TT – Home Learning Task 8

Linear Regression

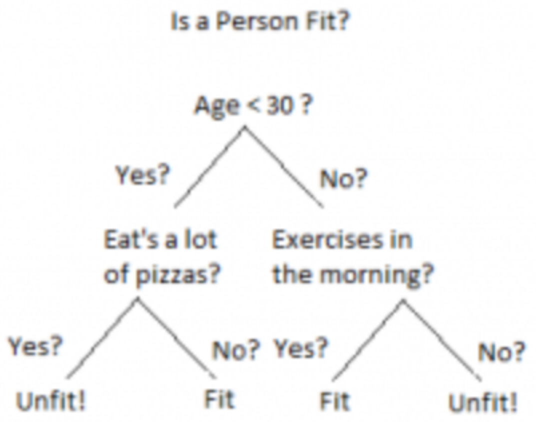
- Linear regression is supervised machine learning
- The algorithm finds the best fit linear line between the independent and dependent variables
- It finds the linear relationship between the dependent and independent variable

Logistic Regression

- Logistic regression is a supervised learning algorithm
- Is used to model the probability of a certain class or event existing such as pass/fail, win/lose
- It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio level independent variables
- It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary)
- It is a predictive analysis
- Some assumptions:
  - The dependent variable should be dichotomous in nature (e.g., presence vs. absent).
  - There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores and removing values below -3.29 or greater than 3.29.
  - There should be no high correlations (multicollinearity) among the predictors. This can be assessed by a correlation matrix among the predictors.
- An example of its use:
  - Do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes vs. no)?

