Quizzes

1. Which data visualization technique is useful for highlighting connectivity patterns in network data?

a. Pie chart
b. Matrix view
c. Scatter plot
d. Bar chart

2. Which of the following is an advantage of using a binary format ("serialization") over JSON for large-scale data? Check all that apply.

a. Binary formats enable more efficient storage of numerical values
b. Binary formats are easier to update
c. Binary formats are human-readable
d. Binary formats often require less expensive parsing

3. 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?

d.

```
<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>
```

5. 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]

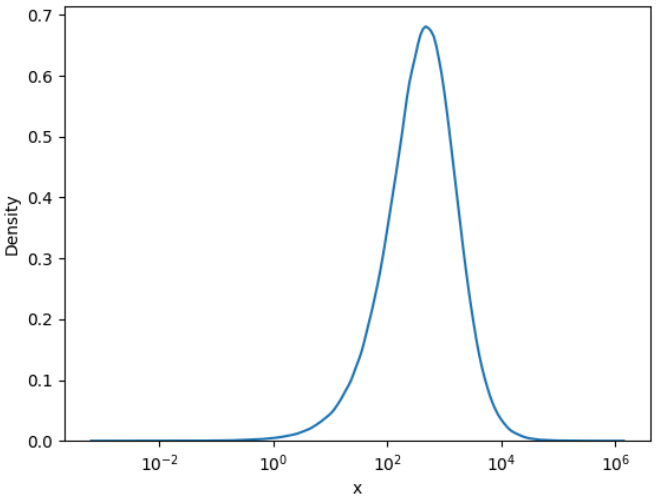
The Orders table contains three columns, order_id, customer_id, and item:

order_id, customer_id, item
[1, 1, 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 5Answer

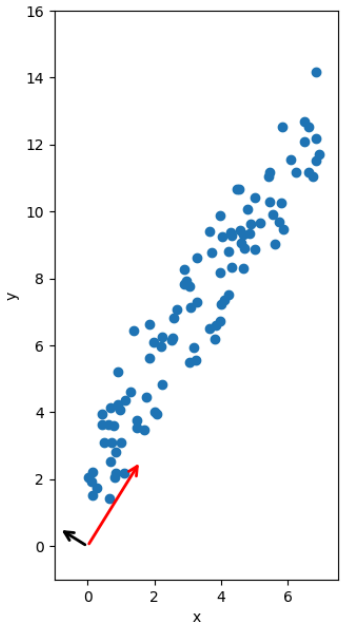
a.
name
[John]
[David]
[Alice]

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



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

2. Given this plot, check all that apply.



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 second principal component for the dataset

c. Black arrow could be returned by the PCA algorithm as the first principal component for the dataset

d. Red arrow could be returned by the PCA algorithm as the first principal component for the dataset

3. 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?

a. The CCDF plot will be monotonically decreasing

b. Binning of data points is required to generate the CCDF plot

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

4. 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?

a. Box plot

b. Histogram with a logarithmic x-axis

c. Histogram with a linear x-axis

d. Cumulative distribution function (CDF) plot

5. 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.

a. Histograms
b. Scatter plot
c. Bar chart
d. Line chart

1. Which of the following statements are true? Assume you are working with non-negative data. Check all that apply.

a. For skewed distribution, median can be smaller, bigger or equal to the mean
b. For skewed distribution, median is smaller than mean
c. For heavy-tailed distribution, median can be smaller than mean
d. For symmetrical distributions, median is always the same as mean

2. 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?

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

3. 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?

a. The true mean may be infinite, so median is better-suited than mean, but variance can still be reported.
b. Both true mean and true variance may be infinite, so only robust statistics like median should be reported.
c. The estimated α is close enough to 2 such that the true mean and variance are probably finite and can be reported.
d. The true variance is likely to be finite, so it's ok to report the estimated variance.

4. 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.

a. If we care about overall performance, micro-average is a better metric than macro-average
b. If we care about performance per each class, micro-average is a better metric than macro-average

c. If we care about performance per each class, macro-average is a better metric than micro-average
d. If we care about overall performance, macro-average is a better metric than micro-average

5. 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?

a. The p-values only suggest how likely each dataset was under H0. More evidence is needed to determine if H0 is 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 contradictory p-values mean H0 is equally likely to be true or false.
d. The first p-value implies that the alternative hypothesis is true. The second p-value implies that H0 is true.

1. 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?

a. Productivity increased by 3 units from time 1 to time 2 in both groups.
b. The treatment group achieved 1/3 of the productivity of the control group.
c. The treatment group was 3 units more productive than the control group after the program.
d. Productivity increased by 3 units more for the treatment group than the control group from time 1 to 2.

