Data science lab: process and methods

Iniziato	mercoledì, 11 gennaio 2023, 09:29
Stato	Completato
Terminato	mercoledì, 11 gennaio 2023, 10:55
Tempo impiegato	1 ora 25 min.
Valutazione	**0,00** su un massimo di 20,00 (**0**%)

Domanda 1

Risposta non data

Punteggio max.: 1,00

1 point (15% penalty for a wrong answer)

The z-score normalization of a feature $x = (x_1, x_2, \ldots, x_n)$ is given by:

$$x' = \frac{x - \frac{1}{n}\sum_i x_i}{\sqrt{\frac{1}{n}\sum_i (x_i - \frac{1}{n}\sum_j x_j)^2}}$$

You are given a feature f with mean μ and standard deviation σ.

Which of the following statements on the z-score normalization is correct?

- (a) It normalizes all points to the [0, +1] range
- (b) It normalizes all points to the $[-\sigma^2, +\sigma^2]$ range
- (c) It normalizes all points to the $[-\sigma, +\sigma]$ range
- (d) None of the other statements is correct
- (e) It normalizes all points to the [-1, +1] range
- (f) It normalizes all points to the $[0, +\sigma]$ range

Domanda 2

Risposta non data

Punteggio max.: 2,00

2 points (no penalty for a wrong answer)

When matching two binary sequences $v_1 = (v_{11}, v_{12}, \ldots v_{1n})$, $v_2 = (v_{21}, v_{22}, \ldots v_{2n})$, the Jaccard similarity is computed as follows:

$$J(v_1, v_2) = \frac{M_{11}(v_1,v_2)}{M_{11}(v_1,v_2)+M_{01}(v_1,v_2)+M_{10}(v_1,v_2)}$$

Where $M_{ij}(v_1, v_2)$ is the number of instances of $k \leq n$ for which $v_{1k} = i$ and $v_{2k} = j$

You are given the following categorical dataset with features x0, x1, x2.

x0	x1	x2
c	b	x
a	a	z
c	c	y
a	a	y

1. What dataset do you obtain by applying 1-hot encoding to all of its features? Specify the column names in the form “attribute_value”

2. Compute J(x1, x2) and J(x3, x4)

Write the answers in the box below using the following syntax:

```
1.
attribute1_value1 attribute1_value2 ... attributeM_valueN
value_1.1 value_1.2 ... value_1.X
value_2.1 value_2.2 ... value_2.X
...
value_Y.1 value_Y.2 ... value_Y.X
2.
J(x1, x2) = value
J(x3, x4) = value
```

For example

```
1.
v1_a v1_b v2_c v2_d
0 1 0 1
1 0 1 0
0 1 0 1
0 1 0 1
2.
J(x1, x2) = 0.9
J(x3, x4) = 0.8
```

Domanda 3

Risposta non data

Punteggio max.: 2,50

2.5 points (no penalty for a wrong answer)

You are given the following list of transactions.

A C D
C D
A C
C
B D
A B C D
C D E
A B C
A C
D
C D
A B
A B D E

Apply the FP-growth algorithm using minsup = 2 (an itemset is frequent if it appears in at

least two transactions).

Write the following:

1. B-CPB (B - Conditional Pattern Base)
2. B-CHT (B - Conditional Header Table)
3. BD-CPB (BD - Conditional Pattern Base)

Write *all* and *only* the correct itemsets, along with their support counts.

Use the following syntax:

```
B-CPB = { element1: support1, element2: support2, ... }
B-CHT = { element1: support1, element2: support2, ... }
BD-CPB = { element1: support1, element2: support2, ... }
```

For example:

```
B-CPB = { a: 1, abc: 2 }
B-CHT = { a: 3, b: 2, c: 1 }
BD-CPB = { a: 1, b: 2, c: 3 }
```

Domanda 4

Risposta non data

Punteggio max.: 1,00

1 point (15% penalty for a wrong answer)

A classification model can be tuned on P hyperparameters. You identify a subset of p < P hyperparameters that require tuning. For each hyperparameter, you identify n candidate values.

For n = 10, p = 5, P = 20, which of the following statements is correct regarding the number of configurations to be assessed?

