Machine Learning:
jnautumn – 17/08/21

Linear Regression:

- Supervised learning
- Performs a regression task: aiming to predict values based on independent variable inputs along a line of best fit. It is mostly used to determine the relationship between two variables and forecasting predictions around the variables.
- Useful in business analytic charts: cost predictions, profit/loss forecasts, etc.

Logistic Regression:

- Supervised learning
- Aims to predict the probability of a target value. Output variables must be binary or binomal (1 or 2, true or false, success or fail). Multinomal and ordinal models also exist.
- Originally developed for ecology and forestry for soil monitoring, it is useful today for other scientific applications such as medicine (detection of diseases) and general technology such as spam filtering.

Decision Tree:

- Supervised learning
- Used for both classification and regression tasks. Data is built using decision nodes where binary 1/0 or true/false is used to determine an output of success or failure. Implemented using the Gini method where the Gini Index is used to calculate binary splits in the decision nodes.
- Most useful with linear datasets and technical or financial applications. Widely used in call centre switchboards to route inbound calls.

SVM (Support Vector Machine):

- Supervised learning. Similar unsupervised model exists in the form of SVC/vector clustering.
- Used primarily for classification tasks and is a popular ML option for handling continuous data and categorical variables. It is built by constructing a hyperplane that aims to segregate data clusters and creates a separating line according to the nearest data points between opposing sets. Wider margins between these points indicates a lower likelihood of sample errors.
- Widely used for the classification of digital media and image search engines. Has been used for OCR and handwriting to data conversions.

Naive Bayes:

- Supervised learning
- It is based on the Bayes Theorem for calculating probabilities and conditional probabilities by implementing the Naive Bayesian Equation.
- It is used primarily for text classification as it performs well in multi class predictions. It is regularly seen in both spam filtering and business sentiment analysis/social media response. Also forms a key system of e-commerce recommendations through the Naive Bayes Classifier

and Collaborative Filtering.

K Means:

- Unsupervised learning
- Also known as the flat clustering algorithm, it performs its calculation based on the squared distance between data points. Shorter distances represent more similar data points/fewer variations in the data class. It follows the expectation maximization approach to solving equations.
- It is mostly used to identify trends in dynamic datasets and is useful in business metrics/KPIs.

KNN / K Nearest Neighbours:

- Supervised learning
- Is an algorithm based on data/cluster similarity to predict the values of new data points. The resulting calculation reflects how closely new points match data used in the initial training dataset and the distance between other points in neighbouring clusters. The distance is most often calculated using the Euclidean method.
- Has uses in the financial industry: determining credit applications, mortgage/loan rates, and current account selection.

Random Forest:

- Supervised learning
- Created by Leo Breiman and Adele Cutler, the algorithm combines the results of two or more decision trees to reach a single conclusion. They are used for both regression tasks (end result calculated by averaging the results of each tree) and classification tasks (end result being the most frequent binary variable/majority vote system)
- It is a preferred algorithm in finance and is especially useful in fraud detection.