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

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

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

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

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

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

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

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

9.

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

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?

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

1. Data collected at regular intervals

2. The gap between the quartiles

3. Use a cross-tab

11. Make a comparison between:

1. Data with nominal and ordinal values

2. Histogram and box plot

3. The average and median

Answer:

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

Answer: The key tasks involved in machine learning are data pre-processing, feature engineering, model training, and model evaluation. Data pre-processing is the process of cleaning and transforming raw data into a form suitable for analysis. It involves tasks such as data cleaning, handling missing values, data normalization, and data transformation.

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

Answer: Quantitative data refers to numerical data that can be measured and expressed in terms of numbers. It includes data such as age, height, weight, and income. Qualitative data, on the other hand, refers to non-numerical data that cannot be measured or expressed in terms of numbers. It includes data such as gender, ethnicity, and marital status. Quantitative data can be further classified into discrete and continuous data, while qualitative data can be classified into nominal and ordinal data.

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

Answer:

| **Attribute** | **Data Type** |
| --- | --- |
| Age | Continuous |
| Gender | Nominal |
| Income | Continuous |
| Marital Status | Ordinal |
| Education Level | Ordinal |
| Blood Type | Nominal |

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

Answer: The various causes of machine learning data issues include missing values, incorrect values, inconsistent data, noisy data, and imbalanced data. The ramifications of these issues are that they can lead to biased models, inaccurate predictions, and poor performance.

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

Answer: Various approaches to categorical data exploration include frequency tables, bar charts, pie charts, and stacked bar charts. For example, a frequency table can be used to show the number of individuals in different age groups, while a bar chart can be used to show the distribution of income levels among different occupations.

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

Answer: The learning activity can be affected if certain variables have missing values as it can lead to biased models, inaccurate predictions, and poor performance. To address missing values, various techniques such as imputation, deletion, and regression can be used.

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

Answer: The various methods for dealing with missing data values include listwise deletion, pairwise deletion, mean imputation, regression imputation, and K-nearest neighbor imputation. Listwise deletion involves removing any record that has a missing value, while pairwise deletion involves using only the available data to compute the missing values. Mean imputation involves replacing the missing value with the mean of the available values, while regression imputation involves using a regression model to estimate the missing value. K-nearest neighbor imputation involves using the values of the nearest neighbors to estimate the missing value.

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

Answer: The various data pre-processing techniques include data cleaning, data transformation, and data reduction. Data cleaning involves removing noise and outliers from the data, while data transformation involves scaling, normalization, and encoding the data. Data reduction involves reducing the dimensionality of the data using techniques such as principal component analysis (PCA) and feature selection. Dimensionality reduction involves reducing the number of features in the data while retaining as much information as possible. Feature selection involves selecting the most relevant features based on their importance to the model.

9.  
i. IQR (Interquartile Range) is a measure of variability that is used to assess the spread of a data set. It is calculated as the difference between the upper quartile (Q3) and the lower quartile (Q1). The IQR is used to identify outliers in a data set, with values that lie more than 1.5 times the IQR below Q1 or above Q3 considered as outliers.

ii. A box plot is a graphical representation of the five-number summary of a data set, which includes the minimum, lower quartile (Q1), median, upper quartile (Q3), and maximum values. The box in the plot represents the IQR, with the median indicated by a horizontal line within the box. The whiskers extend from the box to the minimum and maximum values, but are limited to a length of 1.5 times the IQR. The box plot is a useful tool for identifying outliers, as any values outside of the whiskers are considered potential outliers.

* + Data collected at regular intervals refers to data that is recorded at set time intervals. This type of data is often referred to as time series data and is commonly used in fields such as finance, weather forecasting, and manufacturing. Examples of data collected at regular intervals include stock prices, temperature readings, and production levels.
  + The gap between the quartiles refers to the difference between the upper and lower quartiles in a data set, as measured by the IQR. The size of the gap between the quartiles is a measure of the spread of the data, with a larger gap indicating a more spread-out distribution. This measure can be useful in comparing the spread of different data sets.
  + A cross-tab, short for cross-tabulation, is a table that displays the frequency of occurrences of two or more categorical variables. It is a common tool for exploring relationships between variables in descriptive statistics. Cross-tabs are often used in market research to analyze consumer behavior, and in social sciences to study the relationship between two or more variables.