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

Building a machine learning model involves several steps that are to be followed regardless of the algorithm being used. Here are the key steps that are involved in creation of a model:

1. Gathering data: Collection of the data that will be used for the training and testing of the model. The data is collected from various sources like SQL databases, cloud databases, local storage etc. and stored in a single place to make it easier to work on it
2. Preparing the Data: This step involves working on the collected data for it to give the best possible results while the algorithm is trained on it. This involves steps like data cleaning to remove any bad data like missing or NaN values, selecting features to remove any irrelevant ones etc to optimize it before feeding it to the algorithm.
3. Model selection: Depending on the type of problem we need to solve and the type of data we have, we choose a model best suited for our needs. For a problem, there may be multiple different algorithms that can do the task with varying levels of efficiency, in which case, we need to compare the algorithms and choose the one best suited to our data and the working constraints like time, computing power etc. For a simple regression problem, Complex algorithms like decision trees and neural networks are capable and may even be better than algorithms linear regression but are not suited in cases where the available time and computing power is less so it makes sense to use a relatively inferior model for the job.
4. Model Training: After selecting the right algorithm for the job, we need to feed the data we prepared previously and train it in order to make it capable of delivering results. Most algorithms have some parameters that can be altered to change their performance, these parameters, called hyperparameters, need to be tuned in order to extract the best performance from the model.
5. Model Evaluation: Once the model is trained on the training data, its performance needs to be tested using data previously unseen by the algorithm. The test data already has the expected outcome available with it and the predictions by the algorithm are compared against them to determine its performance. If the results of this are satisfactory then model can be sent for deployment, otherwise the training step is repeated with different parameters or more data.

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

ML is used to deal with a wide spectrum of problem types and hence also has to deal with different types of data. Here we discuss various types of data we can encounter while working on an ML problem:

1. Quantitative Data: This type of data contains numerical values.

Eg. Qualtity of items, profit, age etc

This can further be divided into Discrete and continuous data types.

* Discrete data has values that are represented in whole numbers or fractional representation would not make sense. These values can be counted. Eg. Number of Books, Marbles in a bag, students in a class etc.
* Continuous data has values that are measured not counted and decimal representation of such data is often much more accurate. Eg. Age, height, distance, mass etc.

1. Qualitative Data: Such data cannot be expressed as numbers as it describes the numerically unmeasurable data like true or false, type of car, ranking of an item on shopping site etc.

This can again be categorised into further types. Namely, Structured and unstructured data.

Structured Data contains labels or descriptions for the data that can help us identify the nature of it and its relation to the rest of the data. This can be spreadsheets or organised tables that contain headers describing the type of data.

Unstructured data does not have a known format. This can be text, audio, images or video data.

Another type of classification can be done based on the usage of data. This can be classified into:

1. Nominal: These values can be identification or labels. This is not measurable in numerical sense. Data like gender, country, name of degree come in this category
2. Ordinal: Such data has some order to it but still is not numerically measurable. This can be data like the star rating of a product, customer care feedback etc.
3. Interval: This type of data is based on some scale and zero may not represent the absence of the value. For example, in temperature, zero degrees Celsius does note mean no temperature, it can be measure in negative quantities as well and even absolute zero is still a measurement
4. Ratio: This type of data is measured in relation to some other measurement. Such data types can be seen in height or weight measurements where the measurements are done relative to a scale and can have differencing values on different scales.

3. Distinguish:

1. Numeric vs. categorical attributes

Numerical attributes contain numerical values that may or may not be continuous. These values can be measured numerically, as the name suggests. Categorical data is any other type of data like gender, name, type of car, location etc Such data cannot be numerically measured and arithmetic operations are not possible on such data

1. Feature selection vs. dimensionality reduction

Feature selection refers to the process of determining the features essential to building a model that contribute to the calculation of target variable. Different feature contribute differently to the final prediction and so some are more useful than others that can at time be entirely be discarded to reduce noise and decrease complexity of the model

Dimensionality reduction refers to the method of combining various features into fewer number of them in order to decrease complexity of the model. At times there may be too many features present to account for and some of them may have correlation among them which can affect the prediction model negatively. Such features can be combined using this technique that will allow the model to perform better with less complexity by eliminating the risk of multicollinearity at the cost of losing some information from the data

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

1. The histogram

2. Use a scatter plot

3.PCA (Personal Computer Aid)

1. The histogram

Histogram is a type of data visualization technique that can represent the distribution of data, typically continuous, over a range. This can be used to gain some insight about the data being dealt with like the scale of the data, the nature of its distribution – gaussian, skewed or exponential, outliers in the distribution etc.