- (a) Introducing an additional hyperparameter (with n candidate values) introduces fewer new configurations to be assessed than if introducing an additional value for each hyperparameter
- (b) It is advisable to always choose p = P and the largest possible value for n
- (c) Introducing an additional value for each hyperparameter introduces fewer new configurations to be assessed than if introducing an additional hyperparameter (with n candidate values)
- (d) Introducing an additional candidate value for each hyperparameter does not increase the number of configurations to be assessed
- (e) The number of candidate values n cannot exceed P
- (f) Introducing an additional hyperparameter (with n candidate values) does not increase the number of configurations to be assessed

Domanda 5

Risposta non data

Punteggio max.: 2,00

2 points (no penalty for a wrong answer)

A binary classifier uses the following spherical decision boundary:

$$(x-3)^2+(y-4)^2+(z-1)^2=r^2$$

Points strictly inside the sphere are classified with “Reject”, otherwise they are classified with “Accept”.

Assuming $r^2=15$, report the overall precision and recall for both the classes on the following test set.

x	y	z	Ground truth
2	1	3	Accept
0	2	-1	Accept
-3	1	1	Accept
1	3	-2	Reject
0	0	0	Accept

Use the following notation:

precision(Accept)=value
recall(Accept)=value
precision(Reject)=value
recall(Reject)=value

For example:

precision(Accept)=0.2
recall(Accept)=0.12
precision(Reject)=1.0
recall(Reject)=0.4

Domanda 6

Risposta non data

Punteggio max.: 1,00

1 point (15% penalty for each wrong answer)

The R^2 score is defined as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Where:

- y_i is the i-th target value
- $\hat{y}_i$ is the i-th predicted value
- $\bar{y}$ is the mean over y_i

A linear regression model is trained for a regression problem on a dataset containing the information of the employees of a company. For each employee, their age (in years) and their yearly salary (in euros) are used as input features.

The output feature is a number in the range [0,100], which represents how likely it is for the employee to leave the company.

The following are the coefficients learned for each feature:

- Age: -2.17
- Salary: $1.09 \cdot 10^{-4}$
- Bias (intercept): 117

The performance of the model, as measured in terms of R^2 score on a separate test set, is of 0.37.

Which of the following statements are correct? Select all correct answers (multiple answers may be correct)

Scegli una o più alternative:

- ☐ (a) The R^2 score is low because linear regressions perform poorly when the features have different units of measurement
- ☐ (b) The R^2 score is low because linear regressions perform poorly when the features have different orders of magnitude
- ☐ (c) The coefficients learned would **not** change if the features were normalized with min-max scaling
- ☐ (d) The coefficients learned would change if the features were normalized with min-max scaling
- ☐ (e) The performance of the model would be higher if the features were normalized with min-max scaling
- ☐ (f) The performance of the model would be the same if the features were normalized with min-max scaling
- ☐ (g) The performance of the model would be lower if the features were normalized with min-max scaling

Domanda 7

Risposta non data

Punteggio max.: 1,50

1.5 points (no penalty for a wrong answer)

The coefficients of a linear regression model are found by minimizing the residual sum of squares (RSS) between the predictions $y' = (y'_1, y'_2, \ldots, y'_n)$ and the ground truth $y = (y_1, y_2, \ldots, y_n)$.

$$RSS = \frac{1}{n} \sum_i (y'_i - y_i)^2$$

You are given the following 1-dimensional dataset (feature x) and target variable (y).

x	y
2	-3
3	-4
2	-3
1	-2

You fit this data with a linear regression model that uses no intercept (y' = m * x).

Compute the coefficient *m* learned by the linear regression model.

Write the answer in the box below. Use the following syntax:

m=value

For example:

m=3

Domanda 8

Risposta non data

Punteggio max.: 2,00

2 points (no penalty for a wrong answer)

You are given the following 2-dimensional dataset.

x1	x2
-6.7	-11.4
-3.3	-9.8
9.7	4.1
-4.9	-11.4
-4.1	-7.6
11.2	3.8
-4.2	-9.8
-6.6	-9.8
8.5	3.5
-4.7	-9.1

K-means is trained on this dataset using the following configuration of hyperparameters:

- k = 3
- Max iterations = 20
- Centroids initialization = random
- Number of separate runs = 5

Based on this information, write in the box below:

