Lebanese University
Faculty of information
Section 1

Final Exam: Machine Learning

Time: 120 minutes
Dr. Abbas RAMMAL

Exercise 1.

You have a dataset containing information about different cars and whether they were stolen or not.

Dataset Description:

- The dataset consists of the following features:
- Type of car (e.g., sedan, SUV, sports car)
- Color of the car (e.g., red, blue, black)
- Origin of the car (e.g., Domestic, Imported)

The target variable is whether the car was stolen (binary: Yes, or No).

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	Yes
5	Yellow	Sports	Imported	No
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

1. Using Naive Bayesian classification method, predict a class label (yes or no) for the following unknown's samples:
   - S1: (Red Domestic SUV) No
   - S2: (Yellow Domestic SUV) No
   - S3: (Red Imported SUV) No
   - S4: (Red Domestic Sports) Yes
   - S5: (Yellow Domestic Sports) Yes
   - S6: (Yellow Imported SUV) No

   Show all computation steps.

2. The actual class labels for the unknown samples are as follows:

- S1: Yes
- S2: No
- S3: Yes
- S4: Yes
- S5: No
- S6: No

a) Construct the confusion matrix to compare the predicted class labels with the actual class labels for all samples.

b) Calculate various evaluation metrics such as accuracy, precision, recall, and F1-score to assess the performance of the classifier.

Exercise 2.

You are to cluster eight points: $x_1 = (2,3)$, $x_2 = (5,7)$, $x_3 = (8,1)$, $x_4 = (4,9)$, $x_5 = (1,5)$ and $x_6 = (6,4)$. Suppose, you assigned x_1 and x_6 as initial cluster centers for K-means clustering. The distance matrix based on the Euclidean distance is given in Table 1.

Table 1: Distance matrix for training dataset

$$\text{Distance Matrix} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{bmatrix} 0 & 5 & 2\sqrt{10} & 2\sqrt{10} & \sqrt{5} & \sqrt{17} \\ 5 & 0 & 3\sqrt{5} & \sqrt{5} & 2\sqrt{5} & \sqrt{10} \\ 2\sqrt{10} & 3\sqrt{5} & 0 & 4\sqrt{5} & \sqrt{65} & \sqrt{13} \\ 2\sqrt{10} & \sqrt{5} & 4\sqrt{5} & 0 & 5 & \sqrt{29} \\ \sqrt{5} & 2\sqrt{5} & \sqrt{65} & 5 & 0 & \sqrt{26} \\ \sqrt{17} & \sqrt{10} & \sqrt{13} & \sqrt{29} & \sqrt{26} & 0 \end{bmatrix} \end{array}$$

a) Apply the k-means algorithm to this dataset using the distance matrix we calculated previously for 6 points.

b) Use the Nearest Neighbor clustering algorithm and Euclidean distance to cluster this dataset. Suppose that the threshold t is 4.

c) Use single and complete link agglomerative clustering to group the data described by the distance matrix. Show the dendrograms.

Exercise 3.

You have to study the relationship between the monthly e-commerce sales and the online advertising costs. You have the survey results for 7 online stores for the last year. Your task is to find the equation of the straight line that fits the data best. The following table represents the survey results from the 7 online stores.

Online Advertising Dollars (X)	1.7	1.5	2.8	5	1.3	2.2	1.3
Monthly E-commerce Sales (Y)	368	340	665	954	331	556	376

The sum of squared of x and y values are

$$\sum_{i=1}^{7}(x_i - \bar{x})^2 = 10.5371$$

$$\sum_{i=1}^{7}(x_i - \bar{x})(y_i - \bar{y}) = 1806.76$$

a) Draw a scatter diagram of the data. Does a simple linear regression model seem appropriate here?
b) Fit the simple linear regression model using the method of least squares.
c) Estimate the standard errors of β_0 and β_1. Hint : $SSE = 12484.2246$
d) Estimate the confidence interval of β_0 and β_1.
e) Test H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ using the analysis of variance procedure with $\alpha = 0.05$ and F= 2.4469.
f) Find a 95% prediction interval on y when x=2. T= 1.645.