Data Pre-Processing Comprehensive Question Bank

CSCI316 Big Data Mining Techniques and Implementation

QUESTION 1: Data Understanding and Exploration (12 marks)

Part A (4 marks)

Consider a retail dataset with the following attributes for customer transactions:

- CustomerID (string)
- Age (integer, 18-95)
- Gender (M/F)
- PurchaseAmount (float, $0-$5000)
- ProductCategory (Electronics, Clothing, Books, Home, Sports)
- PaymentMethod (Credit, Debit, Cash, Online)
- TransactionDate (DD/MM/YYYY)

a) Identify the dimension of this dataset and classify each attribute by type (numeric/symbolic, single-valued/multi-valued). (2 marks)

b) Explain how you would convert the ProductCategory attribute using one-hot encoding. Provide the resulting binary representation for a customer who purchased both "Electronics" and "Books". (2 marks)

Part B (4 marks)

You discover that 15% of customers have missing Age values, and the PurchaseAmount distribution shows the following statistics:

- Mean: $245
- Median: $180
- Standard Deviation: $320
- Min: $5, Max: $4850

a) What does the relationship between mean and median suggest about the distribution? Calculate the coefficient of variation and interpret it. (2 marks)

b) Propose and justify two different approaches for handling the missing Age values. Which would you recommend and why? (2 marks)

Part C (4 marks)

a) Create a box plot interpretation for the PurchaseAmount data using the given statistics. Identify potential outliers using the IQR method (k=1.5). Show your calculations. (2 marks)

b) Write Python code to calculate the correlation between Age and PurchaseAmount. Explain what correlation values of +0.73, -0.23, and 0.05 would indicate about customer behavior. (2 marks)

QUESTION 2: Data Quality and Cleaning (10 marks)

Part A (5 marks)

A healthcare dataset contains patient records with the following quality issues:

- Patient ages: [25, 34, 127, 45, -5, 67, 89, 23, 156, 41]
- Blood pressure readings: [120/80, 140/90, "N/A", 200/120, 90/60, "error", 110/70]
- Patient IDs: Some duplicates exist due to system errors

a) Identify outliers in the age data using both the 3σ method and IQR method. Show all calculations step by step. Assume mean=60.2 and std=49.7 for the age data. (3 marks)

b) Classify each issue as outlier, noise, or missing value. Propose specific treatment strategies for each type of problem identified. (2 marks)

Part B (5 marks)

a) Write a Python function called `clean_patient_data()` that:

- Removes age outliers using the IQR method
- Replaces invalid blood pressure readings with the median of valid readings
- Identifies and flags duplicate patient records

Include error handling and document your approach. (3 marks)

b) Explain the difference between data polishing and noise filtering. When would you choose one approach over the other? (2 marks)

QUESTION 3: Data Integration and Aggregation (8 marks)

Part A (4 marks)

You need to integrate data from three sources for a university student performance system:

- **Student Management System**: StudentID, Name, Program, Year
- **Learning Management System**: StudentID, CourseCode, Assignment_Score, Lab_Score
- **Examination System**: StudentID, CourseCode, Exam_Score, Grade

a) Identify potential challenges in integrating these datasets. Discuss attribute matching strategies you would employ. (2 marks)

b) Design a schema for the integrated dataset. What validation checks would you implement to ensure data quality after integration? (2 marks)

Part B (4 marks)

Consider the following sales data that needs aggregation:

ProductID	Location	Date	Sales	Quantity
P001	Sydney	2024-01-15	1200	24
P001	Melbourne	2024-01-15	800	16
P002	Sydney	2024-01-15	1500	30
P001	Sydney	2024-01-16	1000	20

a) Show the result of aggregating this data by ProductID and Date (sum sales and quantity). (2 marks)

b) Discuss the trade-offs of data aggregation. What information might be lost, and when is aggregation beneficial? (2 marks)

QUESTION 4: Data Transformation and Normalization (11 marks)

Part A (5 marks)

Given the following customer income data: [25000, 35000, 45000, 120000, 28000, 52000, 180000, 31000]

a) Apply min-max normalization to scale values to [0,1]. Show calculations for the first three values. (2 marks)

b) Apply z-score normalization to the same dataset. Calculate the mean and standard deviation, then normalize the first three values. Show all steps. (3 marks)

Part B (3 marks)

a) Write Python code to implement both min-max and z-score normalization functions from scratch (without using sklearn). Include input validation. (2 marks)

b) When would you choose z-score normalization over min-max normalization? Provide a specific example scenario. (1 mark)

Part C (3 marks)

Consider transforming the categorical variable "Education Level" with values: [High School, Bachelor, Master, PhD]

a) Show the difference between ordinal encoding and one-hot encoding for this variable. (1 mark)

b) Implement a Python function that performs one-hot encoding from scratch. Handle the case where new categories might appear in test data. (2 marks)

QUESTION 5: Imbalanced Data and Sampling (9 marks)

Part A (4 marks)

A binary classification dataset for fraud detection has the following class distribution:

- Legitimate transactions: 9,850 instances
- Fraudulent transactions: 150 instances

a) Calculate the imbalance ratio and explain why this is problematic for machine learning algorithms. (2 marks)

b) Compare undersampling and oversampling approaches for this dataset. What are the risks and benefits of each? (2 marks)

Part B (5 marks)

a) Implement a Python function for stratified sampling that maintains the original class proportions. The function should take a dataset and sample size as parameters. (3 marks)

b) Design and implement a simple oversampling technique (not SMOTE) that creates synthetic minority class samples. Explain your approach and potential limitations. (2 marks)

QUESTION 6: Feature Engineering and Selection (10 marks)

Part A (5 marks)

You have a dataset of flight information with the following features:

