cap5771sp25
April 29, 2025

Exam 3 version b

Question 1. According to the data-wrangling-v2 lecture (page 46), which of the following scenarios best demonstrates the use of data cube aggregation for multidimensional analysis?

(a) `A financial institution removes redundant columns from its datasets to improve processing speed.`

(b) `A company uses data encryption to ensure customer records are secure during transmission and storage.`

(c) `A retailer aggregates total sales across time, product categories, and geographical regions to enable fast querying at various levels of detail.`

(d) `A university uses clustering to group students based on similar study patterns for academic intervention programs.`

(e) `A hospital applies normalization to standardize patient data across different departments.`

Question 2. According to the lecture on Data Exploration, Which statement correctly describes the cardinality and arity of this data.

Student ID Name Age Major
101 Peter 24 Mathematics
102 Ram 22 Computer Science
103 James 19 Physics

(a) `The Arity of the Student Table is 4, but cardinality refers to the number of relationships between different tables, as we don't have a second table cardinality cannot be defined.`

(b) `The cardinality of the Students table is 4,  and the arity is 3`

(c) `The cardinality of the Student table is 12, and the arity is 3`

(d) `The cardinality and arity refer to the same thing, and the value for this table is 4`

(e) `The cardinality of the Students table is 3, and the arity is 4`

Question 3. According to the lecture on data exploration, which of the following best describes the difference between discrete and continuous attributes?

(a) `Continuous attributes are used in structured data, while discrete attributes are used in unstructured data.`

(b) `Discrete attributes have a finite or countable number of values, while continuous attributes can take any real value within a range.`

(c) `Discrete attributes can only be represented as binary values (0 or 1), while continuous attributes do not.`

(d) `Discrete attributes are always numerical, while continuous attributes are always categorical.`

(e) `There is no real difference between discrete and continuous attributes in data science.`

Question 4. 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 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.`

(b) `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.`

(c) `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.`

(d) `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.`

(e) `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.`

Question 5. According to Classification part I.pdf (slide 32) and the associated lecture, which of the following is an appropriate list of subset sizes when using 5-fold cross-validation on data that has a total of 60 tuples?

(a) `30, 15, 8, 4, 3`

(b) `20, 20, 10, 5, 5`

(c) `22, 17, 12, 7, 2`

(d) `12, 12, 12, 12, 12`

(e) `56, 1, 1, 1, 1`

Question 6. According to the clustering lab, the TF-IDF model is an improvement over the BoW model because instead of just counting word occurrences, it:

(a) `Assigns lower weight to all words in every document`

(b) `Assigns higher weight to words that appear rarely in one document but frequently in other documents`

(c) `Assigns higher weight to words that appear frequently in one document but that appear rarely in other documents`

(d) `Assigns lower weight to words that appear frequently in one document but rarely in other documents`

(e) `Doesn't use a weighting system`

Question 7. 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 nominal data`

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

(c) `Causal relations in nominal data`

(d) `Causal relations in numeric data`

(e) `Correlation in numeric data`

Question 8. 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 9. 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) `Database normalization (SQLite queries) structures data to reduce redundancy, while data preprocessing normalization (Pandas transformations) scales numerical values for consistency.`

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

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

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

Question 10. According to the lecture on data exploration(page 25), which of the following is an example of an asymmetric binary attribute?

(a) `A hair color attribute with values like black, brown, and blonde.`

(b) `A zip code attribute representing different geographic regions.`

(c) `A temperature measurement in Celsius.`

(d) `A gender attribute with values male and female.`

(e) `A medical test result where positive is more significant than negative.`

Question 11. 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(json.dumps(rawtemps)))`

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

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

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

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

Question 12. According to Dr. Grant's lecture on Data Wrangling v2, slide 54, what is the main advantage of using equal-depth partitioning for data discretization?

(a) `Automatically removes outliers when splitting up the data points into bins.`

(b) `Handles dynamic data well due to computations becoming less extensive with each new change in data.`

(c) `Implements very intricate algorithms to ensure that data is handled to perfection.`

(d) `Ensures each bin contains around the same amount of data points, effective for handling skewed data.`

(e) `Splits up the range into intervals of equal size for ease of implementation and understanding.`

Question 13. According to the classification lecture (March 2025), there are several models for classification. Suppose a dataset has two features which are directly graphed against each other in a scatter plot. There are two classes and they are colored blue and orange on the plot. After plotting, it can be seen that the classes form concentric rings. Meaning, the blue dots generally form a small inner ring and the orange dots form a larger outer ring. Given the shape of the data, which model would be best to predict these classes?

(a) `SVM with an RBF kernel`

(b) `Support Vector Machine`

(c) `Logistic Regression with LASSO`

(d) `Logistic Regression`

(e) `Linear Regression with RIDGE`

Question 14. Which of the following is an outlier detection method that uses statistical techniques?

(a) `Data Normalization`

(b) `Feature Engineering`

(c) `Data Binning`

(d) `Principal Component Analysis (PCA)`

(e) `Interquartile Range (IQR)`

Question 15. From the concepts thought in Data Wrangling Lecture, Which of the following clustering techniques is best suited for data smoothing by averaging data points within clusters to reduce variance and enhance patterns?

(a) `OPTICS`

(b) `Gaussian Mixture Models`

(c) `Agglomerative Hierarchical Clustering`

(d) `DBSCAN`

(e) `K-Means Clustering`

Question 16. According to the data visualization part in the data exploration pdf, which of the following best explains why exploratory data analysis can never be truly objective, no matter how rigorously performed?

