Sciences – Section I

الجامعة اللبنانيّة

كلية العلوم - الفرع الأول

FINAL Exam 2023: INFO 437 -E -- Duration: 1h:30m

Exercise IV: We want to apply the "Association rules" model to a TextMining problem. The following table represents the keywords (the most important words) extracted from 7 texts.

N° Text	Keywords
01	Finance, Marché, Budget, Economie
02	Ouverture, Finance, Economie
03	Ouverture, Assemblée, Handball, Sport
04	Directeur, Budget, Finance, Economie
05	Directeur, Assemblée, Handball, Sport
06	Ouverture, Marché, Economie
07	Ouverture, Assemblée, Directeur, Handball, Sport

a) Without doing any calculation, give an association rule of the table whose confidence is equal to 100%. Justify.
b) Rewrite the previous table keeping only the first letter of each keyword (to simplify notation). Apply the a priori algorithm to find all association rules that satisfy **minsup >=40%** and give their confidence. Detail all the steps.

Exercise V: Consider the set D of the following integers: D= {2, 5, 8, 10, 11, 18, 20} We want to divide the data of D into three (3) clusters, using the **Kmeans** algorithm. The distance d between two numbers a and b is calculated as follows: **d(a , b) = |a - b|** (the absolute value of a minus b)

Work to do :

a) Apply **Kmeans** by choosing as initial centers of the 3 clusters respectively: 8, 10 and 11. Show all the calculation steps.
b) Give the final result and specify the number of iterations that were necessary.
c) Can we have a lower number of iterations for this problem? Discuss.