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

a. Low R^2 score indicates that the predictors have no statistically significant correlation with the outcome
b. If we have negative predictors, we should not perform linear regression
c. 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
d. Intercept represents the predicted value of the outcome when all predictor variables are set to zero

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

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

4. 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?

a. A $1 increase in advertising spending predicts a 0.5 absolute increase in y.
b. The transformation implies that traffic is now modeled as a logarithmic, not linear, function of advertising spending
c. The 0.5 coefficient means advertising spending has no relationship with website traffic.
d. A $1 increase in advertising spending predicts a 0.5% increase in website traffic y.

5. 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?

a. The effect of the protein supplement cannot be hypothesized.
b. Taking the protein supplement decreases muscle mass gained.
c. Taking the protein supplement has no effect on muscle mass gained.
d. Taking the protein supplement increases muscle mass gained.

1. 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.

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

2. When performing a randomized experiment (with a treatment group and a control group), which of the following statements is true?

a. Unobserved confounders may threaten the validity of your conclusions
b. 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
c. All participants have the same probability of being assigned to the treatment group
d. Randomized experiments are usually harder to replicate than observational studies

3. Which of the following statements about determining causality from observational data are true? Check all that apply.

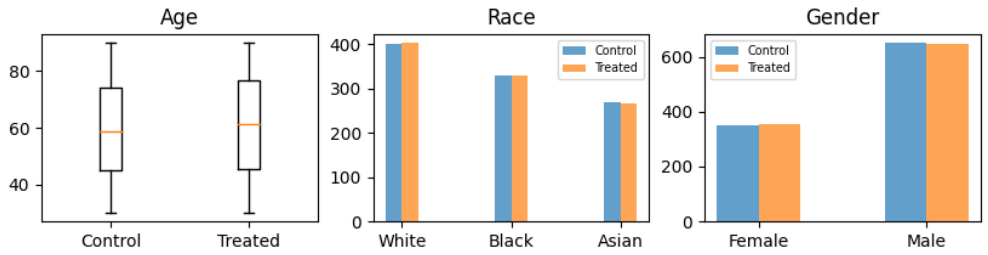
a. Studies should match treated and control units on as many covariates as possible to minimize confounding.

b. Randomized experiments are usually cheaper to conduct than observational studies, as most modern data is found data

c. Sensitivity analysis quantifies the potential impact of unmeasured confounding on study conclusions.

d. Large sensitivity analysis parameter (Γ) means that two subjects with the same unobserved covariates have vastly different probabilities of getting the treatment.

4. 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.



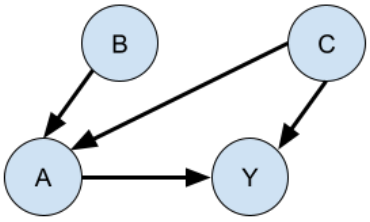
a. If you assume there are no unobserved confounders (as you assume in the "naive model"), to estimate the causal effect of "treated" on "cured", you can directly perform regression analysis, using treatment, age, race, and gender as dependable variables and "cured" as the outcome

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

c. 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"

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

5. 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.



a. A has causal effect on Y

b. C has causal effect on Y

c. B is a confounder when observing effect of A on Y

d. C is a confounder when observing effect of A on Y

1. 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

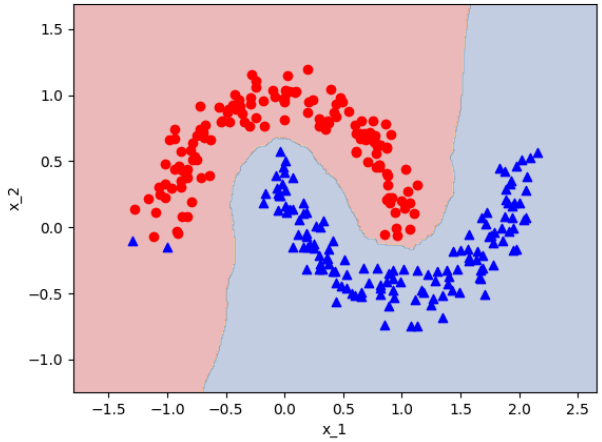
tree and the number of legs appear as a leaf. Which of the following statements are true? Check all that apply.

a. Pruning the leaf corresponding to "number of legs" would result in a tree from a smaller model family, with more bias but less variance.
b. Tail length provides the most information gain compared to other features, so it was selected as the root node.
c. The tree has low variance since all features are considered.
d. Tail length provides no useful information compared to body size, so the tree must be overfitting.

2. 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?

a. Studying 1 more hour increases the log-odds of passing by 0.5.
b. Studying 1 more hour increases the probability p of passing by 0.5.
c. Studying 1 more hour doubles the odds of passing.
d. Studying 1 more hour doubles the probability p of passing.

3. 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.

