1. What is feature engineering, and how does it work? Explain the various aspects of feature engineering in depth.

Feature engineering involves creating new features or modifying existing ones to improve model performance. It includes techniques like transformation, binning, encoding, extraction, and interaction creation, enhancing the model’s ability to understand and predict data patterns.

2. What is feature selection, and how does it work? What is the aim of it? What are the various methods of function selection?

Feature selection identifies the most relevant features for a model, aiming to improve performance and reduce overfitting. Methods include filter (statistical tests), wrapper (model performance evaluation), and embedded (integrated within learning algorithms).

3. Describe the function selection filter and wrapper approaches. State the pros and cons of each approach.

Filter Approach: Uses statistical methods to evaluate features independently of models.

- Pros: Fast, simple, model-agnostic.

- Cons: Ignores feature interactions.

Wrapper Approach: Evaluates feature subsets using model performance.

- Pros: High accuracy, considers interactions.

- Cons: Computationally expensive, risk of overfitting.

4.

i. Describe the overall feature selection process.

The feature selection process involves:

- Data preprocessing.

- Choosing a selection method (filter, wrapper, embedded).

- Evaluating feature relevance.

- Iteratively refining the selected features.

ii. Explain the key underlying principle of feature extraction using an example. What are the most widely used function extraction algorithms?

Feature extraction transforms raw data into a reduced set of features. Example: PCA reduces data dimensionality while preserving variance. Widely used algorithms include PCA, LDA, and t-SNE.

5. Describe the feature engineering process in the sense of a text categorization issue.

For text categorization, feature engineering involves:

- Text cleaning (removing stopwords, punctuation).

- Tokenization (splitting text into words/phrases).

- Encoding (TF-IDF, word embeddings).

- Dimensionality reduction (PCA, LDA).

6. What makes cosine similarity a good metric for text categorization? A document-term matrix has two rows with values of (2, 3, 2, 0, 2, 3, 3, 0, 1) and (2, 1, 0, 0, 3, 2, 1, 3, 1). Find the resemblance in cosine.

Cosine similarity measures text similarity based on direction, not magnitude, handling varying document lengths. Calculation:

\[ \text{Cosine similarity} = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}|| \times ||\vec{B}||} \]

For the given vectors, similarity ≈ 0.59.

7.

i. What is the formula for calculating Hamming distance? Between 10001011 and 11001111, calculate the Hamming gap.

\[ \text{Hamming distance} = \sum (x\_i \neq y\_i) \]

For 10001011 and 11001111, distance = 2.

ii. Compare the Jaccard index and similarity matching coefficient of two features with values (1, 1, 0, 0, 1, 0, 1, 1) and (1, 1, 0, 0, 0, 1, 1, 1), respectively (1, 0, 0, 1, 1, 0, 0, 1).

\[ \text{Jaccard index} = \frac{|A \cap B|}{|A \cup B|} = \frac{4}{7} \approx 0.57 \]

\[ \text{SMC} = \frac{n\_{11} + n\_{00}}{n} = \frac{4 + 1}{8} = 0.625 \]

8. State what is meant by "high-dimensional data set"? Could you offer a few real-life examples? What are the difficulties in using machine learning techniques on a data set with many dimensions? What can be done about it?

High-dimensional data sets have a large number of features. Examples: Genomics data, image data. Challenges include overfitting and computational complexity. Solutions: dimensionality reduction (PCA), feature selection.

9. Make a few quick notes on:

- PCA (Principal Component Analysis): Reduces dimensionality by transforming data into principal components capturing the most variance.

- Use of vectors: Represent data points in a multi-dimensional space, fundamental in machine learning for feature representation.

- Embedded technique: Integrates feature selection within the model training process, e.g., LASSO.

10. Make a comparison between:

- Sequential backward exclusion vs. sequential forward selection: Backward exclusion starts with all features, removes least significant; forward selection starts with no features, adds most significant.

- Function selection methods: filter vs. wrapper: Filters use statistical measures; wrappers use model performance to evaluate features.

- SMC vs. Jaccard coefficient: SMC measures proportion of matching attributes; Jaccard measures similarity by intersection over union of sets.