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

Ans-Feature engineering is the process of using domain knowledge to extract and analyse features from raw data. A feature or column is a property shared by independent units on which analysis or prediction is to be done. Features are used by predictive models and influence results.

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

Ans- Feature selection is the process of selecting a subset of relevant features from a larger set of available features for use in a machine learning model. The aim of feature selection is to improve the model's performance by reducing the number of features used, avoiding overfitting, and reducing training time and computational complexity. There are three main methods for feature selection: filter methods, wrapper methods, and embedded methods. The choice of method depends on the specific problem and the available data.

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

Ans- Feature selection is the process of selecting a subset of relevant features from a larger set of available features to improve the performance of a machine learning model. There are two main approaches to feature selection: filter and wrapper methods.

Filter methods involve evaluating the correlation between each feature and the target variable using statistical measures like mutual information, correlation coefficients, or the chi-squared test. The features are then ranked based on their correlation with the target variable, and a subset of the top-ranked features is selected for use in the machine learning model. The main advantage of filter methods is their speed and scalability, as they can be applied to large datasets with many features. However, filter methods may overlook complex interactions between features and may not consider the relevance of features for a specific machine learning algorithm.

Wrapper methods, on the other hand, involve evaluating the performance of the machine learning algorithm with different subsets of features. The algorithm is trained with different feature subsets, and the subset that produces the best performance is selected. Wrapper methods are more accurate than filter methods but are computationally expensive as they require repeatedly training the algorithm with different feature subsets. Wrapper methods can also overfit the model to the training data, leading to reduced generalization performance.

**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?**

Ans-i. Feature selection is the process of identifying and selecting the most important and relevant features from a given dataset. The overall feature selection process can be summarized as follows:

1. Define the problem and the objectives of the analysis.
2. Gather and preprocess the data, which includes data cleaning, normalization, and transformation.
3. Explore the data through visualization and statistical analysis to gain insights into the data.
4. Select a set of candidate features based on domain knowledge or statistical criteria.
5. Evaluate the candidate features using appropriate metrics, such as accuracy, precision, recall, and F1-score.
6. Choose the final set of features based on the evaluation results and the trade-off between performance and complexity.
7. Test the performance of the model using the selected features on a separate validation dataset.

ii. The key underlying principle of feature extraction is to transform raw data into a more meaningful representation that can be used to train machine learning models. This involves identifying and extracting relevant features from the data, while minimizing noise and removing redundancies.

For example, in the case of image recognition, we might use feature extraction algorithms to identify edges, shapes, and textures in the image, and then use these features as inputs to a machine learning model. Some widely used feature extraction algorithms include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and t-SNE (t-distributed Stochastic Neighbour Embedding). These algorithms can be used to identify the most important features in a dataset, reduce the dimensionality of the data, and improve the accuracy and efficiency of machine learning models.

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

Ans- The feature engineering process for text categorization involves selecting, extracting, and transforming relevant features from raw text data to create a feature vector that can be used as input to a machine learning algorithm. This process includes steps such as data collection, text preprocessing, feature extraction, feature selection, feature transformation, and model training and evaluation. The goal is to identify the most relevant features and create a feature vector that improves the performance of the machine learning algorithm. The process may require multiple iterations and relies on domain knowledge and intuition to select and transform features that are relevant to the text categorization task.

**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.**

Ans- Cosine similarity is a popular metric for text categorization because it measures the similarity between two documents based on the cosine of the angle between their feature vectors. In other words, it measures how similar two documents are based on the direction of their feature vectors, rather than their magnitude. This makes it a good metric for text categorization because it is insensitive to the length of the documents and focuses only on their content.

