cap5771sp25
March 19, 2025

Exam 1 part 1 version d

Question 1. 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) `Ordinal attribute.`

(b) `Nominal attribute.`

(c) `Interval attribute.`

(d) `Ratio attribute.`

(e) `Binary attribute.`

Question 2. 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 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.`

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

(c) `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.`

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

Question 3. According to lecture on January 29th on the Codio Data Exploration Demo, what is the main purpose of using the "with" function when opening a URL request? Ex:

```
req = urllib.request.Request(url, headers=headers)
with urllib.request.urlopen(req) as response:
    data = response.read().decode('utf-8')
...
```

(a) `It continually tries to establish a connection upon failure until success.`

(b) `It allows multiple URL requests to be opened simultaneously without error.`

(c) `It acts as a while loop for as long as the connection remains until terminated within the code block.`

(d) `It is required by the urllib library to be able to open the request.`

(e) `It makes sure that the connection is automatically closed after the code block completes.`

Question 4. According to the Data Exploration PDF (page 13), which of the following best describes the purpose of covariance in statistical analysis?

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

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

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

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

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

Question 5. Which of the following is true about semi-structured data models like XML?

(a) `They are self-describing documents that can represent tree structures and free-text .`

(b) `They are mostly used for numeric data and require precise data types for every field.`

(c) `They rely on predefined schemas that must be applied before data is stored.`

(d) `They do not allow for hierarchical relationships or nesting of elements.`

(e) `They represent data using a set of rigid tables with rows and columns.`

Question 6. Which of the following best describes a data object in the context of data science?

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

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

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

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

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

Question 7. Scenario: A sports analytics company is analyzing player performance statistics from multiple basketball games. Each game is recorded with numerical values such as points scored, assists, and rebounds for each player. According to the lecture on data exploration from slide 27, which type of numeric attribute best describes the number of points a player scores in a game?

(a) `Discrete numeric attribute`

(b) `Nominal attribute`

(c) `Continuous numeric attribute`

(d) `Binary attribute`

(e) `Ordinal attribute`

Question 8. From the the data wrangling demo, what will be the state of df after the following code is executed?

```
import pandas as pd
df = pd.DataFrame({'A': [1, 2, 3], 'B': [4, 5, 6]})
df['C'] = df['A'] + df['B']
df.loc[0, 'A'] = 10
df.drop(columns=['B'])
```

(a) `The assignment to df.loc[0, 'A'] will raise an error.`

(b) `Column 'C' will not be created.`

(c) `The operation df['C'] = df['A'] + df['B'] modifies the original df in place.`

(d) `The column 'B' will still be present in df`

(e) `The column 'B' will be permanently removed.`

Question 9. According to the lecture on data exploration, how do traditional databases stores load data into a table?

(a) `Schemamore.`

(b) `Schema-On-Read.`

(c) `Schema-On-Write.`

(d) `Schemaless.`

(e) `noSQL.`

Question 10. According to the Data Exploration lecture and demo, which of the following best describes the internal structure of a pandas DataFrame as used in Python?

(a) `It is a data field, representing a characteristic or feature of a data object.`

(b) `It is a one-dimensional labeled array capable of holding any data type (integers, strings, floating point numbers).`

(c) `It is implemented as a JSON object, where each key is a column name and each value is a list containing the column's data.`

(d) `It is a dictionary where each key is a column name and each value is a Series object representing that column.`

(e) `It is a list of Excel sheet objects where each column has its own JSON file with a tiny SQLite database running secret queries.`

Question 11. According to data exploration demo, what is the meaning of the cursor.fetchall() in this code block?

```
cursor.execute("SELECT tracks.Name, albums.Title, artists.Name
            FROM tracks JOIN albums ON tracks.AlbumId = albums.AlbumId
              JOIN artists ON albums.ArtistId = artists.ArtistId")
tracks_albums_artists = cursor.fetchall()
for item in tracks_albums_artists:
    print(item)
```

(a) `It forms a new SQL query`

(b) `It creates a new SQL table and stores the data`

(c) `It will crash your whole system`

(d) `It retrieves only one row from the SQL output`

(e) `It retrieves the output as the a of tuples`

Question 12. 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 cardinality of the Students table is 3, and the arity is 4`

(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 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.`

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

Question 13. According to the lecture, what is the primary difference between structured, semi-structured, and unstructured data?

(a) `Unstructured data always consists of multimedia files, while structured data includes only text and numerical values.`

(b) `Semi-structured data lacks any form of organization and cannot be queried efficiently.`

(c) `Structured data can only be stored in relational databases, while semi-structured and unstructured data must be stored in NoSQL databases.`

(d) `Structured data cannot be transformed into unstructured data, whereas semi-structured data can be converted into both structured and unstructured formats.`

(e) `Structured data follows a fixed schema, semi-structured data has a flexible schema, and unstructured data has no predefined schema.`

Question 14. According to the lecture on data exploration from slide 10, which of the following statements about schema-on-read is correct?

