Course: Data science lab: process and methods
Quiz: Exam (2025-02-18)

Politecnico di Torino PORTALE ESAMI

[redacted]	[redacted]
Started on	18 February 2025, 11:11 AM
State	Finished
Completed on	18 February 2025, 12:41 PM
Time taken	1 hour 30 mins
Grade	14,01 out of 20,00 (70%)
Summary of attempt	1 🟥 2 ○ 3 ○ 4 🟥 5 ✓ 6 🟥 7 ▨ 8 ✓ 9 ✓ 10 ○ 11 ✓ 12 ✓ 13 ✓

Question 1

Incorrect

Mark 0,00 out of 1,50

1.5 points (no penalty for a wrong answer)

You are addressing a regression problem. The training set contains 500 1-dimensional points (x) and corresponding targets (y).

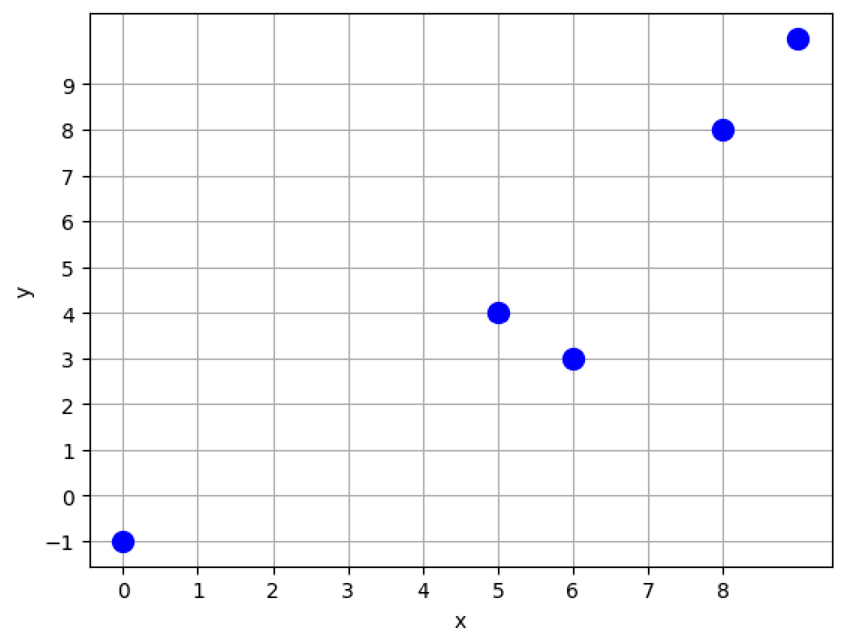
You build the following model:

$$\hat{y} = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

Where $w = [w_0, w_1, w_2, w_3]$ are the weights of the model and $\hat{y}$ is the predicted value. More specifically, the following are the learned weights:

- $w_0 = -0.06$
- $w_1 = 0.13$
- $w_2 = 0.03$
- $w_3 = 0.01$

The following plot shows the 5 test points, as (x, y) pairs.



What is the Mean Squared Error (MSE) of the model on the test set?

Write the answer in the box below. Use at least 4 significant figures.

MSE = 0.94452 ✗

Question 2

Partially correct

Mark 1,33 out of 2,00

2 points (no penalty for a wrong answer)

A decision tree classifier T can be used to compute the probability that a point belongs to a class c (i.e., P(class=c|x, T)) by computing the fraction of training points that belong to class c among all the points that reach the same leaf node reached by x.

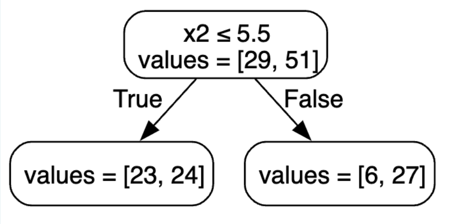
An ensemble of n classifiers $T_1, T_2, \ldots, T_n$ (e.g. a random forest) may use a soft voting policy to compute the probability that a point x belongs to a class c as follows:

