Quiz 2 2023

Question 1

Question text

Which of the following is an advantage of using a binary format ("serialization") over JSON for large-scale data?
Check all that apply.
Question 1 Answer
a.
Binary formats are easier to update
b.
Binary formats enable more efficient storage of numerical values
c.
Binary formats often require less expensive parsing
d.
Binary formats are human-readable

Feedback

The correct answers are: Binary formats often require less expensive parsing, Binary formats enable more efficient storage of numerical values

Question 2

Partially correct
Mark 0.50 out of 1.00

Flag question

Question text

Which of these statements about the flat data model are correct?
Check all that apply.
Question 2 Answer
a.
To separate individual fields in the flat data model, we have to use delimiters such as comma or tab
b.
In the flat data model, data can be represented in plain-text files
c.
For large data, it gets more difficult to update samples stored in the flat data model than in the relational model
d.
It is primarily used to represent complex relationships between data points

Feedback

The correct answers are: In the flat data model, data can be represented in plain-text files, For large data, it gets more difficult to update samples stored in the flat data model than in the relational model

Question 3

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Which data visualization technique is useful for highlighting connectivity patterns in network data?
Question 3 Answer
a.
Bar chart
b.
Matrix view
c.
Scatter plot
d.
Pie chart

Feedback

The correct answer is: Matrix view

Question 4

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You are given two relational tables, Customers and Orders.
The Customers table contains two columns, customer_id and name:
customer_id, name
[1, Alice]
[2, John]
[3, David]
The Orders table contains three columns, order_id, customer_id, and item:
order_id, customer_id, item
[1, 2, Keyboard]
[2, 3, Mouse]
[3, 2, Monitor]

What is the output of the following query, assuming there were no optimization methods affecting the ordering performed:
SELECT name FROM Customers
INNER JOIN Orders ON Customers.customer_id = Orders.customer_id
Question 4 Answer
a.
name
[John]
[David]
[John]
b.
name
[Alice]
[David]
[Alice]
c.
name
[Alice]
[David]
[John]
d.
name
[John]
[David]

Feedback

The correct answer is: name
[John]
[David]
[John]

Question 5

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You are given two relational tables, Customers and Orders.

The Customers table contains two columns, customer_id and name:

customer_id, name

[1, John]

[2, Alice]

[3, David]

The Orders table contains three columns, order_id, customer_id, and item:

order_id, customer_id, item

[1, 2, Keyboard]

[2, 3, Mouse]

[3, 2, Monitor]

How would you represent the customer Alice, taking into consideration both tables, in the XML format?

Question 5 Answer
a.

<customer>

<customer_id>2</customer_id>

<name>Alice</name>

</customer>

<order>

<order_id>1</order_id>

<item>Keyboard</item>

<order_id>3</order_id>

<item>Monitor</item>

</order>

b.

<customer>

```
    <customer_id>2</customer_id>
    <name>Alice</name>
    <order>
        <order_id>1</order_id>
        <item>Keyboard</item>
    </order>
    <order>
        <order_id>3</order_id>
        <item>Monitor</item>
    </order>
</customer>
```

c.

It is not possible to represent the customer Alice in XML format

d.

```
<customer>
    <customer_id>2</customer_id>
    <name>Alice</name>
    <order>
        <order_id>1</order_id>
        <item>Keyboard</item>
    <order>
        <order_id>3</order_id>
        <item>Monitor</item>
</customer>
```

Feedback

Your answer is correct.
The correct answer is:

```
<customer>
    <customer_id>2</customer_id>
    <name>Alice</name>
    <order>
        <order_id>1</order_id>
        <item>Keyboard</item>
    </order>
    <order>
        <order_id>3</order_id>
        <item>Monitor</item>
    </order>
</customer>
```

Quiz 3 2023

Question 1

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You are analyzing data that follows a power law distribution. You decide to visualize the data using the complementary cumulative distribution function (CCDF). Which of the following statements about the CCDF plot are true? Check all that apply.
Question 1 Answer
a.
Binning of data points is required to generate the CCDF plot
b.

The CCDF plot will be monotonically decreasing
c.
The CCDF plot will have the same exponent α as the probability density function (PDF) plot
d.
The CCDF plot will be monotonically increasing

Feedback

The correct answer is: The CCDF plot will be monotonically decreasing

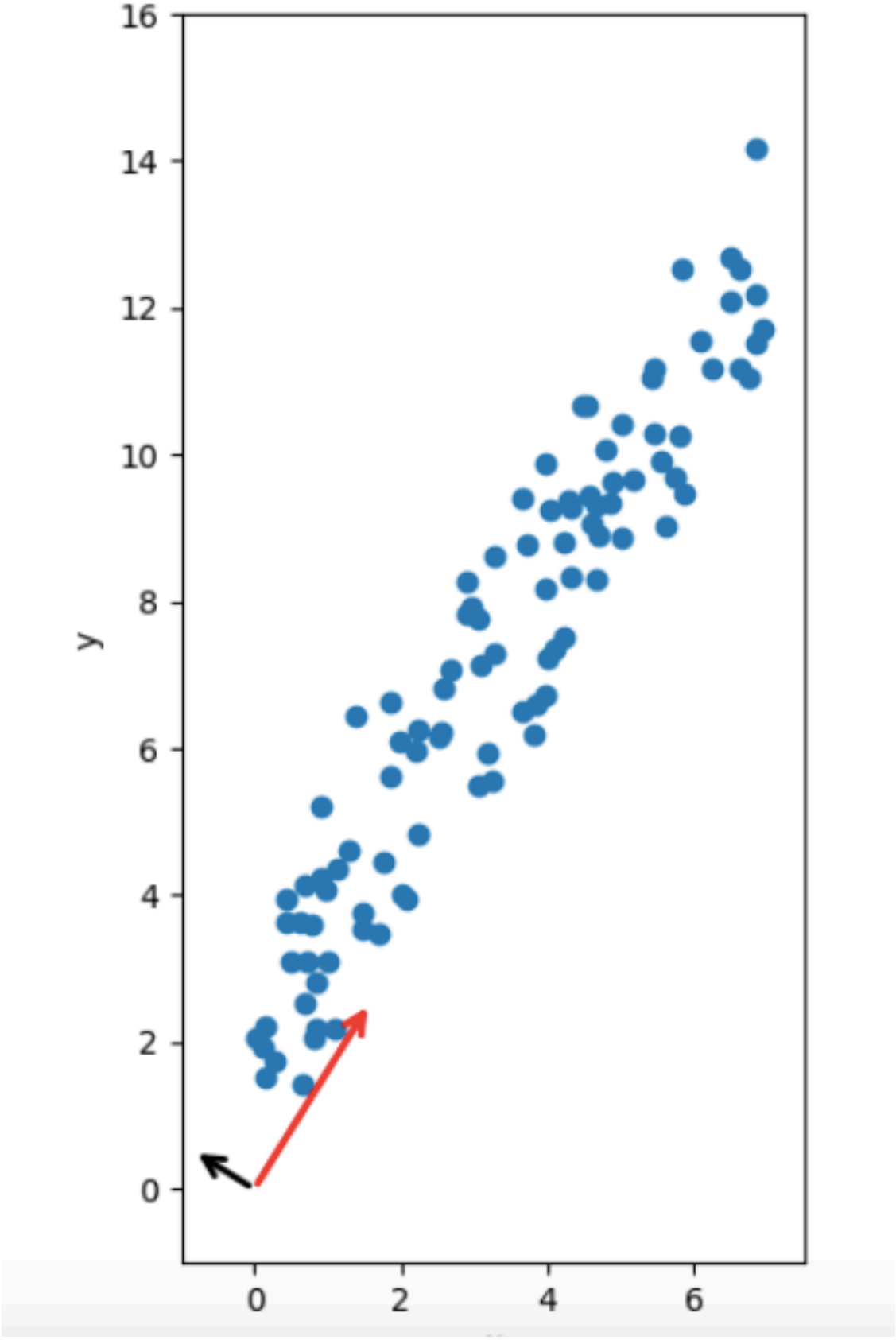
Question 2

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Given this plot, check all that apply.



