Name and surname: ____
Matriculation number: ____

ML4N Exam sample 2024

Rules

- The exam lasts 90 minutes.
- You must bring an ID with a photo and your PoliTO card. Keep them on the desk. The professor will check them during the exam.
- You must bring your own writing papers and tools to write on it.
- You can bring a hand-written single-sided paper page of notes (A4 format size). It is not mandatory.
- You can bring an electronic calculator. It is not mandatory: if there is a computation that you cannot perform on paper, simply substitute the scalar result of the computation with a letter not used elsewhere in the same exercise (e.g.,: $(12*29)^{0.5}$= a).
- Any other electronic device is NOT allowed. No computers, no smartwatches (simple watches are ok), no tablets, no telephones, no earphones, no smart glasses, etc.
- You can bring them in the classroom only if they are completely turned off and sealed in a closed bag, and keep the bag at least 0.5 m away from your seat during the exam.
- Always write your complete name and matriculation number (matricola) on top of each page of the exam you are submitting for evaluation.
- When you finish the exam raise your hand and the professor will collect the exam. If you do not want your exam to be evaluated, tell it to the professor in the classroom. Then, exit immediately from the classroom without talking to anybody. After 90 minutes, all exams will be collected, even if you did not finish.
- Any form of cheating is not tolerated. Any violation of the rules will imply the annulment of your exam and potentially a disciplinary sanction.

Exercise 1 (3 pts)

Which of the following statements are true? Write **True** or **False** alongside each sentence. The number of **True answers ranges from 0 to 5**.

1) A confusion matrix for a binary classification problem is a 2x2 matrix.
2) For a binary classification problem: True Positive + True Negative = False Positive + False Negative.
3) The metric F-measure (also called F-Score) can assume the value 0.
4) For supervised learning, a hypothesis with a small loss on the training set and a large loss on the validation set might indicate overfitting.
5) The metric Precision can be defined for each class of the classification problem.

Exercise 2 (3 pts)

Which of the following statements are true? Write **True** or **False** alongside each sentence. The number of **True answers ranges from 0 to 5**.

1) Outliers of a sample distribution might represent errors in measurements.
2) The mean of a Gaussian distribution might be an outlier of that distribution.
3) Outliers of a sample distribution can be detected through a boxplot.
4) Outliers of a sample distribution can be detected through the GESD test.
5) Outliers of a sample distribution can be detected through the k-Means clustering algorithm.

Name and surname: ____
Matriculation number: ____

Exercise 3 (3 pts)

Which of the following statements are true? Write **True** or **False** alongside each sentence. The number of **True answers ranges from 0 to 5**.

1) k-Means final clusters depend on the initial condition of the algorithm (means or assignments).
2) Gaussian Mixture Model is a soft clustering technique.
3) The silhouette metric is not defined for DBSCAN clustering algorithm.
4) k-Means uses gradient descent to obtain the clusters.
5) k-Means is a technique to find the average of k different samples.

Exercise 4 (2 pts)

Consider the following one-dimensional empirical samples from the same one-dimensional distribution:

[11, 2, 5, 6, 7, -1.5, 5.5]

- Compute the sample mean of the empirical distribution.
- Compute the sample variance of the empirical distribution.

Exercise 5 (2 pts)

Consider the following two samples:
a:[0, 2]
b:[2, 1]

- Compute Euclidean distance $d_E(a,b)$ between the two samples.
- Compute Minkovski distance $d_M(a,b)$ with parameter r=1 between the two samples.

Exercise 6 (2 pts)

Make two examples of non-linear parametrized hypothesis space for 3 dimensions (number of features n=3).

$h_1 : \mathbb{R}^3 \to \mathbb{R}$

$h_2 : \mathbb{R}^3 \to \mathbb{R}$

Write down the two example parametrized models $h_1(\mathbf{x})$ and $h_2(\mathbf{x})$ and make the range for each parameter explicit.

Name and surname: ____
Matriculation number: ____

Exercise 7 (4 pts)

Consider the following 3 samples (training data):

a: [1, -1] with numerical label 1
b: [0, 1] with numerical label -8
c: [-1, 2] with numerical label 4

Consider the following hypothesis: $h_1(x)=x_1+2x_2-10$ and $h_2(x)=x_2^2$

Evaluate the total empirical risk with hypothesis h_1 and h_2 with

- absolute error loss
- 0/1 loss

Which of the two hypotheses is better for ERM when considering these losses?

