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

* Collecting Data:
* Preparing the Data:
* Choosing a Model:
* Training the Model:
* Evaluating the Model:
* Parameter Tuning:
* Making Predictions

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

them.

Numerical Data

Numerical data is any data where data points are exact numbers. Statisticians also might call numerical data, quantitative data. This data has meaning as a measurement such as house prices or as a count, such as a number of residential properties in Los Angeles or how many houses sold in the past year.

Numerical data can be characterized by continuous or discrete data. Continuous data can assume any value within a range whereas discrete data has distinct values.

For example, the number of students taking Python class would be a discrete data set. You can only have discrete whole number values like 10, 25, or 33. A class cannot have 12.75 students enrolled. A student either join a class or he doesn’t. On the other hand, continuous data are numbers that can fall anywhere within a range. Like a student could have an average score of 88.25 which falls between 0 and 100.

Categorical Data

Categorical data represents characteristics, such as a hockey player’s position, team, hometown. Categorical data can take numerical values. For example, maybe we would use 1 for the colour red and 2 for blue. But these numbers don’t have a mathematical meaning. That is, we can’t add them together or take the average.

In the context of super classification, categorical data would be the class label. This would also be something like if a person is a man or woman, or property is residential or commercial.

Time Series Data

Time series data is a sequence of numbers collected at regular intervals over some period of time. It is very important, especially in particular fields like finance. Time series data has a temporal value attached to it, so this would be something like a date or a timestamp that you can look for trends in time.

Text

Text data is basically just words. A lot of the time the first thing that you do with text is you turn it into numbers using some interesting functions like the bag of words formulation.

3. Distinguish:

1. Numeric vs. categorical attributes

Categorical data refers to a data type that can be stored and identified based on the names or labels given to them. A process called matching is done, to draw out the similarities or relations between the data and then they are grouped accordingly. The data collected in the categorical form is also known as qualitative data. Each dataset can be grouped and labelled depending on their matching qualities, under only one category. This makes the categories mutual exclusive.

Numerical data refers to the data that is in the form of numbers, and not in any language or descriptive form. Often referred to as quantitative data, numerical data is collected in number form and stands different from any form of number data types due to its ability to be statistically and arithmetically calculated. It doesn’t involve any natural language description and is quantitative in nature and it is used to measure quantities like a person’s height, age, IQ, etc.

2. Feature selection vs. dimensionality reduction

While both methods are used for reducing the number of features in a dataset

Feature selection is simply selecting and excluding given features without changing them.

Dimensionality reduction transforms features into a lower dimension.

Feature Selection

* Remove features with missing values
* Remove features with low variance
* Remove highly correlated features
* Univariate feature selection
* Recursive feature elimination
* Feature selection using SelectFromModel

Dimensionality Reduction

* PCA

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

1. The histogram

A histogram is a bar graph-like representation of data that buckets a range of classes into columns along the horizontal x-axis. The vertical y-axis represents the number count or percentage of occurrences in the data for each column. Columns can be used to visualize patterns of data distributions.

2. Use a scatter plot

Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.

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

data are explored?

It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. The combination of qualitative and quantitative data can also lead to clashes in the philosophical assumptions behind each approach and therefore recommendations have been made for triangulation to be carried out from a pragmatic, or subtle realist, approach

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

Bell-shaped: A bell-shaped picture, shown below, usually presents a normal distribution. Bimodal: A bimodal shape, shown below, has two peaks. This shape may show that the data has come from two different systems. 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?

* Trimming/removing the outlier
* Quantile based flooring and capping
* Mean/Median imputation

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

certain data sets?

There are three main measures of central tendency: mode. median. mean. Due to the presence of outliers, mean varies too much from the median

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

outliers using a scatter plot?

A large amount of scatter around the line indicates a weak relationship. Little scatter represents a strong relationship. If all points fall directly on a straight line, we have a perfect linear relationship between our two variables. Yes, scatter plot can be used to detect outliers

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

To describe the relationship between two categorical variables, we use a special type of table called a cross-tabulation (or "crosstab" for short). In a cross-tabulation, the categories of one variable determine the rows of the table, and the categories of the other variable determine the columns