Question 2 Answer

a.
Red arrow could be returned by the PCA algorithm as the second principal component for the dataset
b.
Black arrow could be returned by the PCA algorithm as the first principal component for the dataset
c.
Black arrow could be returned by the PCA algorithm as the second principal component for the dataset
d.
Red arrow could be returned by the PCA algorithm as the first principal component for the dataset

Feedback

The correct answers are: Red arrow could be returned by the PCA algorithm as the first principal component for the dataset, Black arrow could be returned by the PCA algorithm as the second principal component for the dataset

Question 3

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You have survey data from 100 people indicating their most preferred color out of red, blue, green, yellow, and purple. You want to visualize the results as counts/frequencies for each color. Which type of plot(s) is (are) suitable in this situation? Check all that apply.
Question 3 Answer
a.
Line chart
b.
Histograms
c.
Scatter plot
d.
Bar chart

Feedback

The correct answer is: Bar chart

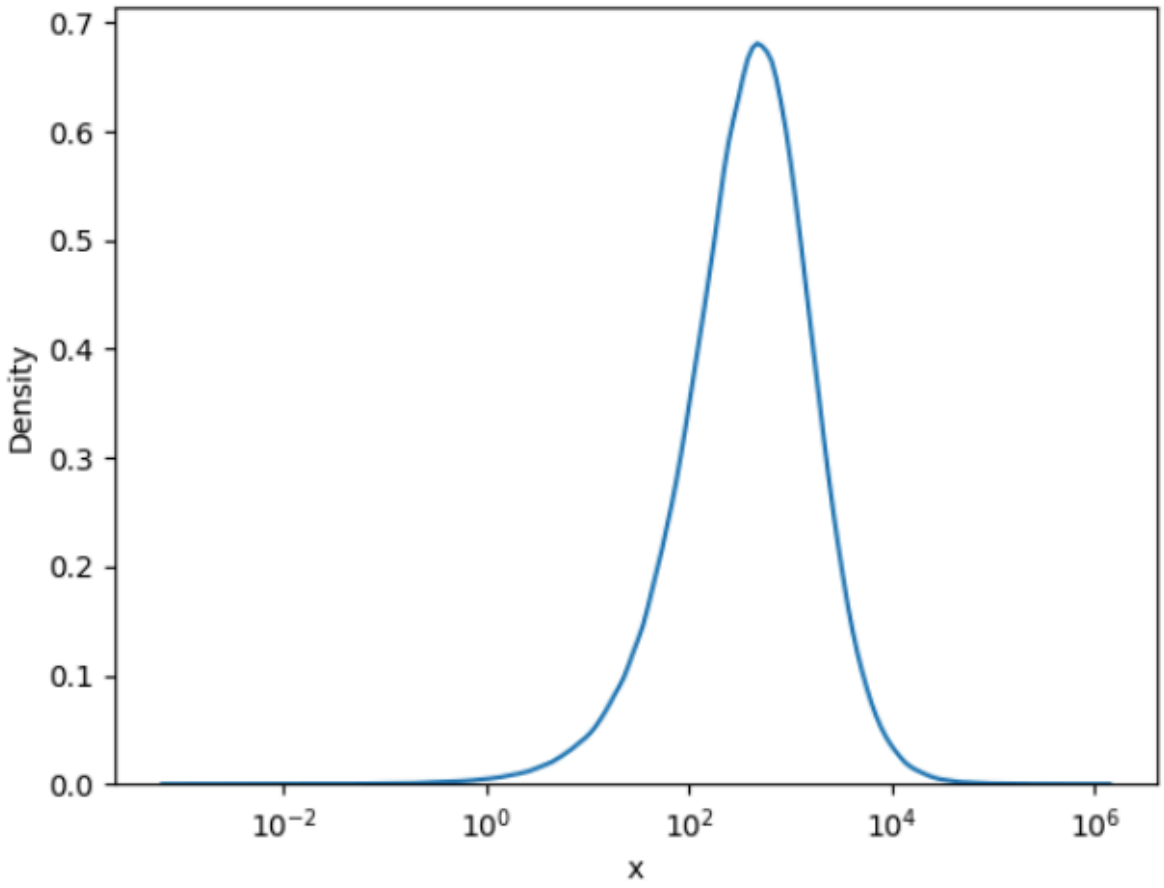
Question 4

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Given this kernel density estimation plot of the data, check the correct statement.



Question 4 Answer
a.
The data could come from a heavy-tailed distribution
b.
The data could come from a Gaussian distribution
c.
The data is two-dimensional
d.
None of the above

Feedback

The correct answer is: The data could come from a heavy-tailed distribution

Question 5

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You have data on the daily revenue of a bakery over the past 5 years. You notice that the daily revenue figures follow a heavy-tailed distribution with many small values and a long tail of rare but extremely high value days. What type of plot would you use to best visualize this distribution?
Question 5 Answer
a.
Cumulative distribution function (CDF) plot
b.
Box plot
c.
Histogram with a linear x-axis
d.
Histogram with a logarithmic x-axis

Feedback

The correct answer is: Histogram with a logarithmic x-axis

Quiz 4 2023

Started on	Friday, 20 October 2023, 13:15
State	Finished
Completed on	Friday, 20 October 2023, 13:25
Time taken	10 mins
Grade	**3.50** out of 5.00 (**70**%)
Feedback	If you have any remarks about the questions or the grading, please submit them using this form!

Question 1

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You collect data that appears to follow a power law distribution. Based on a log-log plot, you estimate the exponent to be $\alpha=1.9$. What can you conclude about the appropriate statistics to describe this data?
Question 1 Answer
a.
The estimated α is close enough to 2 such that the true mean and variance are probably finite and can be reported.
b.
The true variance is likely to be finite, so it's ok to report the estimated variance.
c.
The true mean may be infinite, so median is better-suited than mean, but variance can still be reported.
d.
Both true mean and true variance may be infinite, so only robust statistics like median should be reported.

Feedback

The correct answer is: Both true mean and true variance may be infinite, so only robust statistics like median should be reported.

