cap5771sp25
March 19, 2025

Exam 1 part 2 version b

Question 1. 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) `Automated detection helps flag potential errors, but human verification ensures correctness.`

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

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

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

(e) `Data discrepancies are always intentional.`

Question 2. According to the lecture materials, which method of data cleaning would be MOST appropriate when dealing with a dataset where values are expected to form distinct groups and deviations from these groups are considered errors?

(a) `Chi-square analysis`

(b) `Linear regression smoothing`

(c) `Binning with equal-width partitioning`

(d) `Clustering-based smoothing`

(e) `Z-score normalization`

Question 3. 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) `Binning.`

(b) `One-hot encoding.`

(c) `Regression.`

(d) `Clustering.`

Question 4. 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 at least one column has missing values in the dataset.`

(c) `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.`

(d) `It selects the row indices where the number of missing values is equal to the maximum number of missing values found 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 5. Which of the following statements about Principal Component Analysis (PCA) and Min-Max Normalization is FALSE?

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

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

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

(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 6. According to the Data Wrangling v2 lecture, which of the following is not true of Correlation Analysis?

(a) `Correlation analysis is one of the methods that can be used to identify redundant attributes in datasets.`

(b) `Pearson's product moment coefficient can be used for correlation analysis of numeric data.`

(c) `If our correlation analysis shows a high correlation between attributes, then it also implies causation between these attributes.`

(d) `The smaller the value given by a Chi-Square test the greater the correlation.`

(e) `The Chi-Square test is used for correlation analysis of nominal data.`

Question 7. 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) `A point with a z-score of 5 was marked as an outlier.`

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

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

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

Question 8. 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 always results in better model performance.`

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

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

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

Question 9. 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) `Duplicate the values for machines with less frequent recordings so that we can match all data to the highest logging frequency.`

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

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

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

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

Question 10. 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) `Normalize salary values between 0 and 1 using Min-Max scaling.`

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

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

Question 11. 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 automatically detects and corrects all invalid or incomplete date values without any user intervention.`

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

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

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

Question 12. 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) `Both combining features and data discretization primarily serve the same purpose: reducing redundancy in the dataset.`

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

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

(e) `Combining features primarily reduces dataset size, while data discretization improves model accuracy by increasing feature granularity.`

Question 13. 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) `cd`

(b) `chmod`

(c) `ls -al`

(d) `ls -l`

(e) `mv`

Question 14. 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 filtered Series containing only the rows where the null count equals max_nulls.`

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

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

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

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

Question 15. 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 Cleaning.`

(b) `Dimensionality Reduction.`

(c) `Data Compression.`

(d) `Data Integration.`

(e) `Concept Hierarchy Generation.`

Question 16. 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 store full datasets while reducing redundancy through compression.`

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

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

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

Question 17. 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 rely on the data being normally distributed; linear regression is used to smooth the data, and clustering is used to reduce noise.`

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

(c) `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.`

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

(e) 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.

Question 18. According to lecture slides on data wrangling(Slide 50).Which data transformation technique is primarily used to prepare continuous numerical data for categorical analysis?

(a) Feature Construction

(b) Aggregation

(c) Smoothing

(d) Discretization

(e) Normalization

Question 19. 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 numeric data

(b) Correlation in nominal data

(c) Causal relations in nominal data

(d) The mean of a type of numeric data

(e) Causal relations in numeric data

Question 20. According to the data wrangling lecture, which of the following statements best describes the method of binning for data smoothing?

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

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

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

(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 21. Which of the following is an outlier detection method that uses statistical techniques?

(a) `Feature Splicing`

(b) `Interquartile Range (IQR)`

(c) `Data Normalization`

(d) `Data Binning`

(e) `Principal Component Aggregation (PCA)`

Question 22. 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) `Reducing data volume by using alternative, smaller representations without losing analytical results.`

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

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

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) `Clustering.`

(b) `Data Integration.`

(c) `Regression.`

(d) `Feature Selection.`

(e) `Data Augmentation.`

Question 24. According to the lecture on Data Wrangling, which of the following is NOT considered a measure of data quality?

(a) `Completeness.`

(b) `Consistency.`

(c) `Accuracy.`

(d) `Latency.`

(e) `Timeliness.`

Question 25. According to the data wrangling demo, why is it recommended to handle outliers before imputing missing data?

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

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

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

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

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