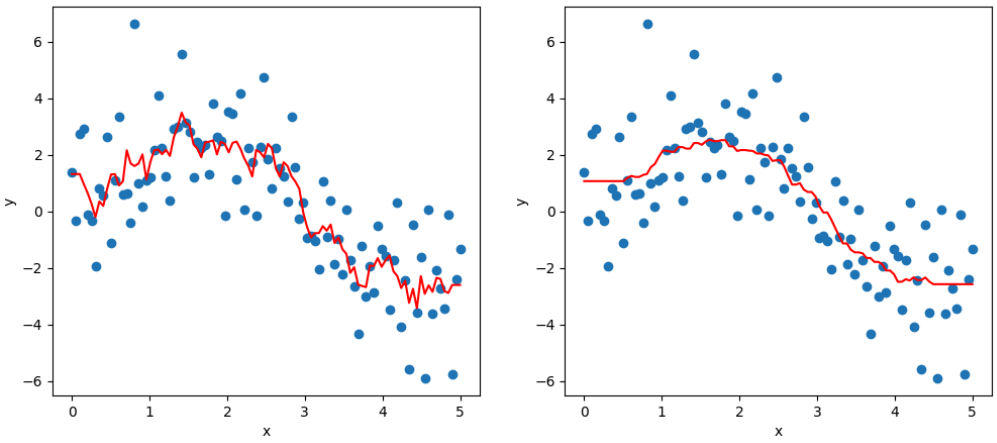


a. Logistic regression
b. K nearest neighbours with K=10
c. K nearest neighbours with K=2
d. None of the above

4. Which of the following statements about the bias-variance tradeoff are true? Check all that apply.

a. Adding more relevant features to a logistic regression model decreases bias but increases variance.
b. Increasing the number of trees in boosted decision trees decreases the bias.
c. Adding depth to a decision tree decreases bias but increases variance.
d. Increasing k in k-NN decreases bias but increases variance.

5. 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.



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

1. 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

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

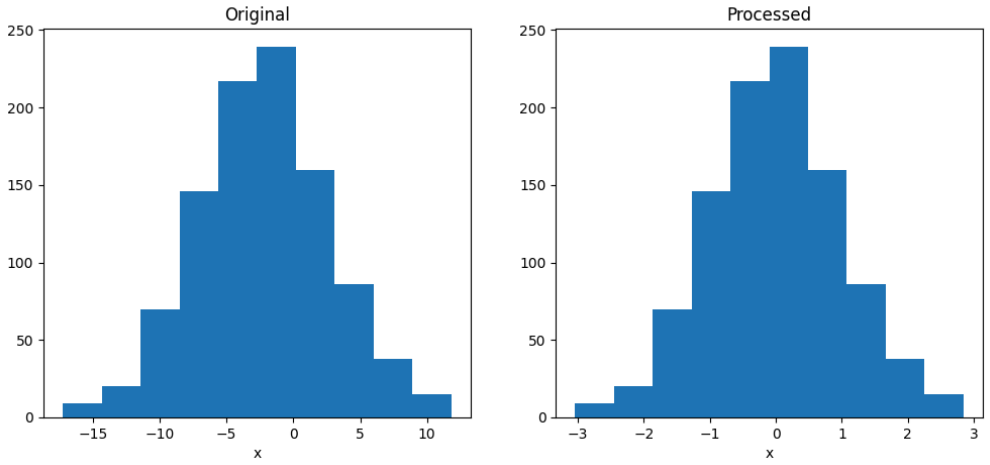
2. Which of these statements about the precision/recall curve are correct? Check all that apply.

a. The precision/recall curve is sensitive to the classification threshold
b. The precision/recall curve for a random classifier corresponds to the identity line (y=x)
c. The area under the precision/recall curve for a random classifier depends on number of positive and negative samples in the dataset
d. The area under the precision/recall curve for a perfect classifier is 1

3. 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.

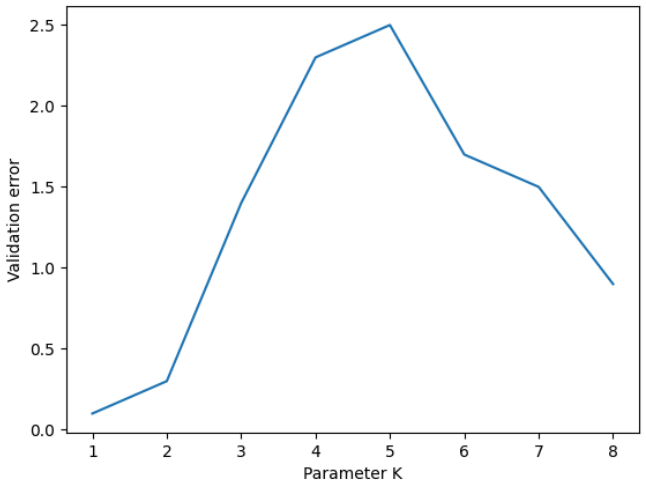
a. While online feature selection methods are fast to apply, they often don't account for the interdependence or relationships between features.
b. Feature selection can lower the chance of overfitting and allows for more efficient training.
c. Feature ranking is often regarded as a superior approach, compared to online feature selection methods. This is because feature ranking considers features as a collective entity rather than evaluating them individually.
d. 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.