In ML, this can be useful in making quick decisions for the processing of various data features. This is especially useful as often times training a model takes a large amount of data and determining its nature is essential to the performance of the data

2. Use a scatter plot

A scatter plot is used to visualize the relationship between two variables. This can be useful in determining conditions like multicollinearity in the data, formation of data clusters, nature of data across a range etc.

In clustering analysis, scatterplot can be used to visualize different data clusters and understand their size and nature. This is not always meaningful but can be useful for a preliminary analysis of the data.

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

For training an ML algorithm, typically a large amount of data is required. This data can be sourced from various places like cloud storage, local archives, monitoring systems etc. It is essential for the efficiency of the model that the data provided to it is of good quality and any bad data or noise is eliminated as much as possible. The data gathered from various sources can contain various issues like missing data, incorrect or illegible information, irrelevant features etc.

Investigating the data can also give insights about the data like it nature of distribution, different scale of distribution across features, collinearity between the features or any unnecessary features that may nor significantly towards the final output. Dealing with such problems allows us to modify the data to a state where it can give the best results when fed to the algorithm.

Qualitative and quantitative data are fundamentally very different. Naturally, the ways of exploring them are also different.

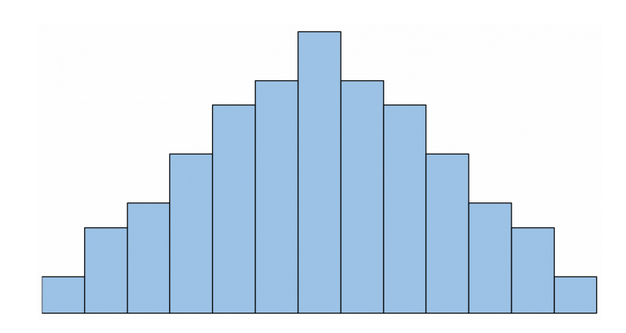
For quantitative data, we can analyse it using statistical analysis to reveal the nature of the data. Mathematical tools can be used to test hypothesis and find relationships in the data

Qualitative data on the other hand cannot be analysed using this method as it makes no sense in such calculations. It is analysed using techniques like content analysis where a researcher can look for the themes and patterns in the data. This data can still be used to form relationships between the features and the target but are more logical rather than mathematical.

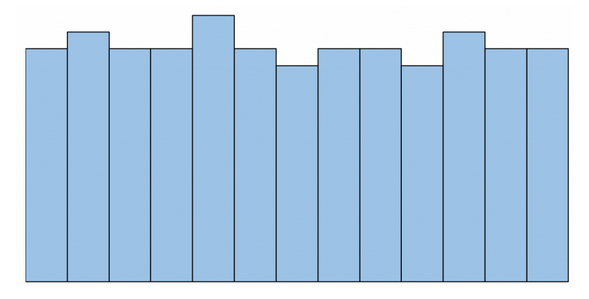
6. What are the various histogram shapes? What exactly are ‘bins'?

A histogram represents the continuous distribution of a variable and the visualization can depict a certain shape that can give us some quick one look information about the nature of the data. The data depiction in a histogram can take one of the following shapes:

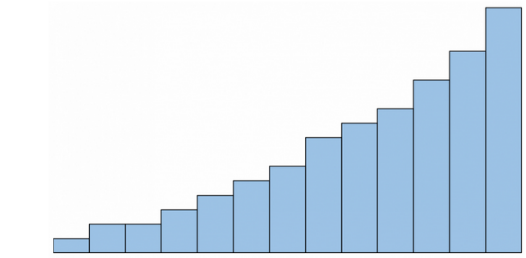
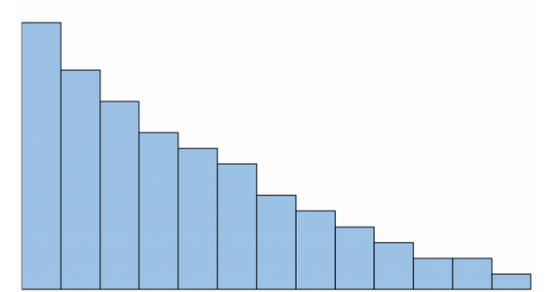
1. Bell curve: This type of curve has a single peak somewhere around the middle and tapers off the sides. This is represents that a data distribution follows the gaussian distribution model and is the most desirable distribution for the data as it gives the best results for the model.



