Sidath Asiri
Undergraduate at Department of Computer Science and Engineering University of Moratuwa | Readaholic | Data Science | Machine Learning | Data Mining
Dec 31, 2017

10 Machine Learning Algorithms You need to Know

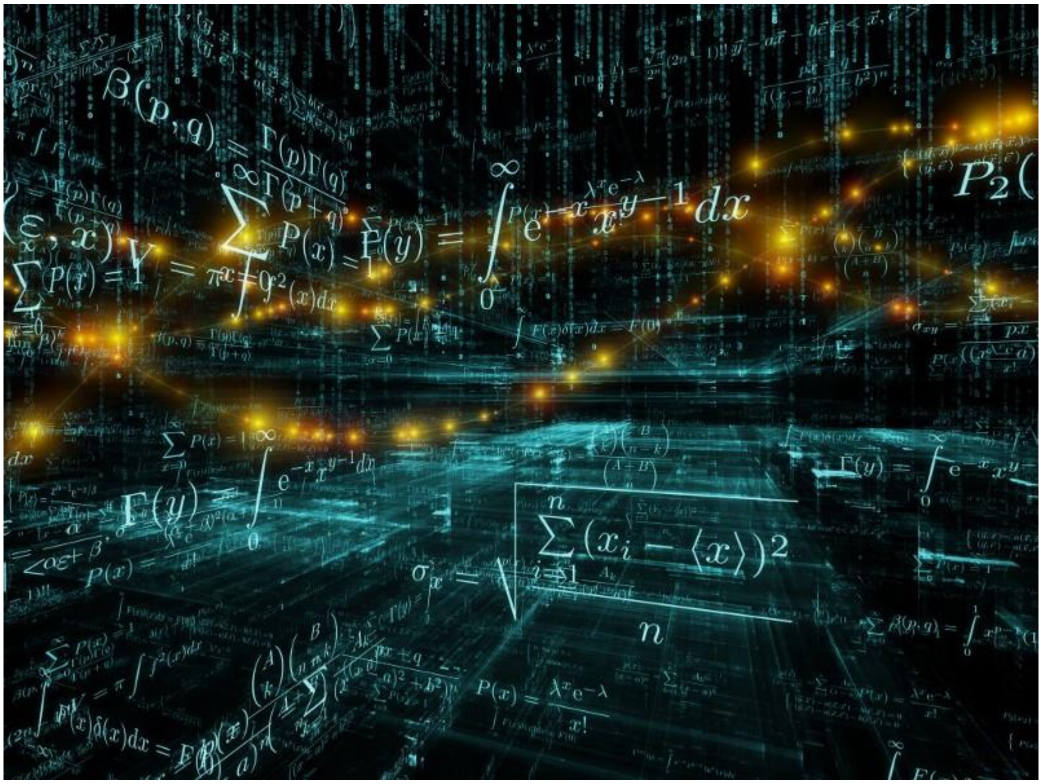


Photo: http://hpc-asia.com/wp-content/uploads/2016/02/equations.jpg

We live in a start of revolutionized era due to development of data analytics, large computing power, and cloud computing. Machine learning will definitely have a huge role there and the brains behind Machine Learning is based on algorithms. This article covers 10 most popular Machine Learning Algorithms which uses currently.

These algorithms can be categorized into 3 main categories.

1. **Supervised Algorithms:** The training data set has inputs as well as the desired output. During the training session, the model will adjust its variables to map inputs to the corresponding output.
2. **Unsupervised Algorithms:** In this category, there is not a target outcome. The algorithms will cluster the data set for different groups.
3. **Reinforcement Algorithms:** These algorithms are trained on taking decisions. Therefore based on those decisions, the algorithm will train itself based on the success/error of output. Eventually by experience algorithm will able to give good predictions.

