**Machine Learning – 5**

**Lokashwar Nallagatla**

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

Machine learning involves tasks such as data collection, data cleaning, feature engineering, model selection, training, and evaluation. Data pre-processing refers to preparing raw data for modeling by cleaning, transforming, and encoding it into a suitable format, including handling missing values, outliers, and normalization.

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

Quantitative data consists of numeric values that can be measured and are used for calculations. Examples include age, income, or temperature.

Qualitative data (also called categorical data) describes attributes or characteristics and cannot be measured numerically. Examples include gender, city of residence, or product type.

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

data = [

{

'id': 1,

'age': 25, # Numerical (integer)

'salary': 55000.50, # Numerical (float)

},

{

'id': 2,

'age': 30,

'salary': 72000.75,

},

{

'id': 3,

'age': 28,

'salary': 80000.00,

},

{

'id': 4,

'age': 35,

'salary': 95000.25,

}

]

for record in data:

print(record)

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

Data issues arise from missing values, outliers, duplicate entries, and inconsistent formatting. These issues can lead to biased models, inaccurate predictions, and overfitting. Incomplete or incorrect data reduces the model’s ability to generalize and perform well on unseen data.

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

Frequency counts: Count how often each category appears. Example: The number of customers from each country.

Bar charts: Visualize the frequency distribution of categories. Example: A bar chart showing the count of different product types sold.

Chi-squared tests: Assess the relationship between two categorical variables. Example: Analyzing if gender is related to purchasing behavior.

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

Missing values can bias the model, reduce its accuracy, or cause the model to ignore important features. Solutions include imputation (replacing missing values with mean, median, or mode), using algorithms that handle missing data, or removing records with missing values.

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

Imputation: Replace missing values with the mean, median, or mode of the column.

Deletion: Remove records with missing values, but this can lead to loss of information.

Prediction: Use other variables to predict and fill in the missing values (e.g., using regression or k-NN).

Use algorithms that handle missing data: Some algorithms, like decision trees, can handle missing values without imputation.

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

Normalization/Standardization: Scaling features to a standard range or distribution.

Encoding: Converting categorical data into numerical form (e.g., one-hot encoding).

Dimensionality reduction: Reducing the number of features while preserving essential patterns, e.g., using PCA.

Feature selection: Identifying and selecting the most important features to improve model performance and reduce overfitting.

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

IQR (Interquartile Range) is the difference between the 75th (Q3) and 25th (Q1) percentiles, showing the middle 50% of the data. It is used to assess the spread of the central data points and to detect outliers (values outside 1.5 \* IQR above Q3 or below Q1).

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 plot displays:

Box: Represents the IQR (Q1 to Q3).

Median line: Shows the middle value (Q2).

Whiskers: Extend to the maximum and minimum values within 1.5 \* IQR from Q1 and Q3.

Outliers: Points outside the whiskers are considered outliers. The lower whisker surpasses the upper whisker when the distribution is skewed to the left. Outliers appear as points beyond the whiskers.

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

1. Data collected at regular intervals

This is time-series data, where observations are recorded at consistent time intervals, such as daily stock prices or monthly sales data. This data type is used for trend analysis, forecasting, and anomaly detection.

2. The gap between the quartiles

The gap between quartiles refers to the Interquartile Range (IQR), which measures the spread of the middle 50% of the data. A larger IQR indicates a wider spread of data, and a smaller IQR indicates less variability.

11. Make a comparison between:

1. Data with nominal and ordinal values

Nominal data is categorical without any order (e.g., color, gender).

Ordinal data has a meaningful order but not a consistent difference between categories (e.g., education level: high school, bachelor’s, master’s).

2. Histogram and box plot

Histogram displays the distribution of a dataset with frequency counts in bins.

Box plot shows the spread, quartiles, and outliers in the data.

3. The average and median

Average (mean) is the sum of all values divided by the number of values and is sensitive to outliers.

Median is the middle value when data is sorted and is less affected by extreme values, providing a better measure of central tendency for skewed data.