**Machine Learning – 4**

**Lokashwar Nallagatla**

1. What are the key tasks involved in getting ready to work with machine learning modelling?

Key tasks include data collection, cleaning, and preprocessing to ensure quality data. Feature engineering helps identify and create relevant variables. The next step is splitting data into training, validation, and test sets. Finally, selecting an appropriate algorithm and tuning its hyperparameters is essential for optimal performance.

2. What are the different forms of data used in machine learning? Give a specific example for each of them.

Numerical data: Data that represents quantities and can be continuous or discrete. Example: Age of a person.

Categorical data: Data representing categories or labels. Example: Type of animal (cat, dog, bird).

Ordinal data: Categorical data with a meaningful order. Example: Education level (high school, bachelor’s, master’s).

Text data: Data represented as unstructured text. Example: Customer reviews.

Image data: Data in the form of images. Example: Medical X-rays.

Time-series data: Data points collected at regular time intervals. Example: Stock prices over time.

3. Distinguish:

1. Numeric vs. categorical attributes

Numeric attributes are quantitative and can be measured, like height or salary.

Categorical attributes represent categories or labels, like gender or colour.

2. Feature selection vs. dimensionality reduction

Feature selection involves choosing the most important features for the model.

Dimensionality reduction reduces the number of features by transforming them into a lower-dimensional space (e.g., using PCA).

4. Make quick notes on any two of the following:

1. The histogram

A histogram is a graphical representation of the distribution of numerical data, with data divided into bins. It helps to visualize the frequency distribution and detect skewness, modality, and outliers.

2. Use a scatter plot

A scatter plot is used to visualize the relationship between two continuous variables. It helps to identify correlations, trends, and outliers by plotting data points on a two-dimensional axis.

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

Investigating data helps identify patterns, outliers, missing values, and correlations, which guide further analysis and model selection. Qualitative data is often explored through categorical summaries or visualizations like bar charts, while quantitative data is typically analysed using statistical measures and histograms.

6. What are the various histogram shapes? What exactly are ‘bins’?

Histograms can have different shapes, such as normal (bell-shaped), skewed (left or right), bimodal (two peaks), or uniform (flat). Bins are intervals or ranges into which data is grouped for visualization. The width of bins affects how the distribution is interpreted.

7. How do we deal with data outliers?

Outliers can be handled by removing them if they are errors, transforming the data (e.g., log transformation), or using models that are robust to outliers, such as decision trees. Alternatively, they can be capped or replaced with a maximum value.

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

The main measures are:

Mean: Average value.

Median: Middle value.

Mode: Most frequent value. The mean can vary significantly from the median in skewed data sets because the mean is sensitive to extreme values (outliers), whereas the median is more robust.

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

A scatter plot is used to visualize the relationship between two variables. If a strong linear or non-linear pattern exists, it indicates correlation. Outliers can be identified as data points that lie far away from the general cluster or trend in the plot.

10. Describe how cross-tabs can be used to figure out how two variables are related.

Cross-tabs (contingency tables) summarize the relationship between two categorical variables by showing the frequency of occurrences for each combination of categories. They help identify associations or dependencies, such as whether two variables are independent or correlated.