4. 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?



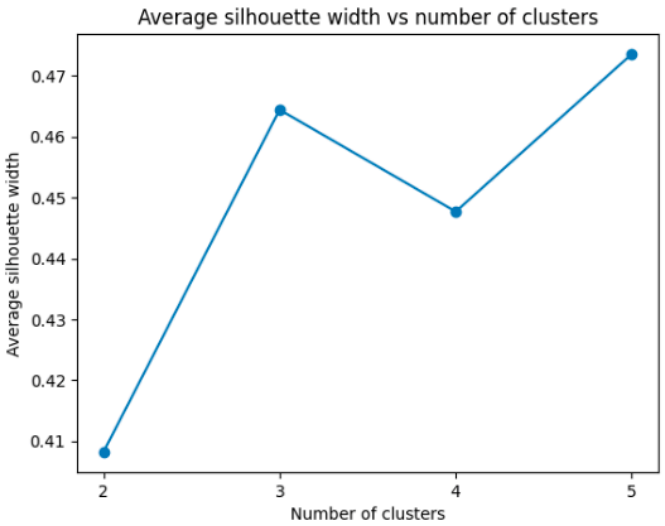
a. Standardization
b. Min-max scaling
c. Logarithmic scaling
d. None of the above

5. 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?



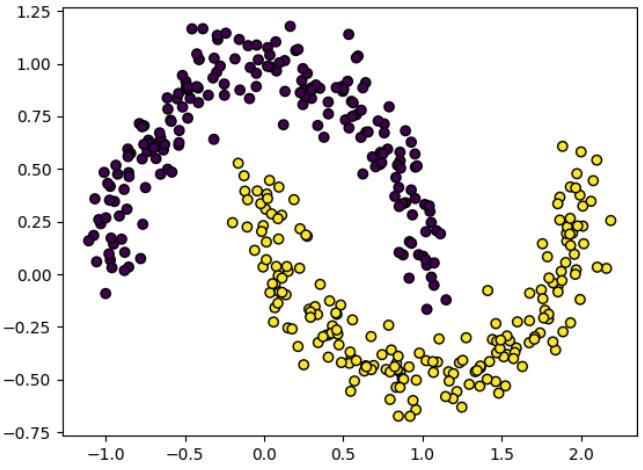
a. K=1
b. K=5
c. K=4
d. K=8

1. Given the silhouette plot in the image below, the best choice of the number of clusters is:



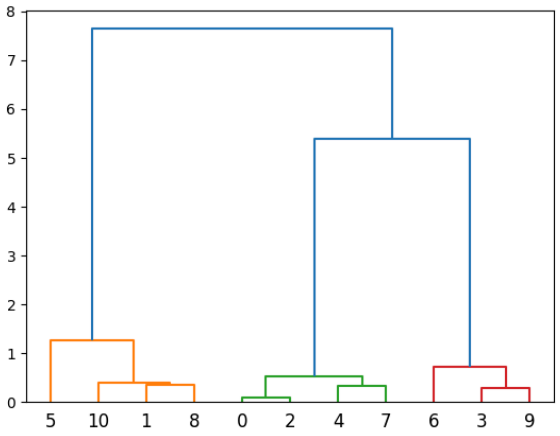
a. 4
b. this cannot be a plot of an average silhouette width.
c. 3
d. 5

2. Which clustering algorithm could have achieved the result on the plot? Assume that different-coloured points belong to different clusters. Check all that apply.



a. DBSCAN
b. K-means
c. K-means++
d. None of the above

3. Which conclusions can be drawn from the given dendrogram? Check all that apply.



a. First two points that could have been grouped were point 0 and point 2
b. Euclidean distance has certainly been used to determine distance between centroids
c. The dataset from the analysis has 11 data points
d. The analysis could have been done by performing agglomerative clustering

4. Which of the following are advantages of the DBSCAN clustering algorithm over k-means clustering? Check all that apply.

a. DBSCAN does not require specifying the number of clusters k in advance.
b. DBSCAN clustering results do not depend on the initialization, unlike k-means.
c. DBSCAN can detect clusters of arbitrary shapes, unlike k-means which assumes convex clusters.
d. DBSCAN has faster runtime than k-means on large datasets.

