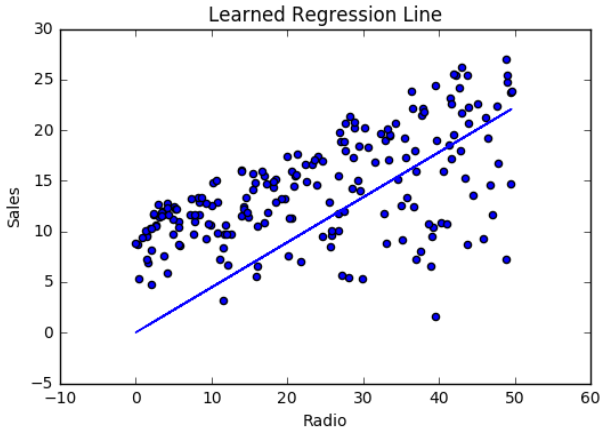
Machine Learning Algorithms **Lisa Duffy**

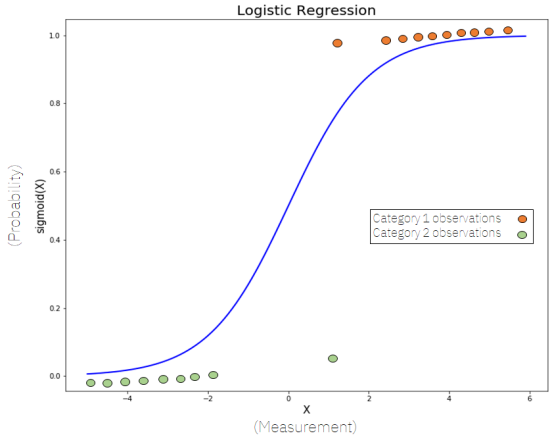
Linear Regression

This is a supervised machine learning algorithm which predicts values based on input data fed into the algorithm. The input data has fixed values which are used to train the algorithm in order to predict future/unknown values. The predicted output, which is based on the input data provided, is continuous and has a constant slope (sometimes called the Line of Best Fit). It is used to predict values within a continuous range such as real estate costs and stock values. An example of a liner regression visualisation showing sales of radio advertising follows:



Logistic Regression

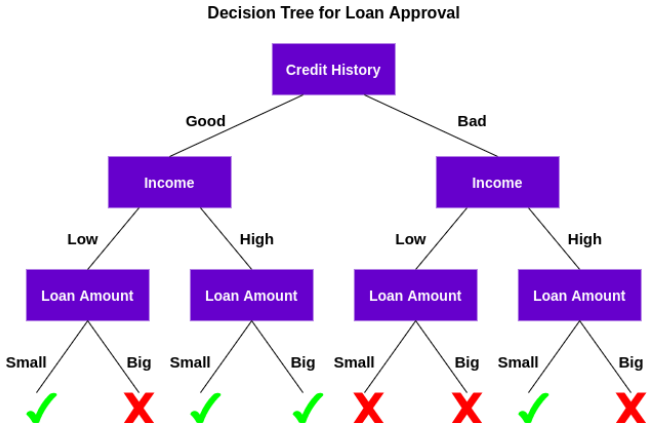
This is also a supervised machine learning algorithm and predicts values based on input data. But the difference to linear regression is that there is not a continuous set of data and it makes use of binary classification. Because of the non-linear nature of the data, a sigmoid (s-shaped curve) is used for predictions. Here follows a curve showing probability in relation to a measurement :



The algorithm requires a fixed number of parameters, which are dependent on the number of input features, and it will output a categorical prediction. This algorithm is particularly useful in categorisations (e.g. determining if a particular plant belongs to a specific species).

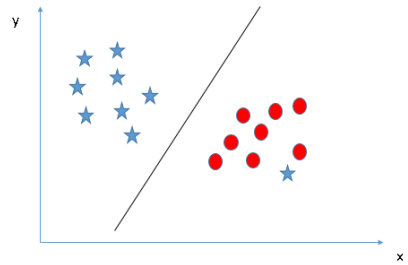
Decision Tree

This is another supervised learning machine learning algorithm which can be used for both classification and regression problems. The model is developed on training data which is fed into the algorithm and the model is then used to predict outcomes for given data. The model makes use of simple decision rules learned from the training data to predict classes / values for the given data. They are used in variety of business situations to help companies predict future outcomes (e.g. what kind of customers should a loan company provide loans to).



SVM (Support Vector Machine)

This is yet another supervised machine learning algorithm and can also be used for both regression and classification problems although it, like decision trees, is mostly used for classification problems. It makes use of features in a set of data and then plots the data as points in a multi-dimensional space (dependent on the number of features). In order to predict the classification of given data, the hyper-plane that differentiates the classes from the training data is determined and from this the given data is classified accordingly.



