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

Features are variables which are use t train a model. There are to type of features, independent and dependent. Algorithm try to find the patters and relation between these independent features and their relation with target or dependent features. for example in case if predicting house price (target feature), number of rooms, bathrooms, parking, space of home and location of home (independent variables) are use by algorithm.

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

- In transformation, if the data is categorical, if may need to transform them into numeric values and also encode them

- In scaling, for example if the data are numeric with various scale, the scaling and bringing all of them in same scale may needed.

- For recognizing type of problem which is classification or regression

3. Describe how nominal variables are encoded.

It depends on the method of encoding, if we use one hot encoding, firstly system recognize number of categories, the create columns per category and fill the related column by 1 and other by 0

In case of label encoding, assign a number to each category and transform them into those numbers.

There are various types of encoding such as hash and binary, encoding

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

We can define different intervals or bins for numerical data and assign a name or number to each interval and all the number in range of selected interval change to the selected label of that bin. For example, age divide to 4 bins (0-20, 20-40, 40-60, above 60)

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

In wrapper methods, the feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset. It is a greedy search approach by evaluating all the possible combinations of features and find features that give highest performance. This method increases cost of modelling and my lead to overfitting but is is accurate and can find the exact best combination of features.

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

When the data has por quality, may have a lot of missing value or noise are irrelevant, also some times the relation between data and target is less than that data cannot give suitable knowledge to model. There are various method such as chi-test, fisher value, correlation coefficient and also some of the algorithm such as random forest can measure the importance of features

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

When two items have same distribution and behaver in same circumstance we call them redundant, in case of features also in by increasing value in a feature another feature value also increase we call them they are redundant or for example if we have location and pin code of a place these two features are almost redundant and we can keep one of them and drop the other one.

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

Euclidean Distance, Manhattan Distance, Minkowski Distance, and Hamming Distance are some of most common use type of distance measurement

9. State difference between Euclidean and Manhattan distances?

Euclidean distance is the shortest path between source and destination which is a straight line between two point but Manhattan distance is sum of all the real distances between source(s) and destination(d) and each distance are always the straight lines.

10. Distinguish between feature transformation and feature selection.

Feature transformation is various method which are uses to transform type of data for example label encoding which use to transform categorical data into numerical data or also they use to scale data such as standard scaler which try to normalize data. in this way the data can be understandable by computer and also control the performance of the model and reduce computational process.

But feature selection is a method to select most relevant feature to the target feature which help to reduce computational process and complexity of the model and also cost of it.

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

1.SVD (Standard Variable Diameter Diameter)

2. Collection of features using a hybrid approach

3. The width of the silhouette

It is a widely used index for assessing the fit of individual objects in the classification, as well as the quality of clusters and the entire classification. The Silhouette Coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample

4. Receiver operating characteristic curve

A ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate and False Positive Rate