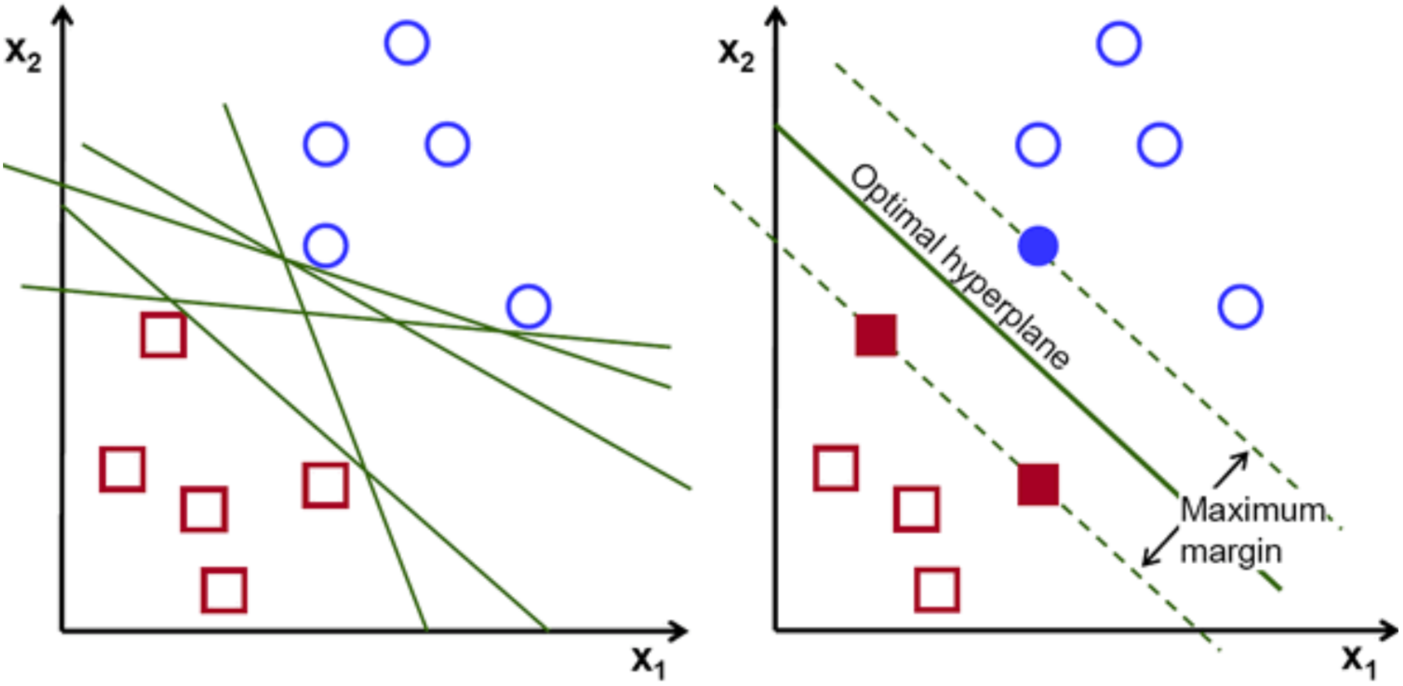
Decision Tree

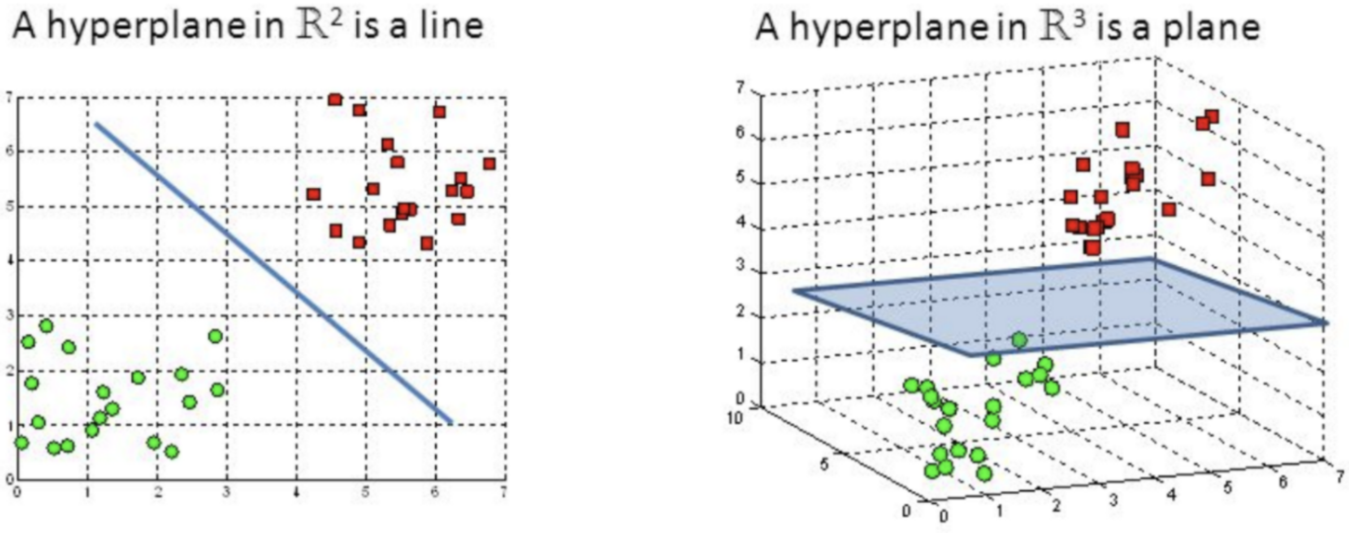
- A type of supervised machine learning
- Where data is continuously split according to a certain parameter
- Can be explained by 2 entities, decision nodes and leaves
- The leaves are the decisions or final outcomes
- The nodes are where the data is split
- Example below:



SVM (Support Vector Machine)

- Is a supervised machine learning algorithm
- It can be used for both classification or regression challenges, but mainly classification problems
- The objective is to find the hyper line in an N-dimensional space (n = the number of features) that distinctly classifies the data points
- To separate the two classes of data points, there are many possible hyperplanes that could be chosen
- The objective is to find a plane that has the maximum margin (the maximum distance between data points of both classes)
- Maximising the margin distance provides some reinforcement so that future data points can be classified with more confidence
- Hyperplanes are decision boundaries that help classify the data points
- Data points falling on either side of the hyperplane can be attributed to different classes
- The dimension of the hyperplane depends on the number of features