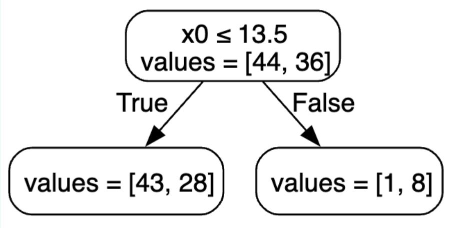
$$P(class = c|x, T_1, T_2, \ldots, T_n) = \frac{1}{n} \sum_{i=1}^{n} P(class = c|x, T_i)$$

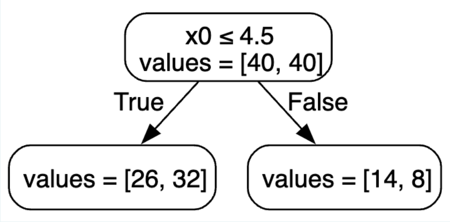
where $P(class = c|x, T_i)$ is the probability that x belongs to class c according to the classifier T_i.

A random forest comprised of 3 decision trees is trained to solve a binary classification problem. The model is trained on a training set and later tested on a separate test set.

The following are the 3 decision trees learned during the training process.





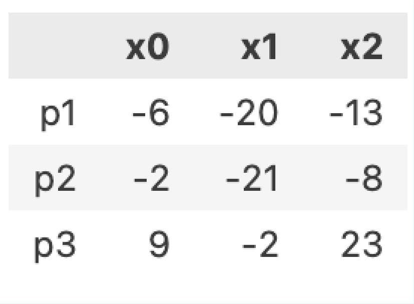


The non-leaf nodes contain, in the first row, the splitting decision, in the form "attribute $\leq$ threshold".

All nodes contain the "values = [value_0, value_1]" row , which represents the number of negative (value_0) and positive (value_1) values. For example, the node "values = [5, 32]" means that 37 points reach the specific node: 5 of them belong to the negative class (class 0), 32 of them belong to the positive class (class 1).

Given an input point p, the random forest uses a soft voting policy to compute P(class=1|p) and thus decide the class label to be predicted.

You are given the following 3 test set points.



	x0	x1	x2
p1	-6	-20	-13
p2	-2	-21	-8
p3	9	-2	23

What are the corresponding values of P(class=1|p)?

