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

A machine learning task is the type of prediction or inference being made, based on the problem or question that is being asked, and the available data. For example, the classification task assigns data to categories, and the clustering task groups data according to similarity. Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms

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

quantitative data

Quantitative data refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it’s quantitative data. Quantitative data can tell you “how many,” “how much,” or “how often”—for example, how many people attended last week’s webinar? How much revenue did the company make in 2019? How often does a certain customer group use online banking?

qualitative data

Unlike quantitative data, qualitative data cannot be measured or counted. It’s descriptive, expressed in terms of language rather than numerical values.

Researchers will often turn to qualitative data to answer “Why?” or “How?” questions. For example, if your quantitative data tells you that a certain website visitor abandoned their shopping cart three times in one week, you’d probably want to investigate why—and this might involve collecting some form of qualitative data from the user. Perhaps you want to know how a user feels about a particular product; again, qualitative data can provide such insights. In this case, you’re not just looking at numbers; you’re asking the user to tell you, using language, why they did something or how they feel.

Qualitative data also refers to the words or labels used to describe certain characteristics or traits—for example, describing the sky as blue or labeling a particular ice cream flavor as vanilla.

* Quantitative data is countable or measurable, relating to numbers. Qualitative data is descriptive, relating to language.
* Quantitative data tells us how many, how much, or how often (e.g. “20 people signed up to our email newsletter last week”). Qualitative data can help us to understand the “why” or “how” behind certain behaviors, or it can simply describe a certain attribute—for example, “The postbox is red” or “I signed up to the email newsletter because I’m really interested in hearing about local events.”
* Quantitative data is fixed and “universal,” while qualitative data is subjective and dynamic. For example, if something weighs 20 kilograms, that can be considered an objective fact. However, two people may have very different qualitative accounts of how they experience a particular event.
* Quantitative data is gathered by measuring and counting. Qualitative data is collected by interviewing and observing.
* Quantitative data is analyzed using statistical analysis, while qualitative data is analyzed by grouping it in terms of meaningful categories or themes.

3. Create a basic data collection that includes some sample records. Have at least one attribute from

each of the machine learning data types.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Gender | Salary | Feedback | Churn |
| 23 | Male | 173829 | It was good | 0 |
| 56 | Female | 234234 | Very bad | 1 |
| 28 | Male | 567634 | Did not like | 1 |

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

* Inadequate Training Data
* Poor quality of data
* Non-representative training data
* Overfitting and Underfitting
* Monitoring and maintenance
* Getting bad recommendations.
* Lack of skilled resources

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

Unique value count – Can be used to analyse how many unique categorical values are present

Frequency Count – can be used to count the number of observations for every class

Cross Tab – to aggregate the numerical values based on the categorical column

6. How would the learning activity be affected if certain variables have missing values? Having said

that, what can be done about it?

The real-world data often has a lot of missing values. If you want your model to work unbiased and accurately then you just can’t ignore the part of “missing value” in your data. One of the most common problems faced in data cleansing or preprocessing is handling missing values

THREE WAYS to treat missing values in dataset are as follows:

* DROPPING
* IMPUTION
* PREDICTIVE MODEL

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

Dropping missing values

This method is commonly used to handle null values. It is easy to implement and there is no manipulation of data required. This varies from case to case on the amount of information you think the variable has. If dataset information is valuable or training dataset has less number of records then deleting rows might have negative impact on the analysis. Deletion methods works great when the nature of missing data is missing completely at random(MCAR) but for non-random missing values can create a bias in the dataset, if a large amount of a particular type of variable is deleted from it.

deleting rows (listwise deletion)

deleting columns

pairwise deletion

Imputing missing values

There exists many approach to missing-data imputation and they usually depend on your problem and how your data algorithm behaves

MEAN AND MEDIAN

If the features are numeric you can use simple approaches, such as mean or median. This is the most common method of imputing missing values of numeric columns. We can make use of ‘SimpleImputer’ from python for this. If there are outliers then the mean will not be appropriate. It’s better to use the median value for imputation in the case of outliers. Median is the middlemost value.

MODE(Frequent Category Imputation)

Mode is the most frequently occurring value. It is used in the case of categorical features. This technique says to replace the missing value with the variable with the highest frequency or replacing the values with the Mode of that column.

8. What are the various data pre-processing techniques? Explain dimensionality reduction and

function selection in a few words.

Predictive Model

It is like running a predictive model to estimate values that will substitute the missing data. you can predict missing value using non-missing data. We just have to divide dataset into two datasets, one with no missing data as a training dataset and second as test dataset having missing values. Then use training dataset to create the model to predict the target variable and predict missing values.

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. This definition is somewhat vague and subjective, so it is helpful to have a rule to apply when determining whether a data point is truly an outlier—this is where the interquartile range rule comes in.

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?

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.

In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.

The Upper quartile (Q3) is the median of the upper half of the data set. The Interquartile range (IQR) is the spread of the middle 50% of the data values. Lower Limit = Q1 – 1.5 IQR. So any value that will be more than the upper limit or lesser than the lower limit will be the outliers.

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

1. Data collected at regular intervals  
Interval data, also called an integer, is defined as a data type which is measured along a scale, in which each point is placed at equal distance from one another. Interval data always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal.

2. The gap between the quartiles

* The quartiles data into three points—a lower quartile, median, and upper quartile—to form four groups of the dataset.
* Along with the minimum and maximum values of the data set, the quartiles divide a set of observations into four sections, each representing 25% of the observations.
* Quartiles are used to calculate the interquartile range, which is a measure of variability around the median.

1. Make a comparison between:

1. Data with nominal and ordinal values

The main differences between Nominal Data and Ordinal Data are: While Nominal Data is classified without any intrinsic ordering or rank, Ordinal Data has some predetermined or natural order. Nominal data is qualitative or categorical data, while Ordinal data is considered “in-between” qualitative and quantitative data.

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

The average is the arithmetic mean of a set of numbers. The median is a numeric value that separates the higher half of a set from the lower half