1. What is feature engineering, and how does it work? Explain the various aspects of feature engineering in depth.

Answer: Feature engineering is the process of transforming raw data into features that represent the underlying problem to predictive models, resulting in an improved model accuracy on unseen data. Some aspects of feature engineering involves handling categorical values, numerical values, decomposing date time column into various columns.

2. What is feature selection, and how does it work? What is the aim of it? What are the various methods of function selection?

Answer: Feature selection is the process of selecting a small subset of features when building a predictive model for a classification or a regression.

It is needed mostly in cases where we want to reduce the complexity of the model, computation cost and improve the accuracy of the prediction made by the model.

3. Describe the function selection filter and wrapper approaches.

Answer: Wrapper methods measure the usefulness of the feature based on model performance. They solve the real problem but are computationally expensive. Filter methods on the other hand pick up the intrinsic features of the model measured via univariate statistics instead of cross validation for performance.

4.

i. Describe the overall feature selection process.

Answer:

1. Calculate the mutual information as score between independent and dependent features.
2. Select the feature with largest score between independent and dependent feature and add it to list of features selected.
3. Repeat steps a and c till a list of features is selected based on our need.

ii. Explain the key underlying principle of feature extraction using an example. What are the most widely used function extraction algorithms?

Answer: The main aim of doing a feature extraction is that a machine learning model does not understand any other data type other than numerical or binary values and so, we try to convert data types with any other data types with numeric values and we feed it to the ML model as per our choosing we try to increase the performance metrics which we decide to measure the model’s performance on.

5. Describe the feature engineering process in the sense of a text categorization issue.

Answer: There are following features which we can use for text categorization issue:

1. Number of characters
2. Count of words
3. Number of punctuations
4. Number of sentences
5. Number of unique words
6. Number of stopwords
7. Calculating average sentence length
8. Stop words / word count feature

6. What makes cosine similarity a good metric for text categorization? A document-term matrix has two rows with values of (2, 3, 2, 0, 2, 3, 3, 0, 1) and (2, 1, 0, 0, 3, 2, 1, 3, 1). Find the resemblance in cosine.

Answer: Cosine similarity is advantageous because even if the 2 similar documents are considererd far away in terms of Euclidean distance, there are chances that they might be closely oriented and the lesser is the angle between the vectors representing the 2 sentences in a multidimensional space, more they are closer.

Q = acos(23/(square root(29)\* square root(39))) = 0.817

7.

i. What is the formula for calculating Hamming distance? Between 10001011 and 11001111, calculate the Hamming gap.

Answer: Hamming distance is nothing but the count of number of different bits present between 2 strings. Here, our hamming distance is 2.

ii. Compare the Jaccard index and similarity matching coefficient of two features with values (1, 1, 0, 0, 1, 0, 1, 1) and (1, 1, 0, 0, 0, 1, 1, 1), respectively (1, 0, 0, 1, 1, 0, 0, 1).

Answer:

Here M00 = 2, M11=4, M10= 1 and M01 = 1

So SMC = 6/8 and Jaccard = 4/6 on comparing the 2 features accordingly.

8. State what is meant by "high-dimensional data set"? Could you offer a few real-life examples? What are the difficulties in using machine learning techniques on a data set with many dimensions? What can be done about it?

9. Make a few quick notes on:

1. PCA is an acronym for Personal Computer Analysis.

Answer : PCA is acronym for Principal component analysis and it is the process of the reducing the number of features to a minimal amount where we still retain the max information about data and stay away from curse of dimensionality.

2. Use of vectors

Answer: A vector is a representation of a data point in a multi-dimensional space where it helps in segregation of one data point from another data point and also helps in comparison with other data points as well.

3. Embedded technique

Answer: Embedded technique such as L1 regression can be understood as adding a penalty against complexity to reduce the possibility of overfitting or variance by adding more bias

Inshort , regularized cost = cost + regularization penalty

10. Make a comparison between:

1. Sequential backward exclusion vs. sequential forward selection

Answer: For sequential forward selection, initially we will have all sets of attributes and then we add the list of variables to the list of variables to be selected once our stopping criteria for selecting variables is met.

For sequential backward exclusion as well, we will all the variables from start and then we start eliminating features with least importance and once we have only those features which retain the max information needed for our modelling purposes, we can proceed further with good features available to us.

2. Function selection methods: filter vs. wrapper

Answer: Wrapper methods measure the usefulness of the feature based on model performance. They solve the real problem but are computationally expensive. Filter methods on the other hand pick up the intrinsic features of the model measured via univariate statistics instead of cross validation for performance.

3. SMC vs. Jaccard coefficient

Answer: SMC is used to compare the diversity and similarity between 2 given datasets. SMC is very similar to Jaccard coefficient and this is more popular than SMC.

SMC = (M00 + M11)/(M00+M11+M01+M10)

Where A and B are 2 objects having binary variable and in our case here, we have 2 binary variables and

M00 = total # of attributes where both A and B have value of 0

M11 = total # of attributes where both A and B have value of 1

M10 = total # of attributes where both A is 1 and B is 0

M01 = total # of attributes where both A is 0 and B is 1

Main difference between SMC and Jaccard is that SMC has M00 in denominator and numerator whereas Jaccard does not have it.