Write the answer in the box below. **Report at least 4 significant figures.**

P(class=1|p1) = 0.4856 ✓

P(class=1|p2) = 0.4719 ✗

P(class=1|p3) = 0.5254 ✓

Question 3

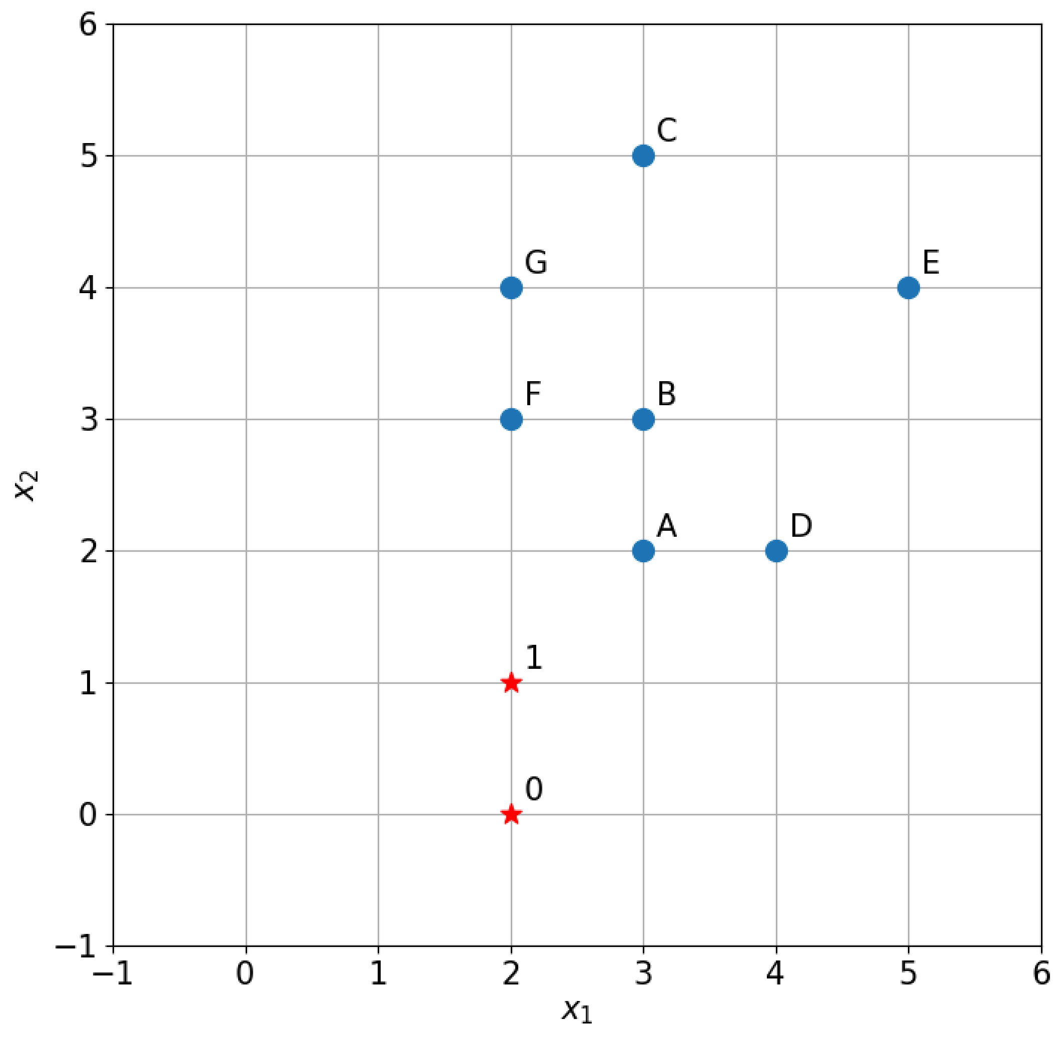
Incorrect

Mark 1,13 out of 1,50

1.5 points (no penalty for a wrong answer)

You are given a dataset containing 7 points in 2 dimensions (x1, x2). Each point is labelled A through G.

You apply K-means clustering with K = 2. The figure below represents the 7 points (blue dots) and the initializations for the two centroids (red stars). Each centroid is labelled with a number (0, 1).



What are the new centroids computed after 1 iteration of the K-means algorithm?

Use the Euclidean distance when computing any distance.

Write your answer in the boxes below. Write one coordinate per box. **Use at least 4 significant figures.**

Centroid 0.

x_1: 0.0000 ✗

x_2: 0.0000 ✓

Centroid 1.

x_1: 3.1429 ✓

x_2: 3.2857 ✓

Question **4**

Incorrect

Mark -0,23 out of 1,50

1.5 points (-15% penalty for a wrong answer)

Which of the following measures can be used to quantify the impurity of a node, for a **decision tree regressor**?

○ a. Variance

○ b. Precision

○ c. None of the other options is correct

○ d. Mean

○ e. Information gain

○ f. Silhouette

○ g. Accuracy

○ h. Entropy

● i. Gini index ✗

Your answer is incorrect.

The correct answer is:
Variance

Question 5

Correct

Mark 2,00 out of 2,00

2 points (no penalty for a wrong answer)

Given a set of documents $D = \{d_1, d_2, ...\}$ and a set of terms $T = \{t_1, t_2, ...\}$, the tf-idf for a term $t \in T$ in a document $d \in D$ is defined as:

$$tf - idf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

A possible definition for $tf(t, d)$ is the raw count of occurrences of t within d.

A possible definition of $idf(t, D)$ is the following:

$$idf(t, D) = log(\frac{|D|}{|\{d \in D : t \in d\}|})$$

You are given the following collection of documents:

<term 1> <term 1> <term 1> <term 1>
<term 3> <term 1> <term 1> <term 1>
<term 2> <term 3> <term 0> <term 0> <term 3>

Where each line represents a document and each term is represented in the form <term X>.

Compute the tf-idf matrix for all terms/documents.

