Commencé le	jeudi 1 février 2024, 10:00
État	Terminé
Terminé le	jeudi 1 février 2024, 11:29
Temps mis	1 heure 29 min
Points	15,00/30,00
Note	**5,00** sur 10,00 (**50**%)

Question **1**

Incorrect

Note de 0,00 sur 1,00

Consider 2 independent standard Gaussian vectors in a space of dimension 10.

What is the distribution of the distance between them?

Veuillez choisir une réponse.

- Exponential
- Chi
- Chi-square ✗
- Gaussian

La réponse correcte est :
Chi

Question **2**

Correct

Note de 1,00 sur 1,00

What is the sub-differential of the function $f(x) = \max(x, 0)$ in 0?

Veuillez choisir une réponse.

- $[1, +\infty)$
- [-1, 1]
- $(-\infty, 1]$
- [0, 1] ✓

La réponse correcte est :
[0, 1]

Question **3**

Correct

Note de 1,00 sur 1,00

The following code is correct for computing the Simple Local Outlier Factor.

```
class SimpleLocalOutlierFactor():
    """Simple Local Outlier Factor, based on the radius of the ball of nearest neighbors.

    Parameters
    ----------
    n_neighbors : int
        Number of nearest neighbors (excluding the target itself).
    metric : str
        Metric to be used (same parameter as in scikit-learn).

    Attributes
    ----------
    factors : array
        Local Outlier Factors
    """
    def __init__(self, n_neighbors=10, metric='cosine'):
        self.nn = NearestNeighbors(n_neighbors=n_neighbors+1, metric=metric)
        self.factors = None

    def fit(self, data):
        """Compute the local outlier factor of each sample.

        Parameters
        ----------
        data : sparse.csr_matrix or np.ndarray, shape (n_samples, n_features)
            Data.
        """
        self.nn.fit(data)
        distances, neighbors = self.nn.kneighbors(data)
        # radius = max distance
        radius = distances[:, -1]
        # remove the nearest neighbor (= the sample itself)
        neighbors = neighbors[:, 1:]
        self.factors = radius * (1 / radius[neighbors]).mean(axis=1)
```

- (●) Vrai ✓
- ( ) Faux

La réponse correcte est « Vrai ».

Question **4**

Incorrect

Note de 0,00 sur 1,00

Let y be the true binary label of a sample and $\hat{y}$ the label predicted by a classifier.

What is called the precision of the classifier?

Veuillez choisir une réponse.

- (●) $P(\hat{y} = 1|y = 1)$ ✗
- ( ) $P(\hat{y} = 0|y = 0)$
- ( ) $P(y = 1|\hat{y} = 1)$
- ( ) $P(y = 0|\hat{y} = 0)$

La réponse correcte est :
$P(y = 1|\hat{y} = 1)$

Question **5**

Correct

Note de 1,00 sur 1,00

In high dimension, it is preferable to use Ball trees rather than KD-trees for nearest neighbor search.

Veuillez choisir une réponse.

- Vrai ✓
- Faux

La réponse correcte est « Vrai ».

Question **6**

Incorrect

Note de 0,00 sur 1,00

Consider a random forest trained over 1,000 samples.

How many trees do you need to guarantee that at least 99% of these samples are used on average during training?

Veuillez choisir une réponse.

- 5
- 10
- 50
- 100 ✗

La réponse correcte est :
5

Question **7**

Incorrect

Note de 0,00 sur 1,00

Data samples are split only along axes in...

Veuillez choisir une réponse.

- KD trees ✗
- Ball trees
- Both

La réponse correcte est :
Both

Question **8**

Correct

Note de 1,00 sur 1,00

When using Extra Trees for classification, it is recommended to scale data so that each feature takes values in [0,1].

Veuillez choisir une réponse.

- Vrai
- Faux ✓

La réponse correcte est « Faux ».

Question **9**

Correct

Note de 1,00 sur 1,00

Which function of Pandas provides the one-hot encoding of categorical data?

Type your answer as a single string, without space and without parentheses.

