Q1. What is feature engineering, and how does it work? Explain the various aspects of feature engineering in depth.

* Feature engineering is the process of creating or modifying features from raw data to improve the performance of machine learning models.
* Aspects: Data Cleaning, Feature Creation, Transformation, Scaling, Normalization, Encoding, Handling Missing Values, Domain Knowledge Incorporation, etc.

Q2. 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 involves choosing a subset of relevant features from the original set to improve model efficiency and prevent overfitting.
* Aim: Improve model performance and reduce complexity.
* Methods: Filter Methods (correlation, statistical tests), Wrapper Methods (search-based techniques), Embedded Methods (model-based techniques).

Q3. Describe the function selection filter and wrapper approaches. State the pros and cons of each approach?

* Filter Approach (Pros): Fast, efficient for high-dimensional data, independent of the learning algorithm.
* Filter Approach (Cons): May not capture feature interaction, less accurate.
* Wrapper Approach (Pros): Captures feature interaction, model-specific performance evaluation.
* Wrapper Approach (Cons): Computationally expensive, may lead to overfitting.

Q4.i. Describe the overall feature selection process.

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

1. Overall Feature Selection Process: Data Preparation, Feature Extraction/Selection, Model Building, Model Evaluation.
2. Feature Extraction Principle: Extracting essential information from data while reducing dimensionality (e.g., using PCA, LDA).

Q5. Describe the feature engineering process in the sense of a text categorization issue.

Convert text to numerical features (bag-of-words, TF-IDF), handle stop words and punctuation, normalize case, apply stemming/lemmatization.

Q6. 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 the cosine of the angle between two vectors, indicating their similarity.
* Cosine Resemblance Calculation: 14 / sqrt(38 \* 20) ≈ 0.119.

Q7.i. What is the formula for calculating Hamming distance? Between 10001011 and 11001111, calculate the Hamming gap.

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).

* 1. Hamming Distance Formula: Number of differing positions between two equal-length strings. Hamming Gap: 10001011 ⊕ 11001111 = 01000100 (Hamming Distance = 2).
  2. ii. Jaccard Index: Intersection of sets / Union of sets (2/6 ≈ 0.333). Jaccard Coefficient: Similarity Matching / (Sum of feature sizes - Similarity Matching) (2 / (8 - 2) = 0.333).

Q8. 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 many features compared to the number of observations.
* Examples: Gene expression data, image data, social network data.
* Difficulties: Curse of dimensionality, increased risk of overfitting, reduced interpretability.
* Solutions: Feature selection, dimensionality reduction.

Q9. Make a few quick notes on:

1. PCA is an acronym for Personal Computer Analysis.
2. 2. Use of vectors
3. 3. Embedded technique

* PCA (Principal Component Analysis) is a dimensionality reduction technique.
* Vectors are used to represent data points in a multi-dimensional space.
* Embedded Technique: Feature selection integrated with model training.

Q10. Make a comparison between:

1. Sequential backward exclusion vs. sequential forward selection
2. Function selection methods: filter vs. wrapper
3. SMC vs. Jaccard coefficient

* Sequential Backward Exclusion vs. Sequential Forward Selection: Backward removes features, forward adds features.
* Filter vs. Wrapper Feature Selection: Filter uses statistical measures, wrapper uses model performance.
* SMC vs. Jaccard Coefficient: Similarity Matching Coefficient is specific to binary attributes, Jaccard is for set comparison.