Question 2

Partially correct
Mark 0.50 out of 1.00

Flag question

Question text

Which of the following statements are true? Assume you are working with non-negative data. Check all that apply.
Question 2 Answer
a.
For heavy-tailed distribution, median can be smaller than mean
b.
For symmetrical distributions, median is always the same as mean
c.
For skewed distribution, median can be smaller, bigger or equal to the mean
d.
For skewed distribution, median is smaller than mean

Feedback

The correct answers are: For heavy-tailed distribution, median can be smaller than mean, For symmetrical distributions, median is always the same as mean

Question 3

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You have a method for classifying data points into 3 classes, A, B, and C. Your dataset is imbalanced: A appears in 70% of the cases, B in 25% of cases, and C in 5% of cases. You evaluate the performance of the method by calculating the fraction of data points classified correctly, for each class. Which statements are true?
Check all that apply.
Question 3 Answer
a.
If we care about performance per each class, micro-average is a better metric than macro-average
b.
If we care about overall performance, macro-average is a better metric than micro-average
c.
If we care about overall performance, micro-average is a better metric than macro-average
d.
If we care about performance per each class, macro-average is a better metric than micro-average

Feedback

The correct answers are: If we care about overall performance, micro-average is a better metric than macro-average, If we care about performance per each class, macro-average is a better metric than micro-average

Question 4

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You perform a hypothesis test of a null hypothesis H0 using Dataset 1 and obtain a p-value of 0.01, leading you to reject H0 at the 5% significance level. You then collect more data, acquiring a new dataset which we call Dataset 2. You perform the hypothesis test again, obtaining a p-value of 0.3, meaning you fail to reject H0 at the 5% significance level. What is the most valid conclusion based on these results?
Question 4 Answer
a.
The contradictory p-values mean H0 is equally likely to be true or false.
b.

The lower p-value indicates H0 is likely false, while the higher p-value indicates H0 is likely true.
c.
The first p-value implies that the alternative hypothesis is true. The second p-value implies that H0 is true.
d.
The p-values only suggest how likely each dataset was under H0. More evidence is needed to determine if H0 is true or false.

Feedback

The correct answer is: The p-values only suggest how likely each dataset was under H0. More evidence is needed to determine if H0 is true or false.

Question 5

Incorrect
Mark 0.00 out of 1.00

Flag question

Question text

You are conducting a study to determine whether a new drug is effective in reducing blood pressure. You collect data from 1000 patients and perform a hypothesis test. The null hypothesis is that the drug has no effect on blood pressure, and the alternative hypothesis is that the drug reduces blood pressure. After conducting the test, you obtain a p-value of 0.03. Which of these statements is correct?
Question 5 Answer
a.
The probability that the drug reduces blood pressure is 97%
b.
If you choose significance level 0.05, this means that the drug has a substantial effect on lowering the blood pressure
c.
If you calculated the p-value on half of the patients only, the p-value would still be smaller than 0.05
d.
None of the above

Feedback

The correct answer is: None of the above

Quiz 5 2023

Started on Friday, 27 October 2023, 13:15

State	Finished
Completed on	Friday, 27 October 2023, 13:25
Time taken	10 mins
Grade	**3.67** out of 5.00 (**73.33**%)
Feedback	If you have any remarks about the questions or the grading, please submit them using this form!

Question 1

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You build a linear regression model to predict website traffic y based on advertising spending x. After assessing the data, you decide to log-transform the outcome and model log(y) instead of y. When evaluating the model, you note the coefficient for x is 0.005. Which interpretation of this coefficient is correct?
Question 1 Answer
a.
The 0.5 coefficient means advertising spending has no relationship with website traffic.
b.
A $1 increase in advertising spending predicts a 0.5% increase in website traffic y.
c.
The transformation implies that traffic is now modeled as a logarithmic, not linear, function of advertising spending
d.
A $1 increase in advertising spending predicts a 0.5 absolute increase in y.

Feedback

The correct answer is: A $1 increase in advertising spending predicts a 0.5% increase in website traffic y.

Question 2

Incorrect
Mark 0.00 out of 1.00

Flag question

Question text

Which of the following transformations can change the R^2 score? Check all that apply.

Question 2 Answer

a.
Applying a quadratic transformation to the predictors

b.
Standarization of predictors

c.
Turning the additive model into a multiplicative model by logarithmically transforming the outcomes

d.
Mean-centering of predictors

Feedback

The correct answers are: Turning the additive model into a multiplicative model by logarithmically transforming the outcomes, Applying a quadratic transformation to the predictors

Question 3

Partially correct
Mark 0.67 out of 1.00

Flag question

Question text

Which of the following statements is correct? Check all that apply.

Question 3 Answer

a.
Intercept represents the predicted value of the outcome when all predictor variables are set to zero

b.
If a predictor has a positive linear regression coefficient, that means that an increase in the predictor is associated with an increase in the outcome

c.
Low R^2 score indicates that the predictors have no statistically significant correlation with the outcome

d.
If we have negative predictors, we should not perform linear regression

Feedback

The correct answer is: Intercept represents the predicted value of the outcome when all predictor variables are set to zero

Question 4

Correct
Mark 1.00 out of 1.00

Flag question

Question text

A study evaluates the impact of a new employee training program on productivity. The study includes a treatment group who received the program and a control group who did not. Productivity (y) was measured before the program (time 1) and after (time 2). The study uses a difference-in-differences regression model with these variables:
Treatment: Binary indicator for treatment group
Time: Binary indicator for time 2
Treatment*Time: Interaction between treatment and time
The coefficient for Treatment*Time is 3. Which interpretation is correct?
Question 4 Answer
a.
The treatment group was 3 units more productive than the control group after the program.
b.
Productivity increased by 3 units from time 1 to time 2 in both groups.
c.
Productivity increased by 3 units more for the treatment group than the control group from time 1 to 2.
d.
The treatment group achieved 1/3 of the productivity of the control group.

Feedback

The correct answer is: Productivity increased by 3 units more for the treatment group than the control group from time 1 to 2.

Question 5

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You want to test if a new protein supplement affects weight lifters' muscle mass. You collect data on muscle mass gained and whether a weight lifter took the protein supplement or not over a 6 month period. You decide to use a linear regression model to predict muscle mass gained based on a binary predictor indicating whether the weight lifter took the protein supplement or not.

What would be the appropriate null hypothesis when testing if the protein supplement has an effect in this regression model?
Question 5 Answer
a.
Taking the protein supplement increases muscle mass gained.
b.
The effect of the protein supplement cannot be hypothesized.
c.
Taking the protein supplement has no effect on muscle mass gained.
d.
Taking the protein supplement decreases muscle mass gained.

Feedback

The correct answer is: Taking the protein supplement has no effect on muscle mass gained.

Quiz 6 2023

Started on	Friday, 3 November 2023, 13:15
State	Finished
Completed on	Friday, 3 November 2023, 13:25
Time taken	10 mins 2 secs
Grade	**3.17** out of 5.00 (**63.33**%)
Feedback	If you have any remarks about the questions or the grading, please submit them using this form!

Question 1

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Population-based Cohort Studies (PBCS) are observational studies that follow a group of individuals over time to understand the relationship between certain factors, such as cardiovascular exercise, and health outcomes, like heart attacks. Unlike Randomized Controlled Trials (RCTs), participants in PBCS are not randomly assigned to groups but rather chosen based on their characteristics or exposure to the factor of interest.
Which of the following statements are correct? Check all that apply.
Question 1 Answer
a.

