AY: 2022-2023
RESIT | **AI-ECUE221**
July 2023

M1-S2: Dept. of Electrical Engineering
Teacher: A. Mhamdi
Time Limit: $1\frac{1}{2}$ h

This document contains **6** pages numbered from **1/6** to **6/6**. As soon as it is handed over to you, make sure it is complete. The **2** 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)*.

❺ **Task N°2:** Each correct answer will grant a mark with no negative scoring.

SELF-REVIEW

Task	1	2	**Total**
Points	9	11	20
Score			

Task N°1

⏳ 40mn | (9 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
$\in$ 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
$\in$	#1	#2	#1	#1	#2	#3

New centroids are:

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

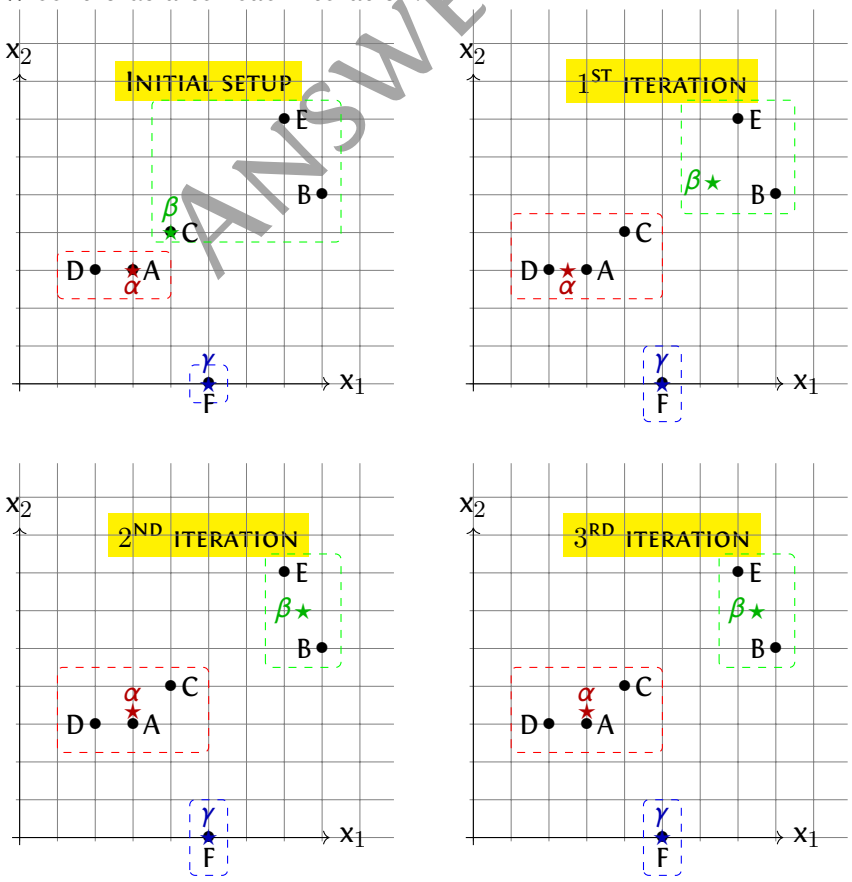
3RD ITERATION

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



AY: 2022-2023	**Full Name:** .....
M1-S2: Dept. of Electrical Engineering	**ID:** .....
Resit | AI-ECUE221	**Class:** .....
July 2023	**Room:** .....
Teacher: A. Mhamdi	**Time Limit:** 1½ h

Answer Sheet

Task N°2 ⌛ 50mn | (11 points)

(a) (½ point) If there is no trend between two variables x and y, we say that there is a "_____" connection between x and y.

○ linear ○ exponential √ random ○ non-random

(b) (½ point) What is the best definition for bias in your data model?

○ Bias is when your predicted values are scattered.

○ Bias is when your data is wrong for different reasons.

○ Bias is when your values are always off by the same percentage.

√ Bias is the gap between your predicted value and the outcome.

(c) (½ point) The data in your model has low bias and low variance. How would you expect the data points to be grouped together on the diagram?

√ They would be grouped tightly together in the predicted outcome.

