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

Feature is nothing but a column representing some information about the case under study

Example

Age, salary, gender, health indicators etc can be used as features to check the health insurance prices

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

1. Noisy data

2. Missing Values

3. Encoding

4. Dimensionality

5. Incorrect data

3. Describe how nominal variables are encoded.

Nominal values are encoded by one hot encoding method, where their presence is marked as 1

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

Numeric features can converted to categorical by using various functions such as Alias, label encoder

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 follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion

Wrapper feature selection also has several drawbacks, including a greater risk of overfitting, longer training times, reliance on a specific classifier

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

One general definition for relevance is that a feature can be regarded as irrelevant if it is conditionally independent of the class labels or it does not influence the class labels; in these cases, it can be discarded.

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

be redundant?

A redundant function is one which has potentially been superseded by another function, and shouldn't be used anymore. For example, if two features {X1, X2} are highly correlated, then the two features become redundant features since they have same information in terms of correlation measure

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

* Cosine Similarity:
* Manhattan distance
* Euclidean distance
* Minkowski distance
* Jaccard similarity

9. State difference between Euclidean and Manhattan distances?

Euclidean distance is the shortest path between source and destination which is a straight line, 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.

The main difference:- Feature Extraction transforms an arbitrary data, such as text or images, into numerical features that is understood by machine learning algorithms. Feature Selection on the other hand is a machine learning technique applied on these (numerical) features.

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

3. The width of the silhouette

Silhouette width is a widely used measure for assessing the fit of individual objects in the classification, as well as the quality of clusters and the entire classification. This index uses two clustering criteria, compactness (average within-cluster distances) and separation (average between-cluster distances), which implies that spherical cluster shapes are preferred over others

4. Receiver operating characteristic curve

The receiver operating characteristic (ROC) curve, which is defined as a plot of test sensitivity as the y coordinate versus its 1-specificity or false positive rate (FPR) as the x coordinate, is an effective method of evaluating the quality or performance of diagnostic tests, and is widely used in radiology to evaluate the performance of many radiological tests. Although one does not necessarily need to understand the complicated mathematical equations and theories of ROC analysis, understanding the key concepts of ROC analysis is a prerequisite for the correct use and interpretation of the results that it provides