Report the result in the box below, listing one document for each line and one term for each column.

For the documents, use the same order in which the documents appear in the above list. For the terms, use their natural order (<term 0>, <term 1>, etc).

You must use the natural logarithm (log-e, or ln) for your computations.

	<term 0>	**<term 1>**	**<term 2>**	**<term 3>**
doc 0	0.0000 ✓	1.6219 ✓	0.0000 ✓	0.0000 ✓
doc 1	0.0000 ✓	1.2164 ✓	0.0000 ✓	0.4055 ✓
doc 2	2.1972 ✓	0.0000 ✓	1.0986 ✓	0.8109 ✓

Question 6

Incorrect

Mark 0,00 out of 1,50

1.5 points (no penalty for a wrong answer)

You are given the following code snippet. What is the output produced by line 21?

```
 1  import numpy as np
 2
 3  a = np.array([
 4      [1, 5, 0, 3],
 5      [1, 4, 0, 8]
 6  ])
 7
 8  b = np.argsort(a, axis=1)
 9  c = b.sum(axis=0)
10
11  d = np.array([
12      [1, 2],
13      [3, 4],
14      [5, 6],
15      [7, 8],
16      [9, 10]
17  ])
18
19  e = d[c]
20
21  print(e.sum(axis=0))
```

Write the answer in the box below.

If an error occurs, write "an error occurs at line X" (X being the line number where the error occurs).

Answer: an error occurs at line 19 ✖

The correct answer is: 28, 32

Question 7

Not answered

Not graded

This is NOT a question

You can use this edit box for any note you may want to take or any draft of your solutions. If you'd like to leave any notes for the professors, you can write them here. No points will be assigned to this question.

Question 8

Correct

Mark 1,00 out of 1,00

1 point (no penalty for a wrong answer)

The lift of an association rule $A \rightarrow B$ is computed as:

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{P(B)}$$

Where $conf(A \rightarrow B)$ is the confidence of the rule, and is defined as:

$$conf(A \rightarrow B) = \frac{P(A \cap B)}{P(A)}$$

You are given the following transactional database.

acde
acde
bc
abcd
bcde
bd
b
abde
bcde
de

You are also given an association rule $c \rightarrow bde$.

Compute the confidence and the lift of the association rule.

Write the answer in the box below, using the following syntax:

confidence: 0.3333 ✓

lift: 1.1111 ✓

Question 9

Correct

Mark 2,00 out of 2,00

2 points (no penalty for a wrong answer)

- Precision(C) is the fraction of correct predictions among the samples predicted with class C
- Recall(C) is the fraction of correct predictions among the samples with actual class C

You want to use a test set with 500 points to test the performance of a binary classifier.

The precision and the recall for the positive class are the following:

precision(positive) = 0.65
recall(positive) = 0.68

You additionally know that the model produces, on this test set, a total of 221 true positives.

What are the precision and recall for the negative class?

precision(negative) = 0.3500 ✓

recall(negative) = 0.3200 ✓

Question 10

Partially correct

Mark 1,28 out of 1,50

1.5 points (-15% penalty for a wrong answer)

Which of the following best describes a key feature of pandas' pivot_table function? Choose all correct answers.

☑ a. It returns a flattened DataFrame by eliminating hierarchical indexing ✗

□ b. It reshapes the DataFrame from wide to long format by unpivoting selected columns

☑ c. It aggregates duplicate entries using a specified aggregation function ✓

□ d. It sorts the resulting DataFrame by its index by default

□ e. It converts DataFrame indices into columns without altering the data structure

□ f. It drops any rows containing missing values

□ g. It automatically fills in missing values based on the median of the data

□ h. It exclusively handles numerical data, automatically converting non-numerical entries to NaN

Your answer is partially correct.

You have selected too many options.
The correct answer is:
It aggregates duplicate entries using a specified aggregation function

Question **11**

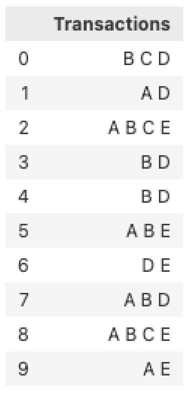
Complete

Mark 2,50 out of 2,50

2.5 points (no penalty for a wrong answer)

A frequent itemset is maximal if it is frequent and none of its supersets is frequent.