Réponse : get_dummies ✓

La réponse correcte est : get_dummies

Question **10**

Correct

Note de 1,00 sur 1,00

You use a Naive Bayes classifier with the Multinomial model and Laplace smoothing on the following training set:

x_1	x_2	x_3	y
0	1	2	0
3	1	1	0
0	1	0	1
0	2	3	1
0	2	0	1
1	1	3	2

What are the parameters of the model for label $y = 1$?

Type your answer as 3 space-separated integers or irreducible fractions (e.g., 0 1/3 2/3).

Réponse : 1/11 6/11 4/11 ✓

La réponse correcte est : 1/11 6/11 4/11

Question **11**

Incorrect

Note de 0,00 sur 1,00

An advantage of XGBoost is that the regression trees used to take the decision can be trained in parallel.

- Vrai ✗
- Faux

La réponse correcte est « Faux ».

Question **12**

Non répondue

Noté sur 1,00

Consider the set $\{0, 1, \ldots, 9\}^2$.

What is the Local Outlier Factor of the sample $(0, 0)$ in this set for the Hamming distance, using 3 nearest neighbors?

Type your answer as an integer or an irreducible fraction (e.g., 2/3).

Réponse : ✗

La réponse correcte est : 2

Question **13**

Incorrect

Note de 0,00 sur 1,00

What is the cosine similarity between these 2 vectors?

$x = (1, 1, 1, 0, 1)$

$y = (1, 1, 0, 1, 1)$

Type your answer as an irreducible fraction (e.g., 2/3).

Réponse : 0.75 ✗

La réponse correcte est : 3/4

Question **14**

Correct

Note de 1,00 sur 1,00

The AUC (ROC Area Under Curve) of a random predictor is 0.

Veuillez choisir une réponse.

- Vrai
- Faux ✓

La réponse correcte est « Faux ».

Question **15**

Correct

Note de 1,00 sur 1,00

You train the Naive Bayes classifier with the Bernoulli model on binary data.

Data samples have dimension 10 and there are 3 labels.

How many parameters are there in the model?

Réponse : 30 ✓

La réponse correcte est : 30

Question **16**

Correct

Note de 1,00 sur 1,00

Dimension reduction by NMF is a projection.

- Vrai
- Faux ✓

La réponse correcte est « Faux ».

Question **17**

Incorrect

Note de 0,00 sur 1,00

You train a random forest to classify the samples of the German credit dataset.

You use the default parameters of scikit-learn.

What is the most important feature?

Type the exact name of the column (e.g., No_of_dependents).

Réponse : Duration ✗

La réponse correcte est : Credit_Amount

Question **18**

Incorrect

Note de 0,00 sur 1,00

You train a Naive Bayes classifier with the Bernoulli model and without Laplace smoothing on the following dataset:

x_1	x_2	x_3	y
1	1	0	0
1	0	1	0
1	1	0	1
0	1	1	1
1	0	1	1
1	1	1	2

The prior is fitted on the training data.

Which label is predicted for the following new sample?

x_1	x_2	x_3	y
1	0	1	?

Veuillez choisir une réponse.

- 0
- 1 ✗
- 2

La réponse correcte est :
0

Question **19**

Correct

Note de 1,00 sur 1,00

Lasso regression refers to a regularization in...

Veuillez choisir une réponse.

- the ℓ_0 pseudo-norm
- the ℓ_1 norm ✓
- the ℓ_2 norm
- the ℓ_∞ norm

La réponse correcte est :
the ℓ_1 norm

Question **20**

Incorrect

Note de 0,00 sur 1,00

You use an Isolation Forest to detect anomalies in the MNIST dataset.

Which digit is the most present in the top-100 most anomalous samples?

Type your answer as an integer (e.g., 2).

Réponse : 8 ✗

La réponse correcte est : 0

Question **21**

Incorrect

Note de 0,00 sur 1,00

You train a Naive Bayes classifier with the categorical model and Laplace smoothing on the following dataset:

x_1	x_2	y
A	a	0
B	a	0
B	b	1
C	a	1
C	a	1
B	b	2

The prior is fitted over data.

What is the posterior distribution of the following new sample?

x_1	x_2	y
B	a	?

Type your answer as 3 space-separated integers or irreducible fractions, for labels 0, 1, 2 (in this order).