- DepartureTime (HH:MM format)
- ArrivalTime (HH:MM format)
- Distance (kilometers)
- Aircraft Type (Boeing737, AirbusA320, etc.)
- Season (Spring, Summer, Fall, Winter)

a) Create three new meaningful features from the existing ones. Justify why each would be valuable for predicting flight delays. (3 marks)

b) Write Python code to extract these features from the raw data. Handle edge cases like flights crossing midnight. (2 marks)

Part B (5 marks)

a) Explain the curse of dimensionality and how it relates to feature selection. Provide a specific example with high-dimensional data. (2 marks)

b) Implement a correlation-based feature selection algorithm that removes features with correlation > 0.95. Show how you would handle both numerical and categorical features. (3 marks)

PRACTICAL CODING QUESTIONS

QUESTION 7: Comprehensive Data Preprocessing Pipeline (15 marks)

You are given a messy customer dataset (`customer_data.csv`) with the following issues:

- Missing values in multiple columns
- Inconsistent date formats
- Outliers in numerical columns
- Mixed data types in some columns
- Duplicate records

Part A (8 marks) Write a complete Python class `DataPreprocessor` that includes methods for:

1. Data exploration and profiling
2. Missing value detection and treatment
3. Outlier detection and handling
4. Data type correction
5. Duplicate removal
6. Feature scaling and encoding

Include proper error handling, logging, and documentation.

Part B (4 marks) Implement a method that generates a comprehensive data quality report including:

- Missing value statistics
- Outlier detection results
- Data type inconsistencies
- Correlation matrix for numerical features

Part C (3 marks) Write unit tests for your key preprocessing methods. Test edge cases and validate that transformations preserve data integrity.

QUESTION 8: Big Data Preprocessing with PySpark (12 marks)

Part A (6 marks)

Write PySpark code to preprocess a large e-commerce dataset with the following requirements:

- Handle missing values using different strategies for different column types
- Detect and remove outliers using statistical methods
- Perform feature engineering to create interaction terms
- Apply one-hot encoding to categorical variables

Show how you would optimize this for large-scale data processing.

Part B (6 marks)

Implement a PySpark pipeline that:

1. Reads data from multiple sources (CSV, JSON, Parquet)
2. Performs data quality checks
3. Applies transformations based on data profiling results
4. Saves the cleaned data in an optimized format

Include error handling and performance optimization strategies.

CASE STUDY QUESTIONS

QUESTION 9: Real-world Data Preprocessing Scenario (13 marks)

You are hired as a data scientist for a streaming service company. You have access to:

- User viewing history (UserID, MovieID, Timestamp, Duration, Rating)
- Movie metadata (MovieID, Title, Genre, ReleaseYear, Director, Duration)
- User demographics (UserID, Age, Gender, Country, SubscriptionType)

The data comes from different sources and has various quality issues typical of real-world scenarios.

Part A (5 marks) a) Perform domain understanding analysis. What business questions could this data answer? (2 marks)

b) Identify potential data quality issues you expect to find and propose solutions. (3 marks)

Part B (4 marks) a) Design a data integration strategy to combine these datasets. Address potential challenges. (2 marks)

b) Create new features that could be valuable for a recommendation system. Justify your choices. (2 marks)

Part C (4 marks) a) Implement a preprocessing pipeline that handles the most critical data quality issues. (2 marks)

b) Design a data validation framework to ensure ongoing data quality as new data arrives. (2 marks)

ADVANCED PROBLEM-SOLVING QUESTIONS

QUESTION 10: Algorithmic Implementation (11 marks)

Part A (6 marks)

Implement the IQR outlier detection method from scratch without using any statistical libraries:

a) Write functions to calculate quartiles (Q1, Q3) and IQR from an unsorted list. (3 marks)

b) Create a function that identifies outliers using the IQR method with customizable k values. (2 marks)

c) Test your implementation with edge cases (empty lists, single values, all identical values). (1 mark)

Part B (5 marks)

a) Implement a robust scaling method that uses median and MAD (Median Absolute Deviation) instead of mean and standard deviation. Explain when this is preferable. (3 marks)

b) Create a visualization function that shows before/after distributions for your scaling method. (2 marks)

INTEGRATION AND EVALUATION QUESTIONS

QUESTION 11: Preprocessing Impact Analysis (8 marks)

Part A (4 marks)

Design an experiment to measure the impact of different preprocessing strategies on machine learning model performance:

a) Outline your experimental design including control variables and metrics. (2 marks)

b) How would you ensure fair comparison between different preprocessing approaches? (2 marks)

Part B (4 marks)

a) Implement a preprocessing evaluation framework that measures data quality improvements using specific metrics. (2 marks)

b) Create visualizations to show the effectiveness of your preprocessing pipeline. (2 marks)

ANSWER GUIDELINES AND MARKING RUBRICS

Theoretical Questions (40-50% of marks)

- Clear conceptual understanding
- Correct terminology usage
- Justified reasoning for approach selection
- Understanding of trade-offs and limitations

Practical Implementation (30-40% of marks)

- Working code with proper syntax
- Efficient algorithms and data structures
- Error handling and edge case management
- Code documentation and readability

Problem-Solving Process (10-20% of marks)

- Step-by-step calculations shown
- Logical approach to complex problems
- Integration of multiple concepts
- Real-world applicability

Mark Distribution Summary:

- **Questions 1-6**: 60 marks (Theory + Basic Implementation)
- **Questions 7-8**: 27 marks (Advanced Coding)
- **Questions 9-11**: 32 marks (Case Studies + Integration)
- **Total**: 119 marks (can be scaled to exam requirements)

This question bank provides comprehensive coverage of data preprocessing topics with varying difficulty levels, practical applications, and assessment formats suitable for a 3-hour examination.