cap5771sp25
April 29, 2025

Exam 3 version a

Question 1. According to the lecture on probability distributions, where do most data points lie in a normal distribution?

(a) `Most data points cluster around the mean, with fewer values appearing as you move further away.`

(b) `Most data points align exactly with the standard deviation, making ÏČ the most frequent value.`

(c) `Most data points are found in the extreme tails, far from the mean.`

(d) `Most data points lie outside the interquartile range (IQR), making the middle range of data less significant.`

(e) `The data points are evenly distributed across all values, with no concentration near the mean.`

Question 2. Which of the following best describes the primary goal of cluster analysis?

(a) `To sequentially label data points using a rule-based decision-making process.`

(b) `To determine the probability distribution of a dataset using statistical models.`

(c) `To partition data points into groups that are as similar as possible within clusters and as dissimilar as possible between clusters.`

(d) `To assign predefined labels to data points based on a training dataset.`

(e) `To reduce data dimensionality by transforming features into a lower-dimensional space.`

Question 3. According to the lecture, when determining the optimal number of clusters in a dataset using the elbow method, what specific pattern does a data scientist look for in the visualization?

(a) `The point where the sum of squared distances doubles compared to the previous number of clusters.`

(b) `The point where the slope of the curve becomes equal to zero.`

(c) `The point where adding more clusters produces only a small reduction in the sum of within-cluster variance.`

(d) `The maximum value on the curve representing the ratio of between-cluster to within-cluster variance.`

(e) `The minimum value on the curve representing total computational complexity for each cluster count.`

Question 4. According to the lecture on data exploration, which type of plot is most effective for identifying outliers in a dataset?

(a) `Box plot`

(b) `Bar chart`

(c) `Pie chart`

(d) `Histogram`

(e) `Scatter plot`

Question 5. According to the lecture on Clustering (slide 14), what is one key advantage of the K-Medoids clustering method over the K-Means clustering method?

(a) `It is faster for large datasets.`

(b) `It requires less computational complexity.`

(c) `It works only for numerical data.`

(d) `It is less sensitive to outliers and noise.`

(e) `It does not require specifying the number of clusters in advance.`

Question 6. In the data exploration files, SQL queries were used to retrieve structured data from a normalized relational database (chinook.db), while Pandas was used to preprocess numerical weather data by transforming temperature values. Considering these two approaches, how does normalization in databases (SQLite queries) differ from normalization in data preprocessing (Pandas transformations)?

(a) `Both types of normalization serve the same purpose: organizing data into structured tables for storage efficiency.`

(b) `Data normalization in Pandas is primarily used to reduce database size and improve storage efficiency.`

(c) `In both databases and data preprocessing, normalization refers exclusively to handling missing values to improve data quality.`

(d) `Normalization in SQL databases is performed using Min-Max scaling and Z-score normalization techniques.`

(e) `Database normalization (SQLite queries) structures data to reduce redundancy, while data preprocessing normalization (Pandas transformations) scales numerical values for consistency.`

Question 7. According to lecture and the slides, are discrete attributes always numbers?

(a) `No, because discrete attributes can only be strings or text.`

(b) `Yes, because discrete attributes are used for analysis, and numbers are easier to analyze.`

(c) `Yes, because discrete attributes must always be numeric.`

(d) `Yes, because all discrete attributes represent countable values.`

(e) `No, because other data types can have an cardinality.`

Question 8. According to Data Wrangling Lecture, what is the purpose of normalization in data preprocessing?

(a) `To scale data values into a common range for analysis`

(b) `To remove all missing values from the dataset`

(c) `To randomly modify data values for security purposes`

(d) `To increase the number of duplicate records`

Question 9. Which of the following best describes a data object in the context of data science?

(a) `A representation of an entity that consists of multiple attributes.`

(b) `A special type of binary data structure used for machine learning models.`

(c) `A single attribute that defines the characteristics of an entity.`

(d) `A temporary dataset that is used only for intermediate calculations.`

(e) `A unique identifier used to differentiate records in a database.`

Question 10. A dataset contains salary information recorded in different currencies (e.g., USD, EUR). How would you standardize this data?