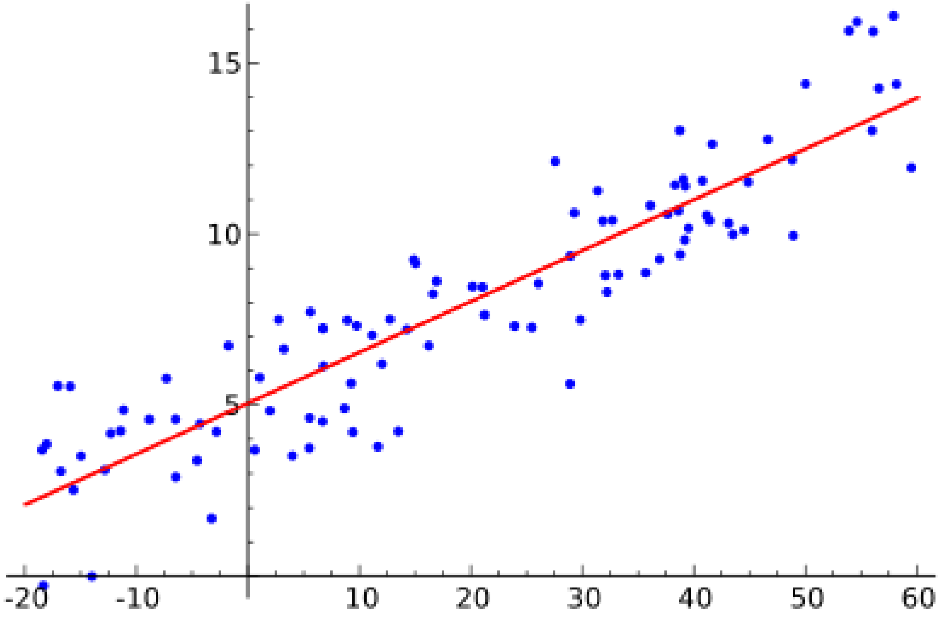
Following algorithms are going to be covered in this article.

- Linear Regression
- SVM (Support Vector Machine)
- KNN (K-Nearest Neighbors)
- Logistic Regression
- Decision Tree
- K-Means
- Random Forest
- Naive Bayes
- Dimensional Reduction Algorithms
- Gradient Boosting Algorithms

1. Linear Regression

Linear Regression algorithm will use the data points to find the best fit line to model the data. A line can be represented by the equation, **y = m*x + c**where **y** is the dependent variable and **x** is the independent variable. Basic calculus theories are applied to find the values for **m** and **c** using the given data set.

Linear Regression has 2 types as **Simple Linear Regression** where only 1 independent variable is used and **Multiple Linear Regression** where multiple independent variables are defined.

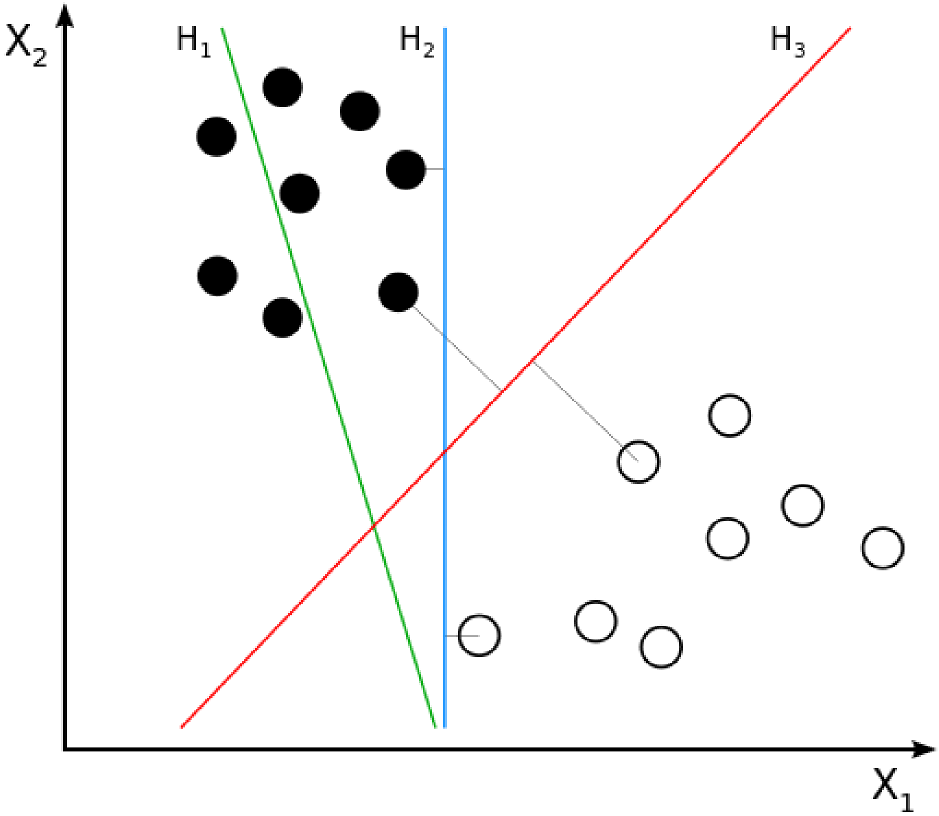


https://upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Linear_regression.svg/400px-Linear_regression.svg.png

