Data Science – Algorithms (Machine Learning)

Task 1

Select one or more choices from the list of common Machine Learning Algorithms, do some investigations and write me a short summary. I am looking for the following:

Linear Regression, Logistic Regression, Decision Tree, SVM (Support Vector Machine), Naive Bayes, KNN (K- Nearest Neighbours), K-Means and Random Forest

- Is it Supervised/Unsupervised/Reinforcement learning?
- What does the algorithm do?
- In which situations will it be most useful?
- (Optional) Can you find any examples of where this algorithm has been used?

(1)

Linear Regression: Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (X) and an output variable (Y) for each example. The model finds the linear relationship between the dependent and independent variable. Linear regressions can be used in business to evaluate trends and make estimates or forecasts.

Advantages - Linear Regression is simple to implement and easier to interpret the output coefficients. When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of its less complexity to compare it to the other algorithms. Linear Regression is susceptible to over-fitting, but it can be avoided using some dimensionality reduction techniques.

Disadvantages - On the other hand in linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique. Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes. But then linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

(2)

K-Means: k-means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabelled data into a predetermined number of clusters based on similarities (k). Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. In other words, the algorithm calculates the average of all the points in a cluster and moves the centroid to that average location. The main aim of this algorithm is to minimise the sum of distances between the data point and their corresponding clusters. This 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 such as customer segmentation, document classification, etc.

Advantages: It is easy to implement k-means and identify unknown groups of data from complex data sets. The results are presented in an easy and simple manner. K-means is suitable for a large number of datasets, and it’s computed much faster than the smaller dataset. It can also produce higher clusters. K-means analysis improves clustering accuracy and ensures information about a particular problem domain is available. Modification of the k-means algorithm based on this information improves the accuracy of the clusters.

Disadvantages: K-means doesn’t allow the development of an optimal set of clusters and for effective results, you should decide on the clusters before. Changing or rescaling the dataset either through normalization or standardization will completely change the final results. K-means clustering technique assumes that we deal with spherical clusters and each cluster has equal numbers for observations. The spherical assumptions have to be satisfied. The algorithm can’t work with clusters of unusual size.