5. Which statements about k-means are true? Check all that apply.

a. K-means clusters are convex shapes, which means it is robust to noisy data and outliers.
b. Lloyd's algorithm is a greedy algorithm for solving k-means that is robust to centroid initialization.
c. The k-means problem is NP-hard but initializations can help avoid bad solutions found by the iterative algorithm.
d. K-means++ improves on random sampling by spreading out initial centers, choosing them more widespread than expected to be found by independent random sampling.

1. 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 logarithm, that is the logarithm with base e. Report the outcome up to 2 decimal points of precision.

a. 0
b. 1.39

c. 1
d. 0.69

2. 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

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

3. Which of the following statements about text tokenization are true? Check all that apply.

a. Lemmatization is most useful for languages with little morphology (i.e., where words aren't inflected in many different ways).
b. Tokenization is more challenging for languages without explicit word delimiters like whitespace, such as Chinese, or those with compound words, like German.
c. Lemmatization maps words to normalized lexicon entries.
d. Stemming reduces sparsity but loses information such as the part of speech (i.e., whether a word is a verb, noun, adjective, etc.).

4. Which of the following statements are true for character encodings:

a. Standard ASCII encoding can be used when you are working with text containing only English alphabet
b. It is safe to read UTF-8 encoded data with ASCII
c. UTF-8 encoding can encode only 256 different characters
d. Latin-1 always encodes each character with 1 byte

5. Which of the following transformations can be used to penalize the words that appear often in the corpus? Check all that apply.

a. Lemmatization
b. Row normalization
c. IDF
d. Stopword removal

1. A document classification model achieves 99% training accuracy but only 60% test accuracy. Which methods could help address this overfitting? Check all that apply.

a. Applying L2 regularization
b. Adding more words to the vocabulary
c. Training on fewer documents

d. Substituting the TF-IDF matrix with less sparse matrix representations obtained from dimensionality reduction techniques like LSA or LDA.

2. When performing topic detection on a collection of long text documents, which of the following statements are true? Check all that apply.

a. Matrix factorization via LSA mitigates the curse of dimensionality by providing features in topic space.
b. TF-IDF weighting helps overcome issues with high frequency non-content words.
c. Decision trees trained on bag-of-words features are scalable to long texts as bag-of-words feature extraction helps normalize for document length
d. Standard clustering methods like k-means struggle due to the curse of dimensionality.

3. Which of the following matrices U and S cannot be obtained by the singular value decomposition of an unknown 2x2 matrix, following the standard conventions? (Here matrices are represented as lists of rows.) Check all that apply.

a. U = [[1, -1/2],
[0, sqrt(3)/2]]
b. S = [[1, 0],
[0, 2]]
c. U = [[1, -2],
[2, 1]]
d. S = [[1, 1],
[1, 1]]

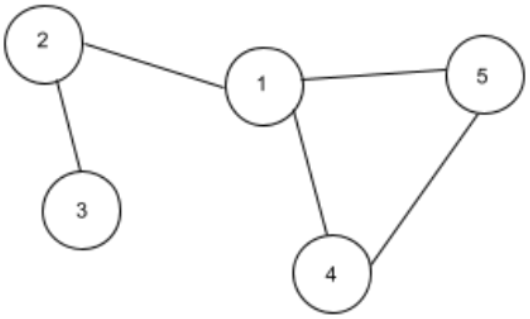
4. Which of these statements are true about document representation for document retrieval task? Check all that apply.

a. Documents can be represented as rows of the TF-IDF matrix
b. Documents cannot be represented as rows of a matrix obtained via LDA
c. Documents cannot be represented as rows of the bag-of-words matrix
d. Documents can be represented as rows of a matrix obtained via LSA

5. Word2vec computes vector representations for words by factoring a large matrix M associating words with contextual windows. Which of the following limitations of representing words as one-hot vectors(with one entry per vocabulary word, all of which are zero, except one) does this approach resolves? Check all that apply.
Question 5Answer

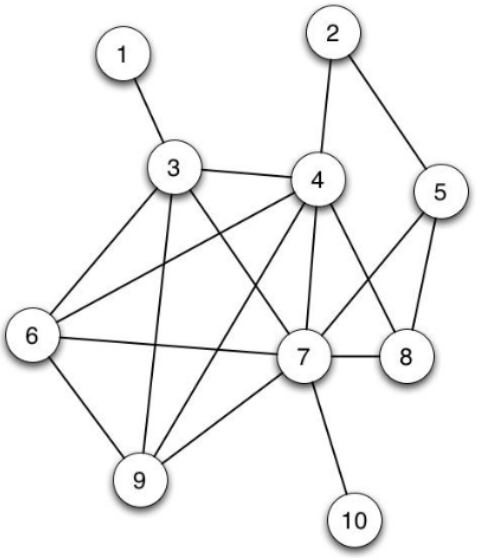
a. Inability of representing word order information in the document to which the method is applied
b. Inability to represent multiple word senses
c. Sparsity resulting from large vocabularies
d. Word representations not capturing semantical closeness of words

