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

Ans: **the keys task that machine learning entails is data gathering, data pre-processing, training the model, evaluating the model’s performance and then repeat cycle until we get the optimal result. Data pre-processing requires feature engineering in which we clean the data, deal with the duplication and missing values.**

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

Ans: **quantitative data is numerical data or data which can be presented in terms of numbers such as age, weight, height or weather temperature. Where as qualitative data is a categorical data which can be categorized into different classes such as color, gender or nationality.**

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

Ans: **there are mainly 3 types of machine learning data types: continuous, categorical, time series. The below is the sample records showing these three data types:**

|  |  |  |
| --- | --- | --- |
| **Date** | **Temp** | **Weather** |
| 1981-01-01 | 44 | Too hot |
| 1981-01-02 | 35 | Hot |
| 1981-01-03 | 30 | Moderately hot |

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

Ans:

|  |  |
| --- | --- |
| **Causes** | **Ramifications** |
| Poor quality of data/irrelevant data | To get the appropriate set of data |
| Inadequate data | To collect adequate data |
| Poor understanding of the data | Data exploration to be done to understand the technical aspects of the data. |
| Underfitting | Employ various technique to get rid of overfitting such as elimination of features or reduce noise |

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

Ans: **we can calculate the mode to see the most frequent value of a categorical variable.**

**We can create a histogram to see the most frequent value of a categorical feature.**

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

Ans: **missing certain values can make the ML model bias or make the model predict incorrectly. Data imputation can be employed to replace those missing values with the most appropriate value such as mean or median or mode.**

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

Ans: **there are two main method to deal with the missing data values in a dataset:**

1. **Deletion: we can delete the rows or a whole column to get rid of the missing values but this step must be taken after a careful consideration as deleting the whole column or deleting rows when the data set is small may impact the machine learning model.**
2. **Data Imputation: In this we replace the missing values with mean or median or mode. Median value is used when we have outliers and mode is used in the case of categorical variable.**

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

Ans: **dimensionality reduction is used to present data from a higher space to a lower space like converting data from 3-D to 2D. this is achieved by eliminating those features from the data set which are least important or don’t contribute to determining the target variable. As a result, it also decreases the complexity of the model.**

**Feature Selection: in this, we select a subset of the original data set by including only those features which may influence the variation in the dependent variable.**

9.

i. What is the IQR? **What criteria are used to assess it?: IQR are inter quartile range which is the difference between 1st and 3rd quartile. IQR are used to assess the minimum and maximum values so the outliers can be determined. IQR= Q3-Q1 where Q1 is the 25th percentile value and Q3 is the 75 percentile value.**

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?:  **box plot has 5 components including min, max, 1st quartile, median and 3rd quartile. Box plot is used to identify outliers as we define the min and max range in the plot and whatever lies outside those range is considered as an outlier. The lower whisker surpasses the length of the upper whisker when there is more data present between the 25th percentile and the minimum value compared to the data present between the maximum value and 75th percentile.**

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

1. Data collected at regular intervals

2. The gap between the quartiles: **quartiles are used in 5 point summary where 5 aspects to be calculated such as minimum value, 1st quartile, median, 3rd quartile and maximum value. The gap between the 1st and 3rd quartile is known as IQR or inter quartile range.**

3. Use a cross-tab: **cross tabulation helps you understand the relationship between two categorical variables. For examples, if you want to see the relationship of gender with interest in sports activities. That would look something like this:**

**What % of male wants to play sports on the weekends**

**What % of female wants to play sports on the weekends**

**What % of male don’t want to play sports on the weekends**

**What % of female don’t want to play sports on the weekends**

1. Make a comparison between:

1. Data with nominal and ordinal values: **nominal data doesn’t have any order whereas ordinal data has order such as rank of someone’s position in an organisation or rank based on age.**

2. Histogram and box plot: **Histogram is used to find the frequency of a particular value for example if we have a dataset in which there is feature called age and we want to find out the maximum number of people in any age category then we can plot a histogram to find the maximum of occurrences. Box plot is used to find the 5 point summary and to check outliers in a dataset.**

3. The average and median: **average value is the mean value which is calculate by adding all the values and dividing by the number of values. Median value is the middle value in a data set which we get after sorting the numbers in the ascending order. Both are measures of central tendency and median is preferably used in the event of outliers as outlier can cause a major shift in the mean value.**