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

Ans1:- Any tangible, measurable data in form of a list, or column that becomes the independent variable for any machine learning modelling problem, is called a feature.

exp = [4,5,3,4,3,7,9,10,3,2,4]

salary = [25000,30000,20000,25000,20000,60000,80000,100000,20000,15000,25000]

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

Ans2:-

When more accurate model has to be created from limited data at hand

When improvement in generalizability is required

When improvement in model interpretability is required.

3. Describe how nominal variables are encoded.

Ans3:- Nominal variables can be encoded in several ways

One Hot Encoding Binary columns equal to the number of class labels present in the nominal variable are created in the form of sparse matrix, such that 0 indicates absence of a class label in a binary column and 1 indicates presence of a class label. Large number of such new variables or dummy variables can lead to the problem of dummy variable trap.

Integer Encoding Values of a nominal variable can be changed into integer values. A column containing class labels New Delhi and Mumbai can be encoded into integer values in the form of 1 and 2.

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

Ans4:- Numeric features or ordinal features can be converted into categorical features by putting values of a numerical feature into bins or intervals. These intervals then become class labels and the feature becomes a categorical feature.

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

Ans5:- Wrapper approach is an iterative approach in feature selection where all combinations of subset of features is used to train the model and a performance metric is used to test the model and check of there is an improvement of score over the previous iteration where a different subset of features were used. Examples of this technique are forward feature selection and backward feature selection.

Advantage is that it works better than filter approach. Disadvantage is that it requires a lot of time to compute feature selection.

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

Ans6:- A feature is considered irrelevant when it does not contribute or is detrimental to the performance of a model. Feature selection methods can be used to quantify a feature's relevance. For example, in a Multiple Linear Regression problem, correlation coefficient can be used to quantify relevance of a feature.

Techniques like forward feature selection can be used to iteratively quantify which features do not contribute to the accuracy of a machine learning model and can be removed.

Feature importance is supported by many machine learning models like Decision Trees and it can be used to quantify relevance of features.

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

Ans7:- A feature is considered redundant when it does not contribute to the performance of a model by providing unique information.

Criteria like Multicollinearity, non-unique information contribution, low feature importance score, little to no improvement in performance during forward and backward feature selection techniques, ANOVA etc. help identify features that could be redundant.

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

Ans8:- 1. Euclidean(L2) Distance:--- The most commonly used measure of distance, it is the length of shortest line between any two points x1 and x2. For any 2 dimensional data, we can write the formula as d = ((p1-q1)^2+(p2-q2)^2)^1/2

2. Minkowski Distance:--- It is a generalization of both Euclidean and Manhattan Distance. d=(sigma(|pi-qi|^p))^1/p where p is the value of order of normalization. When p=1, it represents Manhattan Distance and when p=2, it represents Euclidean Distance.

3. Hamming Distance:--- It is measured as the number of locations at which binary vectors or binary values from two features differ. Hamming distance becomes very important while handling data of gene sequencing and gene code.

4. Manhattan(L1) Distance:--- Sum of absolute differences between points across all dimensions. It can be written as d=sigma(|pi-qi|) where pi and qi are values for i to n data points.

9. State difference between Euclidean and Manhattan distances?

Ans9:- Euclidean distance is the length of the shortest line between any two points x1 and x2. Manhattan distance is the sum of absolute differences between points across all dimensions.

Euclidean distance is mathematically written d = ((p1-q1)^2+(p2-q2)^2)^1/2. Manhattan distance is d=sigma(|pi-qi|) where pi and qi are values for i to n data points.

Euclidean Distance performs poorly on high dimensional data. Manhattan distance works better than Euclidean Distance for high dimensional data.

10. Distinguish between feature transformation and feature selection.

Ans10:- Transforming datapoints into higher or lower dimension, or simply into features that can be used better for training phase and creating new features from existing features is called Feature Engineering. Feature Selection is a method used in which techniques can be used to select features based on their contribution to the model both information and performance-wise.

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

Ans11:-

iii. The width of the silhouette

Estimate of average inter cluster distance to give efficacy/performance of cluster algorithms is called width of the silhouette. It can also be defined as how identical/similar a data point 'x' is to the data points inside the cluster to which x is assigned. Its value ranges from -1 to 1 where 1 means good and -1 means bad.

iv. Receiver operating characteristic curve

Curve plotted between True Positive Rate and False Positive Rate is Receiver Operating Characteristics curve and is used to find the area under the curve for ROC-AUC score for binary classification evaluation. True Positive Rate is calculated as Number of positives that were correctly classified divided by total number of positives. False Positive Rate is calculated as the number of positives incorrectly classified divided by total number of negatives.

True Positive Rate and False Positive Rate are calculated for different thresholds values where thresholds take values starting from the highest probability scores assigned to data points and goes up to the lowest probability score.

The curve is impacted by presence of outliers, and simple models. Extensions can be made to this curve to suit multiclass classification evaluation requirements.