“scikit-learn” is a simple and efficient tool using for machine learning in python. Following is an implementation of Linear Regression using scikit-learn.

2. SVM (Support Vector Machine)

This belongs to classification type algorithm. The algorithm will separate the data points using a line. This line is chosen such that it will be furthermost from the nearest data points in 2 categories.

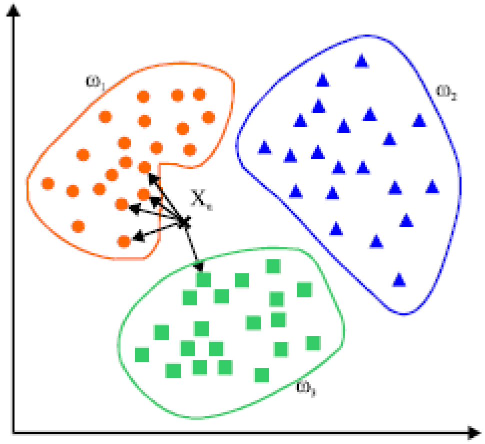


https://upload.wikimedia.org/wikipedia/commons/thumb/b/b5/Svm_separating_hyperplanes_%28SVG%29.svg/512px-Svm_separating_hyperplanes_%28SVG%29.svg.png

In above diagram red line is the best line since it has the most distance from the nearest points. Based on this line data points are classified into 2 groups.

3. KNN (K-Nearest Neighbors)

This is a simple algorithm which predicts unknown data point with its k nearest neighbors. The value of k is a critical factor here regarding the accuracy of prediction. It determines the nearest by calculating the distance using basic distance functions like Euclidean.

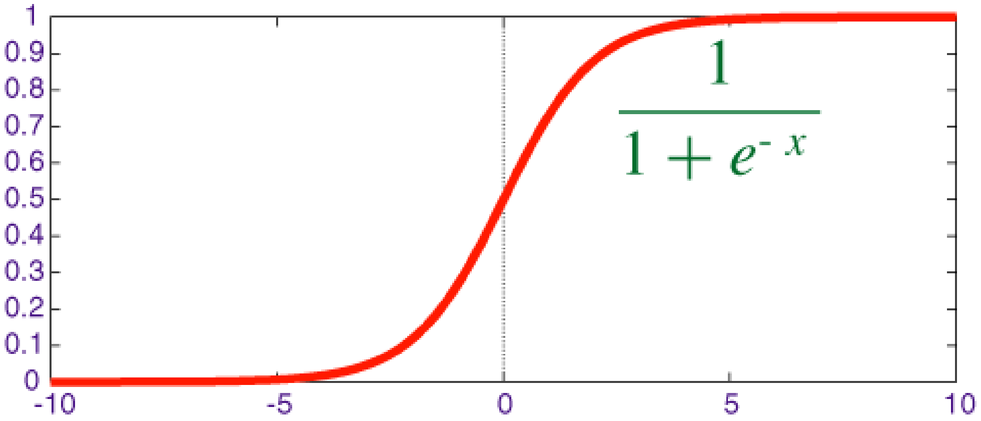


https://au.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/63621/versions/1/screenshot.gif

However, this algorithm needs high computation power and we need to normalize data initially to bring every data point to same range

4. Logistic Regression

Logistic Regression is used where a discreet output is expected such as the occurrence of some event (Ex. predict whether rain will occur or not). Usually, Logistic regression uses some function to squeeze values to a particular range.



Logistic function (https://qph.ec.quoracdn.net/main-qimg-05edc1873d0103e36064862a45566dba)

“Sigmoid” (Logistic function) is one of such function which has “S” shape curve used for binary classification. It converts values to the range of 0, 1 which interpreted as a probability of occurring some event.

