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

1. Make a comparison between:

1. Data with nominal and ordinal values

2. Histogram and box plot

3. The average and median

SOLUTION

1. The key tasks involved in machine learning include data pre-processing, feature selection, model selection, training, and evaluation. Data pre-processing involves cleaning, transforming, and normalizing the data to make it suitable for machine learning algorithms. Data pre-processing includes tasks like data cleaning, data transformation, data integration, and data reduction.
2. Quantitative data is numerical data that can be measured and analyzed using statistical methods. Examples include height, weight, temperature, and age. Qualitative data, on the other hand, is non-numerical data that cannot be measured using statistical methods. Examples include color, shape, and taste.
3. A basic data collection might include records of employees in a company, with attributes like age, salary, job title, department, and years of experience. Age and years of experience are quantitative variables, while job title and department are nominal categorical variables, and salary is a continuous variable.
4. Machine learning data issues can be caused by incomplete data, biased data, noisy data, or irrelevant features. These issues can lead to poor model performance, inaccurate predictions, and biased results.
5. Categorical data exploration can involve creating frequency tables, cross-tabulations, and visualizations like bar charts and pie charts. For example, if we have data on the type of car a person owns, we could create a bar chart to show the frequency of each type of car.
6. Missing values can affect the learning activity by reducing the accuracy of the model and introducing bias into the analysis. To deal with missing values, we can either delete the rows with missing values, fill in the missing values with a mean or median value, or use advanced imputation techniques to estimate the missing values.
7. Methods for dealing with missing data values include deletion, mean or median imputation, regression imputation, multiple imputation, and K-nearest neighbor imputation.
8. Various data pre-processing techniques include data cleaning, data integration, data transformation, and data reduction. Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much of the relevant information as possible. Function selection involves selecting the most appropriate functions to fit the data.
   1. The IQR (Interquartile Range) is a measure of variability that represents the spread of the middle 50% of the data. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). The IQR is used to identify the presence of outliers. A commonly used criterion for identifying outliers is that any data point that is more than 1.5 times the IQR away from the nearest quartile is considered an outlier.
   2. A box plot is a graphical representation of the distribution of a dataset. The various components of a box plot include the median (middle value), the first quartile (Q1), the third quartile (Q3), the interquartile range (IQR), the upper whisker, and the lower whisker. The upper whisker represents the maximum value within 1.5 times the IQR of Q3, and the lower whisker represents the minimum value within 1.5 times the IQR of Q1. The length of the whiskers can vary depending on the data, and the lower whisker can surpass the upper whisker in length if the data is skewed to the left. Box plots can be used to identify outliers by plotting any data points that are more than 1.5 times the IQR away from the nearest quartile as individual points outside the whiskers.
9. Data collected at regular intervals: Data collected at regular intervals, also known as time series data, is data collected over a period of time at fixed intervals. It is commonly used in economics, finance, and weather forecasting. Time series data can be analyzed using techniques such as moving averages, exponential smoothing, and ARIMA models.
10. The gap between the quartiles: The gap between the quartiles, also known as the interquartile range, is a measure of the spread of the data. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). The larger the gap between the quartiles, the more spread out the data is.
11. Use a cross-tab: A cross-tab, also known as a contingency table, is a table that displays the frequency distribution of two categorical variables. It is commonly used in statistics to analyze the relationship between two variables. Cross-tabs can be used to calculate measures such as chi-square and phi-coefficient to determine the strength of the relationship between the variables.
12. Data with nominal and ordinal values: Nominal data is categorical data that has no inherent order, such as gender or race. Ordinal data is categorical data that has a natural order, such as education level or income. Nominal data is typically analyzed using measures such as frequency and percentage, while ordinal data can be analyzed using measures such as median and quartiles.
13. Histogram and box plot: A histogram is a graphical representation of the distribution of a dataset that shows the frequency of values within each bin. A box plot is a graphical representation of the distribution of a dataset that shows the quartiles, median, and outliers. Histograms are useful for visualizing the shape of the distribution, while box plots are useful for identifying outliers and comparing the distributions of multiple datasets.
14. The average and median: The average, also known as the mean, is a measure of central tendency that is calculated as the sum of all values divided by the number of values.