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

Key tasks are:-

- Collect and prepare data
- Choose the model
- Train your machine learning model
- Evaluation
- Parameter tuning
- Predict or draw an inference

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

The general types of data are text, numerical, Categorical and Time series. Further we can classify into training data, testing data and Validation data.

Numeric data can be quantitative like price, age, year etc. It can be also classified as continuous or discrete as height/weight and units sold. Categorical data represents characteristics team, hometown etc. Time series is sequence of numbers collected at regular intervals overall period of time like price of a stock over a period of time. Text data be in any form of text like names, places etc.

3. Distinguish:

1. Numeric vs. categorical attributes

Numeric data can be quantitative like price, age, year etc. It can be also classified as continuous or discrete as height/weight and units sold. But categorical can be as male/female, yes/no, True/false.

2. Feature selection vs. dimensionality reduction

Feature selection means choosing the most important features out of the dataset which helps to predict the output but in case of classification problem often too many factors on the basis of which the final classification is done. The higher the number of features, the harder it gets to visualize the training set and then work on it. For eg in case of SVM if the data is not linearly separated then we increase the dimension to help the model classify the points clearly. Many other model work on the basis of changing the dimension in order to classify the points. It is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

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

1. The histogram

A histogram is a chart that plots the distribution of a numeric variable's values as a series of bars.

2. Use a scatter plot

A scatter plot is a chart type that is normally used to observe and visually display the relationship between variables.

3.PCA (Personal Computer Aid)

Steps to PCA :-

- Standardize the range of continuous initial variables
- Compute the covariance matrix to identify correlations
- Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
- Create a feature vector to decide which principal components to keep
- Recast the data along the principal components axes

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

We need to explore and understand our data set before we define a predictive task and solve it. Whether it's surveying results, sales data, or an email campaign, you've collected data for a specific purpose. By extension, apply this purpose to the questions you're asking of the data itself. Beginning with some specific questions can keep your research focused and allow you to see the forest through the trees. questions may be: "which department brings in the most revenue per year" or "are sales in climbing gear increasing or decreasing this year?" It's important to have a specific question in mind when you begin data analysis so as to provide some structure and avoid stumbling into false positives.

Quantitative data is information about quantities, and therefore numbers, and qualitative data is descriptive, and regards phenomenon which can be observed but not measured, such as language. Contrary to qualitative data, quantitative data is statistical and is typically structured in nature – meaning it is more rigid and defined. This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

Whereas qualitative is open for exploration, quantitative data is much more concise and close-ended. It can be used to ask the questions "how much" or "how many," followed by conclusive information.

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

Bell-Shaped - A histogram is bell-shaped if it resembles a "bell" curve and has one single peak in the middle of the distribution. The most common real-life example of this type of distribution is the normal distribution.

Uniform - A histogram is described as "uniform" if every value in a dataset occurs roughly the same number of times. This type of histogram often looks like a rectangle with no clear peaks.

Bimodal - A histogram is described as "bimodal" if it has two distinct peaks. We often say that this type of distribution has multiple modes – that is, multiple values occur most frequently in the dataset.

Multimodal - A histogram is described as "multimodal" if it has more than two distinct peaks.

Left Skewed - A histogram is left skewed if it has a “tail” on the left side of the distribution. Sometimes this type of distribution is also called “negatively” skewed.

Right Skewed - A histogram is right skewed if it has a “tail” on the right side of the distribution. Sometimes this type of distribution is also called “positively” skewed.

Random - The shape of a distribution can be described as “random” if there is no clear pattern in the data at all.

The towers or bars of a histogram are called bins. The height of each bin shows how many values from that data fall into that range.

7. How do we deal with data outliers?

We either delete the observations, we can replace the observations by mean/mode/median or we can also go ahead with imputation method.

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

A positive inclination is, when the line respectively the plane, in the measuring direction is inclined. The negative inclination is therefore when the line or plane is declined.

Thanks to the reversal measurement, it is possible to make precise absolute measurements (measuring the precise absolute deviation from center of gravity) even by using inclinometers with zero deviation.

The median is generally used for skewed distributions. The mean is used for normal distributions.

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

A scatterplot is a type of data display that shows the relationship between two numerical variables. Each member of the dataset gets plotted as a point whose (x, y)(x,y)left parenthesis, x, comma, y, right parenthesis coordinates relates to its values for the two variables. When the yyy variable tends to increase as the xxx variable increases, we say there is a positive correlation between the variables vice versa.

Yes we can find the data points far away from the rest of the data points in scatter plot which would be the outliers.

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

Cross-tabulation is a mainframe statistical model that follows similar lines. It helps you make informed decisions regarding your research by identifying patterns, trends, and the correlation between your study parameters. For example, consider your college application. First, you needed to look at the academic factor: your grades throughout high school, SAT scores, the field you wanted to major in, and the application essay you would need to write. Second, comes the financial factor, which will look at the tuition fees and possibilities of a scholarship. Lastly, it would be the emotional factor which will consider your distance from home and how far are the universities your friends are considering, so reunions would not be an issue. In other words, cross-tabulating Academics + Finance + Emotions led you to a refined list of universities, one of which is or soon will be your Alma Mater.