1.What exactly is a feature? Give an example to illustrate your point.

Ans.

A feature is an individual measurable property or characteristic of a phenomenon or object that is used to represent it in a dataset.

For example, in a dataset of housing prices, features could include the number of bedrooms, the square footage, the location, and the age of the house.

2.What are the various circumstances in which feature construction is required?

Ans.

Feature construction is required when the available features in a dataset are not sufficient to accurately represent the underlying phenomenon or when there is noise or redundancy in the available features.

Examples of circumstances that may require feature construction include text analysis, image processing, and signal processing.

3.Describe how nominal variables are encoded.

Ans.

Nominal variables are categorical variables that do not have a natural ordering.

They can be encoded using one-hot encoding, where each category is represented by a binary value in a separate column, or using label encoding, where each category is assigned a unique integer value.

4.Describe how numeric features are converted to categorical features.

Ans.

Numeric features can be converted to categorical features by dividing them into intervals or bins and assigning a category label to each bin.

For example, the ages of individuals in a dataset could be converted to categorical features by dividing them into age ranges, such as "0-18", "19-30", "31-50", and "51 and above".

5.Describe the feature selection wrapper approach. State the advantages and disadvantages of this approach.

Ans.

The feature selection wrapper approach involves selecting features based on how well they improve the performance of a specific machine learning algorithm.

Advantages of this approach include that it can result in a more optimized set of features for the specific algorithm being used, and it can help identify important interactions between features.

Disadvantages include that it can be computationally expensive and may not generalize well to other algorithms or datasets.

6.When is a feature considered irrelevant? What can be said to quantify it?

Ans.

A feature is considered irrelevant if it does not provide any useful information for the machine learning task at hand or if it introduces noise or redundancy into the dataset.

One way to quantify the relevance of a feature is to measure its correlation with the target variable or to use statistical tests to determine if it significantly affects the target variable.

7.When is a function considered redundant? What criteria are used to identify features that could be redundant?

Ans.

A function is considered redundant if it can be expressed as a linear combination of other functions or if it does not provide any additional information beyond what is already captured by other functions.

Criteria used to identify potentially redundant features include measuring the correlation between features, checking for linear dependencies, and using techniques such as principal component analysis.

8.What are the various distance measurements used to determine feature similarity?

Ans.

There are several distance measurements that can be used to determine feature similarity, including:

Euclidean distance

Manhattan distance

Cosine similarity

Jaccard distance

Pearson correlation distance

9.State difference between Euclidean and Manhattan distances?

Ans.

The Euclidean distance is the shortest distance between two points in a straight line. It is calculated as the square root of the sum of the squares of the differences between corresponding coordinates. On the other hand, the Manhattan distance is the sum of the absolute differences between corresponding coordinates. It is also known as the "taxicab" distance because it is like calculating the distance between two points on a grid-like city map where you can only travel parallel to the x-axis or y-axis.

10.Distinguish between feature transformation and feature selection.

Ans.

Feature transformation involves transforming the original features into a new set of features using mathematical or statistical operations. The goal is to create a new set of features that better captures the underlying patterns and relationships in the data. Examples of feature transformation include principal component analysis (PCA) and scaling.

Feature selection involves selecting a subset of the original features that are most relevant to the problem at hand. The goal is to remove irrelevant or redundant features to improve the performance of the model and reduce overfitting. Examples of feature selection include filter methods, wrapper methods, and embedded methods.

11.Make brief notes on any two of the following:

Ans.

SVD (Standard Variable Diameter Diameter)

SVD, or Singular Value Decomposition, is a matrix factorization technique that is used to reduce the dimensionality of data. It decomposes a matrix into three matrices: a left singular matrix, a diagonal matrix of singular values, and a right singular matrix. SVD is often used for data compression, noise reduction, and collaborative filtering.

Collection of features using a hybrid approach

A hybrid approach to feature selection involves combining multiple feature selection techniques to identify the most relevant features. This approach can be more effective than using a single technique because different techniques may be better suited to different types of data and feature sets. For example, a hybrid approach may involve using a filter method to remove low-variance features, a wrapper method to select features based on their impact on model performance, and an embedded method to identify features that are important to a specific algorithm.