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

Ans1:- task1: Collect Data.

task2: Prepare the data.

task3: Choose the model.

task4 Train your machine model.

task5: Evaluation.

task6: Parameter Tuning.

task7: Prediction or Inference.

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

Ans2:- Most data can be categorized into 4 basic types from a Machine Learning perspective: numerical data, categorical data, time-series data, and text.

1. Numerical Data:- Numerical data represent values that can be measured and put into a logical order. Examples of numerical data are height, weight, age, number of movies watched, IQ, etc. To graph numerical data, one uses dot plots, stem and leaf graphs, histograms, box plots, ogive graphs, and scatter plots.

2. Categorical Data:- Categorical data is a type of data that can be stored into groups or categories with the aid of names or labels. This grouping is usually made according to the data characteristics and similarities of these characteristics through a method known as matching

3. Time-Series Data:- Time series data is data that is collected at different points in time. This is opposed to cross-sectional data which observes individuals, companies, etc. at a single point in time. Because data points in time series are collected at adjacent time periods there is potential for correlation between observations.

4. Text Data:- Text data usually consists of documents which can represent words, sentences or even paragraphs of free flowing text. The inherent unstructured (no neatly formatted data columns!) and noisy nature of textual data makes it harder for machine learning methods to directly work on raw text data.

3. Distinguish:

1. Numeric vs. categorical attributes

2. Feature selection vs. dimensionality reduction

Ans3:-

Numeric vs. categorical attributes

Numerical data are values obtained for quantitative variable, and carries a sense of magnitude related to the context of the variable (hence, they are always numbers or symbols carrying a numerical value). Categorical data are values obtained for a qualitative variable; categorical data numbers do not carry a sense of magnitude.

Numerical data always belong to either ordinal, ratio, or interval type, whereas categorical data belong to nominal type.

Methods used to analyze quantitative data are different from the methods used for categorical data, even if the principles are the same at least the application has significant differences.

Numerical data are analyzed using statistical methods in descriptive statistics, regression, time series and many more.

For categorical data usually descriptive methods and graphical methods are employed. Some non-parametric tests are also used.

Feature selection vs. dimensionality reduction

The difference is that the set of features made by feature selection must be a subset of the original set of features, and the set made by dimensionality reduction doesn't have to (for instance PCA reduces dimensionality by making new synthetic features from linear combination of the original ones, and then discarding the less important ones).

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

The histogram

Use a scatter plot

PCA (Personal Computer Aid)

Ans4:-

The Histogram:-

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

Histograms are commonly used in statistics to demonstrate how many of a certain type of variable occurs within a specific range. For example, a census focused on the demography of a country may use a histogram to show how many people are between the ages of 0 - 10, 11 - 20, 21 - 30, 31 - 40, 41 - 50, etc.

Histograms can be customized in several ways by the analyst. The first is to change the interval between buckets.

The other consideration is how to define the y-axis. The most basic label is to use the frequency of occurrences observed in the data, but one could also use percentage of total or density instead.

The Scatter Plot:-

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

Scatter plots’ primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

Identification of correlational relationships are common with scatter plots. In these cases, we want to know, if we were given a particular horizontal value, what a good prediction would be for the vertical value. You will often see the variable on the horizontal axis denoted an independent variable, and the variable on the vertical axis the dependent variable.

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

Ans5:- If your data set is messy, building models will not help you to solve your problem. What will happen is “garbage in, garbage out.” In order to build a powerful machine learning algorithm. We need to explore and understand our data set before we define a predictive task and solve it.

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

Ans6:-

Bell-shaped: A bell-shaped picture, usually presents a normal distribution.

Bimodal: A bimodal shape, has two peaks. This shape may show that the data has come from two different systems. If this shape occurs, the two sources should be separated and analyzed separately.

Skewed right: Some histograms will show a skewed distribution to the right. A distribution skewed to the right is said to be positively skewed. This kind of distribution has a large number of occurrences in the lower value cells (left side) and few in the upper value cells (right side). A skewed distribution can result when data is gathered from a system with has a boundary such as zero. In other words, all the collected data has values greater than zero.

