**CLASSIFICATION ALGORITHMS**

**What is Classification?**

Classification is the process of recognizing, understanding and grouping idea and objects into pre-set categories or “subpopulations”. Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. It can be referred to as a form of “pattern recognition”, with classification algorithms applied to the training data to find the same pattern in the future sets of data.

**What is Classification Algorithm?**

Classification algorithm is a supervised learning technique used to identify the category of new observations based on training data. It is used to categorize data into a class or category and can be performed on both structured and unstructured data. There are three forms of classification algorithms:

1. Binary Classification – spam or not, true or false, yes or no, male or female
2. Multiclass Classification – classify colour: red or green, blue, yellow or brown
3. Multilabel Classification – classify the object and colour: red car, green car or blue bus

Classification algorithms can be further divided into mainly two categories:

1. **Linear Models** – a) Logistic Regression

b) Support Vector Machines (SVM)

1. **Non-Linear Models** – a) K-Nearest Neighbours

b) Kernel SVM

c) Naïve Bayes

d) Decision Tree Classification

e) Random Forest Classification

**Linear Models**

1. Logistic Regression: is a binary classification algorithm for categorical variables. There are three types:

* Binary or Binomial – a dependent variable will have only two possible types either 1 or 0. E.g., success or failure, win or lose, yes or no.
* Multinomial – a dependent variable can have three or more possible unordered types or types with no quantitative significance. E.g., type A, type B, or type C; red or blue or green.
* Ordinal – a dependent variable can have three or more possible ordered types or types with a quantitative significance. E.g., poor or good or very good or excellent and each category can have scores like 0, 1, 2, 3.
* Real life applications:
* Predicting a probability of a person having a heart attack.
* Predicting a customer’s propensity to purchase a product or halt a subscription.
* Predicting the probability of failure of a given process or product.

1. Support Vector Machines (SVM): uses algorithms to train and classify data within degrees of polarity, taking it to a degree beyond X/Y prediction. An SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

* Real life applications:
* Face recognition
* Protein-fold and remote homo-logy detection
* Hand-writing recognition
* Text and hypertext categorization
* Classification of images
* Bioinformatics
* Generalized Prediction Control (GPC)
* Geo and environmental sciences

**Non-Linear Models**

1. K-Nearest Neighbour (K-NN): Is a pattern recognition algorithm that uses training datasets to find the K closes relatives in future examples. It classifies new data points depending upon the class of most data points among the K neighbour, where K is the number of neighbours to be considered. The K-NN algorithm assumes that similar things exist in proximity. It is a lazy learning algorithm and a non-parametric method. Its is sensitive to outliers and works best on low dimension data set.

* Real life applications:
* Credit card fraud detection
* Text mining
* Agriculture
* Finance
* Facial recognition
* Recommendation systems (Amazon, Netflix, etc.)

1. Kernel SVM: Kernels are sets of mathematical functions used by SVM algorithms. The function of kernel is to take data as input and transform it into the required form. SVM Kernel functions can be linear, non-linear, polynomial, radial basis function (RBF) and sigmoid. Works on dataset having many features, provides a clear margin of separation.
2. Naïve Bayes: Calculates the possibility of whether a data point belongs within a certain category or does not. In text analysis, it can be used to a preset “tag” (classification) or not. E.g., Text “A great game” = Tag “Sports”. It works very fast and it is based on Bayes’ Theorem which gives an assumption of independence among predictors.

* Real life applications:
* Solving multi-class prediction problems
* E-mail services (spam filtering)
* Real-time predictions
* Sentiment analysis

1. Decision Tree Classification: Is a support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs and utility. The observations about an item are represented in the branches and conclusions are represented in the leaves. It is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g., whether a coin flip comes up heads or tails), each branch represents the outcome of the test, each leaf node represents a class label (decision taken after computing all attributes). There are three types of nodes: decision nodes(squares), chance nodes(circles) and end nodes (triangles). There are two main types of decision tree: Categorical variable decision tree and Continuous variable decision tree.

* Real life applications:
* Assessing prospective growth opportunities
* Using demographic data to find prospective clients
* Serving as a support tool in several fields including operations in planning logistics and strategic management, engineering, education, law, business, healthcare and finance.

1. Random forest classification: Is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. The general idea of the bagging method is that a combination of learning models increases the overall result. It is a fast, simple, flexible and predictive modelling tool.

* Real life application:
* Banking to detect fraudsters out to scam the bank
* Stock market
* E-commerce to determine if a customer will like the product or not.
* Medicine to analyse a patient’s medical history to identify diseases.