PBCS cannot be used for causal studies.
b.
PBCSs may be applied in scenarios where RCTs are not feasible.
c.
The results of PBCSs are only valid if the participants are informed of the group (control or treatment) that they are assigned to.
d.
In PBCS the reserachers do not have control over the treatment assignment.

Feedback

The correct answers are: In PBCS the reserachers do not have control over the treatment assignment., PBCSs may be applied in scenarios where RCTs are not feasible.

Question 2

Incorrect
Mark 0.00 out of 1.00

Flag question

Question text

When performing a randomized experiment (with a treatment group and a control group), which of the following statements is true?
Question 2 Answer
a.
All participants have the same probability of being assigned to the treatment group
b.
Randomized experiments are usually harder to replicate than observational studies
c.
Unobserved confounders may threaten the validity of your conclusions
d.
For every participant, the probability of being assigned to the treatment group is the same as the probability of being assigned to the control group

Feedback

The correct answer is: All participants have the same probability of being assigned to the treatment group

Question 3

Partially correct
Mark 0.50 out of 1.00

Flag question

Question text

Which of the following statements about determining causality from observational data are true? Check all that apply.
Question 3 Answer
a.
Large sensitivity analysis parameter (Γ) means that two subjects with the same unobserved covariates have vastly different probabilities of getting the treatment.
b.
Studies should match treated and control units on as many relevant covariates as possible to minimize confounding.
c.
Randomized experiments are usually cheaper to conduct than obsevational studies, as most modern data is found data
d.
Sensitivity analysis quantifies the potential impact of unmeasured confounding on study conclusions.

Feedback

The correct answer is: Sensitivity analysis quantifies the potential impact of unmeasured confounding on study conclusions.

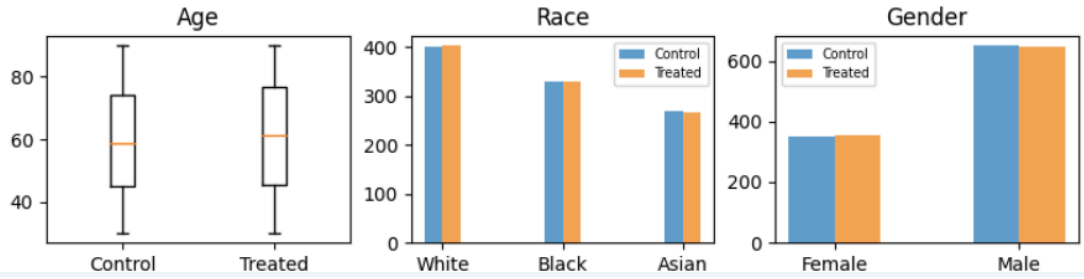
Question 4

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You are given data with patients containing their age, gender, race, an indicator whether they were treated with a drug or not ("treated"), and an indicator whether they were cured or not ("cured"). In the plot, you see the distributions of the attributes: age, race and gender for treated and control group (control group - "treated" = 0). You want to investigate the effect of the drug. You may assume the attributes are independent of one another. What steps should you take? Check all that apply.



Question 4 Answer

a.

Prior to the regression analysis, you should do exact matching of patients using age, gender, and race as covariates

b.

In the presence of unobserved covariates, regression analysis using treatment, age, race, and gender as dependable variables, and "cured" as the outcome, may not let you identify the causal effect of "treated" on "cured"

c.

If you assume there are no unobserved confounders (as you assume in the "naive model"), to estimate the causal effect of "treated" on "cured", you can directly perform regression analysis, using treatment, age, race, and gender as dependable variables and "cured" as the outcome

d.

Prior to the regression analysis, you should estimate propensity scores for each patient and do matching of patients based on that

Feedback

The correct answers are: If you assume there are no unobserved confounders (as you assume in the "naive model"), to estimate the causal effect of "treated" on "cured", you can directly perform regression analysis, using treatment, age, race, and gender as dependable variables and "cured" as the outcome, In the presence of unobserved covariates, regression analysis using treatment, age, race, and gender as dependable variables, and "cured" as the outcome, may not let you identify the causal effect of "treated" on "cured"

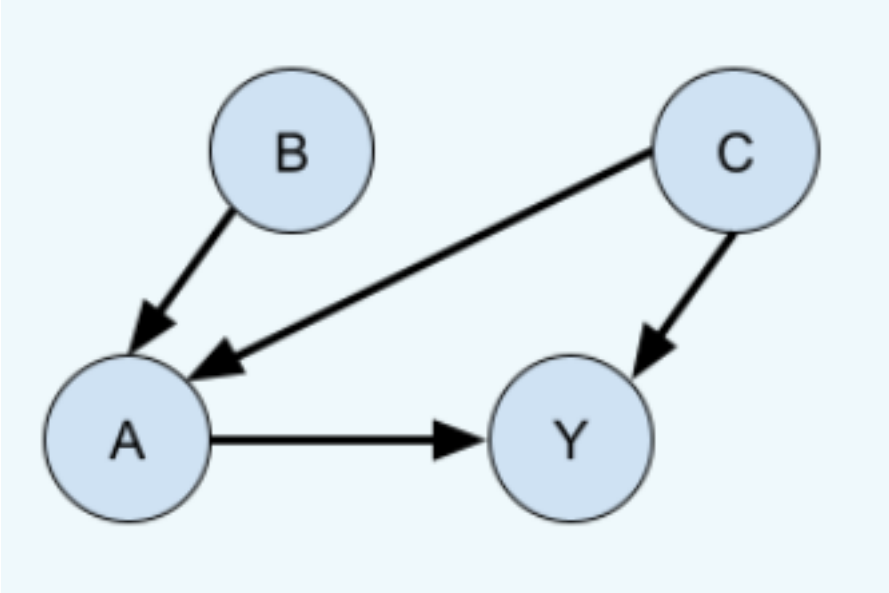
Question 5

Partially correct
Mark 0.67 out of 1.00

Flag question

Question text

You are given the causal diagram in the image below. All the variables are one-dimensional and there are no unobserved variables affecting A, B, C and Y. Check all that apply.



Question 5 Answer

a.

C has causal effect on Y

b.

B is a confounder when observing effect of A on Y

c.

C is a confounder when observing effect of A on Y

d.

A has causal effect on Y

Feedback

The correct answers are: C is a confounder when observing effect of A on Y, A has causal effect on Y, C has causal effect on Y

-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-

1. CS-401
2. Week 8
3. Quiz 7 2023

Quiz 7 2023

Started on	Friday, 10 November 2023, 13:16
State	Finished
Completed on	Friday, 10 November 2023, 13:26
Time taken	10 mins
Grade	**2.67** out of 5.00 (**53.33**%)

Feedback If you have any remarks about the questions or the grading, please submit them using this form!

Question 1

Correct
Mark 1.00 out of 1.00

Flag question

Question text

A logistic regression model predicts the probability p of students passing an exam based on the number X of hours studied. The model contains a coefficient β1 = 0.5 for the study hours variable. Which interpretation of this coefficient is correct?
Question 1 Answer
a.
Studying 1 more hour doubles the probability p of passing.
b.
Studying 1 more hour increases the probability p of passing by 0.5.
c.
Studying 1 more hour increases the log-odds of passing by 0.5.
d.
Studying 1 more hour doubles the odds of passing.

Feedback

The correct answer is: Studying 1 more hour increases the log-odds of passing by 0.5.