1. The number of distances computed at each iteration of k-means
2. The maximum number of distances computed for a single run of k-means
3. The maximum number of distances that need to be computed to train the model

Use the following syntax:

1. value1
2. value 2
3. value 3

For example:

1. 3
2. 6
3. 9

Domanda 9

Risposta non data

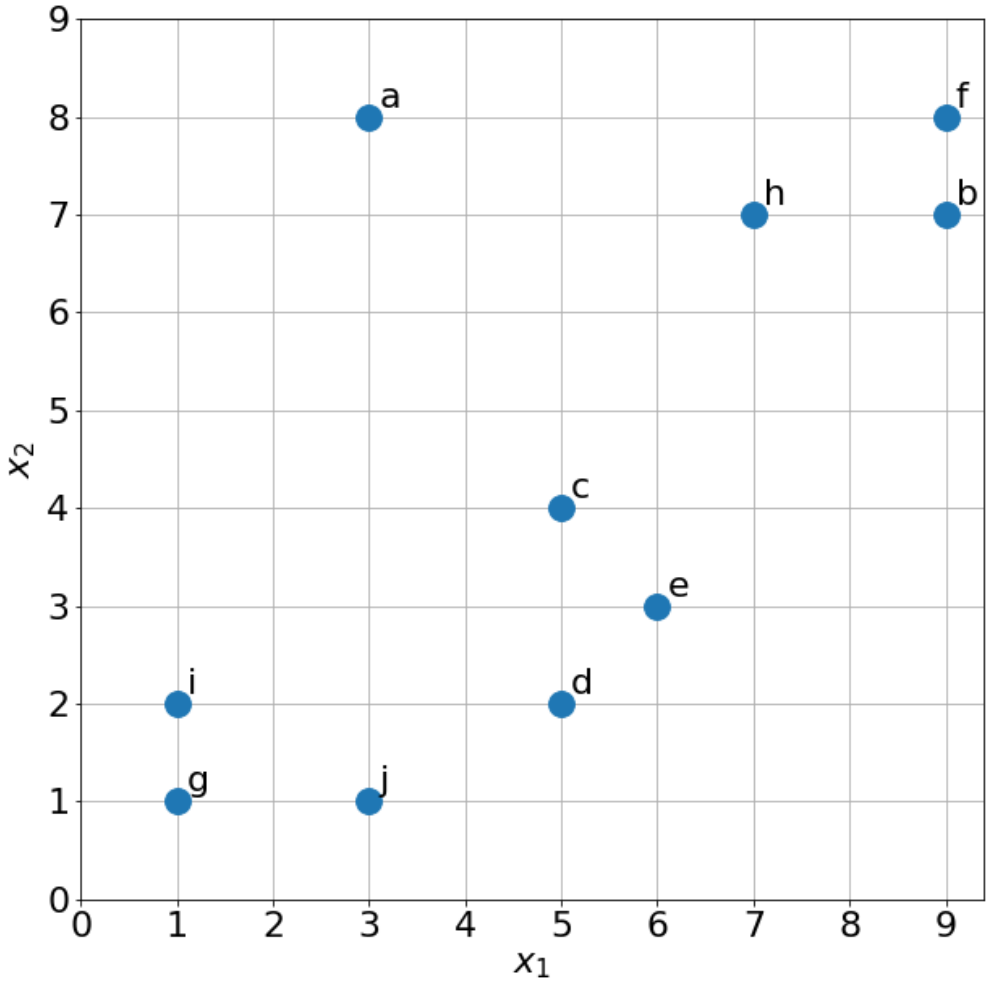
Punteggio max.: 2,00

2 points (no penalty for a wrong answer)

For two n-dimensional points $P_1 = (P_{11}, P_{12}, \ldots P_{1n})$ and $P_2 = (P_{21}, P_{22}, \ldots P_{2n})$ the Manhattan distance is defined as follows:

$dist(P_1, P_2) = \sum_i |P_{1i} - P_{2i}|$

Using the Manhattan distance, apply the DBSCAN algorithm to the following points in the bidimensional space.