(a) `Objectivity in data analysis is guaranteed if the dataset is large enough.`

(b) `A truly objective dataset would render EDA obsolete.`

(c) `The questions asked during exploration inherently shape the results and interpretations.`

(d) `Data collection methods are universally standardized, ensuring objectivity.`

(e) `Data visualization eliminates all subjectivity by presenting facts as they are.`

Question 17. According to the Clustering lecture (Slide 6), what is the main reason Euclidean distance may be inappropriate for high-dimensional data in clustering tasks?

(a) Because Euclidean distance always produces non-convex clusters, even in low-dimensional space.

(b) Because in high dimensions, distances between points become increasingly similar, reducing the effectiveness of separation.

(c) Because Euclidean distance ignores data normalization, which is only addressed by cosine similarity.

(d) Because Euclidean distance in high dimensions causes your computer to become self-aware.

(e) Because Euclidean distance requires labeled data to compute similarity.

Question 18. According to the lecture on Data Wrangling (Slide 7- 9), what is the primary advantage of leveraging metadata for detecting data discrepancies?

(a) Metadata eliminates the need for domain experts because it interprets and validates data on its own.

(b) Metadata replaces raw data values with statistical averages, ensuring that all discrepancies are removed.

(c) Metadata provides structured information that can reveal hidden mismatches or inconsistencies between recorded and actual conditions, thereby enhancing the accuracy of the data.

(d) Metadata relies entirely on user intuition, making it inherently prone to subjective biases.

(e) Metadata automatically corrects all errors in a dataset by comparing values across unrelated sources.

Question 19. According to the lecture on classification (Slides 30-33), which of the following statements about model evaluation metrics is correct?

(a) The ROC curve plots false negative rate against true negative rate to visualize classifier performance.

(b) The F1-score is the arithmetic mean of precision and recall, providing equal weight to both metrics.

(c) Precision measures the percentage of tuples labeled as positive that are actually positive, while recall measures the percentage of positive tuples correctly identified.

(d) `In a confusion matrix, accuracy is computed as (FP + FN) / (TP + TN + FP + FN).`

(e) `Specificity calculates the ratio of true negatives to all negative predictions, while sensitivity measures the ratio of false positives to all positive predictions.`

Question 20. According to lecture (Text Retrieval and Extraction, slide 12), why is cosine similarity used in the Vector Space Model to rank the relevance of documents to a query?

(a) `Because it measures the difference in page rank scores between two documents to rank them.`

(b) `Because it ensures that stopwords and stemming are completely ignored during ranking.`

(c) `Because it directly counts how many times each query word appears in the document without normalization.`

(d) `Because it computes the maximum term frequency across all documents to identify the most common document.`

(e) `Because it measures the angle between document and query vectors, allowing relevance to be assessed independently of document length.`

Question 21. According to the lecture on Data Exploration, which type of attribute is represented by the following scenario? In a race, five athletes finish at different times. You rank them from 1st place to 5th place based on who finishes first to last. What type of attribute is used when ranking athletes from 1st to 5th?

(a) `Interval attribute.`

(b) `Binary attribute.`

(c) `Ordinal attribute.`

(d) `Nominal attribute.`

(e) `Ratio attribute.`

Question 22. According to the lecture on database querying, what is the primary reason for using a cursor in SQL operations?

(a) `A cursor allows for row-by-row processing of query results, making it useful for handling large datasets.`

(b) `A cursor prevents SQL injection attacks by default, without needing parameterized queries.`

(c) `A cursor is required to execute any SQL query, including CREATE TABLE statements.`

(d) `A cursor automatically optimizes SQL queries to improve performance.`

(e) `A cursor stores query results permanently in the database for future use.`

Question 23. According to slide 10 of the Intro to Data Science lecture, which of the following best describes the difference between Machine Learning and Data Science?

(a) `Machine learning is the study of building neural networks to generate new data, such as images and text. Data science uses simpler models to analyze existing data and learn from it.`

(b) `Machine learning deals with the creation of new models to predict based on training data. Data science deals with exploring data and finding trends, in part by using those models.`

(c) `Data science encompasses the process of gathering and cleaning data, while machine learning only focuses on the specific algorithms used to analyze that data.`

(d) `Machine learning is about working with structured data, such as from databases and spreadsheets. Data science often deals with that data, but can deal with unstructured data as well, such as images and videos.`

(e) `Machine learning is purely theoretical, while data science is purely practical.`

Question 24. According to the lecture on data integration, which of the following is a key challenge when integrating data from multiple sources?

(a) `Using only structured data and ignoring unstructured data sources.`

(b) `Identifying real-world entities from different data sources, such as matching "Bill Clinton" with "William Clinton."`

(c) `Ensuring that all data sources are stored in the same physical location.`

(d) `Converting all data into a single file format, such as CSV.`

(e) `Ensuring that all data sources use the same programming language for data storage.`

Question 25. According to the Data Wrangling Part 2 Lecture slides 14 and 15, what are the similarities and differences in the application of linear regression and clustering methods for data smoothing?

(a) `Both methods assume that the residuals of the data must follow a normal distribution, but linear regression is used for modeling relationships, and clustering is used for data normalization.`

(b) `Both methods rely on the data being normally distributed; linear regression is used to smooth the data, and clustering is used to reduce noise.`

(c) `Both methods are used to handle group characteristics of data, but linear regression is used to find linear relationships between data, while clustering is used to identify distinct groups within the data.`

(d) `Both methods require strong computational power; linear regression is used to predict future data, and clustering is used to generate new variables.`

(e) `Both methods create new datasets; linear regression is used for data standardization, and clustering is used to simplify data dimensions.`