**Machine Learning Assignment 5**

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

Ans-) Key tasks involved in machine learning include data collection, data pre-processing, feature selection, model selection, and model evaluation. Data pre-processing refers to the steps taken to prepare data for use in a machine learning model. This may involve cleaning the data, dealing with missing values, encoding categorical variables, and scaling or normalizing the data.

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

Ans-) Quantitative data refers to numerical measurements or counts, while qualitative data refers to non-numerical or categorical data that is descriptive in nature. Quantitative data can be further classified into discrete or continuous data types. Discrete data refers to numerical data that can only take on specific values, while continuous data refers to numerical data that can take on any value within a range.

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-) Sample data collection:

| **Attribute** | **Type** |
| --- | --- |
| Age | Numeric |
| Gender | Categorical |
| Income | Numeric |
| Education Level | Categorical |
| Occupation | Categorical |

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

Ans-) Various causes of machine learning data issues include missing data, incorrect or inconsistent data, irrelevant or redundant data, and data that is biased or unrepresentative. These issues can lead to inaccurate or unreliable results, and can also affect the performance of the model.

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

Ans-) Approaches to categorical data exploration include frequency tables, bar charts, and stacked bar charts. For example, a frequency table can be used to display the number or proportion of observations in each category, while a stacked bar chart can be used to show the distribution of categories across different groups.

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 values can affect the performance of a machine learning model, as many algorithms cannot handle missing data. One approach to dealing with missing values is to remove any observations with missing values, but this can lead to a loss of information. Another approach is to impute the missing values using methods such as mean imputation or regression imputation.

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

Ans-) Methods for dealing with missing data include mean imputation, median imputation, mode imputation, regression imputation, and multiple imputation. Mean imputation involves replacing missing values with the mean of the observed values for that variable. Median imputation replaces missing values with the median of the observed values, while mode imputation replaces missing values with the mode of the observed values. Regression imputation involves using a regression model to predict the missing values, while multiple imputation involves creating multiple imputed datasets and combining the results.

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

Ans-) Data pre-processing techniques include normalization, standardization, feature scaling, dimensionality reduction, and function selection. Dimensionality reduction involves reducing the number of features in the dataset, while function selection involves selecting the most important features for the model.

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?

Ans-) The IQR, or interquartile range, is a measure of the spread of a distribution. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) of the data. The IQR can be used to identify outliers, which are typically defined as values that are more than 1.5 times the IQR below Q1 or above Q3.

A box plot is a graphical representation of a distribution that displays the median, quartiles, and outliers of the data. The components of a box plot include the median (a horizontal line inside the box), the first quartile (Q1) and third quartile (Q3) (the lower and upper boundaries of the box), and the whiskers (lines that extend from the box to the highest and lowest observations that are within 1.5 times the IQR of the upper and lower quartiles). The lower whisker will surpass the upper whisker in length if the data is skewed to the left. Box plots can be used to identify outliers, which are typically represented as individual points outside the whisk

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

1. Data collected at regular intervals

Ans-) Data collected at regular intervals is also referred to as time-series data. It is a type of data that is collected at equally spaced time intervals, such as every minute, hour, day, or month. Time-series data is commonly used in various fields, such as economics, finance, meteorology, and engineering, to analyze trends, patterns, and anomalies over time. Some examples of time-series data include stock prices, weather conditions, and website traffic.

2. The gap between the quartiles

Ans-) The gap between the quartiles is also known as the interquartile range (IQR), which is a measure of the spread or dispersion of a data set. The IQR is calculated by subtracting the first quartile (Q1) from the third quartile (Q3). It represents the middle 50% of the data, excluding the outliers. The IQR is a robust measure of dispersion since it is not influenced by extreme values or outliers in the data set. The larger the IQR, the more dispersed the data is. It is commonly used in box plots to identify outliers and to compare the dispersion of different data sets.