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

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

3. Distinguish:

1. Numeric vs. categorical attributes

2. Feature selection vs. dimensionality reduction

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

1. The histogram

2. Use a scatter plot

3.PCA (Personal Computer Aid)

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

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

7. How do we deal with data outliers?

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

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

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

Answer:

1. The key tasks involved in getting ready to work with machine learning modeling include:

a) Data collection and cleaning - collecting relevant data and cleaning it to remove any inconsistencies or errors.

b) Data exploration and visualization - exploring the data to gain insights and identify patterns through visualization.

c) Feature engineering - selecting and transforming relevant features to improve the performance of the model.

d) Data splitting and sampling - splitting the data into training and testing sets and sampling techniques like cross-validation.

e) Choosing an appropriate model - selecting an appropriate model based on the type of problem and data.

f) Training and fine-tuning the model - training the model on the training set and fine-tuning it using hyperparameter tuning.

1. The different forms of data used in machine learning are:

a) Numerical data - data that consists of numbers, such as age or temperature.

Example: A dataset containing the heights and weights of individuals.

b) Categorical data - data that consists of categories or labels, such as gender or type of product.

Example: A dataset containing the types of products sold by a store.

c) Text data - data that consists of words and sentences.

Example: A dataset containing customer reviews of a product.

d) Image data - data that consists of images or visual data.

Example: A dataset containing images of cats and dogs.

1. Distinguishing:

* Numeric vs. categorical attributes - Numeric attributes are those that contain numerical values and are used for mathematical operations, while categorical attributes are those that contain categories or labels.
* Feature selection vs. dimensionality reduction - Feature selection involves selecting a subset of relevant features from the data, while dimensionality reduction involves reducing the number of features while retaining as much information as possible.

1. Quick notes:

* The histogram - A histogram is a type of chart that displays the distribution of numerical data. It consists of a series of bars, where each bar represents a range of values, and the height of the bar represents the frequency of data points in that range.
* Scatter plot - A scatter plot is a type of chart that displays the relationship between two variables. It consists of a series of data points plotted on a two-dimensional coordinate system.

1. Investigating data is necessary to gain insights into the underlying patterns and relationships between variables, as well as to identify outliers and errors. There is no discrepancy in how qualitative and quantitative data are explored, as the same techniques can be used for both types of data.
2. The various histogram shapes include normal distribution, skewed left, skewed right, and bimodal distribution. 'Bins' are the intervals or ranges of values used to group the data in a histogram.
3. Outliers can be dealt with by removing them from the dataset, imputing them using statistical techniques, or treating them as a separate category in the analysis.
4. The various central inclination measures include mean, median, and mode. Mean varies too much from median in certain datasets when there are extreme values or outliers that pull the mean in one direction.
5. A scatter plot can be used to investigate bivariate relationships by plotting the two variables on a two-dimensional coordinate system and looking for patterns or relationships between them. Outliers can be identified using a scatter plot as they appear as data points that lie far from the general pattern.
6. Cross-tabs can be used to figure out how two variables are related by creating a table that shows the frequency of each combination of the variables. This table can be used to calculate conditional probabilities and test for independence between the variables.