cap5771sp25
March 19, 2025

Exam 1 part 2 version d

Question 1. According to data-wrangling-v2 page 17. Which of the following processes is often used to combine data from different sources (virtual or actual) and provide users with a unified view of the data?

(a) `Data Augmentation.`

(b) `Feature Selection.`

(c) `Clustering.`

(d) `Regression.`

(e) `Data Integration.`

Question 2. According to the lecture Data Wrangling Part 2 (Slides 41, 42), which of the following statements about sampling methods is correct?

(a) `Systematic sampling selects random samples by grouping data points and selecting an equal number from each group.`

(b) `Cluster sampling is identical to stratified sampling because both methods divide the data into groups and select samples randomly from each group.`

(c) `Sampling without replacement means that once a data point is selected, it may still be chosen again in the next iteration.`

(d) `Stratified sampling is useful when the data has distinct subgroups, ensuring proportional representation from each group.`

(e) `Simple random sampling ensures that each data point is selected with equal probability, and once selected, it cannot be chosen again.`

Question 3. According to lecture on data mining principles, which of the following is NOT a data reduction strategy?

(a) `Feature selection`

(b) `Dimensionality reduction`

(c) `Data compression`

(d) `Numerosity reduction`

(e) `Data migration`

Question 4. In which scenario would stratified sampling be the more appropriate choice over simple random sampling?

(a) 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

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

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

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

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

Question 5. 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 create new datasets; linear regression is used for data standardization, and clustering is used to simplify data dimensions.

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

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

Question 6. From the Data Wrangling Demo,

```
import pandas as pd
from scipy.stats import zscore
def correcting_datatypes(df, date_cols=None, categorical_cols=None, float_cols=None):
    if date_cols:
        for col in date_cols:
            if col in df.columns:
            try:
                df[col] = pd.to_datetime(df[col], format="mixed", errors="coerce")
                print(f'Converted 'col' to datetime')
            except ValueError as e:
                print(f'Warning: Could not convert column 'col' to datetime. Error: e')
            else:
                print(f'Warning: Column col does not exist in the DataFrame.')
```

In the function correcting_datatypes, we see the code snippet:

df[col] = pd.to_datetime(df[col], format='mixed', errors='coerce')

What is the significance of setting format='mixed' when converting a column to a datetime in this context?

(a) `It indicates that the column may contain dates in multiple formats, prompting Pandas to attempt parsing each entry flexibly.`

(b) `It ensures that dates are strictly interpreted with a single format, causing any values that don't match to be rejected.`

(c) `It forces the parser to treat every value as UTC time, regardless of the original timezone or date string.`

(d) `It automatically detects and corrects all invalid or incomplete date values without any user intervention.`

Question 7. According to the Data Wrangling lecture, which of the following best describes numerosity reduction in data preprocessing?

(a) `Eliminating missing values by imputing mean or median values.`

(b) `Scaling numerical attributes to a common range for consistency.`

(c) `Aggregating multiple datasets into a single unified schema.`

(d) `Removing unimportant attributes to reduce dataset complexity.`

(e) `Reducing data volume by using alternative, smaller representations without losing analytical results.`

Question 8. According to lecture about data wrangling, which of the following is NOT a common technique used for data smoothing in the data transformation process?

(a) `Clustering.`

(b) `Regression.`

(c) `Binning.`

(d) `One-hot encoding.`

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

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

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

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

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

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

Question 10. 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 combining features and data discretization primarily serve the same purpose: reducing redundancy in the dataset.

(e) Both techniques are exclusively used for reducing overfitting in machine learning models.

Question 11. According to the lecture slides data-wrangling-v2, which of the following BEST describes the concept of "Independence of Semantic Variations" in the context of data integration?

(a) The ability of the system to recognize and reconcile differences in data meanings and usage across systems, such as understanding "DOB" and "DateOfBirth" as equivalent terms.

(b) The use of correlation analysis to identify redundant attributes across different databases.

(c) The ability of the system to access data regardless of its physical storage location.

(d) The ability of the system to operate with any data structure or syntax, handling databases like SQL or NoSQL.

(e) The process of physically moving all data into a single data warehouse.

Question 12. According to the lecture and readings on Data Wrangling, what is the key difference between metadata-based detection and rule-based detection in identifying data discrepancies?

(a) In the context of metadata-based detection the only capable action is finding query duplicate records, in contrast the rule based detection can do data cleaning in range of every type of recorded errors.

(b) Rule based detection focuses on discerning quantitative information while metadata detection goes beyond numbers.

(c) Metadata-based detection utilizes stored data attributes such as domain, range, and dependencies, whereas rule-based detection enforces predefined conditions like numerical ranges or logical constraints.

(d) Metadata-based detection requires AI models to function, while rule-based detection is manually performed.

(e) Metadata-based detection modifies data directly, while rule-based detection only flags errors without modification.

Question 13. Based on the lecture and readings on Data Wrangling, what is the primary difference between incomplete data and intentional missing data in a real-world dataset?

