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

engineering in depth.

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

methods of function selection?

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

approach?

4.

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?

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

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.

7.

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).

8. State what is meant by &quot;high-dimensional data set&quot;? 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?

9. Make a few quick notes on:

PCA is an acronym for Personal Computer Analysis.

2. Use of vectors

3. Embedded technique

10. Make a comparison between:

1. Sequential backward exclusion vs. sequential forward selection

2. Function selection methods: filter vs. wrapper

3. SMC vs. Jaccard coefficient

### **1. What is Feature Engineering, and How Does It Work?**

**Feature Engineering:**

* **Definition:** Feature engineering is the process of using domain knowledge to create new features or modify existing features in a dataset to improve the performance of machine learning models.
* **Aspects:**
  1. **Feature Creation:** Generating new features from existing data. For example, creating a "BMI" feature from "weight" and "height" features in a health dataset.
  2. **Feature Transformation:** Scaling, normalizing, or applying mathematical transformations to features. For example, applying logarithmic transformation to skewed data.
  3. **Feature Extraction:** Extracting meaningful features from raw data, such as using Principal Component Analysis (PCA) to reduce dimensionality.
  4. **Feature Selection:** Identifying and selecting the most relevant features for the model to improve performance and reduce complexity.

### **2. What is Feature Selection, and How Does It Work?**

**Feature Selection:**

* **Definition:** Feature selection is the process of selecting a subset of relevant features for use in model construction, with the aim of improving model performance, reducing overfitting, and decreasing computational cost.
* **Aim:** To improve the efficiency of the model by removing irrelevant or redundant features and to enhance the model's ability to generalize to unseen data.
* **Methods:**
  1. **Filter Methods:** Evaluate features based on statistical measures such as correlation with the target variable (e.g., chi-square test, mutual information).
  2. **Wrapper Methods:** Use a predictive model to evaluate feature subsets (e.g., forward selection, backward elimination).
  3. **Embedded Methods:** Perform feature selection as part of the model training process (e.g., Lasso regression, decision trees).

### **3. Feature Selection Filter and Wrapper Approaches**

**Filter Approach:**

* **Definition:** Evaluates features independently of the learning algorithm using statistical techniques or domain knowledge to select features before model training.
* **Pros:**
  1. **Computationally Efficient:** Less computationally expensive since it doesn’t require model training.
  2. **Scalability:** Can handle large datasets well.
* **Cons:**
  1. **No Interaction:** Does not consider interactions between features.
  2. **Suboptimal Feature Sets:** May select features that are individually significant but not collectively useful.

**Wrapper Approach:**

* **Definition:** Evaluates feature subsets based on model performance by training and validating the model with different subsets of features.
* **Pros:**
  1. **Model-Specific:** Tailored to the specific model being used, potentially leading to better performance.
  2. **Interaction Consideration:** Considers feature interactions during evaluation.
* **Cons:**
  1. **Computationally Expensive:** Requires training multiple models, which can be resource-intensive.
  2. **Risk of Overfitting:** May lead to overfitting, especially with complex models and small datasets.

### **4. Feature Selection and Extraction**

**i. Overall Feature Selection Process:**

1. **Define Objective:** Determine the goal of feature selection (e.g., improving model performance, reducing dimensionality).
2. **Choose Method:** Select an appropriate feature selection method (filter, wrapper, embedded).
3. **Evaluate Features:** Apply the chosen method to evaluate and select features.
4. **Train Model:** Use the selected features to train the model.
5. **Assess Performance:** Evaluate the model’s performance using metrics such as accuracy, precision, recall, and F1-score.

**ii. Feature Extraction Principle:**

* **Principle:** Feature extraction involves transforming raw data into a set of new features that capture the most important aspects of the data. This is often achieved using dimensionality reduction techniques or creating composite features.
* **Example:** Using PCA to reduce the dimensionality of image data while preserving the most significant variance.
* **Common Algorithms:**
  1. **Principal Component Analysis (PCA):** Reduces dimensionality by projecting data onto the principal components.
  2. **Linear Discriminant Analysis (LDA):** Projects data onto a lower-dimensional space while preserving class separability.
  3. **t-Distributed Stochastic Neighbor Embedding (t-SNE):** Reduces dimensionality for visualization while preserving the structure of data points.

### **5. Feature Engineering in Text Categorization**

**Feature Engineering for Text Categorization:**

1. **Text Preprocessing:** Tokenization, stemming, lemmatization, and removing stop words.
2. **Feature Extraction:** Converting text into numerical features using techniques such as bag-of-words, TF-IDF (Term Frequency-Inverse Document Frequency), or word embeddings (e.g., Word2Vec).
3. **Feature Selection:** Selecting the most relevant terms or n-grams to improve model performance and reduce dimensionality.