Question 2

Partially correct
Mark 0.67 out of 1.00

Flag question

Question text

Which of the following statements about the bias-variance tradeoff are true? Select all that apply.
Question 2 Answer
a.
Increasing k in k-NN decreases bias but increases variance.
b.
Adding more features to a logistic regression model decreases bias but increases variance.
c.
Adding depth to a decision tree decreases bias but increases variance.

d.
Increasing the number of trees in boosted decision trees decreases the bias.

Feedback

The correct answers are: Adding depth to a decision tree decreases bias but increases variance., Adding more features to a logistic regression model decreases bias but increases variance., Increasing the number of trees in boosted decision trees decreases the bias.

Question 3

Partially correct
Mark 0.50 out of 1.00

Flag question

Question text

A decision tree classifier is trained to classify mammals based on body size, tail length, and number of legs.
Tail length is the root node, while the body size appears deeper in the tree and the number of legs appear as a leaf.
Which of the following statements are true? Check all that apply.
Question 3 Answer
a.
Tail length provides no useful information compared to body size, so the tree must be overfitting.
b.
The tree has low variance since all features are considered.
c.
Tail length provides the most information gain compared to other features, so it was selected as the root node.
d.
Pruning the leaf corresponding to "number of legs" would result in a tree from a smaller model family, with more bias but less variance.

Feedback

The correct answers are: Pruning the leaf corresponding to "number of legs" would result in a tree from a smaller model family, with more bias but less variance., Tail length provides the most information gain compared to other features, so it was selected as the root node.

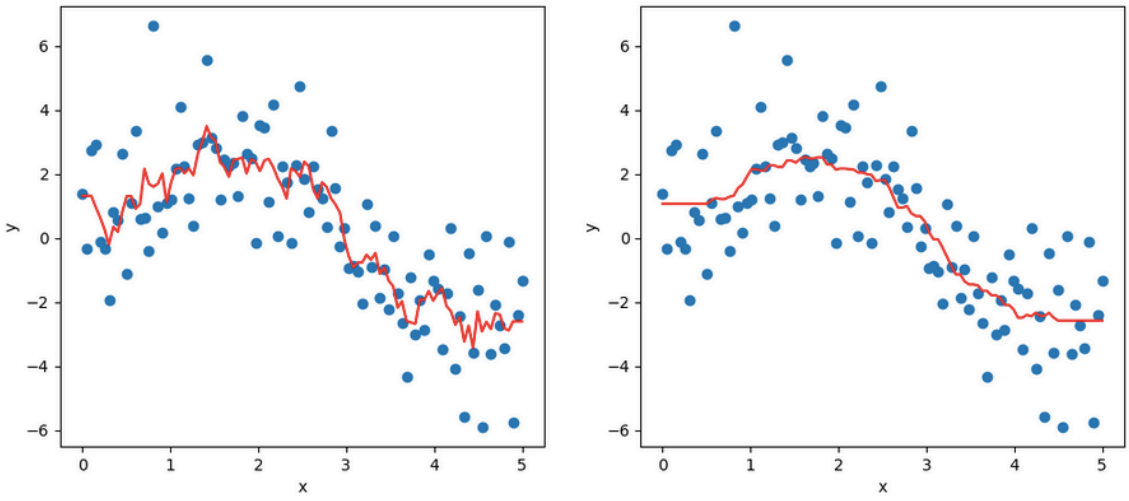
Question 4

Partially correct
Mark 0.50 out of 1.00

Flag question

Question text

You are given the plots below, and you were told that the fits (red line) were obtained using k nearest neighbours. What can you conclude from the plots? Check all that apply.



Question 4 Answer

a.

If the data points are uniformly weighted, the left fit was generated with a smaller parameter k than the right plot

b.

I was lied to, the red fits cannot have been generated using k nearest neighbours

c.

If the data points are uniformly weighted, the right fit was generated with a smaller parameter k than the left plot

d.

If the data points are weighted differently, you cannot rule out that the two fits were generated with the same parameter k

Feedback

The correct answers are: If the data points are uniformly weighted, the left fit was generated with a smaller parameter k than the right plot, If the data points are weighted differently, you cannot rule out that the two fits were generated with the same parameter k

Question 5

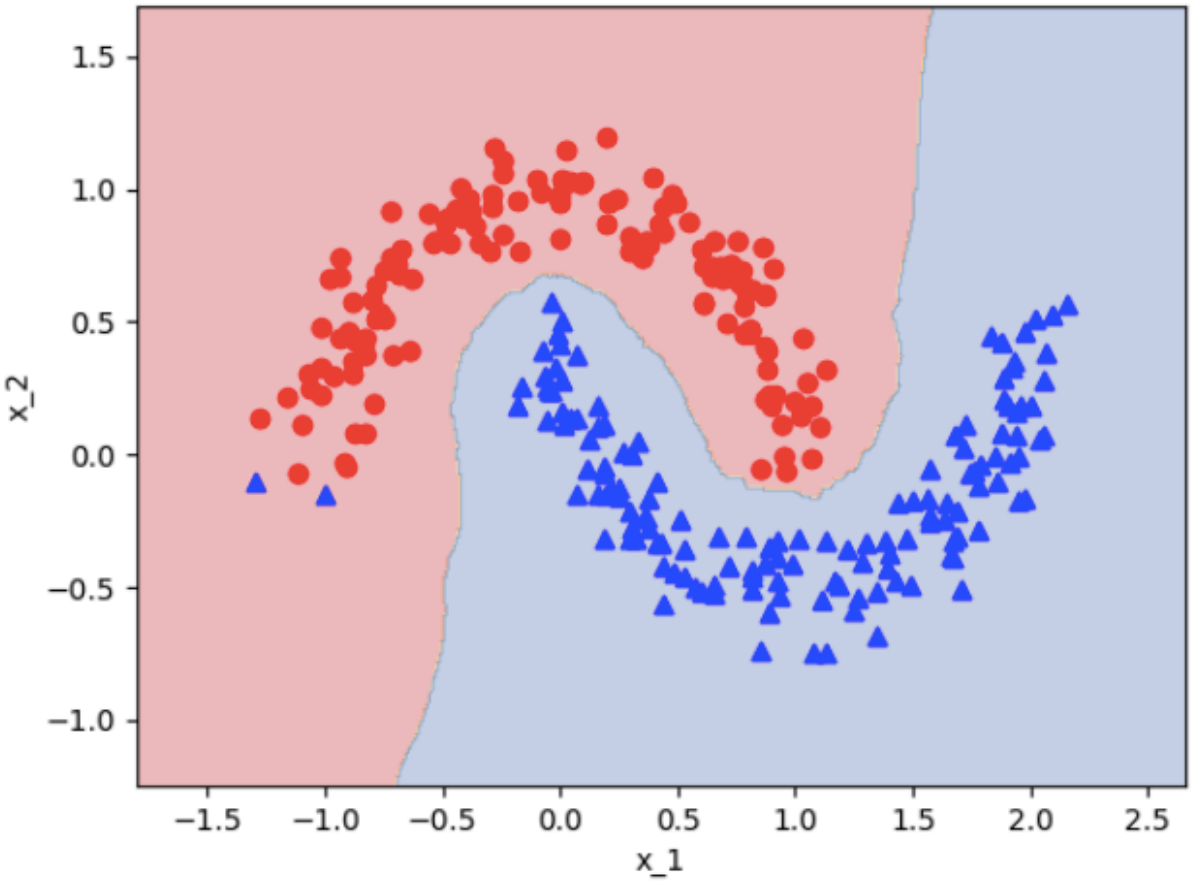
Incorrect

Mark 0.00 out of 1.00

Flag question

Question text

The plot shows data points from two classes (red circles and blue triangles) as well as the decisions made by some classifier: points on red background are classified as red, and points on blue background as blue. Which classifier could have been used to make these decisions (using x_1 and x_2 as features)? Check all that apply.



