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

* Collect Data
* Prepare the data
* Choose the model
* Train your machine model
* Evaluation
* Parameter Tuning
* Prediction or Inference

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

Numerical Data – Examples- house prices or as a count

Categorical Data- hockey player’s position, team

Time Series Data- finance,

Text- letters, numbers

3. Distinguish:

1. Numeric vs. categorical attributes

A categorical variable is a category or type. For example, hair color is a categorical value or hometown is a categorical variable. Species, treatment type, and gender are all categorical variables.

A numerical variable is a variable where the measurement or number has a numerical meaning. For example, total rainfall measured in inches is a numerical value, heart rate is a numerical value, number of cheeseburgers consumed in an hour is a numerical value.

2. Feature selection vs. dimensionality reduction

Dimensionality reduction

Dimensionality reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction.

Feature selection yields a subset of features from the original set of features, which are best representatives of the data. It is an exhaustive search.

-In text data, features might be size of characters or some global features of the text. Feature selection will keep only certain features of those.

-Feature selection is done in the context of an optimization problem.

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

1. 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.

2. Use a scatter plot

Use a scatter plot to determine whether or not two variables have a relationship or correlation

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?

Quantitative research deals with numbers and statistics, while qualitative research deals with words and meanings. Quantitative methods allow you to systematically measure variables and test hypotheses. Qualitative methods allow you to explore concepts and experiences in more detail.

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

A histogram displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin. Bins are also sometimes called "intervals", "classes", or "buckets".

7. How do we deal with data outliers?

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.

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.

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.

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.

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?

Mean, Median, and Mode. A measure of central tendency is an important aspect of quantitative data. It is an estimate of a “typical” value. Three of the many ways to measure central tendency are the mean, median and mode.

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 shows the association between two variables. A scatter plot matrix shows all pairwise scatter plots for many variables. 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. There is at least one outlier on a scatter plot in most cases, and there is usually only one outlier.

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

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.