Use the following hyperparameters: ϵ = 2.5, minpoints=2 (at least 2 points as neighbors)

For each point write:

- The assigned label (N=noise, B=border, C=core)
- The assigned cluster id (order of cluster ids is not important, use -1 for noise points)

Use the following format: “point label cluster”
Example:

```
a N -1
b C 0
c C 0
d C 1
e B 1
```

Domanda 10

Risposta non data

Punteggio max.: 2,00

2 points (no penalty for a wrong answer)

Consider the following Python code. Write down what it prints to the standard output (write "an error occurs" if, at any point, an error occurs during the execution of the code).

```
import numpy as np

a = np.random.random((32, 224, 224, 3))
b = np.random.random((32, 224, 224, 3))
b = b.mean(axis=0)
print(b.shape)
c = a * b
c = (c - c.mean()) / c.std()
out = c.sum(axis=-1).reshape(32, -1)
print(out.shape)
```

Domanda 11

Risposta non data

Punteggio max.: 2,00

2 points (no penalty for a wrong answer)

The following is the pandas' documentation for the unstack() method.

pandas.DataFrame.unstack

`DataFrame.unstack(level=-1, fill_value=None)` [source]

Pivot a level of the (necessarily hierarchical) index labels.

Returns a DataFrame having a new level of column labels whose inner-most level consists of the pivoted index labels.

If the index is not a MultiIndex, the output will be a Series (the analogue of stack when the columns are not a MultiIndex).

Parameters: **level** : *int, str, or list of these, default -1 (last level)*
Level(s) of index to unstack, can pass level name.

fill_value : *int, str or dict*
Replace NaN with this value if the unstack produces missing values.

Returns: **Series or DataFrame**

You are given the following Python snippet.

```
import pandas as pd

X = [
   [ 4, 2, 4 ],
   [ 4, 2, 4 ],
   [ 3, 2, 3 ],
   [ 1, 3, 2 ],
   [ 3, 3, 3 ],
   [ 3, 2, 1 ],
   [ 3, 3, 3 ],
   [ 4, 1, 4 ]
]

df = pd.DataFrame(data=X, columns=["height", "width", "weight"])
out = df.groupby(["height", "width"]).count().unstack()
```

What does *out* contain? Write your answer in the box below, specifying:

- Type (DataFrame or Series)
- Columns (for DataFrames only)
- Index
- Data (the contents of the Series/DataFrame)

If the execution of the code results in an error, write “an error occurs” instead.

For "Data", you can separate multiple columns (in the case of DataFrames) using a space. You can separate rows using newlines.

In the case of a multi-level index/column, you must represent them in the same way as with the data (separating with spaces and newlines). You need to repeat the outermost index/column whenever needed.

The following are some examples of the representation you should use.

Example 1

DataFrame

	a	b	c
0	3	3	2
1	2	4	4

Expected representation

Type
DataFrame

Columns
a b c

Index
0 1

Data
3 3 2
2 4 4

Example 2

Multi-index/column DataFrame

		W		X
		a	b	c
y	**0**	2	3	3
	1	4	3	4
z	**0**	3	1	4
	1	3	3	4

Expected representation

Type
DataFrame

Columns
W W X
a b c

Index
y y z z
0 1 0 1

Data
2 3 3
4 3 4
3 1 4
3 3 4

Example 3

Series

```
a     11
b      7
c      5
d     -1
```

Expected representation

Type
Series

Index
a b c d

Data
11 7 5 -1

Domanda 12

Risposta non data

Punteggio max.: 1,00

1 point (15% penalty for a wrong answer)

Consider the following Python code.

```
import numpy as np
import sklearn as sk

X = np.random.random((3, 4))
pipe = sk.pipeline.make_pipeline(
        sk.preprocessing.PolynomialFeatures(degree=3),
        sk.linear_model.Lasso(alpha=0.8)
)
pipe.fit(X, np.arange(3))
```

Which of the following statements is correct regarding the *.fit()* method?

- (a) None of the other answers is correct
- (b) It transforms the original features into polynomial ones (up to degree 3) and trains a linear regressor on them
- (c) It concatenates all polynomial features up to degree 3 to the original ones and trains a polynomial regressor on them
- (d) It trains a polynomial regressor followed by a linear regressor with weights regularization
- (e) It transforms the original features into polynomial ones (up to degree 3) and trains a polynomial regressor on them
- (f) It concatenates all polynomial features up to degree 3 to the original ones and trains a linear regressor on them