Name and surname: ____________________________________________
Matriculation number: __________________________________________

# ML4N Exam – January 30th, 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 (2 pts)

What are the main three components of ML? State them and briefly describe them in max 2 lines each.

Name and surname: ___________________________________________
Matriculation number: __________________________________________

# 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) If the loss function is differentiable, then also the empirical risk is differentiable.
2) In supervised learning, gradient descent on the parameters is the only method to find the best hypothesis h in H.
3) The learning rate is a hyper-parameter of an ML model H.
4) In gradient-descent methods, the empirical risk function of the trainable parameters is locally approximated.
5) The 0/1 loss function is continuous in all its domain.

# 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) Naive Bayes supervised method uses the Hinge loss.
2) The number of trees is a hyper-parameter of decision tree models.
3) The logistic regressor is a regressor model and technique for supervised learning.
4) The SVM technique uses as model the space of linear maps.
5) Random forests can be used for regression problems and for classification problems.

# Exercise 4 (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) Dimensionality reduction creates a compressed representation of data.
**2)** After applying PCA, the new data with the new features will have a diagonalized covariance matrix.
3) t-SNE stands for total-Sample Normalized Estimation.
4) You cannot perform standardization before applying PCA to the dataset.
5) In PCA, a principal component can be defined as a non-linear combination of observed variables.

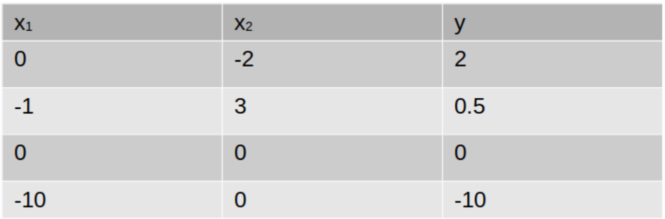
Name and surname: ______________________________________________
Matriculation number: ____________________________________________

# Exercise 5 (5 pts)

Consider the following regression model:

$h(\mathbf{x})=w_1x_1+w_2x_2^2$

And the following training data (with features $x_1$ and $x_2$ and label y):



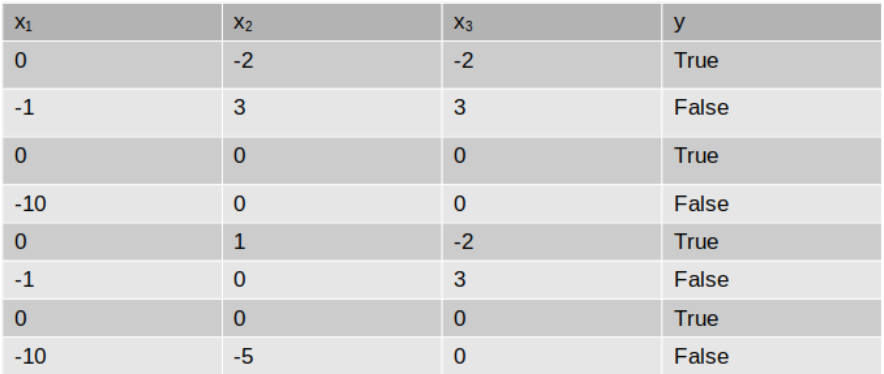
| $x_1$ | $x_2$ | y |
|---|---|---|
| 0 | -2 | 2 |
| -1 | 3 | 0.5 |
| 0 | 0 | 0 |
| -10 | 0 | -10 |

- Write the empirical risk minimization problem for the training data, using absolute error loss. This should be a function of only $w_1$ and $w_2$.
- Now assume that $w_1$ and $w_2$ are given, $w_1=1$, and $w_2=1$.
  Compute the updated value of weight $w_2$ by applying a single step of gradient descent algorithm on the whole training data. Use a learning rate equal to 0.25
  **Hint**: Since $w_1=1$, you can rewrite the ERM as a function of only $w_2$. You need to compute the partial derivative only with respect to $w_2$, and evaluate it only on $w_2=1$. There is no need to write the general formulation of the full gradient with respect to all possible values of $w_1$ and $w_2$.

Name and surname: ____________________________________________
Matriculation number: ___________________________________________

# Exercise 6 (3 pts)

Consider the following data:



| $x_1$ | $x_2$ | $x_3$ | y |
|---|---|---|---|
| 0 | -2 | -2 | True |
| -1 | 3 | 3 | False |
| 0 | 0 | 0 | True |
| -10 | 0 | 0 | False |
| 0 | 1 | -2 | True |
| -1 | 0 | 3 | False |
| 0 | 0 | 0 | True |
| -10 | -5 | 0 | False |

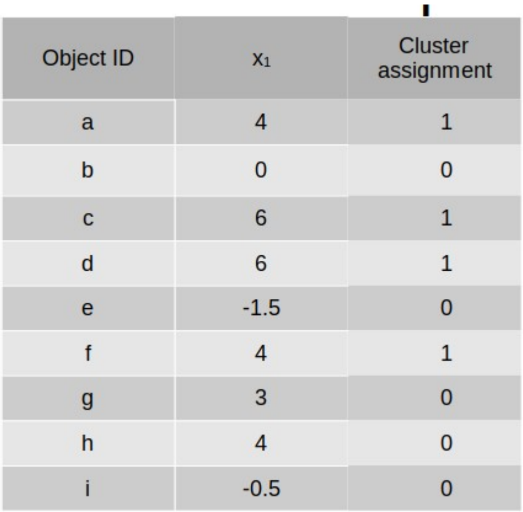
- Split all data into training, validation and test set. Choose the proportion as you wish and state it.
- Split all data according to a 4-fold cross validation. State the content of the different training and validation sets.