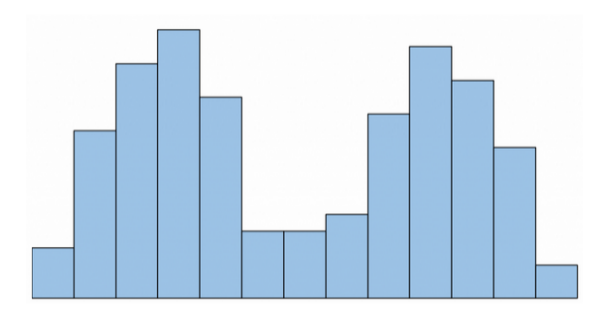
1. Uniform: This distribution does not show a clear peak of the data and is mostly like a rectangular shape with no distinguishing peaks



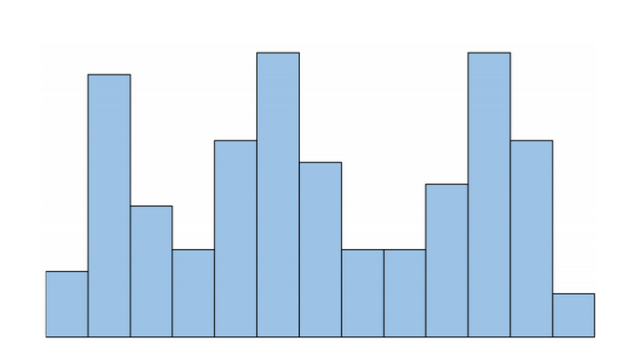
1. Skewed: Data can sometime show a skewed shape where it can show a concentration of observations in one of the sides of the graph. It can be right or left type depending on the data. The peaks here are found on one of the extremes of the graph



1. Bimodal: Such type of data displays two peaks on the distribution instead of just one. This displays the presence of more than one frequently occurring value of the feature



1. Multimodal: As the name suggests, such distributions can have more than two peaks and hence more frequent variables



1. Random: Here if no clear pattern is visible to the data, its said to be randomly distributed

Bins in a histogram are the bars whose height represents the frequency of occurrence of the observation

7. How do we deal with data outliers?

Outliers are the data points that do not follow the general pattern the rest of the data follows. These can have a significant impact on the model training and hence need to be dealt with to reduce or remove their impact. There are several ways of dealing with outliers:

1. Removal of outliers: This is probably the most obvious way of dealing with such data. Removal of outliers from the data will certainly hemp in eliminating the impact on the model but at the cost of some information loss
2. Data Transformation: Outlier data can occur due to the nature of data as well. If the data is skewed, datapoints depicted on the tail will be visible as outlier. This can be resolved using standardization or normalization where the data is converted into a normal distribution. This can help in eliminating the outliers without information loss and make the model perform better.
3. Using statistical methods: Methods like median and IQR to eliminate some data can help in dealing with outliers of there are a significantly large number of them. Such methods are less sensitive to the outlier data but can also result in significant data loss if not used properly
4. Using a robust regression analysis: For regression analysis, using a robust regression model like the huber estimator can reduce the impact the outlier on the model creation

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

For statistical analysis, there are three measures of central tendency that are widely used:

1. Mode: Mode is the most frequently occurring value in a distribution. This has one advantage over the other two in that it can be used for both categorical and numerical distributions. However, mode may not be an accurate representation of the central tendency all the time. There can also be multiple models for multimodal data.
2. Median: Median is simply the middle value of the data that divides it into equal halves. This is usually the preferred measure over mean and mode since it is more accurate and less affected by outliers and skewness in the data. However, this cannot be used for categorical data.
3. Mean: Arithmetically, mean is the sum of the observation values divided by the number of observations. This can be used for continuous and categorical numeric data. Mean is less accurate than median as it is easily affected by any skewness and outliers in the data

Mean takes into account the values of all the observations and is hence affected heavily by any skewness and outliers in the data. An outlier with a value very large or small compared to the resto of the data can cause a big change in the value of the mean. Median on the other hand is not affected by the outliers or skewness as it is just the middle observation of the data when in order. Due to this, in datasets having skewed nature or large outliers, mean and median values can vary greatly

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 can be used to observe the relationship between variables. It can represent various patterns or shapes that can emerge with enough datapoints plotted on the graph.

A positive relationship between the variables can be seen between the variables by a slope increasing from left to right. Similarly, negative can be visualized seeing an opposite slope. Density of the datapoints can also indicate the formation of clusters in the data allowing for formation of various classes and groups if required

Outliers can easily be visualised in a scatterplot as they will take a visibly different position in the graph than the pattern followed by the rest of the data.

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

Cross tabs are a way of organizing data and showing the relationships between two variables. This can be created using a table that contains the frequency or the number of observations that fall in each of the categories. This can help us determine the relationships between two variables.

For example, considering a comparison between gender and choice of ice cream flavours, a cross tab will be able to quickly represent what flavours are more preferred by a specific gender and by how much. It can allow for quick analysis of any preference depending on the gender