Higher Institute of Technological Studies of Bizerte

AY: 2023-2024 | **L3-S5: Dept. of Electrical Engineering**
Exam | **Machine Learning** | **Teacher: A. Mhamdi**
Jan. 2024 | **Time Limit: $1\frac{1}{2}$ h**

This document contains **5** pages numbered from **1/5** to **5/5**. As soon as it is handed over to you, make sure it is complete. The **3** tasks are independent and can be treated in the order that suits you.

The following rules apply:

❶ **No document** is allowed in the examination room.

❷ **Any electronic material**, except basic calculator, is prohibited.

❸ **Mysterious or unsupported answers** will not receive full credit.

❹ **Round results** to the nearest thousandth *(i.e., third digit after the decimal point)*.

Task Nº1

⌛ 20mn | (4 points)

You are given a dataset as in Tab. 1 that contains information about the number of hours studied (x) and the corresponding scores (y) obtained by a group of students. Your task is to apply the normal equation to find the parameters of a linear regression model.

Table 1: Number of hours studied and scores

Hours	2	3	4	5	6	7	8	9	10
Scores	2	3.5	5	7	12	15	16.5	17.5	18

(a) (2 points) Apply the normal equation to solve for the parameters in:

$$y = \theta_0 + \theta_1 x \tag{1}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{bmatrix}$$

and

$$y^T = \begin{bmatrix} 2 & 3.5 & 5 & 7 & 12 & 15 & 16.5 & 17.5 & 18 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 9 & 54 \\ 54 & 384 \end{bmatrix} \quad \text{and} \quad X^T y = \begin{bmatrix} 96.5 \\ 716.0 \end{bmatrix}$$

Applying the normal equation yields:

$$\theta = \left(X^T X\right)^{-1} X^T y \quad \therefore \quad \begin{cases} \theta_0 &= -2.978 \\ \theta_1 &= 2.283 \end{cases}$$

(b) (2 points) Use the previous equation to predict the score for a student who studied for 6 hours. Calculate the mean squared error (MSE) between the predicted and actual scores.

Let's denote by $\hat{y}$ the predicted score: $\hat{y} = 10.72$. The actual score y is 12.

$$\text{MSE} = 1.638 \because \text{MSE} = (\hat{y} - y)^2$$

Task №2 ⌛ 30mn | (6 points)

Consider the dataset in TAB. 2 of cars with four features: horsepower, acceleration, fuel efficiency (in miles per gallon), and weight (in kilograms). Each car belongs to one of four classes: sedan, sports car, suv, or truck. We want to classify a new car based on its standardized horsepower, acceleration, fuel efficiency and weight.

Table 2: Features of cars

Car	Horsepower	Acceleration	MPG	Weight	Class
1	200	8	25	1200	sedan
2	300	6	20	1500	sport car
3	150	9	30	1400	suv
4	250	7	15	2000	truck
5	180	8.5	22	1300	sedan
6	280	6.5	18	1600	sport car
7	160	9.5	28	1500	suv
8	240	7.5	16	1800	truck

(a) (2 points) Standardize the data in TAB. 2. *(These are the standard deviation for each feature: Horsepower* 52.2*, Acceleration* 1.15*, MPG* 5.17 *and Weight* 244.63*)*

At first, we compute the mean μ and the standard-deviation σ for each feature.

Feature	$\underbrace{\text{Horsepower}}_{hp}$	$\underbrace{\text{Acceleration}}_{a}$	$\underbrace{\text{MPG}}_{mpg}$	$\underbrace{\text{Weight}}_{w}$
μ	220.0	7.75	21.75	1537.5
σ	52.2	1.15	5.17	244.63

For each feature value, we subtract μ and divide it by σ.

$$hp = \begin{bmatrix} -0.383 & 1.533 & -1.341 & 0.575 & -0.766 & 1.149 & -1.149 & 0.383 \end{bmatrix}$$

$$a = \begin{bmatrix} 0.217 & -1.522 & 1.087 & -0.652 & 0.652 & -1.087 & 1.522 & -0.217 \end{bmatrix}$$

$$mpg = \begin{bmatrix} 0.629 & -0.338 & 1.596 & -1.306 & 0.048 & -0.725 & 1.209 & -1.112 \end{bmatrix}$$

$$w = \begin{bmatrix} -1.38 & -0.153 & -0.562 & 1.891 & -0.971 & 0.255 & -0.153 & 1.073 \end{bmatrix}$$

(b) (4 points) Let's say we have a new car with the following specifications: horsepower of 220, acceleration of 7.2, fuel efficiency of 24 MPG, and weight of 1400 Kg. We want to classify this car using k-**NN** with $k = 3$ and the Euclidean distance on the standardized data.