**Hint:** To simplify your work, you can assign identifiers/names to samples (e.g., a, b, ...)

Name and surname: ____________________________________________
Matriculation number: __________________________________________

# Exercise 7 (3 pts)

You are given the following dataset with 1 feature and its corresponding cluster assignment after a hard-clustering algorithm is run. As distance measure, consider Euclidean distance (absolute value of the difference, in 1D).



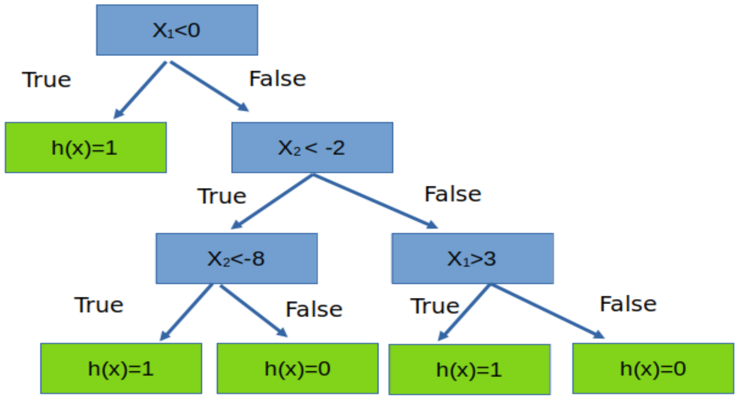
| Object ID | $x_1$ | Cluster assignment |
|---|---|---|
| a | 4 | 1 |
| b | 0 | 0 |
| c | 6 | 1 |
| d | 6 | 1 |
| e | -1.5 | 0 |
| f | 4 | 1 |
| g | 3 | 0 |
| h | 4 | 0 |
| i | -0.5 | 0 |

Compute the silhouette score for point a.
Compute the silhouette score for point c.
Which of the two points (a or c) has a better Silhouette metric?

Name and surname: ______________________________________________
Matriculation number: ____________________________________________

# Exercise 8 (5 pts)

Consider the following flowchart schema of a decision tree hypothesis h for classification trained on some training data.



Consider now the following validation and test data

Validation data

| $x_1$ | $x_2$ | y |
|---|---|---|
| 1 | 1 | 0 |
| 3 | -9 | 1 |
| 4 | -1 | 1 |
| -1 | 10 | 1 |

Test data

| $x_1$ | $x_2$ | y |
|---|---|---|
| 33 | 3 | 1 |
| 0 | 1 | 0 |
| -3 | 3 | 1 |
| 3 | -3 | 0 |

- Draw the decision boundary of the hypothesis h on a two dimensional Cartesian plane. Mark each defined area with the corresponding output of the hypothesis.
- What is the output of the decision tree h(**x**) on the validation data?
- What is the accuracy on test data?

Name and surname: ____________________________________________
Matriculation number: _________________________________________

# Exercise 9 (3 pts)

You are given a dataset X with target classes y. Both X and y are represented as NumPy arrays. The contents of X and y are unknown, you only know the following:

- X.ndim == 2
- y.ndim == 1
- X.shape[0] == y.shape[0]
- X.shape[1] > 10
- X and y do not contain any missing values
- y contains integers from 0 to 9, representing one of 10 classes

Consider the following snippet of Python code:

```
import numpy as np
X = ...
y = ...

W = np.random.random((X.shape[1], 10))
b = np.random.random(10)
print((W+b).shape)
```

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

1) (10, 10)
2) (10, X.shape[1])
3) (10, X.shape[0])
4) (X.shape[1], 10)

The code continues as follows:

```
y_pred = np.argmax(X * W + b, axis=1)
accuracy = np.mean(y_pred == y)
print(accuracy)
```

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

1) ~0.1
2) ~0.5
3) ~0.0
4) ~1.0
5) An error occurs
6) None of the other answers is correct

Name and surname: ___________________________________________
Matriculation number: _________________________________________

# Exercise 10 (2 pts)

You are given the following snippet of Python code:

```
import numpy as np
a = list(range(0,4,2))
b = np.ones(len(a), dtype=int)
c = (1 + b) - a

M = pd.DataFrame([
[4, 2, 0, 3, 3],
[4, 1, 4, 0, 0],
[0, 2, 3, 4, 1],
[4, 3, 0, 2, 2]
], columns=range(0,5))
X = M.loc[a,c].values
print(X)
```

1) What is the content of variable X?
2) What is the type of X?

- 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 np.ndarray, separate each row with a new line, and each column with a space. For example, for `np.array([[1,2,3,4],[5,6,7,8],[9,10,11,12]])`, write:
  1 2 3 4
  5 6 7 8
  9 10 11 12