(a) `Group salaries by currency type without converting them.`

(b) `Remove all salary records that are not in USD for simplicity.`

(c) `Convert all salaries into a single currency using an exchange rate table before analysis.`

(d) `Normalize salary values between 0 and 1 using Min-Max scaling.`

Question 11. In the context of data visualization, which of the following best describes the key difference between a histogram and a bar chart?

(a) `A bar chart always has bars touching each other, whereas a histogram never does.`

(b) `A bar chart is only used for time-series data, while a histogram is used for all types of data.`

(c) `A histogram requires an equal number of values in each bin, while a bar chart does not.`

(d) `A histogram represents categorical data, whereas a bar chart represents numerical data.`

(e) `A histogram groups data into bins to show distribution, while a bar chart compares distinct categories.`

Question 12. According to Lecture 1 on Data Exploration, which of the following is NOT an attribute type in data classification?

(a) `Nominal`

(b) `Ratio`

(c) `Ordinal`

(d) `Relational`

Question 13. According to the lecture on Data Wrangling by Dr. Christan Grant and Dr. Laura Melissa Cruz Castro, referencing Data Wrangling Part 1 slides (slide 7), which method is most effective for addressing inconsistencies among categorical data entries such as "NYC" and "New York City"?

(a) `Implementing a random forest classifier to resolve categorical discrepancies.`

(b) `Using clustering algorithms to group similar categorical entries.`

(c) `Employing regression analysis to predict correct categorical labels.`

(d) `Maintaining discrepancies to preserve the raw data format.`

(e) `Standardizing categorical values through mapping to ensure uniformity across the dataset.`

Question 14. According to the Clustering Lecture, Slide 22 states that hierarchical clustering is generally 'more deterministic' than partitioning methods like k-means. Considering the typical process of agglomerative hierarchical clustering (AGNES, Slide 24), what procedural characteristic is the main reason for this determinism?

(a) `The final number of clusters must be specified beforehand, which rigidly guides the merging process to a predetermined outcome.`

(b) `Hierarchical clustering always explores all possible merging sequences, guaranteeing convergence to the single, globally optimal deterministic solution.`

(c) `The initial calculation of the full dissimilarity matrix provides an unchanging foundation, ensuring all subsequent merge decisions are predetermined by this matrix.`

(d) `The dendrogram output format itself forces a deterministic structure, as there's only one way to represent the hierarchy visually.`

(e) `The algorithm makes greedy merging decisions based on the chosen linkage criteria at each step, and these decisions are typically fixed and not revisited later in the process.`

Question 15. According to Data Wrangling v2 slide 4, which Data Preprocessing task is applied when a university categorizes student enrollment data by class level (Freshman, Sophomore, Junior, Senior) and then further groups it into Undergraduate and Graduate programs for reporting purposes?

(a) `Dimensionality Reduction.`

(b) `Data Cleaning.`

(c) `Data Compression.`

(d) `Data Integration.`

(e) `Concept Hierarchy Generation.`

Question 16. According to the data wrangling discussion, in the context of data cleaning, which of the following is an example of inconsistent data ?

(a) `A negative salary value of negative 10 dollars`

(b) `A blank occupation field`

(c) `Missing values in the salary field`

(d) `Different rating systems used (1,2,3) vs (A,B,C)`

(e) `Duplicate customer records`

Question 17. According to the Data Exploration PDF (page 13), which of the following best describes the purpose of covariance in statistical analysis?

(a) `Covariance is the same as correlation.`

(b) `Covariance always provides a standardized value between -1 and 1.`

(c) `Covariance measures the direction of the relationship between two variables.`

(d) `Covariance is used to calculate the mean of a dataset.`

(e) `A covariance value of zero always indicates that two variables are independent.`

Question 18. According to the data-exploration lecture, which of the following statements best describes the differences between schema-on-write and schema-on-read approaches in data management?