Question 5 Answer
a.
Logistic regression
b.
K nearest neighbours with K=10
c.
K nearest neighbours with K=2
d.
None of the above

Feedback

The correct answer is: K nearest neighbours with K=10

The correct answer is: K nearest neighbours with K=10

Quiz 8 2023

Started on	Friday, 17 November 2023, 13:15
State	Finished
Completed on	Friday, 17 November 2023, 13:25
Time taken	10 mins
Grade	**3.67** out of 5.00 (**73.33**%)
Feedback	If you have any remarks about the questions or the grading, please submit them using this form!

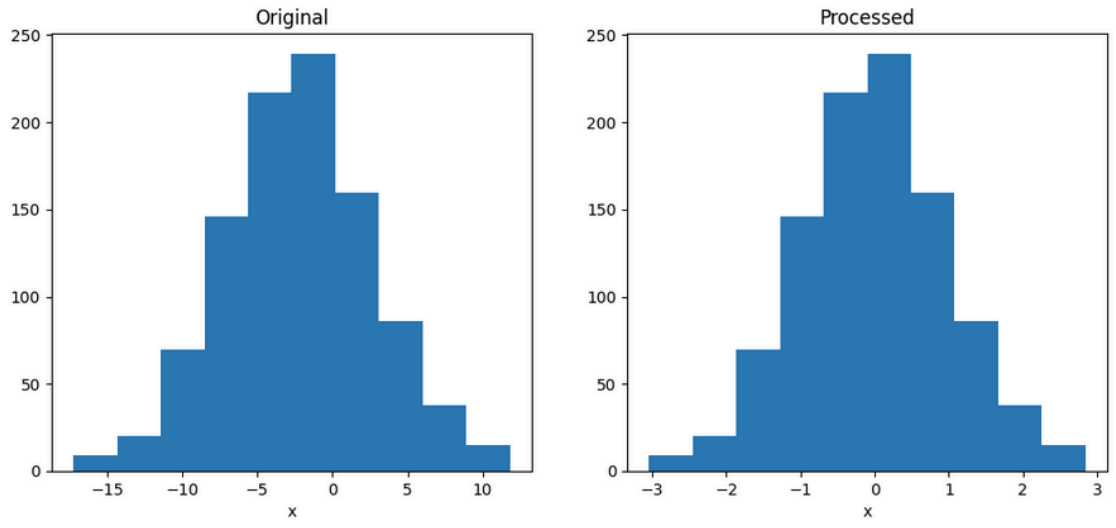
Question 1

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You have one-dimensional data whose distribution is shown in the left plot. After feature pre-processing, you obtain the data with the distribution shown in the right plot. Which feature transformation could have been performed?



Question 1 Answer
a.

Standardization
b.
Min-max scaling
c.
Logarithmic scaling
d.
None of the above

Feedback

The correct answer is: Standardization

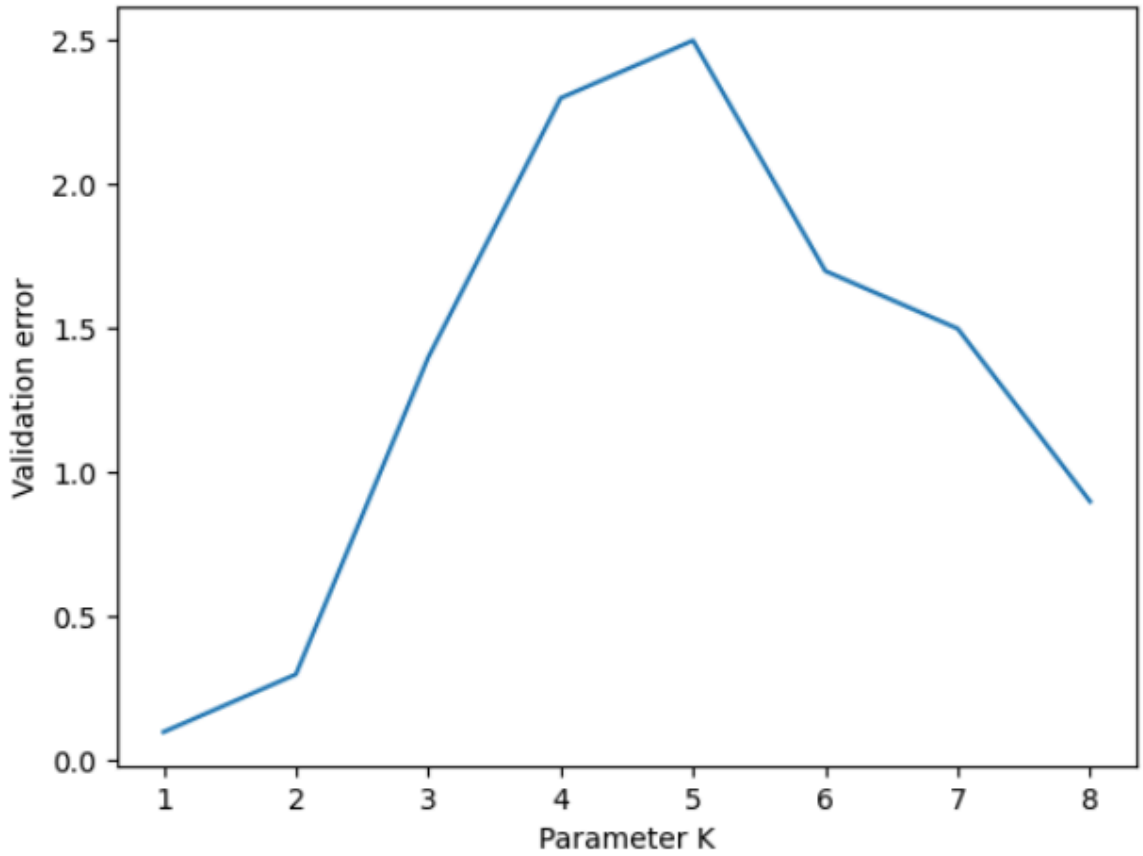
Question 2

Correct
Mark 1.00 out of 1.00

Flag question

Question text

You perform cross-validation to determine the best choice for a parameter K. You obtain the plot below. Which value of the parameter K should you choose?



