**MACHINE LEARNING ALGORITHMS**

**Linear Regression**

Linear Regression is a **machine learning algorithm based on supervised learning**. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting

**•Is it Supervised/Unsupervised/Reinforcement learning?**

Supervised

**•What does the algorithm do?**

* Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.
* Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.
* Linear Regression is a supervised machine learning algorithm where the ***predicted output*** is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

**•In which situations will it be most useful?**

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship.

**•(Optional) Can you find any examples of where this algorithm has been used?**

Predicating sales – sales being the outcome and advertising being the predictor variable. It will estimate the sales dependant on how much the advertising is. So if advertising is increased it will predict what sales might be.

**Logistic Regression**

What is logistic regression in Machine Learning?

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1.

**•Is it Supervised/Unsupervised/Reinforcement learning?**

***Supervised*** learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variables (x) and an target variable (Y) when you train the model

**•What does the algorithm do?**

**•In which situations will it be most useful?**

When to use logistic regression. Logistic regression is applied to predict the categorical dependent variable. In other words, it's used when the prediction is categorical, for example, yes or no, true or false, 0 or 1

**•(Optional) Can you find any examples of where this algorithm has been used?**

Predicting if an email is spam or not spam

Whether a tumour is malignant or benign

Whether a mushroom is poisonous or edible.

**Decision Tree**

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree

**•Is it Supervised/Unsupervised/Reinforcement learning?**

Decision Trees are a type of ***Supervised*** Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

**•What does the algorithm do?**

Goes through each branch and gives it a yes/no outcome and moves onto the next branch until you reach a conclusion.

A decision tree is an upside-down tree that makes decisions based on the conditions present in the data

A decision tree follows a set of if-else conditions to visualize the data and classify it according to the conditions.

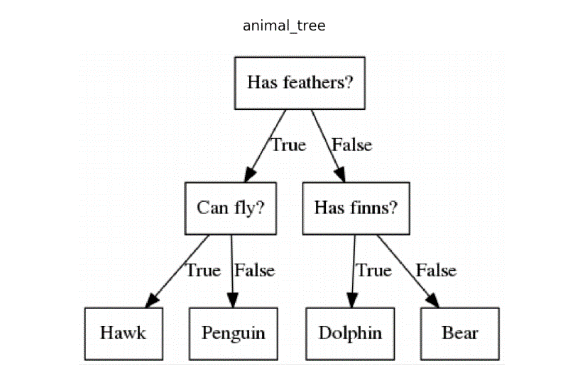
**•In which situations will it be most useful?**

When wanting to go through all the possible outcomes for a problem.

**•(Optional) Can you find any examples of where this algorithm has been used?**

An example of a decision tree can be explained using above binary tree. Let’s say you want to predict whether a person is fit given their information like age, eating habit, and physical activity, etc. The decision nodes here are questions like ‘What’s the age?’, ‘Does he exercise?’, ‘Does he eat a lot of pizzas’? And the leaves, which are outcomes like either ‘fit’, or ‘unfit’. In this case this was a binary classification problem (a yes no type problem).

Identifying an animal



**SVM (Support Vector Machine)**

SVM is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems.

**•Is it Supervised/Unsupervised/Reinforcement learning?**

***supervised***

**•What does the algorithm do?**

The algorithm creates a line or a hyperplane which separates the data into classes.

**•In which situations will it be most useful?**

used for classification, regression and outlier’s detection

**•(Optional) Can you find any examples of where this algorithm has been used?**

Facial Expression Classification.

Text Classification.

Speech Recognition.

handwriting recognition

**Naive Bayes**

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

**•Is it Supervised/Unsupervised/Reinforcement learning?**

***Supervised*** learning algorithm, which is based on Bayes theorem and used for solving classification problems

**•What does the algorithm do?**

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature

**•In which situations will it be most useful?**

**Real time Prediction**: Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.

**Multi class Prediction**: This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.

**Text classification/ Spam Filtering/ Sentiment Analysis**: Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)

**Recommendation System**: Naive Bayes Classifier and Collaborative Filtering together builds a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not

**•(Optional) Can you find any examples of where this algorithm has been used?**

a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’.

**KNN (K-Nearest Neighbours)**

The k-nearest neighbours (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

**•Is it Supervised/Unsupervised/Reinforcement learning?**

***Supervised***

**•What does the algorithm do?**

use a set of input values to predict output values. KNN is one of the simplest forms of machine learning algorithms mostly used for classification. It classifies the data point on how its neighbour is classified

**•In which situations will it be most useful?**

if the data consists of more than two labels or in simple words if you are required to classify the data in more than two categories then KNN can be a suitable algorithm

**•(Optional) Can you find any examples of where this algorithm has been used?**

**Banking System** - KNN can be used in banking system to predict weather an individual is fit for loan approval? Does that individual have the characteristics similar to the defaulters one?

**Calculating Credit Ratings** - KNN algorithms can be used to find an individual’s credit rating by comparing with the persons having similar traits.

**Politics** - With the help of KNN algorithms, we can classify a potential voter into various classes like “Will Vote”, “Will not Vote”, “Will Vote to Party ‘Congress’, “Will Vote to Party ‘BJP’.

**K-Means**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. ... To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.” A cluster refers to a collection of data points aggregated together because of certain similarities.

**•Is it Supervised/Unsupervised/Reinforcement learning?**

***unsupervised***

**•What does the algorithm do?**

K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum.

The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

* Determines the best value for K centre points or centroids by an iterative process.
* Assigns each data point to its closest k-centre. Those data points which are near to the particular k-centre, create a cluster.

**•In which situations will it be most useful?**

The K-means clustering algorithm is used to find groups which have not been explicitly labelled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

**•(Optional) Can you find any examples of where this algorithm has been used?**

Let’s consider the data on drug-related crimes in Canada. The data consists of crimes due to various drugs that include, Heroin, Cocaine to prescription drugs, especially by underage people. The crimes resulted due to these substance abuse can be brought down by starting de-addiction centres in areas most afflicted by this kind of crime. With the available data, different objectives can be set. They are:

Classify the crimes based on the abuse substance to detect prominent cause.

Classify the crimes based on age groups.

Analyse the data to determine what kinds of de-addiction centre is required.

Find out how many de-addiction centres need to be setup to reduce drug related crime rate.

The K-means algorithm can be used to determine any of the above scenarios by analysing the available data.

**Random Forest**

Random Forests is a Machine Learning algorithm that tackles one of the biggest problems with Decision Trees: *variance*.

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees.

Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction

**•Is it Supervised/Unsupervised/Reinforcement learning?**

*Supervised*

**•What does the algorithm do?**

Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

**•In which situations will it be most useful?**

It provides higher accuracy through cross validation. Random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data

**•(Optional) Can you find any examples of where this algorithm has been used?**

Random Forest is used in **banking** to detect customers who are more likely to repay their debt on time. It’s also used to predict who will use a bank’s services more frequently. They even use it to detect fraud. Talk about the robin hood of algorithms!

**Stock traders** use Random Forest to predict a stock’s future behaviour. It’s used by retail companies to recommend products and predict customer satisfaction as well.

In **healthcare**, Random Forest can be used to analyse a patient’s medical history to identify diseases. Pharmaceutical scientists use Random Forest to identify the correct combination of components in a medication or predict drug sensitivity. Sometimes Random Forest is even used for computational biology and the study of genetics.