(a) `Extra data existing constitutes as incomplete data, whereas missing intentional data results from the deletion of data.`

(b) `Incomplete data is always incorrect, while intentional missing data is always correct.`

(c) `Incomplete data is always caused by privacy concerns, while intentional missing data occurs due to data collection issues.`

(d) `Incomplete data occurs due to unintentional errors or missing records, whereas intentional missing data is deliberately left blank or replaced with a default value.`

(e) `Intentional missing data can be ignored, whereas incomplete data must always be corrected.`

Question 14. According to data-wrangling-v2 slide 51, which of the normalization techniques mentioned is most suitable in the following scenario to ensure fair treatment of all features?

A machine learning model is being trained on a dataset where the features have significantly different ranges. One feature represents annual income in the range of thousands to millions and has a non Gaussian distribution. Another feature represents age in the range of 18 to 90. The model relies on distance-based calculations, such as k-nearest neighbors (KNN).

(a) `Binarization.`

(b) `Decimal Scaling.`

(c) `Min-Max Scaling.`

(d) `Z-Score Normalization.`

(e) `Standard Scaler.`

Question 15. According to the lecture Data Wrangling Part 2 (Slides 20, 21), which of the following statements about outlier detection methods is correct?

(a) `Box plots visually identify outliers as points beyond 3 standard deviations from the mean.`

(b) `Clustering techniques like DBSCAN consider outliers as data points that do not belong to any cluster.`

(c) `Correlation analysis is unaffected by outliers in the dataset.`

(d) `The IQR method defines outliers as values more than 2.5 times the IQR above the third quartile.`

(e) `The Z-score method uses the mean and standard deviation to identify outliers.`

Question 16. According to the data integration part in data wrangling lecture slides(18-24), a national healthcare network is merging patient records from two independent hospital systems into a unified database. One system, used by Hospital A, stores Social Security Numbers (SSNs) in the format "XXX-XX-XXXX", while the other system, used by Hospital B, records SSNs as a continuous 9-digit number ("XXXXXXXXX").

After integration, duplicate patient records emerge, insurance claims are mismatched, and critical medical histories are fragmented across multiple profiles. Additionally, some patients appear to have two separate billing accounts, while others become unidentifiable due to inconsistencies in the identification process.

What is the primary data integration issue here?

(a) `Some patient records reference SSNs that do not exist in the integrated database.`

(b) `The same patient data is duplicated in different formats, causing unnecessary storage.`

(c) `The SSN field is formatted differently across databases, requiring transformation.`

(d) `The same patient exists in both systems but is not recognized as a single entity.`

(e) `The same SSN has been recorded with inconsistent values, leading to identity mismatches.`

Question 17. According to the lecture Data Wrangling Part 2 (Slide 8), why is data cleaning considered an iterative process?

(a) `Data cleaning is a linear process that follows a strict set of predefined steps without revisiting previous stages.`

(b) `Data cleaning involves continuous improvement by identifying and correcting errors over time.`

(c) `Once a dataset is cleaned, further cleaning is unnecessary unless new data is added .`

(d) `Data cleaning aims to eliminate all variations in the dataset, ensuring that every value is uniform.`

(e) `Data cleaning primarily focuses on formatting inconsistencies rather than detecting outliers or resolving missing values.`

Question 18. According to the lecture on Data Wrangling, which of the following is NOT considered a measure of data quality?

(a) `Consistency.`

(b) `Latency.`

(c) `Accuracy.`

(d) `Completeness.`

(e) `Timeliness.`

Question 19. According to lecture on data preprocessing, what is the primary purpose of data discretization?

(a) `To normalize data by scaling values between 0 and 1.`

(b) `To combine data from multiple sources into a unified format.`

(c) `To randomly sample data for efficient storage.`

(d) `To transform continuous data into categorical bins.`

(e) `To eliminate missing values from a dataset.`

Question 20. A dataset contains salary information recorded in different currencies (e.g., USD, EUR). How would you standardize this data?

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

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

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

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

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

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

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

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

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

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

Question 22. According to Data Wrangling lecture, what is a key difference between lossless and lossy compression?

(a) `Lossless compression reduces data size by removing outliers.`

(b) `Lossy compression can reconstruct data with 100% accuracy.`

(c) `Lossless compression preserves all original data, while lossy compression removes some information.`

(d) `Lossy compression is always preferred in data science applications.`

(e) `Lossy compression always results in better model performance.`

Question 23. Principal Component Analysis (PCA) as a data reduction method minimizes error by:

(a) `maximizing the number of features in the dataset.`

(b) `capturing the largest amount of variation in data.`

(c) `identifying the best cluster of a data set.`

(d) `removing all correlations between variables.`

(e) `capturing the smallest amount of variation in data.`

Question 24. According to Data Wrangling v2, slide 38, which of the following is a key characteristic of parametric data reduction methods?

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

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

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

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

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

Question 25. What does it mean to use a democratic method for outlier identification in data analysis?

(a) `A method that uses voting from human experts to determine if a data point is an outlier.`

(b) `A method that combines multiple outlier detection techniques and identifies an outlier based on majority agreement.`

(c) `A method that assigns equal weights to all data points to ensure fair representation in outlier detection.`

(d) `A method that eliminates outliers by removing the smallest and largest values in the dataset without further analysis.`

(e) `A method that selects outliers based on the most frequently occurring values in the dataset.`