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



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

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



a. Clustering coefficient of node 9 is 1
b. Clustering coefficient of node 2 is 0
c. Clustering coefficient of node 8 is 0.7
d. Clustering coefficient of node 6 is higher than clustering coefficient of node 7

3. 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.

a. The graph does not have self-loops
b. This graph does not have multi-edges (multiple edges connecting the same nodes)
c. This is a bipartite graph
d. This is a weighted graph

4. Which of the following statements about real-world networks and Erdös-Rényi random graphs are correct? Check all that apply.

a. Short paths typically exist in real-world networks, but not in Erdös-Rényi random graphs.
b. Short paths are greedily discoverable in real-world networks.
c. Short paths are easily discoverable in both real-world networks and Erdös-Rényi random graphs with decentralized algorithms.
d. In an Erdös-Rényi random graph with n nodes, where edges exist with probability p independently from one another, the average clustering coefficient of a node is linearly proportional to p.

5. Which of these statements about bipartite graphs are true?
Check all that apply.

a. The sum of degrees of nodes in one partition is always equal to the sum of degrees of nodes in the other partition
b. 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)
c. Projection of a bipartite graph cannot be a bipartite graph
d. Bipartite graph can be directed

Mock Exam

1. You are given two relational tables, Authors and Books.

The Authors table contains three columns, 'author_id', 'name' and 'country':

author_id name country
1 Haruki Murakami Japan
2 George Orwell UK
The Books table contains three columns, 'book_id', 'title', and 'author_id':

book_id title author_id
101 Kafka On the Shore 1
102 1984 2
103 Norwegian Wood 1
How would you best represent this data, taking into consideration both tables, in the JSON format? Select one:

b.

```
{
  "authors": [
    {
      "author_id": 1,
      "name": "Haruki Murakami",
      "country": "Japan",
      "books": [
        { "book_id": 101, "title": "Kafka on the Shore" },
        { "book_id": 103, "title": "Norwegian Wood" }
      ]
    },
    {
      "author_id": 2,
      "name": "George Orwell",
      "country": "UK",
      "books": [
        { "book_id": 102, "title": "1984" }
      ]
    }
  ]
}
```

2. You are building a classifier and want to rank features based on how informative they are for predicting the target variable. One of the methods you consider is mutual information. Which of the following statements are true regarding mutual information? Select one or more:

a. Mutual information ranks features based only on linear relationships, similar to Pearson correlation, which makes it unsuitable for detecting nonlinear dependencies
b. Mutual information can only be computed for numerical variables and cannot be applied to categorical data.
c. Mutual information is zero if and only if the variables are statistically independent.
d. Mutual information measures both linear and nonlinear dependencies between features and the target, unlike methods such as Pearson correlation, which only capture linear relationships.

3. Which of the following statements are true? Select one or more:

a. The K-means algorithm is inherently limited in the shapes of clusters it can identify, as it can only successfully find clusters that are convex.
b. Clustering is an unsupervised task used to learn classification y based on labeled training data (X,y).
c. Soft Clustering assigns a single, absolute cluster label to a data point, while Hard Clustering assigns a probability distribution over all clusters.
d. Dimensionality Reduction is an unsupervised task where the goal is to compute a function f(X) that results in a continuous, simpler representation y.

4. Which of the following statements correctly describes the concepts used to represent clusters or define their "nearness"? Select one or more:

a. A cluster can be represented by its Centroid (an existing data point closest to others, useful in non-Euclidean spaces) or by a Clustroid (an artificial point calculated as the average of all data points).
b. One way to define "nearness" (Intercluster Distance) of two clusters is by calculating the minimum distance between any two points, one from each cluster (known as single-linkage).
c. Agglomerative Hierarchical Clustering is highly resistant to the curse of dimensionality because the cluster distance calculation is based only on a select few cluster representatives (centroids/clustroids).
d. The Agglomerative approach begins with all data points as individual clusters and iteratively merges them, while the Divisive approach starts with one cluster and recursively splits it.

5. You want to remove very frequent words (stopwords) using a simple heuristic "remove words that appear in at least p% of documents", as these words do not provide a meaningful signal for your ML model. Thus, you implement the following function:

```
from typing import List

def remove_stopwords_by_docfreq(docs: List[str],

                                tokenizer: Callable,
```

```
                        p: float = 0.75):
    N = len(docs)
    tokenized_docs = [tokenizer(doc) for doc in docs]

    # compute document frequencies
    docfreq = Counter()
    for doc in tokenized_docs:
        for w in set(doc):
            docfreq[w] += 1

    allowed = {w for w, df in docfreq.items() if df / N >= p}

    new_docs = []
    for doc in tokenized_docs:
        new_docs.append([w for w in doc if w in allowed])
    return new_docs
```

What is the mistake in the implementation above? Select one:

a. It should define docfreq as the total count of each word, not per-document presence.
b. It uses >=p instead of <=p, so it keeps the very frequent words instead of removing them.
c. There is no bug.
d. It should use raw term frequency instead of document frequency.

