cap5771sp25
April 29, 2025

Exam 3 version c

Question 1. According to data-wrangling-v2 slide 25 and associated lectures, which of the following can be determined by a chi-square test in data wrangling?

(a) `Correlation in numeric data`

(b) `The mean of a type of numeric data`

(c) `Correlation in nominal data`

(d) `Causal relations in numeric data`

(e) `Causal relations in nominal data`

Question 2. What is the support of an itemset in frequent pattern mining according to the lecture slides?

(a) `The fraction of transactions in the dataset that contain the itemset.`

(b) `The number of times an itemset appears in a single transaction.`

(c) `The probability that an itemset appears in at least one transaction.`

(d) `The correlation coefficient between two itemsets.`

(e) `The percentage increase in item purchases after introducing a new item.`

Question 3. According to Lecture 1 on Data Exploration, which of the following is NOT an attribute type in data classification?

(a) `Ratio`

(b) `Ordinal`

(c) `Relational`

(d) `Nominal`

Question 4. According to Data wrangling lecture, In a machine learning project, two different techniques were used for attribute construction:

Combining features: Created a new feature by merging "purchase frequency" and "average transaction value" to better capture customer spending behavior. Data discretization: Transformed a continuous "age" variable into categorical bins such as "young," "middle-aged," and "senior."

Considering these techniques, how do combining features and data discretization differ in their impact on a machine learning model?

(a) `Data discretization creates more detailed numerical features, while combining features simplifies the dataset.`

(b) `Combining features enhances predictive power by creating richer representations, while data discretization simplifies models and improves interpretability.`

(c) `Combining features primarily reduces dataset size, while data discretization improves model accuracy by increasing feature granularity.`

(d) `Both techniques are exclusively used for reducing overfitting in models.`

(e) `Both combining features and data discretization primarily serve the same purpose: reducing redundancy in the dataset.`

Question 5. According to the Data Wrangling lecture slides (Slide 22), what is an advantage of using a data lake over a traditional data warehouse?

(a) `Data warehouses are faster than data lakes for all types of queries.`

(b) `Data lakes require predefined schemas before storing data.`

(c) `Data lakes store data in raw form, allowing flexible analysis.`

(d) `Data lakes store only structured data.`

Question 6. According to the lecture on classification, what is the main drawback of using accuracy as the sole metric for evaluating classification models?

(a) `Accuracy is the only metric needed to evaluate model performance.`

(b) `High accuracy always indicates a well-performing model.`

(c) `Accuracy automatically adjusts for class imbalance.`

(d) `Accuracy can be misleading when dealing with imbalanced datasets, as it may favor the majority class.`

(e) `Accuracy directly measures how well a model generalizes to new data.`

Question 7. According to the lecture on data exploration, which one of the following is not apart of "The 5 Vs of Big Data"?

(a) `Variety`

(b) `Veracity`

(c) `Volume`

(d) `Visualization`

(e) `Velocity`

Question 8. According to lecture, what is NOT an advantage to having a dataset with low dimensions?

(a) `Feature selection.`

(b) `Data is less sparse.`

(c) `Clustering is more meaningful.`

(d) `Visualization is simpler.`

(e) `Reduces space and time complexity of data mining.`

Question 9. According to the introduction to data science lecture, what is defined as the velocity of big data?

(a) `How fast data is created.`

(b) `42 nm/ps.`

(c) `How fast you receive the data.`

(d) `The speed of the data.`

(e) `Change over time.`

Question 10. According to the lecture on data preprocessing, which of the following is NOT mentioned as a major task?

(a) `Data integration.`

(b) `Data encryption.`

(c) `Data cleaning.`

(d) `Data transformation and discretization.`

(e) `Data reduction.`

Question 11. Which of the following is an outlier detection method that uses statistical techniques?

(a) `Interquartile Range (IQR)`

(b) `Feature Engineering`

(c) `Data Binning`

(d) `Principal Component Analysis (PCA)`

(e) `Data Normalization`

Question 12. According to the lecture on Data Exploration, what is the trouble with relying solely on summary statistics when analyzing a dataset?