The standardized new input is:

$$\begin{bmatrix} 0.0 & -0.478 & 0.435 & -0.562 \end{bmatrix}$$

The distance of the observation to each car is:

Car #1	Car #2	Car #3	Car #4	Car #5	Car #6	Car #7	Car #8
1.156	2.051	2.366	3.068	1.477	1.925	2.467	2.298

The k nearest neighbors with $k = 3$ are:

- Car #1 (Distance = 1.156)
- Car #5 (Distance = 1.477)

- Car #6 (Distance = 1.925)

Among the k nearest neighbors, "Sedan" appears twice, and "sport car" appears once. Therefore, based on the k-NN algorithm with Euclidean distance on standardized data, the new car with specifications of horsepower 220, acceleration 7.2, fuel efficiency 24 MPG, and weight 1400 kg is predicted to belong to the "Sedan" class.

Task N°3

⌛ 40mn | (10 points)

Use the K-means algorithm and Manhattan distance ($p = 1$) to cluster the following 6 points into 3 clusters.

Point	A	B	C	D	E	F
x_1	3	8	4	2	7	5
x_2	3	5	4	3	7	0

(a) (6 points) Perform K-means clustering and show all the calculations performed at each iteration. *(Initial centroids α, β and γ are set at* A, C *and* F *respectively.)*

1ST ITERATION

Datum point	A	B	C	D	E	F
Feature x_1	3	8	4	2	7	5
Feature x_2	3	5	4	3	7	0
Distance to α	0	7	2	1	8	5
Distance to β	2	5	0	3	6	5
Distance to γ	5	8	5	6	9	0
∈ Cluster	#1	#2	#2	#1	#2	#3

New centroids are:

$$\alpha \begin{pmatrix} 2.5 \\ 3 \end{pmatrix}; \beta \begin{pmatrix} 19/3 \\ 16/3 \end{pmatrix}; \gamma \begin{pmatrix} 5 \\ 0 \end{pmatrix}$$

2ND ITERATION

Datum	A	B	C	D	E	F
x_1	3	8	4	2	7	5
x_2	3	5	4	3	7	0
$d(\_, \alpha)$	0.5	7.5	2.5	0.5	8.5	5.5
$d(\_, \beta)$	17/3	6/3	12/3	20/3	7/3	20/3
$d(\_, \gamma)$	5	8	5	6	9	0
∈	#1	#2	#1	#1	#2	#3

New centroids are:

$$\alpha \begin{pmatrix} 3 \\ 10/3 \end{pmatrix}; \beta \begin{pmatrix} 7.5 \\ 6 \end{pmatrix}; \gamma \begin{pmatrix} 5 \\ 0 \end{pmatrix}$$

3RD ITERATION

Datum	A	B	C	D	E	F
x_1	3	8	4	2	7	5
x_2	3	5	4	3	7	0
$d(\_, \alpha)$	$1/3$	$20/3$	$5/3$	$4/3$	$23/3$	$16/3$
$d(\_, \beta)$	7.5	1.5	5.5	8.5	1.5	8.5
$d(\_, \gamma)$	5	8	5	6	9	0
$\in$	#1	#2	#1	#1	#2	#3

Centroids are:

$$\alpha \begin{pmatrix} 3 \\ 10/3 \end{pmatrix}; \ \beta \begin{pmatrix} 7.5 \\ 6 \end{pmatrix}; \ \gamma \begin{pmatrix} 5 \\ 0 \end{pmatrix}$$

(b) (4 points) Draw a 2-d space with all the 6 points. Show the clusters and the new centroids after each iteration.