Example: 0 1/3 2/3

Réponse : 2/9 5/9 2/9 ✕

La réponse correcte est : 18/41 18/41 5/41

Question **22**

Incorrect

Note de 0,00 sur 1,00

Let x be a binary vector.

Testing whether two distinct random bits of x are equal is a Locally Sensitive Hashing scheme.

- Vrai
- Faux ✕

La réponse correcte est « Vrai ».

Question **23**

Correct

Note de 1,00 sur 1,00

Consider the following matrix:

$$\begin{matrix}0&1&0\\0&0&1\\1&1&0\end{matrix}$$

Assume this matrix is stored in the sparse CSR format of Scipy.

What is the index pointer vector (indptr)?

Type your answer as space-separated integers.

Example: 1 2 3 4

Réponse : 0 1 2 4 ✓

La réponse correcte est : 0 1 2 4

Question **24**

Correct

Note de 1,00 sur 1,00

A matrix X with Frobenius norm $||X|| = 8$ has top singular values 5, 4, 4.

What is the minimum square error of a rank-3 approximation of X?

Type your answer as an integer.

Réponse : 7 ✓

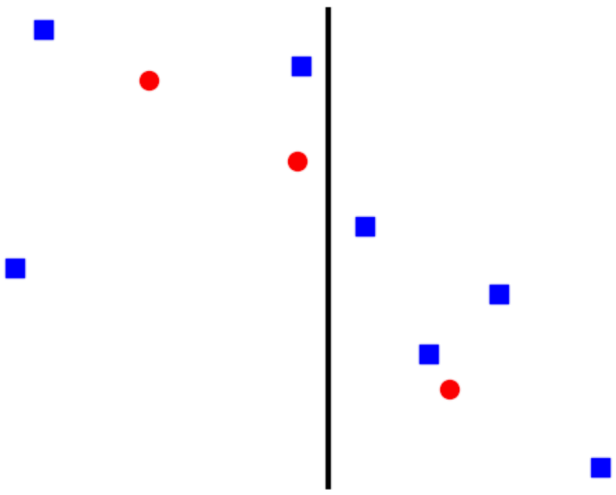
La réponse correcte est : 7

Question **25**

Incorrect

Note de 0,00 sur 1,00

What is the gain of the following split in terms of Gini impurity?



Type your answer as an irreducible fraction (e.g., 2/3).

Réponse : 9/100 ✗

La réponse correcte est : 1/50

Question **26**

Correct

Note de 1,00 sur 1,00

Non-negative matrix factorization is applicable to binary data.

Veuillez choisir une réponse.

- Vrai ✓
- Faux

La réponse correcte est « Vrai ».

Question **27**

Incorrect

Note de 0,00 sur 1,00

Consider the following 7 data samples, with anomaly scores:

$0.3, 0.2, 0.1, 0.1, 0.6, 0.2, 0.1$

Only the first one is a true anomaly.

What is the corresponding ROC AUC score?

Type your answer as an irreducible fraction (e.g., 2/3).

Réponse : 37/50 ✗

La réponse correcte est : 5/6

Question **28**

Correct

Note de 1,00 sur 1,00

Consider the hashing function $h_z(x) = 1_{z^T x > 0}$ where x is the data sample and z is the random vector whose components are i.i.d. uniform on $(-1, 1)$.

This defines a Locally Sensitive Hashing scheme.

- (●) Vrai ✓
- ( ) Faux

La réponse correcte est « Vrai ».

Question **29**

Correct

Note de 1,00 sur 1,00

What is the Jaccard distance between these two vectors?

$(0, 1, 1, 1, 0, 0, 1, 1)$

$(1, 1, 1, 0, 0, 1, 1, 0)$

Type your answer as an integer or an irreducible fraction (e.g., 1/2).

Réponse : 4/7 ✓

La réponse correcte est : 4/7

Question **30**

Incorrect

Note de 0,00 sur 1,00

Consider an Extra Tree applied to the following 2D samples:

$(1, 0), (-2, 0), (0, 1)$

What is the average depth of sample $0, 1$?

Type your answer as an integer or an irreducible fraction (e.g., 1/2).

Réponse : 4/3 ✕

La réponse correcte est : 3/2

◄ Notebook on anomaly detection

Aller à...