Common Machine Learning Algorithms

1 Linear Regression

Linear regression is a statistical and machine learning algorithm which is a type of supervised learning. The algorithm is a linear model and shows the correlation between a set of input values and a single output value. Linear regression can be used to fit a predictive model to observed data by extrapolating the observed data. The model is unbounded so is not suitable for classification problems.

2 Logistic Regression

Logistic regression is a type of supervised machine learning and is used when the dependent variable is categorical. In this algorithm, a logistic function acts upon data within a linear regression model in order to predict the target categorical dependent variable. There are different types of logistic regression: binary logistic regression is when the categorical response has two possible outcomes; multinomial logistic regression is when there are three or more unordered categories; ordinal logistic regression has three or more ordered categories (eg ratings 1-10). Decision boundaries can be set to determine which class specific data belongs to, and this boundary can be linear or non-linear. Logistic regression was used in numerous social science applications in the early 20th Century.

3 Decision Tree

A decision tree is a supervised machine learning model. It is non-parametric and uses a tree-like model of decisions and the resulting possibilities of each decision. They are used to solve both classification and regression problems by splitting the data set into smaller data sets. They can both visually and explicitly represent decision making. The algorithm learns simple decision rules from previous data. A training data set is input to the decision tree, and from this a set of rules is formulated in order to make predictions on the decisions likely to be made.

4 SVM (Support Vector Machine)

SVMs are a set of supervised learning methods which are used to solve both classification and regression problems, as well as detect outliers. SVMs are effective in high dimensional spaces and in cases where the number of dimensions is greater than the number of samples, however is less effective when the data set is large or when it has more noise (such as when there are overlaps in the classes). In the algorithm, each item of data is plotted as a point in n-dimensional space where n is the number of features. We then identify a hyper-plane which differentiates the two classes by segregating the data with equal margins between the hyper-plane and each class of data. The algorithm ignores outliers and finds the hyper-plane that has the maximum margin.

5 Naive Bayes

Naive Bayes are a set of supervised learning algorithms which are used to solve classification problems. The algorithm is based on the Bayes Theorem for calculation probabilities and conditional probabilities but works on the naive assumption of independence among predictors; that the presence of a feature in a class is unrelated to the presence of any other feature. The model is fast and simple and is particularly useful for large data sets, however due to the assumption of independence can produce inaccurate results.

6 KNN (K- Nearest Neighbours)

The KNN algorithm is a supervised machine learning algorithm. It can be used to solve classification and regression problems and is easy to implement, however it becomes significantly slower as the size of the data grows. The algorithm is based on the assumption that similar data points are close to each other and calculates the distance between data points, orders the indices by size and selects the first K entries of smallest distance. For regression problems the algorithm returns the mean of the K labels, and for classification problems the mode of the K labels. KNN is typically used for solving problems which require identifying similar objects, such as recommender systems which recommend products or services based on those a customer has previously used.

7 K-Means

K-means clustering is a method of unsupervised machine learning which is used to identify clusters of data objects in a data set. The algorithm works to group data into a fixed number K of clusters based on the similarity of features of the data. It does this by using iterative refinement, randomly generating initial estimates for the K centroids, which represent the centres of the clusters. Data points are each allocated to a cluster, and iterative calculations optimise the positions of the centroids, ceasing either when the centroids have stabilised or when the defined number of iterations has been performed. This method is typically used to confirm business assumptions about unlabelled groups within data, or to identify unknown groups in complex data sets.

8 Random Forest

Random forest is a supervised learning algorithm which solves classification and regression problems by averaging decision outcomes from multiple independent decision trees. Since the accuracy of a decision tree depends heavily on the strategic splits that are made, random forests produce more accurate and stable predictions.