Exercise 8 (3 pts)

Train data

Feature 1	Feature 2	Feature 3	Label
0	1	0	-1
-1	0	1	1
4	1	1	-1
0	0.5	0	1

Test data

Feature 1	Feature 2	Feature 3	Label
-1	1	1	1
3	1	0	-1

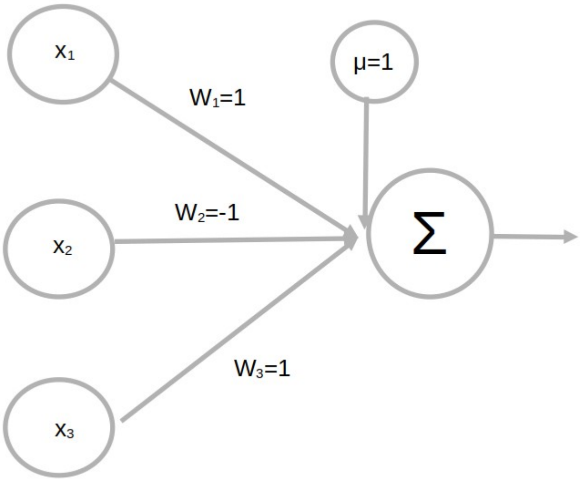
Given training data, what would predict a 3-Nearest Neighbor (3-NN) classifier on test data, if using Manhattan distance (Minkovski distance with parameter r=1) and majority voting?

What is the accuracy on test data of the 3-NN?

Name and surname: ____
Matriculation number: ____

Exercise 9 (4 pts)

The following neural network processes a 3 dimensional input and provides a single output. There is a single neuron and no activation function.



Consider the sample

x=[1, 2, 0]

with label y=0

Compute the output of the neural network h(**x**) for sample **x.**

Then, compute the squared error loss for sample **x**.

Finally, do a backward pass to update the weight w_1 (only this single weight) with a single step of the standard stochastic gradient descent, with acstep size (learning rate) of 0.1.

$$\underbrace{\mathbf{w}^{(k+1)}}_{\text{new guess}} = \underbrace{\mathbf{w}^{(k)}}_{\text{current guess}} - \underbrace{\alpha^{(k)}}_{\text{step size}} \mathbf{g}^{(k)} \text{ with } \mathbf{g}^{(k)} \approx \nabla f\left(\mathbf{w}^{(k)}\right)$$

What is the new value of w_1?

Name and surname: ____
Matriculation number: ____

Exercise 10 (3 pts)

Consider the following snippet of Python code:

```
import numpy as np
a = np.linspace(0, 50, 100)
b = np.ones((75,2))
c = np.random.random((25,2))
print(a.shape, b.shape)
```

What is the output of the code? There is only **one correct answer**. Write **True** alongside the correct sentence:

1) (100, 2), (75, 2)
2) (100, 2), (25, 2)
3) (100,), (75, 2)
4) (100,), (25, 2)

The code continues as follows:

```
d = np.concatenate([b,c], axis=0) * a.reshape(-1, 1)
e = d.reshape(25, -1, 2)
print(e.shape)
```

What is the output of this second part of the code? There is only **one correct answer**. Write **True** alongside the correct sentence:

1) (25, -1, 2)
2) (25, 4, 2)
3) (25, 100, 2)
4) (75, 2)
5) (25, 2)
6) An error occurs
7) None of the other answers is correct

Name and surname: ____
Matriculation number: ____

Exercise 11 (2 pts)

Consider the following Pandas DataFrame (df):

| Age | Gender | Occupation |

0 | 23 | M | student |
1 | 27 | M | consultant |
2 | 34 | F | manager |
3 | 44 | F | manager |
4 | 54 | M | manager |
5 | 29 | M | freelancer |
6 | 21 | F | student |

What is the output of the following operation?

```
import pandas as pd
df = pd.DataFrame({
   "Age": [23, 27, 34, 44, 54, 29, 21],
   "Gender": ["M", "M", "F", "F", "M", "M", "F"],
   "Occupation": ["student", "consultant", "manager", "manager", "manager", "freelancer", "student"]
})

print(df.loc[df["Gender"]=="F"].
     groupby("Occupation")['Age'].
     mean())
```

- If an error occurs, write "an error occurs at [LINE NUMBER]".
- If the answer is a DataFrame, use | to separate columns and newline to separate rows as above. Remember that the first column represents the index.
- If the answer is a Series, use newline to separate rows and | to separate index and values. For example, for `pd.Series([1,2,3], index=['a','b','c'])`, write:
  a | 1
  b | 2
  c | 3