AY: 2023-2024
MIDTERM | **Machine Learning**
Nov. 2023

L3-S5: Dept. of Electrical Engineering
Teacher: A. Mhamdi
Time Limit: 1h

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 **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)*.

❺ **Task №3:** Each correct answer will grant a mark with no negative scoring.

SELF-REVIEW

Task	1	2	3	**Total**
Points	4	4	12	20
Score				

Task №1

⌛ 20mn | (4 points)

Given the code shown hereafter:

[1]:
```
import numpy as np
```

[2]:
```
np.set_printoptions(precision=3)
```

[3]:
```
X = np.random.randint(40, 80, (7, 3))
X
```

[3]:
```
array([[60, 49, 50],
       [61, 73, 74],
       [45, 44, 75],
       [70, 73, 51],
       [76, 75, 74],
       [66, 52, 63],
```

```
       [50, 45, 47]])
```

[4]:
```
X.mean(axis=0)
```

[4]:
```
array([61.143, 58.714, 62.   ])
```

[5]:
```
X.std(axis=0)
```

[5]:
```
array([10.063, 13.188, 11.637])
```

[6]:
```
from sklearn.preprocessing import StandardScaler
```

[7]:
```
sc = StandardScaler()
```

[8]:
```
X = sc.fit_transform(X)
X[1:-3, -2]
```

[9]:
```
y = ['Tunis', 'Kasserine', 'Bizerte', 'Ariana', 'Gafsa', 'Sfax',␣
↪'Bizerte']
```

[10]:
```
from sklearn.preprocessing import LabelEncoder
```

[11]:
```
le = LabelEncoder()
```

[12]:
```
y = le.fit_transform(y)
y
```

(a) (2 points) What will be the output of cell #8?

[8]:
```
array([ 1.083, -1.116,  1.083])
```

(b) (2 points) What will be the output of cell #12?

[12]:
```
array([5, 3, 1, 0, 2, 4, 1])
```

Task N°2

⏳ 15mn | (4 points)

Consider a binary classification problem where a model is trained to classify emails as spam or non-spam. The model is evaluated on a test dataset consisting of 200 emails, and the confusion matrix is as follows:

	Predicted Class	
	Spam	Non-Spam
Actual Class		
Spam	120	10
Non-Spam	20	50

We are considering spam as the positive case. This means that we are interested in how well the model identifies spam emails correctly.

Perform the following calculations using the provided confusion matrix:

(a) (1½ points) Precision

Precision is the ratio of true positives to the predicted positives.

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ &= \frac{120}{120 + 20} \\ &= 0.857 \end{aligned}$$

(b) (1½ points) Recall

Recall is the ratio of true positives to the actual positives.

$$\begin{aligned} \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ &= \frac{120}{120 + 10} \\ &= 0.923 \end{aligned}$$

(c) (1 point) F1-score

The f1-score is the harmonic mean of precision and recall.

$$\begin{aligned} \text{f1-score} &= \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \\ &= \frac{2}{\frac{1}{0.857} + \frac{1}{0.923}} \\ &= 0.889 \end{aligned}$$

AY: 2023-2024	**Full Name:** .....
L3-S5: Dept. of Electrical Engineering	**ID:** .....
Midterm | Machine Learning	**Class:** .....
Nov. 2023	**Room:** .....
Teacher: A. Mhamdi	**Time Limit:** **1h**

Answer Sheet

Task N°3 ⌛ 25mn | (12 points)

(a) (1 point) Predicting whether a customer responds to a particular advertising campaign or not is an example of what?.

√ Classification problem ○ Regression ○ None of the above

(b) (1 point) Multiple linear regression is appropriate for:

○ Predicting the sales amount based on month.

○ Predicting whether a drug is effective for a patient based on her characteristics.

√ Predicting tomorrow's rainfall amount based on the wind speed and temperature.

