Home Learning Task Week 6

Machine Learning Algorithms

Linear Regression

* This is a form of supervised learning and is used to predict the behaviour of data based on a set of variables.
* Tries to find a straight line that best fits the plotted data.
* As the name suggests, the two variables which are on the x-axis and y-axis should be linearly correlated.
* This is most useful for finding a relationship between two variables.
* This can be used to understand the relationship between say spending on advertising and the revenue generated, linear regression can also be used to predict the grade that a student will attain based on the number of hours they have put in for preparation.

Logistic Regression

* This is also a form of supervised learning and is used to predict a relationship between features and the probability of a categorical outcome, which are in two or more categories.
* The predictions are in the form of probabilities of an event occurring.
* Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category.
* An example of this is in determining whether a tumour is malignant or not, another example is to analyse whether a customer will purchase a product or not.

Decision Trees

* This is another form of supervised learning and are used to create a model that predicts the value of a target variable by learning simple decision rules gathered from the data features.
* They can be used for solving both regression and classification problems.
* A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; the edges represent the answers to the question and the leaves represent the actual output or class label.
* Decision trees can be very useful tools for individuals, investors, and business professionals that need a visual way to break down a complex decision
* An example of a decision tree being used could be a farmer using rainfall forecasts, predictions of commodity prices and yields per acre to decide between planting wheat, corn or applying for a government subsidy by not planting anything.

SVM (Support Vector Machine)

* This is another form of supervised learning and used to classify inputs into a predefined class such as Yes / No. This is also based on the idea of finding a hyperplane (a decision boundary that helps classify the data points) that divides a dataset into to groups.
* This can be used for solving both regression and classification problems, but mostly for classification problems.
* This is most useful when we are dealing with types of data such as images, text and audio.
* Examples of where the algorithm has been used is facial recognition, detecting spam e-mails and response analysis.

Naive Bayes

* This is a form of supervised learning and uses probabilistic classification algorithm based on the Bayes Theorem: it uses probability to make predictions for the purpose of classification.
* if you use Naive Bayes for sentiment analysis, the sentence ‘I like to eat cake’, the algorithm will look at the individual words and not the full sentence. For this algorithm, phrases like ‘I like to eat cake’, ‘to eat cake like I’, and ‘cake I like eat to’ are the same.
* This is most useful when we are dealing with a large text based dataset and a small training set.
* Examples of where the algorithm has been used include sentiment analysis, filtering spam and document/image classification systems.

KNN (K Nearest Neighbours)

* This is a form of supervised and pattern classification learning algorithm which helps you to find which class the (new input) test value belongs to when k nearest neighbours are chosen using the distance between them.
* This can be used for solving both regression and classification problems.
* The algorithm can be used for attributing the missing value of both categorical and continuous variables and to predict a particular outcome.
* Examples of where the algorithm has been used include recommendation systems, semantic searching, and anomaly detection.

K-Means

* This is a form of unsupervised learning and is used to group similar data points together into clusters and discover underlying patterns. To do this, K-means looks for a fixed number (k) of clusters in a dataset.
* The data used by this algorithm is unlabelled (data without defined categories or groups).
* This algorithm can be useful for finding groups which have not been labelled in the data and detect anomalies in a set of data.
* Examples of where the algorithm has been used include document classification, customer segmentation and detection of insurance fraud.

Random Forest

* This is a form of supervised learning and this builds an ensemble of decision trees and then merges them together to attain a more accurate and stable prediction.
* A random forest is also seen as an example of an ensemble, which is a combination of predictions from different models. In an ensemble, predictions could be combined either by majority-voting or by taking averages.
* This can be used for solving both regression and classification problems and can handle large data sets. You should note that the final/leaf node will be either the majority class for classification problems or the average for regression problems
* The algorithm can me modelled for prediction and behaviour analysis.
* This can be used for Fraud Detection, Credit Scoring and value prediction.