○ They would be grouped tightly together but far from the predicted.

○ They would be scattered around the predict outcome.

○ They would be scattered far away from the predicted outcome.

(d) (½ point) You are working on a project that involves clustering together images of different dogs. You take image and identify it as your centroid image. What type of machine learning algorithm are you using?

○ Centroid reinforcement

○ k-nearest neighbors

○ Binary classification

√ K-means clustering

(e) (½ point) What is ensemble modeling?

○ When you create an ensemble of your training and test data set.

○ When you create an ensemble of different servers to run the algorithms.

DO NOT WRITE ANYTHING HERE

✂----------

√ When you use several ensembles of machine learning algorithms.

○ When you find the one best algorithm for your ensemble.

(f) (½ point) The dataset you have scraped seems to exhibit lots of missing values. What action will help you minimizing that problem?

○ Wise fill-in of controlled random values.

○ Replace missing values with averaging across all samples.

○ Remove defective samples.

√ Imputation.

(g) (½ point) Which of the following methods can be used either as an unsupervised learning or as a dimensionality reduction technique?

○ SVM

√ PCA

○ LDA

○ TSNE

(h) (½ point) The error function most suited for gradient descent using logistic regression is

○ The entropy function.

○ The squared error.

√ The cross-entropy function.

(i) (½ point) Someone on your data science team recommends that you use decision trees, naive Bayes and k-nearest neighbors, all at the same time, on the same training data, and then average the results. What is this an example of?

○ Regression analysis

○ Unsupervised learning

○ High -variance modeling

√ Ensemble modeling

(j) (½ point) You are using k-nearest neighbors and you have a k of 1. What are you likely to see when you train the model?

○ Low bias & low variance

√ Low bias & high variance

✂- -

- ○ High bias & low variance
- ○ High bias & high variance

(k) (½ point) "______________" refers to a model that can neither model the training data nor generalize to unseen data.

- √ Underfitting
- ○ Good fitting
- ○ Overfitting

(l) (½ point) You created a machine learning system that interacts with its environment and responds to errors and rewards. What type of machine learning system is it?

- ○ Supervised learning
- ○ Semi-supervised learning
- √ Reinforcement learning
- ○ Unsupervised learning

(m) (½ point) You work for a website that helps match people up for lunch dates. The website boasts that it uses more than 500 predictors to find customers the perfect date, but many costumers complain that they get very few matches. What is a likely problem with your model?

- ○ Your training set is too large.
- ○ You are underfitting the model to the data.
- √ You are overfitting the model to the data.
- ○ Your machine is creating inaccurate clusters.

(n) (½ point) What is the difference between unstructured and structured data?

- ○ Unstructured data is always text.
- ○ Unstructured data is much easier to store.
- √ Structured data has clearly defined data types.
- ○ Structured data is much more popular.

(o) (½ point) Your data science team is often criticized for creating reports that are boring or too obvious. What could you do to help improve the team?

- √ Suggest that the team is probably underfitting the model to the data.

DO NOT WRITE ANYTHING HERE

✂----------

- ◯ Suggest that unsupervised learning will lead to more interesting results.
- ◯ Make sure that they are picking the correct machine learning algorithms.
- ◯ Encourage the team to ask more interesting questions.

(p) (½ point) What syntax do you use to import 'DataFrames' package to your Julia session after installing it?

- ◯ `Pkg.add("DataFrames")`
- ◯ `add DataFrames`
- ◯ `use DataFrames`
- √ `using DataFrames`

(q) (1 point) Write the line of code that will import a CSV file named '`data.csv`' as a **Julia** `DataFrame`, called `df`.

```
df = CSV.read("data.csv", DataFrame)
```

(r) (1 point) Suppose you have a function `f`, defined as follows:

```
function f(x)
    return x+2
    2x
end
```

What is the value of `f(1)`? Justify.

`f(1) = 3`. The '`return`' keyword causes the function `f` to exit once `x+2` is computed.

(s) (1 point) What would be the theoretical mean of the following random values?

```
randn(12345)
```

The `randn` function returns random values from the standard normal distribution. Hence, the mean of 12345 random values would be 0.