(c) (1 point) The key difference between simple and multiple regression is:

○ Multiple linear regression introduces polynomial features

√ To estimate a single dependent variable, simple regression uses one independent variable whereas multiple regression uses multiple.

○ Simple regression assumes a linear relationship between variables, whereas this assumption is not necessary for multiple regression.

○ Simple linear regression compresses multidimensional space into one dimension.

(d) (1 point) In a dataset, what do the rows represent?

○ Dependent variable

○ Independent variables

√ Observations

○ Features

(e) (1 point) In a machine learning model, the input variables have to be independent.

√ True ○ False

DO NOT WRITE ANYTHING HERE

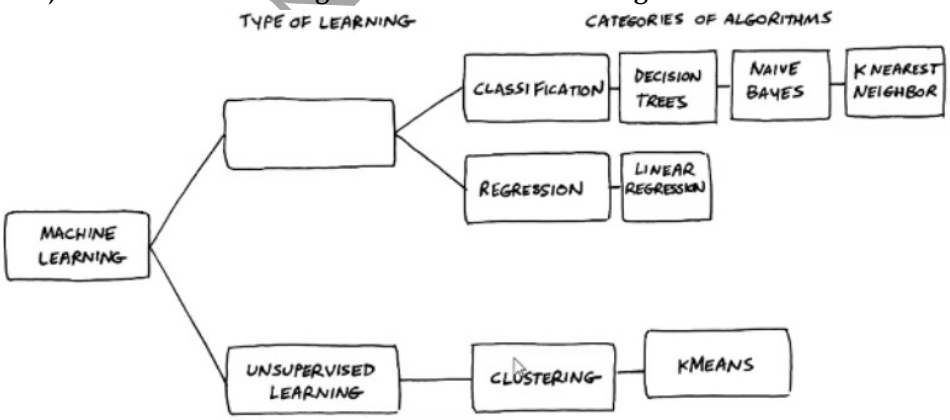
(f) (1 point) Which one is a sample application of regression?

- ○ Predicting whether a patient has cancer or not.
- ○ Grouping of similar houses in an area.
- √ Forecasting rainfall amount for next day.
- ○ Predicting if a team will win or not.

(g) (1 point) What is the correct order for using a model?

- ○ Clean the data, fit the model on the entire dataset, split the data into training and test sets, evaluate model accuracy.
- √ Clean the data, split the data into training and test sets, fit the model on the training set, evaluate model accuracy.
- ○ Split the data into training and test sets, fit the model on the train set, evaluate model accuracy.
- ○ Split the data into training and test sets, fit the model on the train set, clean the data, evaluate model accuracy.

(h) (1 point) What is the missing information in this diagram?



- ○ Training Set
- ○ Unsupervised Data
- √ Supervised Learning
- ○ Binary Classification

(i) (1 point) In traditional computer programming, you input commands. What do you input with machine learning?

○ patterns ○ programs ○ rules √ data

DO NOT WRITE ANYTHING HERE

(j) (1 point) How do machine learning algorithms make more precise predictions?

- ○ The algorithms are typically run more powerful servers.
- √ The algorithms are better at seeing patterns in the data.
- ○ Machine learning servers can host larger databases.
- ○ The algorithms can run on unstructured data.

(k) (1 point) For linear regression, the model is

$$h_{\theta,\, b}(x) = x^T\theta + b$$

Which of the following are the inputs, or features, that are fed into the model and with which the model is expected to make a prediction?

○ θ and b ○ m ○ (x, y) √ x

The input features x are fed into the model to generate a prediction $h_{\theta,\, b}(x)$.

(l) (1 point) For linear regression, if you find parameters θ and b so that $\mathcal{J}(\theta,\, b)$ is very close to zero, what can you conclude?

- ○ This is never possible – there must be a bug in the code.
- ○ The selected values of the parameters θ and b cause the algorithm to fit the training set really poorly.
- √ The selected values of the parameters θ and b cause the algorithm to fit the training set really well.

The model fits the training set well when the cost $\mathcal{J}$ is small.