(a) `Sometimes two datasets can have identical summary statistics but very different distributions.`

(b) `Summary statistics are only valid if the dataset follows a normal distribution.`

(c) `Summary statistics always give an incomplete picture of the data.`

(d) `Please don't pick me`

(e) `Summary statistics are only useful for small datasets.`

Question 13. Based on the lecture on clustering, imagine this scenario: A data analyst is working on a dataset containing customer information and wants to segment customers into meaningful groups based on their purchasing behavior. Which of the following clustering methods is the best for this task?

(a) `Utilizing hierarchical methods to build a tree-like structure of nested clusters, allowing flexible exploration of customer segments.`

(b) `Implementing density methods to detect clusters of varying shapes and sizes by identifying dense regions in the data.`

(c) `Applying partitioning methods to divide the dataset into \(k\) distinct clusters, ensuring each data point belongs to exactly one cluster.`

(d) `Using basic concepts of cluster analysis to understand different clustering techniques before selecting an appropriate method.`

(e) `Using density-based and grid-based methods to improve clustering efficiency by organizing data into grids before detecting dense regions.`

Question 14. You are given a dataset containing customer purchase information of a store. Your boss wants to compare how many customers fall into different age groups (e.g., 18-25, 26-35, 36-45, etc.) and also analyze the distribution of total purchase amounts. Based on your learnings from the data exploration lecture, determine which visualization techniques are the best ones for your use case.

(a) `Histogram for purchase amounts and a bar chart for age groups`

(b) `Pie chart for purchase amounts and a histogram for age groups`

(c) `Bar chart for both purchase amounts and age groups`

(d) `Histogram for both purchase amounts and age groups`

(e) `Bar chart for purchase amounts and a histogram for age groups`

Question 15. In a dataset with both continuous and categorical variables, which visualization method is most suitable for examining the relationship between a continuous variable and a categorical variable with more than two categories?

(a) `Line chart for each category of the categorical variable.`

(b) `Violin plot for the continuous variable and split by the categorical variable.`

(c) `Heatmap of correlations between all variables.`

(d) `Scatter plot with different colors for each category of the categorical variable.`

(e) `Box plot for the continuous variable and color-coded by the categorical variable.`

Question 16. According to data-wrangling-v2 page 17. Which process is often used to combine data from different sources (virtual or actual) and provide users with a unified view of the data?

(a) `Data Integration.`

(b) `Data Augmentation.`

(c) `Feature Selection.`

(d) `Regression.`

(e) `Clustering.`

Question 17. According to the lecture materials, which method of data cleaning would be MOST appropriate when dealing with a dataset where values are expected to form distinct groups and deviations from these groups are considered errors?

(a) `Linear regression smoothing`

(b) `Clustering-based smoothing`

(c) `Binning with equal-width partitioning`

(d) `Chi-square analysis`

(e) `Z-score normalization`

Question 18. According to the lecture on Pattern Mining Part 1, slide 10, which of the following best defines the structural and mathematical properties of a closed pattern within a transaction database?

(a) `A closed pattern is defined as the largest possible frequent itemset within a given dataset, ensuring that all of its subsets are also frequent and can be directly derived from its occurrence statistics.`

(b) `A closed pattern is any frequent itemset where the sum of its subset supports is equal to its own support count, indicating a perfect correlation between the items it contains.`

(c) `A closed pattern is a subset of a maximal pattern, where the subset retains only the most statistically significant co-occurring items while discarding those that do not contribute to higher confidence in association rules.`

(d) `A frequent itemset \(X\) is considered a closed pattern if and only if there exists no proper superset of \(X\) with an identical support count, thereby ensuring that all significant item relationships are captured without unnecessary redundancy.`

(e) A frequent itemset qualifies as a closed pattern when its absolute support exceeds a dataset-specific entropy threshold, thereby distinguishing itself as an essential pattern for association rule mining.

