الجامعة اللبنانية
كلية الاعلام
الفرع الاول

Universite Libanaise
Faculté d'Information
Br I

امتحان الفصل
للعام الجامعي 2025/2024

المرحلة:
السنة المنهجية: الثالثة
المادة: إستخراج البيانات وتخزينها
الأستاذ: عباس رمال
الدورة: الأولى

Exercise 1: Clustering

Consider the following 6 data points in a 2-dimensional feature space:

$x_1 = (0, 0)$; $x_2 = (0, 1)$; $x_3 = (-1, 2)$; $x_4 = (2, 0)$; $x_5 = (3, 0)$; $x_6 = (4, -1)$

Let's calculate the squared distances between each pair of points using the Euclidean distance formula:

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	1	5	4	9	17
x_2	1	0	2	5	10	20
x_3	5	2	0	12	20	34
x_4	4	5	12	0	1	5
x_5	9	10	20	1	0	2
x_6	17	20	34	5	2	0

a) Perform K means clustering on this dataset. Use the first and last data points as initial centers ($K = 2$). Given the final parameters, which cluster would $x^* = (1, 1)$ belong to?

b) Perform agglomerative Hierarchical Clustering using single linkage as the cluster distance measure. Draw the associated tree.

c) Perform agglomerative Hierarchical Clustering using complete linkage as the cluster distance measure. Draw the associated tree.

Exercise 2: Decision Tree Classifier

We consider a real dataset with three categorical attributes (Weather, Temperature, Day) and a binary target variable (Play: Yes or No). Here are 10 records in the dataset:

Color	Size	Prize	Label
Red	Small	10 < 35	Class A
Blue	Large	15 < 35	Class B
Green	Medium	20 < 35	Class A
Red	Large	25 < 35	Class B
Blue	Medium	30 < 35	Class A
Green	Small	35 ≥ 35	Class B
Red	Medium	40 ≥ 35	Class A
Blue	Small	45 ≥ 35	Class B
Green	Large	50 ≥ 35	Class A
Red	Large	55 ≥ 35	Class B

a) Take the optimal split for the "Prize" attribute in your decision tree: < 35 and ≥ 35 Construct a decision tree based on this training data. For splitting, use information gain as measure for impurity. Build a separate branch for each attribute.

b) Split the Prize attribute in two way using Gini Index.

c) Create the contingency table for Color and Size attributes. Test whether the two variables (color and size) are related to each other. Take the critical value of chi-square is 9.488

Exercise 3: Evaluation of classifiers

1. Given a data set $D = \{o_1, \ldots, o_n\}$ with known class labels $C(o_i) \in C = \{A, B, C\}$ of the objects. In order to evaluate the quality of a classifier K, each object $o_i \in D$ is additionally classified using K, yielding class label $K(o_i)$. The results are given in the table below.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$C(o_i)$	A	B	A	C	C	B	A	A	A	B	B	C	C	C	B
$K(o_i)$	A	A	C	C	B	B	A	A	A	C	A	A	C	C	B

a) Setup the confusion matrix.
b) Compute the accuracy / classification error
c) For each class $i \in C$ compute precision and recall.
d) Compute the F1-measure for all classes.

4+2+3+2+5

2. **Explain how Leave One Out Cross-Validation is implemented.**
3. **What are the advantages of K-Fold cross validation relative to the validation set approach?**
4. To construct a Receiver Operating Characteristic (ROC) curve, we need to calculate the True Positive Rate (TPR) and False Positive Rate (FPR) at different cutoff points. The cutoff point represents the threshold used to classify samples as positive or negative. Construct the ROC curve on cutoff point = 0.3 and 0.7.

Instance	Actual	Probability Yes	Probability No
S1	Yes	0.80	0.20
S2	No	0.25	0.75
S3	Yes	0.40	0.60
S4	No	0.65	0.35
S5	Yes	0.70	0.30
S6	No	0.50	0.50