cap5771sp25
March 19, 2025

Exam 1 part 2 version e

Question 1. According to the lecture Data Wrangling Part 2 (Slide 8), why is data cleaning considered an iterative process?

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

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

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

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

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

Question 2. 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 ensures that dates are strictly interpreted with a single format, causing any values that don't match to be rejected.`

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

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

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

Question 3. Based on the lectures on Data Wrangling, answer the following. Consider a dataset of temperature readings from various machines/sensors collected at varying time intervals. Few of the machines log data every second, while few does it every minute. What is the best approach to preprocess this data in order to ensure consistency in the dataset?

(a) `First normalize the dataset, and then adjust all the frequencies to match the highest recorded value.`

(b) `Duplicate the values for machines with less frequent recordings so that we can match all data to the highest logging frequency.`

(c) `Use interpolation methods to estimate/generate the missing values for the less frequent recordings.`

(d) `Remove the temperature readings from machines that do not log at the same frequency , and maintain uniformity.`

(e) `Aggregate the readings over a fixed time window, like considering average temperature per minute, in order to standardize timestamps.`

Question 4. According to lecture on data mining principles, which of the following is NOT a data reduction strategy?

(a) `Feature selection`

(b) `Data compression`

(c) `Dimensionality reduction`

(d) `Numerosity reduction`

(e) `Data migration`

Question 5. 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) `Concept Hierarchy Generation.`

(b) `Data Cleaning.`

(c) `Dimensionality Reduction.`

(d) `Data Compression.`

(e) `Data Integration.`

Question 6. 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 use of correlation analysis to identify redundant attributes across different databases.`

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

(c) `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.`

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

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

Question 7. In which scenario would stratified sampling be the more appropriate choice over simple random sampling?

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

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

(c) `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`

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

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

Question 8. 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 primarily reduces dataset size, while data discretization improves model accuracy by increasing feature granularity.`

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

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

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

Question 9. 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) `Intentional missing data can be ignored, whereas incomplete data must always be corrected.`

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

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

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

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

Question 10. According to the lecture on data wrangling demo, If null_counts_rows is a Pandas Series containing the number of null values in each row, what does the following command return?
null_counts_rows[null_counts_rows == max_nulls]

(a) `A count of how many rows contain max_nulls null values.`

(b) `A NumPy array of null counts where they equal max_nulls.`

(c) `A list of row indices where the number of null values is equal to max_nulls.`

(d) `A boolean mask indicating which rows have exactly max_nulls null values.`

(e) `A filtered Series containing only the rows where the null count equals max_nulls.`

Question 11. According to Data Wrangling v2, slide 38, which of the following is a key characteristic of parametric data reduction methods?

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

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

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

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

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

Question 12. According to the data wrangling lecture, which of the following statements best describes the method of binning for data smoothing?

(a) `It applies clustering algorithms to group similar data points and uses cluster centroids to represent and smooth data.`

(b) `It uses regression models to predict and replace noisy data values based on the relationship with auxiliary variables.`

(c) `It transforms data by scaling values to fall within a smaller, specified range, such as 0.0 to 1.0.`

(d) `It identifies outliers by calculating the Interquartile Range (IQR) and removing data points that fall outside 1.5 times the IQR from the quartiles.`

(e) `It involves partitioning sorted data into equal-frequency or equal-width bins and then smoothing by bin means or boundaries.`

Question 13. 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) `Data warehousing is only suitable for structured data, while virtual integration supports both structured and unstructured data.`

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

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

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

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

Question 14. According to lecture on data preprocessing, what is the primary purpose of data discretization?

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

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

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

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

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

Question 15. Data quality issues can arise due to various reasons, such as human errors, system faults, or intentional manipulation. Which of the following correctly describes an example of a noisy data issue?

(a) `A dataset where different records have conflicting birthday and age values.`

(b) `A company setting all customers' birthdates to January 1st for convenience.`

(c) `A salary value recorded as ''-10'' due to an error in data entry.`

(d) `A survey where some participants' responses are missing.`

(e) `An employee record where the occupation field is left blank.`

Question 16. Principal Component Analysis (PCA) as a data reduction method minimizes error by:

(a) `removing all correlations between variables.`

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

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

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

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

Question 17. 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) `Correlation in numeric data`

(c) `Causal relations in numeric data`

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

(e) `Causal relations in nominal data`

Question 18. 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) `80`

(c) `150`

(d) `50`

(e) `72`

Question 19. According to the data wrangling demo, why is it recommended to handle outliers before imputing missing data?

(a) `Because outliers cannot be identified after missing data is imputed.`

(b) `To avoid introducing new outliers during the imputation process.`

(c) `To prioritize outlier removal over addressing missing data.`

(d) `To reduce computational overhead during the imputation process.`

(e) `To ensure that summary statistics like the mean or median are not skewed by extreme values.`

Question 20. According to the Data Wrangling lecture (v2, CAP 5771), which would be an example of data discrepancy detection using rules?

(a) `Heights are usually around 5 ft and a height of 10 ft was found.`

(b) `After regression, one point deviated significantly more than the others.`

(c) `A height of 10 ft was deemed odd because the study only allowed heights of 5 ft to 7 ft.`

(d) `A point with a z-score of 5 was marked as an outlier.`

(e) `K-means clustering found several points that did not belong to any cluster.`

Question 21. 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) `The same SSN has been recorded with inconsistent values, leading to identity mismatches.`

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

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

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

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

Question 22. According to the lecture on Data Wrangling (Slide 7- 9), what is the primary advantage of leveraging metadata for detecting data discrepancies?

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

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

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

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

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

Question 23. 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) Data Integration.

(c) Regression.

(d) Clustering.

(e) Feature Selection.

Question 24. According to slide 15 of the Data Wrangling Part 2 lecture, how can clustering be used to help identify outliers?

(a) Clustering puts outliers in the largest cluster, since clusters are always part of the majority group.

(b) Outliers do not affect the clustering process and clusters cannot identify them.

(c) Clustering looks for data points that are close to the average value of the data set, revealing outliers in the process.

(d) Clusters create distinct groups, so any points outside of those groups can be easily identified as a potential outlier.

(e) `Clustering reduces the data set to only the most central data points, eliminating outliers.`

Question 25. According to the lecture on Data Wrangling part-2 from Slide 14: Regression For Data Smoothing. Let's assume you have a dataset where the `Annual_Income` attribute has missing values for some records. After analysis, you discover that `Annual_Income` is highly correlated with extttEducation_Level and `Years_of_Work_Experience`.

Which imputation method should you apply to most accurately estimate the missing `Annual_Income` values?

(a) `Replace missing incomes with the overall mean income.`

(b) `Fill in missing incomes with zero.`

(c) `Use regression imputation using Education_Level and Years_of_Work_Experience.`

(d) `Discard all records with missing Annual_Income.`

(e) `Apply random imputation using values from observed incomes.`