Skewed left: Some histograms will show a skewed distribution to the left.A distribution skewed to the left is said to be negatively skewed. This kind of distribution has a large number of occurrences in the upper value cells (right side) and few in the lower value cells (left side). A skewed distribution can result when data is gathered from a system with a boundary such as 100. In other words, all the collected data has values less than 100.

Uniform: A uniform distribution, as shown below, provides little information about the system. An example would be a state lottery, in which each class has about the same number of elements. It may describe a distribution which has several modes (peaks). If your histogram has this shape, check to see if several sources of variation have been combined. If so, analyze them separately. If multiple sources of variation do not seem to be the cause of this pattern, different groupings can be tried to see if a more useful pattern results. This could be as simple as changing the starting and ending points of the cells, or changing the number of cells. A uniform distribution often means that the number of classes is too small.

Random: A random distribution, as shown below, has no apparent pattern. Like the uniform distribution, it may describe a distribution that has several modes (peaks). If your histogram has this shape, check to see if several sources of variation have been combined. If so, analyze them separately. If multiple sources of variation do not seem to be the cause of this pattern, different groupings can be tried to see if a more useful pattern results. This could be as simple as changing the starting and ending points of the cells, or changing the number of cells. A random distribution often means there are too many classes.

7. How do we deal with data outliers?

Ans7:- 1. Deleting the values: You can delete the outliers if you know that the outliers are wrong or if the reason the outlier was created is never going to happen in the future. For example, there is a data set of peoples ages and the usual ages lie between 0 to 90 but there is data entry off the age 150 which is nearly impossible. So, we can safely drop the value that is 150.

2. Changing the values: We can also change the values in the cases when we know the reason for the outliers. Consider the previous example for measurement or instrument errors where we had 10 voltmeters out of which one voltmeter was faulty. Here what we can do is that we can take another set of readings using a correct voltmeter and replace them with the readings that were taken by the faulty voltmeter.

3. Data transformation: Data transformation is useful when we are dealing with highly skewed data sets. By transforming the variables, we can eliminate the outliers for example taking the natural log of a value reduces the variation caused by the extreme values. This can also be done for data sets that do not have negative values.

4. Using different analysis methods: You could also use different statistical tests that are not as much impacted by the presence of outliers – for example using median to compare data sets as opposed to mean or use of equivalent nonparametric tests etc.

5. Valuing the outliers: In case there is a valid reason for the outlier to exist and it is a part of our natural process, we should investigate the cause of the outlier as it can provide valuable clues that can help you better understand your process performance. Outliers may be hiding precious information that could be invaluable to improve your process performance. You need to take the time to understand the special causes that contributed to these outliers. Fixing these special causes can give you significant boost in your process performance and improve customer satisfaction. For example, normal delivery of orders takes 1-2 days, but a few orders took more than a month to complete. Understanding the reason why it took a month and fixing this process can help future customers as they would not be impacted by such large wait times.

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

Ans8:- An essential statistical concept is the “measure of central tendency“. This measure is an important way to summarize the dataset with one representative value. This measure provides a rough picture of where data points are centered. The commonly used measures of central tendency are:

Mean

Median

Mode

The median is generally used for skewed distributions. The mean is not a robust tool since it is largely influenced by outliers. A mean is computed by adding up all the values and dividing that score by the number of values. The Median is the number found at the exact middle of the set of values.

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

Ans9:- Bivariate analysis:- is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding the relationship between them. Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation. If the line goes from a high-value on the y-axis down to a high-value on the x-axis, the variables have a negative correlation . A perfect positive correlation is given the value of 1.

If there is a regression line on a scatter plot, you can identify outliers. An outlier for a scatter plot is the point or points that are farthest from the regression line. If one point of a scatter plot is farther from the regression line than some other point, then the scatter plot has at least one outlier.

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

Ans10:- Cross tabulation is a method to quantitatively analyze the relationship between multiple variables. Also known as contingency tables or cross tabs, cross tabulation groups variables to understand the correlation between different variables. It also shows how correlations change from one variable grouping to another.