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

A: A feature is nothing but an input variable used to train a model. Predictions are also made using features. For example, consider a bank is trying to predict whether a potential borrower will default or not. Here, income can be a good feature to use as input.

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

A: When some variables or observations have information that is not easily interpretable, it would be useful to divide them into small easily consumable variables. In other cases, it might be useful to combine two or more variables to create a more meaningful variable.

3. Describe how nominal variables are encoded.

A: For nominal variables, there is no order for the values. So, each value has to be treated equally. In order to do so, dummy variables are assigned to each value and each of these dummy variables have values 0 and 1.

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

A: Numeric features can be divided into ranges and each range is considered as a category.

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

A: In this approach a subset of features are used to train a model. Based on the performance of this model, it is decided whether to add or remove features from this subset.

This method increases the accuracy but it might be prone to overfitting.

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

A: A feature is said to be irrelevant when it does not contribute to predictions. Relevance of a feature is termed as feature importance. Feature importance is quantified in different ways for different models. For example, coefficients can be used for linear regression.

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

A: A feature is said to be redundant when it depends on another feature. Measures such as correlation values, VIF scores can be used to identify redundant features.

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

A: Euclidean, Manhattan and Minkowski distances.

9. State difference between Euclidean and Manhattan distances?

A: Euclidean distance is the geometrical distance between two points. Manhattan distance represents the sum of differences between X and Y coordinates of two points respectively.

10. Distinguish between feature transformation and feature selection.

A: Feature transformation implies changing values within a feature while feature selection deals with features themselves.

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

A: Width of silhouette represents how close a point is to its current cluster compared to the closest cluster.

4. Receiver operating characteristic curve

A: True positive rate vs false positive rate is plotted by varying threshold probability in a classification model. The resultant curve is called Receiver operating characteristic curve or ROC curve in short.