An important feature of the SVM algorithm to note is that it has a feature to ignore outliers and is therefore more robust. It is most commonly used in the classification of genes and also in handwriting recognition.

Naïve Bayes

This is a set of algorithms that belong to the supervised learning class. They make use of Bayes Theorem to calculate probabilities of a hypothesis given prior knowledge (which is obtained through the training data). Bayes' Theorem is as follows:

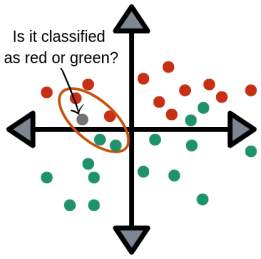
$$P(h|d) = (P(d|h) * P(h)) / P(d) \text{ where}$$

- **P(h|d)** is the probability of hypothesis h given the data d. This is called the posterior probability.
- **P(d|h)** is the probability of data d given that the hypothesis h was true.
- **P(h)** is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- **P(d)** is the probability of the data (regardless of the hypothesis).

A fundamental assumption made in this model is that each feature in the data makes an equal and independent contribution to the outcome (this is seldom true in real-world situations, but this model still seems to work well in practice).
After calculating the posterior probability for a number of different hypotheses, the hypothesis with the highest probability is selected and called the maximum a posteriori (MAP). It is primarily a classification algorithm and can be used for binary and multi-class classification problems. Examples of application include text classification and spam filtering.

KNN (K-Nearest Neighbours)

This is another supervised learning machine learning algorithm that can be used to solve both classification and regression problems. It classifies data based on the theory that it will most likely be similar to the data that exists in close proximity. The training data is used to plot data points and then a prediction is made for the given data based on where it lies in relation to those data points. K is used to refer to number of nearest neighbours and can be set by the algorithm. The image below shows a visual representation of this algorithm:



It is often used in recommendation systems such as Netflix and Amazon, but its practical use is limited as the algorithm becomes significantly slower as the amount of data increases.

K-Means

This is an unsupervised machine learning algorithm which means that the training data does not include specific desired solutions and therefore looks for trends / patterns to predict outcomes. It tries to group similar items in the form of clusters and K refers to the number of groups represented. The first step in this algorithm is determining the K-number and this can be done either by the Elbow or Silhouette method. Once this has been determined the algorithm can be used. The figures below show a visual explanation of how the algorithm works:

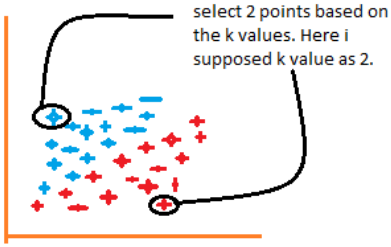
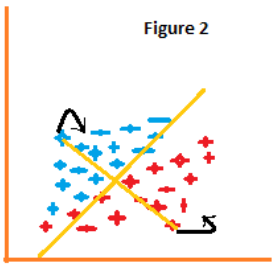


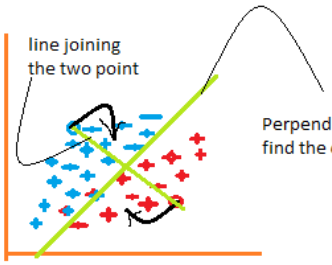
Figure 1



F2: Find the average of all the blue points and red points and move the selected points to **centroid.**

= original points

= the original points moved to centroid.



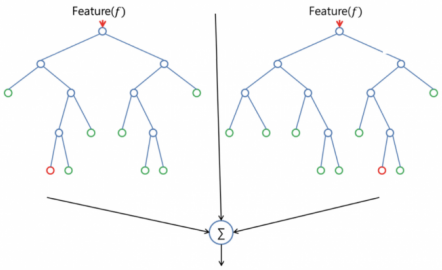
F3: Some of the red points changed to blue points, that means they belong to the group blue now. Again the repeat the same process.

F4: The same process has been applied here. This process will be continued until we get the **two complete different cluster.**

K-Means clustering is used in search engines to in order to group the search results.

Random Forest

This is again a supervised machine learning algorithm and it builds multiple decision trees (see above) and then merges them together to get a more accurate and stable prediction. An important difference to note between decision trees and random trees is that while decision trees formulate a set of rules from the training data to make predictions, a random forest algorithm selects observations and features from the training data to build several decision trees and then averages the results to make a prediction. The figure below shows a random forest with 2 decision trees:



It is a very versatile algorithm and is used in a variety of applications and is quick to train. But as the size of the forest grows, the speed at which predictions are made can be slow. It is commonly used in e-commerce to determine whether a customer will enjoy a product or not and also in medicine to analyse a patient's medical history to identify diseases.