Question 2 Answer
a.
K=5
b.
K=8
c.
K=1
d.
K=4

Feedback

The correct answer is: K=1

Question 3

Partially correct
Mark 0.67 out of 1.00

Flag question

Question text

Which statements are correct regarding feature selection in the context of a classification task when dealing with a dataset with both continuous and categorical features? Check all that apply.
Question 3 Answer
a.
In the context of the χ^2
-test, the p-value serves as an indicator of the strength of the association between the class and the feature under consideration.
b.
While online feature selection methods are fast to apply, they often don't account for the interdependence or relationships between features.
c.
Feature ranking is often regarded as a superior approach, compared to online feature selection methods. This is because feature ranking considers features as a collective entity rather than evaluating them individually.
d.
Feature selection can lower the chance of overfitting and allows for more efficient training.

Feedback

The correct answer is: Feature selection can lower the chance of overfitting and allows for more efficient training.

Question 4

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Given the following confusion matrix, which statements about the classifier are true? Check all that apply.

		Class	
		TRUE	FALSE
Predicted	TRUE	30	10
	FALSE	10	50

Question 4 Answer

a.
The precision is 0.75
b.
The accuracy is 0.8
c.
The F1 score equals the precision
d.
The classifier has higher recall than precision

Feedback

The correct answers are: The accuracy is 0.8, The precision is 0.75, The F1 score equals the precision

Question 5

Incorrect
Mark 0.00 out of 1.00

Flag question

Question text

Which of these statements about the precision/recall curve are correct?
Check all that apply.
Question 5 Answer
a.
The area under the precision/recall curve for a random classifier depends on number of positive and negative samples in the dataset
b.
The precision/recall curve for a random classifier corresponds to the identity line (y=x)
c.
The precision/recall curve is sensitive to the classification threshold
d.
The area under the precision/recall curve for a perfect classifier is 1

Feedback

The correct answers are: The area under the precision/recall curve for a perfect classifier is 1, The area under the precision/recall curve for a random classifier depends on number of positive and negative samples in the dataset

Quiz 10 2023

Started on	Friday, 1 December 2023, 13:16
State	Finished
Completed on	Friday, 1 December 2023, 13:26
Time taken	10 mins 1 sec
Grade	**5.00** out of 5.00 (**100**%)
Feedback	If you have any remarks about the questions or the grading, please submit them using this form!

Question 1

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Given the bag-of-words matrix below, which was obtained without any preprocessing of the text, what conclusions can you draw? Check all that apply.

Bag-of-words matrix

2	3	0	0
1	0	2	0
2	0	0	0
0	1	2	1
3	1	1	2

Question 1 Answer
a.
All documents consist of at most 4 words
b.
The corpus has 5 documents
c.
The corpus has 4 documents
d.
All documents consist of at most 4 unique words

Feedback

The correct answers are: The corpus has 5 documents, All documents consist of at most 4 unique words

Question 2

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Which of the following statements are true for character encodings:
Check all that apply.
Question 2 Answer
a.
It is safe to read UTF-8 encoded data with ASCII
b.
Standard ASCII encoding can be used when you are working with text containing only English alphabet
c.
Latin-1 always encodes each character with 1 byte
d.
UTF-8 encoding can encode only 256 different characters

Feedback

The correct answers are: Latin-1 always encodes each character with 1 byte, Standard ASCII encoding can be used when you are working with text containing only English alphabet

Question 3

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Which of the following statements about text tokenization are true? Check all that apply.
Question 3 Answer
a.
Stemming reduces sparsity but loses information such as the part of speech (i.e., whether a word is a verb, noun, adjective, etc.).
b.
Lemmatization maps words to normalized lexicon entries.
c.
Lemmatization is most useful for languages with little morphology (i.e., where words aren't inflected in many different ways).

d.
Tokenization is more challenging for languages without explicit word delimiters like whitespace, such as Chinese, or those with compound words, like German.

Feedback

The correct answers are: Lemmatization maps words to normalized lexicon entries., Stemming reduces sparsity but loses information such as the part of speech (i.e., whether a word is a verb, noun, adjective, etc.)., Tokenization is more challenging for languages without explicit word delimiters like whitespace, such as Chinese, or those with compound words, like German.

Question 4

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Assuming we do bigram (2-gram) tokenization with whitespace delimiters in the following corpus, what would be the IDF of the token (a a):
corpus = {
'Doc0': "a a b d c c",
'Doc1': "d b c a a c",
'Doc2': "c d b a a a",
'Doc3': "c c a a d b"
}
For the calculation, use the natural logartihm, that is the logarithm with base e. Report the outcome up to 2 decimal points of precision.
Question 4 Answer
a.
0.69
b.
1.39
c.
1
d.
0

Feedback

The correct answer is: 0

Question 5

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Which of the following transformations can be used to penalize the words that appear often in the corpus? Check all that apply.
Question 5 Answer
a.
Stopword removal
b.
Lemmatization
c.
IDF
d.
Row normalization

Feedback

The correct answers are: IDF, Stopword removal

Quiz 12 2023

Started on	Friday, 15 December 2023, 13:15
State	Finished
Completed on	Friday, 15 December 2023, 13:25
Time taken	10 mins
Grade	**2.67** out of 5.00 (**53.33**%)
Feedback	If you have any remarks about the questions or the grading, please submit them using this form!

Question 1

Incorrect
Mark 0.00 out of 1.00

Flag question

Question text

Which of the following statements about real-world networks and Erdös-Rényi random graphs are correct? Check all that apply.
Question 1 Answer
a.
Short paths typically exist in real-world networks, but not in Erdös-Rényi random graphs.
b.
In an Erdös-Rényi random graph with n nodes, where edges exist with probability p independetly from one another, the average centrality measure of a node is linearly proportional to p.
c.
Short paths are greedily discoverable in real-world networks.
d.
Short paths are easily discoverable in both real-world networks and Erdös-Rényi random graphs with decentralized algorithms.

Feedback

The correct answers are: In an Erdös-Rényi random graph with n nodes, where edges exist with probability p independetly from one another, the average centrality measure of a node is linearly proportional to p., Short paths are greedily discoverable in real-world networks.

Question 2

Partially correct
Mark 0.67 out of 1.00

Flag question

Question text

What can you say about the graph that is represented by the following set of edges:
(1, 3)
(2, 3)
(1, 5)
(4, 5)
Check all that apply.
Question 2 Answer
a.
This graph does not have multi-edges (multiple edges connecting the same nodes)
b.
The graph does not have self-loops
c.
This is a weighted graph
d.
This is a bipartite graph

Feedback

The correct answers are: This is a bipartite graph, The graph does not have self-loops, This graph does not have multi-edges (multiple edges connecting the same nodes)

Question 3

Incorrect
Mark 0.00 out of 1.00

Flag question

Question text

Which of these statements about bipartite graphs are true?
Check all that apply.
Question 3 Answer
a.
Bipartite graph can be directed
b.
Projection of a bipartite graph cannot be a bipartite graph
c.
The sum of degrees of nodes in one partition is always equal to the sum of degrees of nodes in the other partition
d.
Projection of a bipartite graph can be a complete graph (i.e. a graph without self-loops in which every pair of nodes is connected with an edge)

Feedback

The correct answers are: Bipartite graph can be directed, The sum of degrees of nodes in one partition is always equal to the sum of degrees of nodes in the other partition, Projection of a bipartite graph can be a complete graph (i.e. a graph without self-loops in which every pair of nodes is connected with an edge)