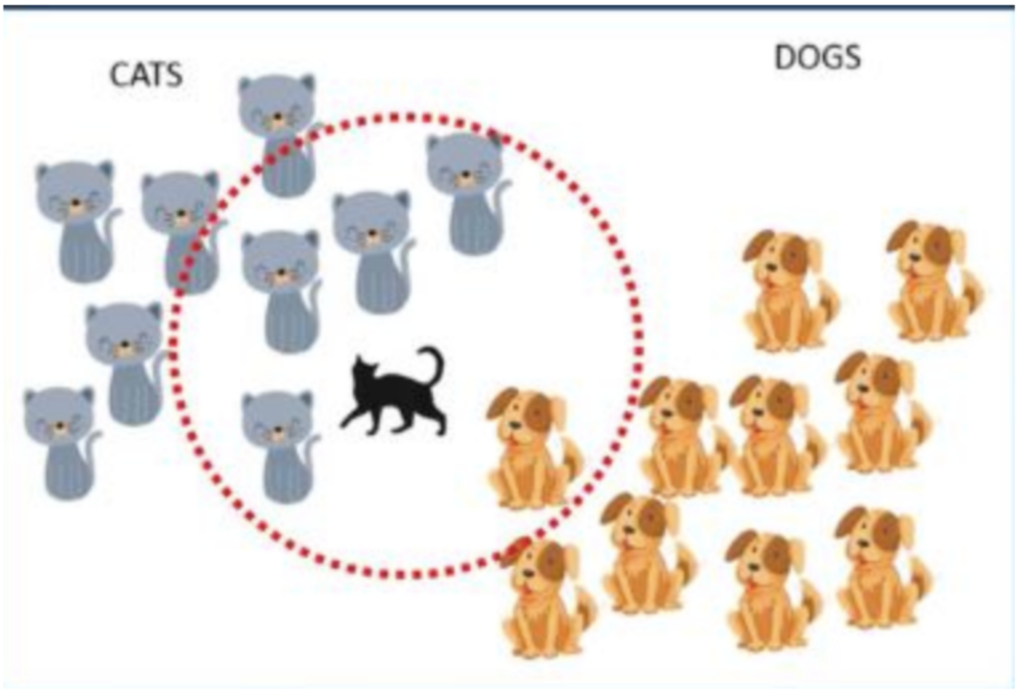


Naive Bayes

- Supervised learning algorithm based on the Bayes theorem
- It is used for solving classification problems
- Mainly used in text classification that includes a high dimensional training dataset
- Simple and effective classification algorithm that helps in building fast machine learning models that can make quick predictions
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object
- It assumes that the occurrence of a certain feature is independent of the occurrence of other features
- Bayes theorem, is used to determine the probability of a hypothesis with prior knowledge
- It depends on conditional probability
- An example is spam filtration

KNN (K- Nearest Neighbours)

- Supervised machine learning
- Can be used to solve classification and regression problems
- It finds the number of nearest neighbours to a new unknown variable
- The variable that has to be predicted or classified is denoted by the symbol 'K'
- Its aim is to locate all of the closest neighbours around a new unknown data point in order to figure out what class it belongs to.
- KNN calculates the distance from all points in the proximity of the unknown data and filters out the ones with the shortest distances to it.
- It is often referred to as a distance-based algorithm.



K-Means

- Unsupervised machine learning
- This algorithm explores for a pre-planned number of clusters in an unlabelled multi-dimensional dataset
- Its output is how an optimised cluster can be expressed
- First the cluster centre is the arithmetic mean of all the data points associated with the cluster
- Then each point is adjoint to its cluster centre in comparison to other cluster centres
- The 2 interpretations are the foundation of the k-means clustering model
- It enables the user to cluster the data into several groups by detecting the distinct categories of groups in the unlabelled datasets by itself, without the necessity of training the data
- It is very smooth in terms of interpretation and resolution.
- For a large number of variables present in the dataset, K-means operates quicker than Hierarchical clustering.
- While redetermining the cluster centre, an instance can modify the cluster.
- K-means reforms compact clusters.
- It can work on unlabelled numerical data.
- It is fast, robust and uncomplicated to understand and yields the best outcomes when datasets are well distinctive (thoroughly separated) from each other.

Random Forest

- Supervised machine learning
- Classifier that combines a number of decision trees on different subsets of a dataset and averages the results to increase the dataset's predicted accuracy
- It creates a forest out of an ensemble of decision trees, which are commonly trained using the 'bagging' approach
- The 'bagging' method's basic premise is that combining several learning models improves the final output
- The random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions
- The random forest method has the following steps:
    - **Step 1:** In a Random forest, n random records are chosen at random from a data collection of k records.
    - **Step 2:** For each sample, individual decision trees are built.
    - **Step 3:** Each decision tree produces a result.
    - **Step 4:** For classification and regression, the final output is based on Majority Voting or Averaging, accordingly.
- Examples:
    - Banking
        - In banking, a random forest is used to estimate a loan applicant's creditworthiness. This assists the lending organization in making an informed judgment about whether or not to grant the loan to the consumer. The random forest technique is often used by banks to detect fraudsters.