cap5771sp25
March 19, 2025

Exam 1 part 2 version c

Question 1. 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 survey where some participants' responses are missing.`

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

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

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

Question 2. 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) `One-hot encoding.`

(b) `Binning.`

(c) `Clustering.`

(d) `Regression.`

Question 3. According to the lecture Data Wrangling Part 2 (Slide 8), why is data cleaning considered an iterative process?

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

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

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

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

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

Question 4. 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) `Causal relations in nominal data`

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

(c) `Correlation in numeric data`

(d) `Correlation in nominal data`

(e) `Causal relations in numeric data`

Question 5. According to slide data-wrangling-v2, what is the primary purpose of the chi-squared (ÏĞÂš) test in data analysis?

(a) `To calculate the exact probability of an event occurring in a dataset.`

(b) `To identify the correlation coefficient between two continuous variables.`

(c) `To determine whether there is a significant relationship between categorical variables.`

(d) `To ensure that all datasets conform to a normal distribution.`

(e) `To replace missing values in datasets by estimating expected frequencies.`

Question 6. According to Data Wrangling lecture, what is a key difference between lossless and lossy compression?

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

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

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

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

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

Question 7. 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) `72`

(d) `50`

(e) `150`

Question 8. According to Data Wrangling lecture slides (Slide 32), what is the curse of dimensionality, and how does it impact machine learning models?

(a) `More dimensions always improve model accuracy by providing additional features for learning.`

(b) `The curse of dimensionality only affects text-based datasets, not numerical or categorical data.`

(c) `The curse of dimensionality is when datasets become too large to fit into memory, causing slow computations.`

(d) `Increasing the number of dimensions makes feature selection unnecessary since all features contribute equally.`

(e) `As the number of dimensions increases, data points become more sparse, making clustering and distance-based models less meaningful.`

Question 9. 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) `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.`

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

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

(d) `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.`

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

Question 10. Which of the following statements about Principal Component Analysis (PCA) and Min-Max Normalization is FALSE?

(a) `PCA reduces dimensionality by projecting data onto orthogonal components that capture maximum variance.`

(b) `PCA requires input data to be normalized or standardized before application.`

(c) `Min-Max Normalization scales data to a fixed range, e.g., using the formula` $x' = \frac{x - \texttt{min}(x)}{\texttt{max}(x) - \texttt{min}(x)}$

(d) `PCA is effective in handling multicollinearity by transforming correlated variables into uncorrelated components.`

(e) `Min-Max Normalization is robust to outliers in the dataset.`

Question 11. 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 selecting a sample from a homogeneous population where all members have similar characteristics`

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

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

(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 12. 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) `Data Compression.`

(b) `Concept Hierarchy Generation.`

(c) `Data Integration.`

(d) `Data Cleaning.`

(e) `Dimensionality Reduction.`

Question 13. According to the Data Wrangling lecture, If a hospital is using an automated system to detect anomalies in patient records. The system flags a case where a patient's recorded heart rate is 250 bpm, which is far beyond normal human limits. However, upon human inspection, it was found that the value was incorrectly entered due to a versioning error by the new device.

Which option below best communicates the role of combined computer and human inspection in data cleaning?

(a) `Computers can always detect and correct errors without human intervention.`

(b) `Automated detection helps flag potential errors, but human verification ensures correctness.`

(c) `Data cleaning should only rely on human expertise, as machines cannot detect anomalies.`

(d) `Data discrepancies are always intentional.`

(e) `Human inspection is unnecessary when automated tools are used.`

Question 14. According to the data wrangling lecture, which of the following statements best describes the method of binning for data smoothing?

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

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

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

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

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

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

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

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

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

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

Question 16. 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) `Both techniques are exclusively used for reducing overfitting in machine learning models.`

(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) `Data discretization creates more detailed numerical features, while combining features simplifies the dataset.`

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

Question 17. 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 18. According to the lecture on Data Wrangling, what is the primary purpose of data reduction in data preprocessing?

(a) `To integrate data from multiple sources into a single database.`

(b) `To obtain a reduced representation of the data set that is much smaller in volume but produces similar analytical results.`

(c) `To create new attributes that capture information more effectively than the original ones.`

(d) `To transform all data into a standardized format.`

(e) `To eliminate all noise and inconsistencies from the data set.`

Question 19. Think back to the Data Wrangling demo - when you first downloaded the data from Kagglehub, the data was stored in a hidden directory. What command could be used to find this folder?

(a) `ls -al`

(b) `cd`

(c) `mv`

(d) `ls -l`

(e) `chmod`

Question 20. According to slide 15 of the Data Wrangling Part 2 lecture, how can clustering be used to help identify outliers?

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

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

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

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

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

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 SSN field is formatted differently across databases, requiring transformation.`

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

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

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

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

Question 22. According to the Data Wrangling lecture, which of the following best describes numerosity reduction in data preprocessing?

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

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

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

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

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

Question 23. 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) `Min-Max Scaling.`

(b) `Binarization.`

(c) `Standard Scaler.`

(d) `Z-Score Normalization.`

(e) `Decimal Scaling.`

Question 24. According to the lecture on the Data Wrangling Demo, what does this line of code do in descriptive_utils.py?

null_counts_rows[null_counts_rows == max_nulls].index.tolist()

(a) `It selects the row indices where the number of missing values is equal to the minimum number of missing values found in the dataset.`

(b) `It selects the row indices where the total count of missing values is greater than or equal to the maximum number of missing values found in the dataset.`

(c) `It selects the row indices where the number of missing values is equal to the maximum number of missing values found in the dataset.`

(d) `It selects the row indices where at least one column has missing values in the dataset.`

(e) It selects the row indices where the number of missing values is equal to the average number of missing values across all rows.

Question 25. Principal Component Analysis (PCA) as a data reduction method minimizes error by:

(a) removing all correlations between variables.

(b) capturing the smallest amount of variation in data.

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

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

(e) maximizing the number of features in the dataset.