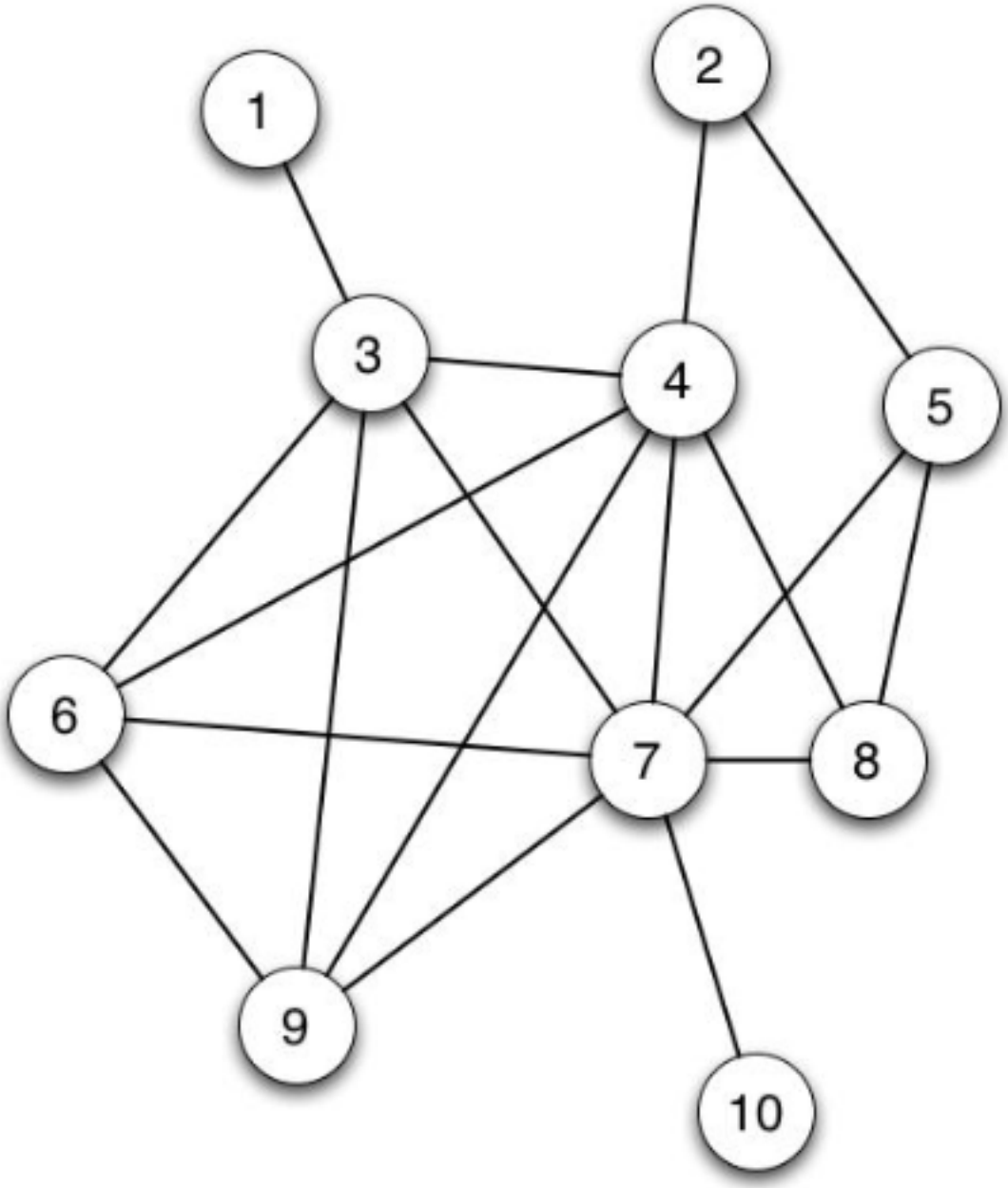
Question 4

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Given the unweighted graph in the image, which statements are true? Check all that apply.



Question 4 Answer

a.

Clustering coefficient of node 8 is 0.7

b.

Clustering coefficient of node 6 is higher than clustering coefficient of node 7.

c.

Clustering coefficient of node 2 is 0

d.

Clustering coefficient of node 9 is 1

Feedback

The correct answers are: Clustering coefficient of node 6 is higher than clustering coefficient of node 7., Clustering coefficient of node 2 is 0, Clustering coefficient of node 9 is 1

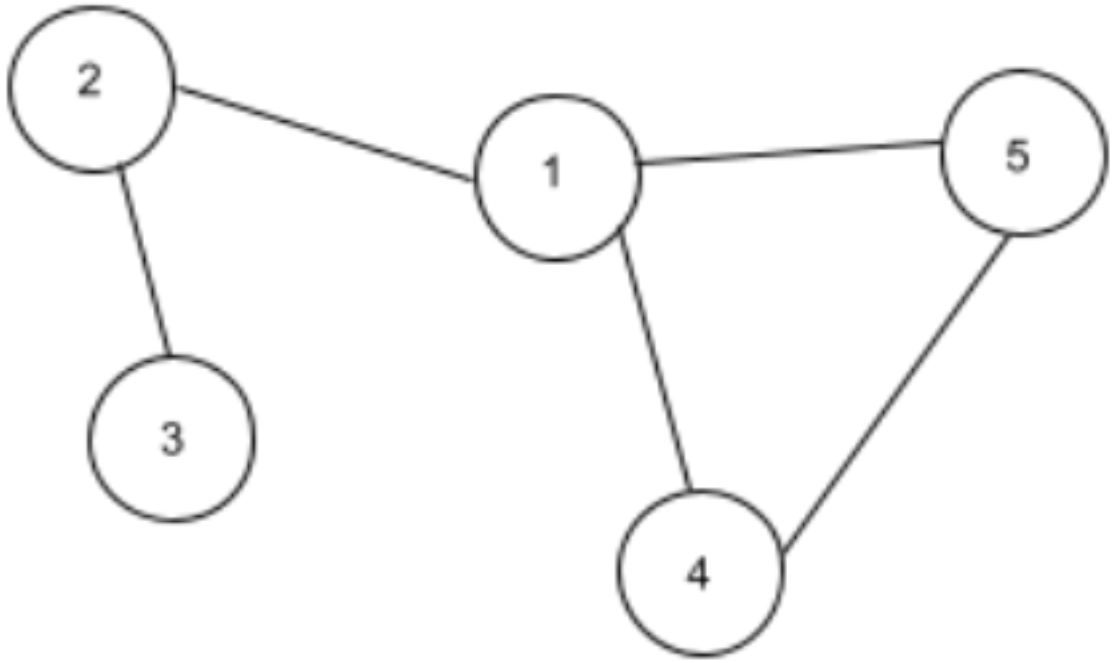
Question 5

Correct
Mark 1.00 out of 1.00

Flag question

Question text

Given the unweighted graph in the image below, what can you say about centrality measures? Check all that apply.



Question 5 Answer
a.
Closeness centrality of node 3 is smaller than the one of node 1
b.
Closeness centrality of node 5 is 0.125
c.
Betweenness centrality of node 3 is 0
d.
Betweenness centrality of node 1 is bigger than the one of node 4

Feedback

The correct answers are: Betweenness centrality of node 1 is bigger than the one of node 4, Betweenness centrality of node 3 is 0, Closeness centrality of node 3 is smaller than the one of node 1