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

Ans:

1. Data Collection: Gathering relevant and reliable data from various sources.
2. Data Splitting: Dividing the data into training, validation, and testing sets for model evaluation.
3. Data Pre-processing: Cleaning and transforming the data to handle missing values, outliers, and inconsistencies.
4. Feature Engineering: Selecting or creating meaningful features that capture relevant information for the model.
5. Model Selection: Choosing the appropriate machine learning algorithm or model based on the problem and data characteristics.
6. Model Training: Fitting the selected model to the training data to learn the underlying patterns.
7. Model Evaluation: Assessing the performance of the trained model using appropriate evaluation metrics.
8. Model Tuning: Fine-tuning the model parameters to optimize its performance.
9. Deployment: Implementing the model in a production environment for real-world applications.

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

Ans:

Different forms of data used in machine learning:

a. Numerical Data: Represents quantitative measurements or values. Example: Temperature readings, age, income.

b. Categorical Data: Represents discrete and non-numeric variables. Example: Gender, color, product categories.

c. Text Data: Represents unstructured text information. Example: Customer reviews, emails, social media posts.

d. Image Data: Represents visual information in the form of images or pixels. Example: Digital photographs, medical scans.

e. Time Series Data: Represents data collected over a sequence of time points. Example: Stock prices, weather data.

3. Distinguish:

1. Numeric vs. categorical attributes

2. Feature selection vs. dimensionality reduction

Ans:

1. Numeric vs. categorical attributes:

Numeric attributes represent continuous or discrete numerical values that can be measured or counted. Example: Age, temperature.

Categorical attributes represent distinct categories or labels. They cannot be measured or ordered numerically. Example: Gender, color.

Feature selection vs. dimensionality reduction:

Feature selection is the process of selecting a subset of relevant features from the original feature set. It aims to improve model performance, reduce overfitting, and enhance interpretability.

Dimensionality reduction is the process of reducing the number of features by transforming feature onto new space while preserving important information and selecting features which preserves most information as in PCA. It aims to eliminate redundant or irrelevant features and reduce computational complexity.

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

1. The histogram

2. Use a scatter plot

3.PCA

Ans:

1. The histogram: A histogram is a graphical representation of the distribution of a dataset. It consists of a series of bars, where each bar represents a range of values and the height represents the frequency or count of data points within that range.
2. Scatter plot: A scatter plot is a two-dimensional plot that displays the relationship between two continuous variables. Each data point is represented as a dot on the plot, and the position of the dot corresponds to the values of the two variables.

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

Ans:

It is necessary to investigate data to understand its characteristics, identify patterns, and gain insights. Exploring data helps in detecting outliers, understanding the distribution, assessing relationships between variables, and making informed decisions. The exploration process may differ for qualitative and quantitative data. Qualitative data may be explored through methods like content analysis, thematic analysis, or coding schemes, while quantitative data can be explored using statistical techniques and visualizations.

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

Ans:

Various histogram shapes include:

* Normal Distribution: Bell-shaped curve with a symmetrical distribution.
* Skewed Distribution: Asymmetrical distribution with a longer tail on one side.
* Bimodal Distribution: Distribution with two distinct peaks.
* Uniform Distribution: Flat and evenly distributed data.

Bins in a histogram are the intervals or ranges into which the data is divided. They represent the width of each bar in the histogram and determine the granularity of the data representation.

7. How do we deal with data outliers?

Ans:

Outliers are data points that significantly deviate from the overall pattern or distribution of the data. They can arise due to various reasons such as measurement errors or rare events. Dealing with outliers depends on the context and the specific analysis. Here are some common approaches:

* Remove outliers: If the outliers are due to errors or data quality issues, they can be removed from the dataset. However, this should be done carefully, considering the impact on the overall analysis.
* Transform data: Applying transformations such as logarithmic or power transformations can help reduce the impact of outliers and make the data more suitable for analysis.
* Winsorization: Winsorization replaces extreme values with less extreme values to minimize their influence on the analysis. For example, capping the outliers at a certain percentile value.

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

Ans:

Various central inclination measures:

* Mean: The arithmetic average of a set of values. It is calculated by summing all the values and dividing by the number of observations. The mean is sensitive to outliers because it takes into account all values, including extreme ones.
* Median: The middle value in a sorted list of values. It is less affected by outliers compared to the mean since it is based on the order of values rather than their magnitude.
* Mode: The most frequently occurring value in a dataset. It is useful for categorical or discrete data.

The mean can vary significantly from the median in certain data sets, especially when there are outliers present. Outliers, being extreme values, can pull the mean towards their direction, resulting in a deviation from the central position of the data. The median, on the other hand, is not influenced by the actual values of the outliers but rather by their relative position within the sorted dataset.

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

Ans:

A scatter plot is a graphical representation of the relationship between two continuous variables. It plots each data point as a dot on the graph, with one variable represented on the x-axis and the other on the y-axis. Scatter plots are useful for exploring the correlation or association between the two variables. They can help identify patterns, trends, and potential outliers. Outliers can appear as data points that significantly deviate from the general pattern of the scatter plot, lying far away from the bulk of the data points. However, the identification of outliers solely based on a scatter plot might not be definitive, and further analysis is often required.

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

Ans:

Cross-tabs for understanding variable relationships:

Cross-tabulation, or cross-tabs, is a technique used to explore the relationship between two categorical variables. It creates a table that displays the frequency or count of observations for each combination of categories from the two variables. Cross-tabs allow us to examine how the distribution of one variable varies across different categories of another variable. They can reveal associations, dependencies, or differences between the variables. Measures such as chi-square tests or measures of association like Cramer's V can be used to quantify the strength and significance of the relationship between the variables. Cross-tabs are commonly used in market research, social sciences, and business analytics to analyze customer behavior, product preferences, survey responses, and more.