Given the transactional dataset below, apply the Apriori algorithm to extract all frequent itemsets.



	Transactions
0	B C D
1	A D
2	A B C E
3	B D
4	B D
5	A B E
6	D E
7	A B D
8	A B C E
9	A E

The value of minsup is 2 (an itemset is frequent if it appears in at least 2 transactions).

An itemset is considered to be frequent if its support count is equal to or higher than the minsup.

Question 1) List all frequent itemsets having length 2, along with their support count.

Question 2) List all itemsets of length 3 that have been generated by Apriori after the join and prune steps, before counting their support in the database.

Question 3) List all maximal frequent itemsets, along with their support count.

Use the following notation:

A1) (... list of itemsets w/ support count ...)
A2) (... list of itemsets ...)
A3) (... list of itemsets w/ support count ...)

For example
A1) (ab: 2, ac: 3, ad: 2)
A2) (abc, abd, abe)
A3) (ab: 2, ad: 2, bce: 3)

```
A1) (ab:4, ac:2, ad:2, ae:4, bc:3, bd:4, be:3, ce:2)
A2) (abc, abd, abe, ace, bce)
A3) (ad:2, bd:4, abce:2)
```

A1) (ab:4, ac:2, ad:2, ae:4, bc: 3, bd: 4, be:3, ce: 2)
A2) (abc, abd, abe, ace, bce)
A3) (ad: 2, bd: 2, abce: 2)

Comment: ok

Question **12**

Correct

Mark 1,50 out of 1,50

1.5 points (-15% penalty for each wrong answer)

Which of the following statements on the curse of dimensionality is correct? Select all correct options.

□ a. The curse of dimensionality is only relevant for low-variance datasets, as high-variance datasets naturally avoid sparsity issues

□ b. Feature selection and dimensionality reduction techniques have no impact on mitigating the curse of dimensionality

☑ c. As the number of dimensions increases, data points become more sparse, making it harder to generalize models ✓

☑ d. Distance metrics become less meaningful in high-dimensional spaces, affecting clustering and nearest neighbor methods ✓

□ e. The curse of dimensionality primarily affects small datasets, while larger datasets (i.e., with a larger number of points) are not impacted

□ f. If a dataset has a large number of points, high-dimensional spaces remain well-populated: as a consequence, machine learning models are not affected by the curse of dimensionality in this case

□ g. In high-dimensional spaces, most points remain well-separated, making classification and clustering more effective

□ h. The curse of dimensionality is only a theoretical issue; in practice, modern machine learning algorithms generally adjust for high-dimensional data

Your answer is correct.

The correct answers are:
As the number of dimensions increases, data points become more sparse, making it harder to generalize models,

Distance metrics become less meaningful in high-dimensional spaces, affecting clustering and nearest neighbor methods

Question **13**

Correct

Mark 1,50 out of 1,50

1.5 points (-15% penalty for a wrong answer)

For a data point x assigned to a cluster C, we define:

$$a(x) = \frac{1}{|C|-1} \sum_{\substack{y \in C \\ y \neq x}} d(x,y)$$

$$b(x) = \min_{C' \neq C} \left(\frac{1}{|C'|} \sum_{y \in C'} d(x,y) \right)$$

The silhouette coefficient for the point x is then defined as:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

Why is the $b(x)$ term defined using the min function?

○ a. None of the other options is correct

○ b. To maximize the separation between clusters, thus increasing clustering stability

○ c. To ensure that all clusters contribute equally to the silhouette score

○ d. To assign more weight to clusters with the largest number of points when computing the silhouette score

● e. To select the nearest neighboring cluster, which best represents the alternative grouping of the point ✓

○ f. To minimize the distance between a point and its assigned cluster

○ g. To smooth out variations in distance across different clusters

○ h. To assign more weight to clusters with the smallest number of points when computing the silhouette score

Your answer is correct.

The correct answer is: To select the nearest neighboring cluster, which best represents the alternative grouping of the point