6. You are working with a text dataset that contains three sentences:

The weather today is sunny!
It rained yesterday.
I am going shopping today.
You want to analyze the similarity between the documents so you perform simple preprocessing consisting of punctuation removal, lowercasing all letters and splitting by space to obtain words. Then, you construct the TF-IDF matrix and use document

features to compute the cosine similarity between the documents. Select the correct statement(s). Select one or more:

a. Cosine similarity between (2) and (3) is equal to zero
b. Cosine similarity between (1) and (2) is equal to zero
c. Cosine similarity between (1) and (2) is non-zero
d. Cosine similarity between (1) and (3) is equal to zero

7. You compute rdd2 once and use it for multiple actions:

```
rdd2 = rdd1.map(f1)

list1 = rdd2.filter(f2).collect()

list2 = rdd2.filter(f3).collect()
```

What is the best fix to avoid recomputing map(f1) multiple times? Select one:

a. Call rdd1.persist() instead of rdd2.persist()
b. Add rdd2.persist() after the map
c. Replace collect() with count()
d. Replace both filters with one filter

8. Consider the following operation on an RDD:

map(f),
filter(g),
reduce(h),
count(),
take(10).

Which option correctly classifies them into transformations and actions? Select one:

a. Transformations: map, filter, reduce — Actions: count, take
b. Transformations: map, filter — Actions: reduce, count, take
c. Transformations: map, filter, count — Actions: reduce, take
d. Transformations: map, filter, take — Actions: reduce, count, take

9. In a Spark application, you are tasked with counting the total number of log entries containing the string "ERROR" across a large, distributed RDD named log_rdd. You write the following two PySpark code snippets to attempt this task:

Snippet 1 (Using Standard Variable):

```
error_count_1 = 0
```

```
def count_errors_1(line):
    # This closure attempts to update the global variable
    if "ERROR" in line:
        error_count_1 += 1

log_rdd.foreach(count_errors_1)
print(error_count_1)
```

Snippet 2 (Using Spark Accumulator):

```
sc is the SparkContext
error_count_2 = sc.accumulator(0)

def count_errors_2(line):
    # This closure uses the Spark-provided tool for aggregation
    if "ERROR" in line:
        error_count_2.add(1)

log_rdd.foreach(count_errors_2)
print(error_count_2.value)
```

Which of the following statements correctly describe the final result obtained from these snippets in a distributed Spark cluster environment? Select one or more:

a. Snippet 1 will always print the correct total number of "ERROR" entries because Python's global keyword ensures variable updates are synchronized across all worker nodes.
b. Snippet 2 will always print the correct total number of "ERROR" entries.
c. Snippet 1 will typically print zero (0) or an incomplete count because the error_count_1 variable is copied to each worker process, and the updates made on the workers are not communicated back to the driver program.

d. Snippet 2 correctly uses an Accumulator, which is designed specifically as a write-only, distributable variable for efficiently aggregating values (like counters and sums) from the worker processes back to the driver.

10. Below is a simplified cluster-centroid table (z-scored units) for the 4 post-mortem curve features:

Cluster	Premortem Mean	Short-Term Boost	Long-Term Boost	Halving Time
C1	-0.2	+1.8	-0.1	-1.2
C2	-1.1	-0.4	-0.7	+0.1
C3	+0.3	+0.9	+1.7	+0.9
C4	+1.9	-0.8	-0.6	+0.2

Which interpretation is most accurate? Select one:
a. C2 shows people who were extremely famous before death
b. C4 represents people with very fast decay of attention after death
c. C3 represents "rise" curves: moderate short-term boost and unusually high long-term boost
d. C1 represents people with low short-term reaction but persistent long-term interest

11. A linear classifier for sentiment has weight vector β (in $\mathbb{R}^3$) for three words {"great", "awful", "movie"}. You observe the following in your training set:

Word	Appears only in documents with sentiment:
great	5
awful	1
movie	1-5

The learned weights without regularization are:

$\beta = (\beta_1 = 12.0, \beta_2 = -10.0, \beta_3 = 0.1)$
You now add L2 regularization with $\lambda = 0.5$ to the loss:

