Linear regression

This is a supervised machine learning algorithm. It used to find a cause-and-effect relationship between variables. It is a model that assumes a linear relationship between the independent and dependent variable(s), independence between residuals (each observation) and that data is normally distributed. You can have more than one input (independent) variable. It works on producing a line of best fit that is based on a linear equation. It uses data that has been labelled to train the ML algorithm. The model gets the best fit by finding the best values for the y-intercept and coefficient of x. The model aims to predict the dependent (y) variable such that the error between the predicted and true value is minimum. This algorithm is useful for making predictions and forecasts, however outliers can have a significant effect.

Logistic regression

This is a supervised machine learning algorithm. It is a classification algorithm and used when the dependent variable is categorical. The algorithm is used to predict the probability of a binary outcome using a logistic function, which gives a value between 0 (impossible) and 1(certain). The dependent variable is categorical in nature and fits into one of two categories. The independent variables can be continuous or discrete (nominal or ordinal). This is a model that assumes dependent variable being binary and little or no collinearity between independent variables. Examples of this include whether an email is spam or not and if a bank should accept an application for a credit card.

Decision Tree

Decision Trees (DTs) are a supervised learning algorithms used for both classification and regression. They are a flowchart based on previous experiences to make a decision. The internal nodes of the trees represent a test or a question on an attribute; each branch is the possible outcome of the question asked, and the terminal node, which is also called as the leaf node, denotes a class label. It is useful for solving decision related problems, however it can have many layers which lead to more complexity.

SVM (Support Vector Machine)

SVM is a supervised learning method used for classification and regression. SVM maps data points for different classes and separates them with hyperplanes. The aim is to find a hyperplane that has a maximum distance between data points of both classes (maximum margin). The constructed hyperplanes can be both linear and nonlinear to help classify the data points. The advantages of SVM is they work well in high dimensional space but are not suitable for large datasets. Another issue with SVM is when there is overlap between classes. An example of its use is in predicting whether cancer is malignant or benign

Naive Bayes

This is a classifier algorithm based on conditional probability. It is used to predict values (y) based on features (x). The parameters are assumed to have an equal weighting and independent of each other. Having many parameters makes it difficult to calculate the conditional probability of an event. To make this simpler we assume all features on independent, which makes classification simpler with fewer parameters. Typical applications include filtering spam and classifying documents. The advantages of are the algorithm performs well considering the assumptions are rarely true. It can handle multi-dimensional data and perform well with a small set of data. However other algorithms may perform better with lots of data.

KNN (K- Nearest Neighbours)

KNN is a supervised machine learning algorithm and used for both classification and regression. It works by classifying new data set based on distance. It finds KNN to the test data and then classification is performed. The main assumption is that similar items are in close proximity to each other. You need to first decide the value of K. K indicates the count of the nearest neighbours. If it is too small, it can give inaccurate solutions. Larger values of K will have smoother decision boundaries which mean lower variance but increased bias The value of K has a powerful effect and can result in test data being classified in different classes depending on the value chosen. After choosing K, the distance of K number of neighbours is calculated. Among these k neighbours, count the number of the data points in each category and that is predicted class for test data. It is simple to implement and with enough data can do well. The challenges are selecting a suitable ‘K’ value and the many computations that need to take place.

K-Means

K-Means is a unsupervised clustering algorithm. Data is analysed looking for meaningful groups based on similarity. First you need to choose the ‘K’ value which represents the number of clusters. For each cluster a centroid is chosen and based on the proximity to the centroid, the data points will belong to the corresponding cluster. Once grouped into clusters the geometrical centre for each cluster can be calculated to find the centroid for each. One of the challenges is selecting the number of clusters initially. If it is too small the centroid may not lie in the clusters and if too big the clusters may be split. An example is clustering patients into different subgroups and build a model for each subgroup to predict the probability of the risk of having heart attack.

Random Forest

This is a supervised machine learning algorithm. It combines multiple DTs to form a ‘forest’. It can be used for both classification and regression. DTs are merged together for a more accurate prediction. The theory is that whilst individual DTs may have errors, as a group the majority will be correct and thereby giving a positive outcome overall. DTs are mixed together randomly. Features are selected features to build DTs and then outcomes are averaged. It is used in banking to find customers likely to repay on time. It can also be used by retail companies to recommend products and predict customer satisfaction as well.