(a) `Schema-on-write and schema-on-read are interchangeable terms referring to the same data loading approach in traditional databases.`

(b) `Schema-on-write provides more flexibility than schema-on-read by allowing data to be loaded without a predefined structure.`

(c) `Schema-on-read allows for more flexibility by deferring schema application until data is read, while schema-on-write requires a predefined structure before data can be loaded, resulting in faster read times but less agility.`

(d) `Schema-on-read requires data transformation before loading, while schema-on-write allows data to be copied directly to the file store without transformation.`

(e) `Schema-on-read is only applicable to structured data formats like relational databases, while schema-on-write is used for semi-structured data like XML.`

Question 19. According to the lecture "Classification part II.pdf" regarding the principles of multi-layer neural networks, what is the primary role of the hidden layer?

(a) `To cycle back information to previous layers for better weight adjustments.`

(b) `To store intermediate predictions that are directly used as final outputs.`

(c) `To randomly adjust weights without learning from training samples.`

(d) `To directly pass the input values to the output layer without modification.`

(e) `To transform weighted inputs non-linearly, allowing the network to learn complex patterns.`

Question 20. According to Dr. Grant's lecture on Data Wrangling v2, slide 38, which of the following is a key characteristic of parametric data reduction methods?

(a) `They rely exclusively on clustering techniques to reduce dimensionality.`

(b) `They do not require any predefined assumptions about data distribution.`

(c) `They store full datasets while reducing redundancy through compression.`

(d) `They always result in data loss due to lossy compression techniques.`

(e) `They assume a specific mathematical model to approximate data.`

Question 21. According to the lecture, In the context of structured, semi-structured, and unstructured data, which of the following best describes a key characteristic of structured data?

(a) `Structured data allows for flexible and evolving data formats without predefined organization.`

(b) `Structured data is always stored as plain text without any specific format.`

(c) `Structured data completely eliminates the need for data validation.`

(d) `Structured data follows a predefined schema, ensuring consistency and integrity.`

(e) `Structured data does not require any indexing or constraints for efficient querying .`

Question 22. In which scenario would stratified sampling be the more appropriate choice over simple random sampling?

(a) When selecting a sample in which each individual has an equal probability of being chosen, without considering subgroups

(b) When performing sampling without replacement to ensure no individual is selected more than once

(c) When selecting a sample from a homogeneous population where all members have similar characteristics

(d) When ensuring that specific subgroups in a population are proportionally represented in the sample

(e) When every individual in the population has an equal chance of being selected, but some individuals may appear multiple times due to sampling with replacement

Question 23. A dataset contains student test scores: 50, 55, 60, 62, 65, 68, 72, 75, 80, 150. Using the Interquartile Range (IQR) method, which of the following values would be classified as an outlier?

(a) No outliers exist in this dataset.

(b) 72

(c) 50

(d) 150

(e) 80

Question 24. According to the Data Wrangling lecture(v2-31,32,35), In a data science project, If you encounter a dataset with significant noise and dimensionality issues. Which sequence of preprocessing steps would be MOST appropriate to prepare this data while preserving its essential characteristics?

(a) Use equal-frequency binning first, followed by dimensionality reduction through PCA then apply decimal scaling.

(b) Apply data cleaning to handle noise, then use PCA for dimensionality reduction, followed by clustering-based discretization.

(c) Start with z-score normalization, then apply clustering-based discretization, and finish with dimensionality reduction using PCA.

(d) Apply PCA first, then use equal-width binning, followed by min-max normalization.

(e) `Begin with decimal scaling normalization, use histogram analysis for discretization then apply correlation analysis.`

Question 25. According to the Data Exploration lecture and as seen in the slides, what percentage of the data is within 2 standard deviations of the mean in a normal distribution curve?

(a) `92%`

(b) `96%`

(c) `95%`

(d) `68%`

(e) `99.7%`