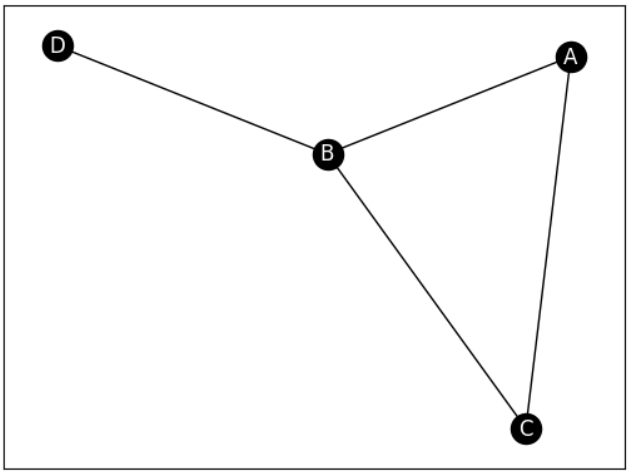
$L(\beta) = \Sigma_i(y_i - x_i^T\beta)^2 + \lambda||\beta||^2_2$
Question: Which of the following outcomes is impossible after retraining? (Multiple answers) Select one:

a. β_1 decreases in magnitude
b. The training error increases
c. β_2 decreases in magnitude
d. The test error decreases
e. β_3 becomes exactly zero

12. A researcher is plotting the distribution of city populations, which follows a heavy-tailed power law. She decides to use log-log axes. Why is this transformation helpful? Select one:

a. It normalizes the data so all values are between 0 and 1
b. It makes the distribution appear bell-shaped and easier to interpret
c. It removes outliers automatically from the dataset
d. It linearizes the power-law relationship, making the slope represent the exponent

13. You work for a logistics company and you are tasked with analyzing their distribution network. You are given a graph containing nodes (cities) A, B, C and D and edges (roads) which connect them and you want to analyze the load of the traffic through the cities in the network. You know that the company always uses the shortest path between two cities when planning the transfer of products so you opt to use the betweenness centrality. Given the graph in the picture, what is the betweenness centrality of the nodes in the network (answers follow order [node A, node B, node C, node D])? Select one:



a. [1/3, 2/3, 1/3, 0]
b. [0, 2/3, 0, 0]
c. [2/3, 1, 2/3, 1/6]
d. [0, 3, 0, 0]

14. We use a logistic regression to model the probability p of a disease being present. The result is:
$\text{logit}(p) = -3 + 0.02 \times X_1 + 0.5 \times X_2$
where X_1 is the person's age in years, and X_2 indicates if the person is a smoker ($X_2 = 1$) or not ($X_2 = 0$). Which of the following statements is true? Select one:

a. The odds of having the disease for smokers are approximately three times that for non-smokers.
b. Aging by one year increases the odds by approximately 2%.
c. The log-odds increase by 50% for smokers.
d. Being a non-smoker increases the log-odds by 0.5.

15. In propensity score matching, two subjects with equal propensity scores will have: Select one:

a. Equal outcomes
b. Identical treatment effects

c. Identical covariates
d. Similar distribution of observed covariates

16. Which of the following are valid examples of confounders in an observational study? Select one:

a. A variable that affects both the probability of receiving the treatment and the outcome
b. A variable that affects treatment assignment but has no effect on the outcome
c. A variable that affects the outcome but is completely unrelated to treatment assignment
d. A variable measured after the treatment is applied

17. Which of the following statements about logistic regression are correct? Select one:

a. The cost function of logistic regression is concave.
b. Logistic regression assumes that each class's points are generated from a Gaussian distribution.
c. Models the log-odds as a non-linear function.
d. Logistic regression is equivalent to a neural network without hidden units and using cross-entropy loss.

18. When is accuracy NOT an appropriate evaluation metric? (Select all that apply) Select one or more:

a. When false positives and false negatives have different costs
b. When all errors are equally important and classes are balanced
c. When you have a binary classification problem
d. When classes are highly imbalanced/skewed
e. When detecting rare events like fraud

19. Which of the following statement(s) are true regarding cross-validation? Select one or more:

a. Reporting performance from the same cross-validation used to tune hyperparameters yields an unbiased estimate of generalization error
b. It is guaranteed to prevent overfitting
c. It is useful when there is not enough data for a train, validation, and test split
d. It is often used to select hyperparameters

20. You deploy a medical classifier for a rare disease (1% positives). On a test set of 1,000 patients, your model outputs:

Predicted positive: 50
True positive (TP): 5

False positive (FP): 45
Predicted negative: 950
False negative (FN): 5
True negative (TN): 945
Which statement is more accurate? Select one:

a. Precision and recall are both very high; the model is excellent.
b. F1 is 0.01; the model is close to random.
c. Accuracy is low; the model is useless.
d. Accuracy is high, but precision is low; the model is risky if false alarms are costly.