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

**Ans:** Data Gathering -> Data Preprocessing -> Split train and test data -> ML algorithm selection -> Model evaluation -> Validation -> Complexity & Improvement if any ->Model deployment.

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

**Ans:** Quantitative data are data about numeric variables (e.g. how many; how much; or how often).

Qualitative data are data about categorical variables (e.g. what type).

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

**Ans:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Name | Age | Sex | Marks | Percentage | Is\_Scholar | Course | University | State |
| 1 | Abc | 23 | Female | 92 | 78.5% | Yes | A | Univ\_A | X |
| 2 | Xyz | 25 | Male | 83 | 92.45% | No | B | Univ\_B | Y |

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

**Ans:**

Not enough training data - Need to collect as much data as possible, or can every time include test data as training data and tune your model continuously.

Poor Quality of data - Put in every ounce of effort in cleaning up your training data

Irrelevant Features - “Garbage in, garbage out (GIGO).” Feature selection, extraction, and by creating new features, we can remove irrelevant features as much as possible.

Nonrepresentative training data - Use representative data during training, so your model won’t be biased among one or two classes when it works on testing data.

Overfitting and Underfitting – To overcome overfitting we can simplify the model by selecting one with fewer parameters, by reducing the number of attributes in training data, constraining the model, gather more training data, reduce the noise.

To overcome underfitting - Select a more advanced model, one with more parameters, Train on better and relevant features, Reduce the constraints.

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

**Ans:** Below are various approaches:

Replacing values - Just replacing the categories with the desired numbers.

Label Encoding - Allows you to convert each value in a column to a number.

One-Hot encoding - convert each category value into a new column and assign a 1 or 0 (True/False) value to the column.

Binary Encoding - In this technique, first the categories are encoded as ordinal, then those integers are converted into binary code, then the digits from that binary string are split into separate columns. This encodes the data in fewer dimensions than one-hot.

Backward Difference Encoding - This technique falls under the contrast coding system for categorical features. A feature of K categories, or levels, usually enters a regression as a sequence of K-1 dummy variables. In backward difference coding, the mean of the dependent variable for a level is compared with the mean of the dependent variable for the prior level. This type of coding may be useful for a nominal or an ordinal variable.

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

**Ans:** If certain variables have missing values, then learning cannot be done correctly, since major parameters will be calculated wrongly and a clear result sets cannot be achieved. To fix it we need to handle these missing values, by imputing data into these missing values.

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

**Ans:** Below are various methods:

Deleting Rows - 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.

Replacing With Mean/Median/Mode - We can calculate the mean, median or mode of the feature and replace it with the missing values.

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.

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

**Ans:** Finding and handling missing values, Encoding categorical data, Feature scaling, standardize & normalize data.

Function selection yields a subset of features from the original set of features, which are best representatives of the data.

Dimensionality reduction is usually performed either by selecting a subset of the original dimensions or/and by constructing new dimensions.

9.

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

**Ans:** IQR stands for inter quartile range, it is used to identify outliers.

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?

**Ans:** Boxplot has minimum, maximum, median, first quartile and third quartile in the data set.

The longer whisker on the lower side suggests that there may be larger variance among the smaller values, since there is a greater distance from the 1st quartile to the lower extreme than from the median to the 1st quartile.

The points that lie outside the whiskers, that is, (1.5 x IQR) in both directions are generally considered as outliers.

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

1. Data collected at regular intervals

2. The gap between the quartiles – This is called interquartile range, it is the range of the middle half of the data that shows how spread out the data is.

3. Use a cross-tab - A contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. They provide a basic picture of the interrelation between two variables and can help find interactions between them.

1. Make a comparison between:

1. Data with nominal and ordinal values - Nominal data is defined as data that is used for naming or labelling variables, without any quantitative value. It is sometimes called “named” data - a meaning coined from the word nominal.

Ordinal data is a type of categorical data with an order. The variables in ordinal data are listed in an ordered manner. The ordinal variables are usually numbered, so as to indicate the order of the list.

2. Histogram and box plot - A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data.

Boxplot has minimum, maximum, median, first quartile and third quartile in the data set.

3. The average and median – Average(mean) is the sum of values divided by count of values, whereas median is the middle value in an ordered data set.