(a) `Schema is applied while reading the data.`

(b) `Schema-on-read enforces strict data structure constraints.`

(c) `Schema is applied before data is stored.`

(d) `Schema-on-read is commonly used in traditional relational databases.`

(e) `Schema-on-read makes data loading slower than schema-on-write.`

Question 15. 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` σ `the most frequent value.`

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

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

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

Question 16. In data classification, how are attributes like ID numbers and ZIP codes categorized?

(a) `Neither qualitative nor quantitative, as they do not fit into standard data classification categories.`

(b) `Qualitative attributes, because they represent categorical information without inherent numerical meaning.`

(c) `Quantitative attributes, because they consist of numerical values that can be used in mathematical operations.`

(d) `Both qualitative and quantitative attributes, depending on the context in which they are used.`

(e) `Only ZIP codes are quantitative, while ID numbers are qualitative.`

Question 17. According to the data exploration lecture, what is the primary reason for detecting and handling missing data during the data preparation phase?

(a) `To make the dataset look more professional in presentations.`

(b) `To improve data visualization aesthetics by eliminating blank spaces.`

(c) `To ensure the accuracy and reliability of statistical analyses and machine learning models.`

(d) `To automatically generate new data to replace the missing values without validation.`

(e) `To reduce the file size of the dataset for faster processing.`

Question 18. According to the lecture, in the context of data attribute types, which of the following statements best describes a key difference between interval-scaled and ratio-scaled attributes?

(a) `Both interval-scaled and ratio-scaled attributes are used only for nominal data types.`

(b) `Ratio-scaled attributes have a true zero point, allowing meaningful multiplicative comparisons, whereas interval-scaled attributes do not.`

(c) `Interval-scaled attributes are always discrete, while ratio-scaled attributes are always continuous.`

(d) `Interval-scaled attributes can only be used in qualitative analysis, whereas ratio-scaled attributes are used in quantitative analysis.`

(e) `Ratio-scaled attributes cannot be negative, while interval-scaled attributes must always be positive.`

Question 19. What do you need to be aware of when imputing missing values using the mean of an attribute?

(a) `Mean imputation works well for categorical attributes.`

(b) `Mean imputation is always preferable to median imputation.`

(c) `Mean imputation can distort the distribution and underestimate variance.`

(d) `Mean imputation ensures missing values do not affect correlation calculations.`

(e) `Mean imputation preserves all statistical properties of the data.`

Question 20. According to the lecture on data exploration, which of the following types of data is considered semi-structured?

(a) `Video files`

(b) `Server log files`

(c) `Customer service phone call recordings`

(d) `SQL application database`

(e) `Excel spreadsheets`

Question 21. According to lecture materials on exploratory data analysis, a data scientist observes two histograms with identical five-number summaries but differing bin structures during a quality control audit of pharmaceutical batch processing times. Which fundamental analytical limitation of boxplots does this scenario best demonstrate?

(a) `Boxplots obscure distribution modality and density variations across value ranges`

(b) `Boxplots inaccurately represent continuous data as discrete quartiles`

(c) `Boxplots require minimum sample sizes exceeding 100 observations`

(d) `Boxplots confuse stakeholders by using medians instead of means`

(e) `Boxplots exaggerate outlier impacts through 1.5xIQR whisker thresholds`

Question 22. According to lecture data exploration demo, when using Python to interact with a database, the cursor.execute(...) method is commonly used to execute SQL queries or commands. Which of the following statements correctly describes the usage of cursor.execute(...)?

(a) `cursor.execute(...) is used to commit changes to the database.`

(b) `cursor.execute(...) takes an SQL query as an argument and executes it.`

(c) `cursor.execute(...) is used to create new database tables.`

(d) `cursor.execute(...) returns a database connection object.`

(e) `cursor.execute(...)  is used to close the database connection.`

Question 23. According to the intro to data science lecture, which characteristic of big data refers to the trustworthiness, accuracy, and overall quality of the data being analyzed?

(a) `Veracity`

(b) `Value`

(c) `Volume`

(d) `Variety`

(e) `Velocity`

Question 24. 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, 46-99) 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) `Bar chart for purchase amounts and a histogram for age groups`

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

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

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

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

Question 25.

with open('sample.json','r') as f:
    data = json.load(f)

What is the most likely data type of data after executing this code?

(a) `A hierarchical tree structure where each node is a JSON key-value pair.`

(b) `An instance of a user-defined class dynamically inferred from the JSON schema.`

(c) `A serialized JSON object stored as a Python string.`

(d) `A Pandas DataFrame if the JSON contains tabular data.`

(e) `A Python dictionary or a list, depending on the JSON structure.`