Question 19. According to data-wrangling-v2 P25,26 A city shows a strong correlation between the number of ice cream shops and the crime rate. According to the lecture's discussion of chi-square test and correlation, what is the most likely explanation for this relationship?

(a) There is a hidden third variable (such as population density or temperature) that affects both variables independently.

(b) The correlation is coincidental and has no statistical significance.

(c) The relationship must be inversely causal - crime rates influence ice cream shop numbers.

(d) Ice cream consumption directly causes higher crime rates.

(e) The high chi-square value proves that ice cream shops cause crime.

Question 20. According to the lecture and materials, During the data exploration phase, which of the following methods is suitable for identifying patterns of missing values in a dataset?

(a) Calculate the mean and median for each variable.

(b) Use a classification model to predict missing values.

(c) Use a missing value matrix or heatmap to display the locations of missing values.

(d) Perform Principal Component Analysis (PCA) to reduce dimensionality.

(e) Use a scatter plot to visualize the relationship between two variables.

Question 21. Which of the following best explains the distinction between feature selection and feature engineering in the context of building a predictive model?

(a) Feature selection focuses on identifying the most relevant and non-redundant features from the existing dataset, while feature engineering involves transforming or creating new features to improve model performance.

(b) Feature selection and feature engineering are identical and both refer to reducing the size of the dataset by removing noise.

(c) Feature engineering is only applicable when using deep learning, while feature selection is used for traditional machine learning models.

(d) Feature selection creates new variables by combining raw inputs, while feature engineering removes irrelevant data to simplify the model.

Question 22. From chapter 2 of our textbook, "Data Mining Concepts and Techniques," which of the following statements about data preprocessing is false?

(a) `Data cleaning is an iterative process of discrepancy detection and transformation, handling missing values, noise, outliers, and inconsistencies.`

(b) `Data quality includes accuracy, completeness, consistency, timeliness, believability, and interpretability, evaluated based on intended data use.`

(c) `Data integration combines multiple sources into a coherent store, addressing metadata and tuple duplication.`

(d) `Data transformation via sampling and compression create reduced representations while preserving all statistical properties and complete information content.`

Question 23. According to the Clustering lecture (Slide 9 of Clustering.pdf), which clustering method is most appropriate when you assume the data is generated from a known statistical distribution such as a mixture of Gaussians?

(a) `Probabilistic and generative models, as they estimate parameters to fit the data to a statistical model.`

(b) `Hierarchical clustering methods, since they require Gaussian assumptions to form dendrograms.`

(c) `Partitioning algorithms, since they divide the data into predefined groups based on nearest centroid.`

(d) `Density-based clustering methods, as they assume clusters are shaped like Gaussian blobs.`

(e) `Grid-based methods, which are tailored for statistical distribution modeling through kernel density estimation.`

Question 24. According to the lecture on data wrangling part2 (slide no 21), Prof Laura, Which of the following best describes the primary difference between data warehousing and virtual data integration?

(a) `Data warehousing is only suitable for structured data, while virtual integration supports both structured and unstructured data.`

(b) `Virtual integration requires all data sources to follow the same schema, while data warehousing does not.`

(c) `Data warehousing physically consolidates data, whereas virtual integration accesses data in real-time without replication.`

(d) `Data warehousing allows for real-time querying of distributed data, while virtual integration requires batch processing.`

(e) `Virtual integration stores data permanently, while data warehousing deletes data after each query.`

Question 25. According to the class notes on Data Exploration Demo, which of the following is the correct syntax for loading a JSON object (represented as a Python dict) called "rawtemps" into a pandas dataframe?

(a) `df = pd.read_json(io.StringIO(rawtemps))`

(b) `df = pd.read_json(io.BytesIO(json.dumps(rawtemps).encode('utf-8')))`

(c) `df = pd.read_json(rawtemps)`

(d) `df = pd.read_json(json.dumps(rawtemps))`

(e) `df = pd.read_json(io.StringIO(json.dumps(rawtemps)))`