### **6. Cosine Similarity in Text Categorization**

**Cosine Similarity:**

* **Definition:** Measures the cosine of the angle between two vectors in a multidimensional space. In text categorization, it is used to determine the similarity between two documents based on their term frequency vectors.
* **Example Calculation:**
  + **Vectors:** v1=(2,3,2,0,2,3,3,0,1)\mathbf{v1} = (2, 3, 2, 0, 2, 3, 3, 0, 1)v1=(2,3,2,0,2,3,3,0,1) and v2=(2,1,0,0,3,2,1,3,1)\mathbf{v2} = (2, 1, 0, 0, 3, 2, 1, 3, 1)v2=(2,1,0,0,3,2,1,3,1)
  + **Cosine Similarity Formula:** Cosine Similarity=v1⋅v2∥v1∥∥v2∥\text{Cosine Similarity} = \frac{\mathbf{v1} \cdot \mathbf{v2}}{\|\mathbf{v1}\| \|\mathbf{v2}\|}Cosine Similarity=∥v1∥∥v2∥v1⋅v2​
  + **Dot Product:** 2∗2+3∗1+2∗0+0∗0+2∗3+3∗2+3∗1+0∗3+1∗1=4+3+0+0+6+6+3+0+1=232\*2 + 3\*1 + 2\*0 + 0\*0 + 2\*3 + 3\*2 + 3\*1 + 0\*3 + 1\*1 = 4 + 3 + 0 + 0 + 6 + 6 + 3 + 0 + 1 = 232∗2+3∗1+2∗0+0∗0+2∗3+3∗2+3∗1+0∗3+1∗1=4+3+0+0+6+6+3+0+1=23
  + **Magnitudes:** ∥v1∥=22+32+22+02+22+32+32+02+12=4+9+4+0+4+9+9+0+1=40≈6.32\|\mathbf{v1}\| = \sqrt{2^2 + 3^2 + 2^2 + 0^2 + 2^2 + 3^2 + 3^2 + 0^2 + 1^2} = \sqrt{4 + 9 + 4 + 0 + 4 + 9 + 9 + 0 + 1} = \sqrt{40} \approx 6.32∥v1∥=22+32+22+02+22+32+32+02+12​=4+9+4+0+4+9+9+0+1​=40​≈6.32
  + **∥v2∥\|\mathbf{v2}\|∥v2∥:** 22+12+02+02+32+22+12+32+12=4+1+0+0+9+4+1+9+1=29≈5.39\sqrt{2^2 + 1^2 + 0^2 + 0^2 + 3^2 + 2^2 + 1^2 + 3^2 + 1^2} = \sqrt{4 + 1 + 0 + 0 + 9 + 4 + 1 + 9 + 1} = \sqrt{29} \approx 5.3922+12+02+02+32+22+12+32+12​=4+1+0+0+9+4+1+9+1​=29​≈5.39
  + **Cosine Similarity:** 236.32×5.39≈0.68\frac{23}{6.32 \times 5.39} \approx 0.686.32×5.3923​≈0.68

### **7. Distance and Similarity Measures**

**i. Hamming Distance:**

* **Definition:** Measures the number of differing positions between two strings of equal length.
* **Formula:**d=number of differing positionsd = \text{number of differing positions}d=number of differing positions
* **Example Calculation:**
  + **Strings:** 10001011 and 11001111
  + **Hamming Distance:** Count differing positions: 2nd, 3rd, 6th, and 7th positions differ. Hence, the distance is 4.

**ii. Jaccard Index and Similarity Matching Coefficient (SMC):**

* **Jaccard Index:**
  + **Definition:** Measures the similarity between two sets by comparing the size of their intersection to the size of their union.
  + **Formula:** J(A,B)=∣A∩B∣∣A∪B∣J(A, B) = \frac{|A \cap B|}{|A \cup B|}J(A,B)=∣A∪B∣∣A∩B∣​
  + **Example:** For vectors (1, 1, 0, 0, 1, 0, 1, 1) and (1, 1, 0, 0, 0, 1, 1, 1), intersection size = 4, union size = 6, so Jaccard Index = 4/6 ≈ 0.67.
* **Similarity Matching Coefficient (SMC):**
  + **Definition:** Measures the similarity based on the proportion of matching values.
  + **Formula:** SMC=n11+n00nSMC = \frac{n\_{11} + n\_{00}}{n}SMC=nn11​+n00​​
  + **Example:** For vectors (1, 1, 0, 0, 1, 0, 1, 1) and (1, 1, 0, 0, 0, 1, 1, 1), where n11n\_{11}n11​ = number of matches where both are 1, n00n\_{00}n00​ = number of matches where both are 0, and nnn = total number of pairs.

### **8. High-Dimensional Data Sets**

**High-Dimensional Data Set:**

* **Definition:** Data sets with a large number of features (dimensions), often resulting in challenges related to sparsity and computational complexity.
* **Examples:**
  1. **Text Data:** Document-term matrices where each term represents a feature.
  2. **Genomics:** Gene expression data where each gene represents a feature.
* **Challenges:**
  1. **Curse of Dimensionality:** Increased risk of overfitting and increased computational cost.
  2. **Sparsity:** Many features may have zero values, leading to sparse matrices.
* **Solutions:**
  1. **Dimensionality Reduction:** Techniques such as PCA or t-SNE.
  2. **Feature Selection:** Using filter, wrapper, or embedded methods to reduce the number of features.

### **9. Quick Notes**

**1. PCA (Principal Component Analysis):**

* **Definition:** A dimensionality reduction technique that transforms data into a lower-dimensional space while preserving as much variance as possible.

**2. Use of Vectors:**

* **Definition:** Vectors represent data points in a multidimensional space and are used in various algorithms to measure distances, similarities, or perform transformations.

**3. Embedded Technique:**

* **Definition:** Feature selection performed as part of the model training process. Examples include Lasso regression and decision tree-based methods.

### **10. Comparisons**

**1. Sequential Backward Exclusion vs. Sequential Forward Selection:**

* **Sequential Backward Exclusion:** Starts with all features and iteratively removes the least significant ones based on model performance.
* **Sequential Forward Selection:** Starts with no features and iteratively adds the most significant ones based on model performance.

**2. Feature Selection Methods:**

* **Filter vs. Wrapper:**
  + **Filter:** Independent of the model, faster but may miss feature interactions.
  + **Wrapper:** Considers feature interactions, more accurate but computationally expensive.

**3. SMC vs. Jaccard Coefficient:**

* **SMC (Similarity Matching Coefficient):** Measures proportion of matching values.
* **Jaccard Coefficient:** Measures the size of the intersection relative to the size of the union, focusing on the presence of features.