1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two.

Quantitative data refers to any information that can be quantified, counted or measured, and given a numerical value. Qualitative data is descriptive in nature, expressed in terms of language rather than numerical values.

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

Articulate the problem early

* Classification.
* Clustering
* Regression
* Ranking

Establish data collection mechanisms

* Data Warehouses and ETL
* Data Lakes and ELT
* Handling human factor

Check your data quality

Format data to make it consistent

Reduce data

Complete data cleaning

Create new features out of existing ones

Join transactional and attribute data

Rescale data

4. What are the various causes of machine learning data issues? What are the ramifications?

Understanding Which Processes Need Automation

Lack of Quality Data

Inadequate Infrastructure

Implementation

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

Unique value count

Frequency Count

Variance

Histogram

6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

1. Deleting Rows

This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set.

2. Replacing With Mean/Median/Mode

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set.

3. Assigning An Unique Category

A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values.

7. Describe the various methods for dealing with missing data values in depth.

1. KNN is a machine learning algorithm which works on the principle of distance measure. This algorithm can be used when there are nulls present in the dataset.
2. Predicting The Missing Values - Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance.

8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

Dimensionality Reduction:

This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Component Analysis).

9.

i. What is the IQR? What criteria are used to assess it?

The interquartile range rule is useful in detecting the presence of outliers. Outliers are individual values that fall outside of the overall pattern of a data set.

ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

In its simplest form, the boxplot presents five sample statistics - the minimum, the lower quartile, the median, the upper quartile and the maximum - in a visual display. The box of the plot is a rectangle which encloses the middle half of the sample, with an end at each quartile.

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

1. Data collected at regular intervals

Time series analysis is a method of analyzing a series of data points collected over a period of time. In time series analysis, data points are recorded at regular intervals over a set period of time, rather than intermittently or at random

2. The gap between the quartiles

The interquartile range or IQR is the range of the middle half of a set of data. It is the difference between the upper quartile and the lower quartile.

3. Use a cross-tab

Cross tabulation is a method to quantitatively analyze the relationship between multiple variables

1. Make a comparison between:

1. Data with nominal and ordinal values

Nominal data is a group of non-parametric variables, while Ordinal data is a group of non-parametric ordered variables.

2. Histogram and box plot

Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data.

3. The average and median

Mean: This measures the average of the interval data sample provided.

Median: This is used to determine which variable in the data set is at the centre. It is easily calculated because interval data is equidistant.