To calculate the cosine similarity between the two rows given in the document-term matrix, we first need to calculate the dot product of the two vectors, which is the sum of the products of the corresponding elements:

(22) + (31) + (20) + (00) + (23) + (32) + (31) + (03) + (1\*1) = 25

Also need to calculate the magnitude (or length) of each vector, which is the square root of the sum of the squares of its elements:

sqrt(2^2 + 3^2 + 2^2 + 0^2 + 2^2 + 3^2 + 3^2 + 0^2 + 1^2) = sqrt(30) = 5.48

sqrt(2^2 + 1^2 + 0^2 + 0^2 + 3^2 + 2^2 + 1^2 + 3^2 + 1^2) = sqrt(30) = 5.48

Now, we can calculate the cosine similarity by dividing the dot product by the product of the magnitudes of the vectors:

cosine\_similarity = dot\_product / (magnitude1 \* magnitude2) = 25 / (5.48 \* 5.48) = 0.817

Thus, the resemblance in cosine between the two rows is 0.817. This indicates a relatively high degree of similarity between the two documents.

**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).**

Ans-i.The formula for calculating Hamming distance is:

Hamming distance = number of positions where symbols are different

Between the strings 10001011 and 11001111, the Hamming distance is 2, since the symbols in positions 3 and 7 are different.

**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?**

Ans-A high-dimensional dataset is a collection of data points that have many features or attributes, making it difficult to visualize or analyse directly. In other words, a dataset is considered high-dimensional when the number of features or variables is much larger than the number of observations or data points.

Real-life examples of high-dimensional datasets such ,

* Financial data that measures many different economic indicators, such as stock prices, interest rates, and exchange rates.
* Image or video data that have a large number of pixels, colors, or frames.
* Text data that have a large number of words, topics, or sentiments.

The difficulties in using machine learning techniques on a high-dimensional dataset are mainly due to the curse of dimensionality, which refers to the fact that the sample space grows exponentially with the number of dimensions. As a result, the data become sparse, and the risk of overfitting increases, making it challenging to generalize the results to new data. Also, the computational cost of many algorithms increases rapidly with the number of dimensions, making it difficult to process large datasets in a reasonable amount of time.

To address these challenges, several techniques can be used to reduce the dimensionality of the data, such as Feature selection, Feature extraction and Regularization etc.

**9. Make a few quick notes on:**

**i. PCA is an acronym for Principal component Analysis.**

**ii. Use of vectors**

**iii. Embedded technique**

Ans- i. PCA is an acronym for Principal Component Analysis, which is a technique used for dimensionality reduction and feature extraction.

ii. Vectors are often used in machine learning for representing data points and features. Vectors can be used to represent numerical values or categorical values as one-hot encodings.

iii. Embedding techniques are used to map high-dimensional data into a lower-dimensional space while preserving the most relevant information. Embeddings are commonly used in natural language processing for representing words as vectors in a lower-dimensional space.

**10. Make a comparison between:**

**i. Sequential backward exclusion vs. sequential forward selection**

**ii. Function selection methods: filter vs. wrapper**

**iii. SMC vs. Jaccard coefficient**

Ans- i. Sequential backward exclusion and sequential forward selection are two common feature selection methods used in machine learning. Sequential backward exclusion starts with all the features and then removes one feature at a time, in a backward direction, until the desired number of features is reached. On the other hand, sequential forward selection starts with one feature and adds one feature at a time, in a forward direction, until the desired number of features is reached. The main difference between these methods is the direction of the search. Sequential backward exclusion starts with all the features and gradually eliminates them, while sequential forward selection starts with a single feature and gradually adds more.

ii. Function selection methods are used in feature selection to determine which features should be included in a model. Filter methods evaluate the relevance of features to the target variable and rank them based on a predefined metric. Wrapper methods, on the other hand, use a model to evaluate the performance of a subset of features and select the one that performs the best. Filter methods are faster and less computationally expensive than wrapper methods, but they may not always capture the complex relationships between features.

iii. SMC (Simple Matching Coefficient) and Jaccard coefficient are both similarity measures used in data analysis. SMC measures the proportion of matched features between two data points, while Jaccard coefficient measures the proportion of intersecting features to the union of features between two data points. While both measures are useful for analyzing data, SMC is more appropriate when the data is binary and the number of features is relatively small, while Jaccard coefficient works well with non-binary data and a larger number of features.