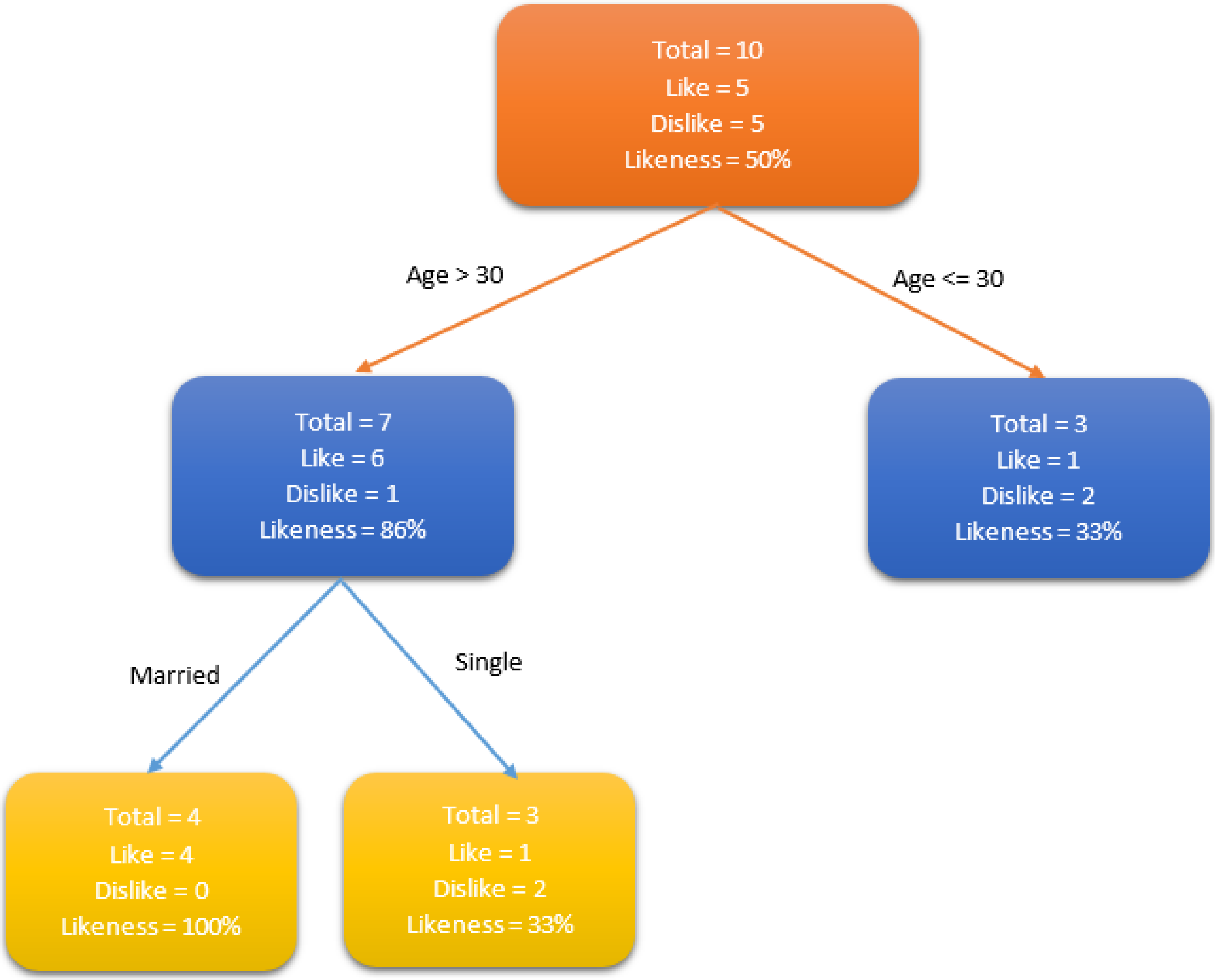
y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

Above is a simple logistic regression equation where b0, b1 are constants. While training values for these will be calculated such that the error between prediction and actual value become minimum.

5. Decision Tree

This algorithm categorizes the population for several sets based on some chosen properties (independent variables) of a population. Usually, this algorithm is used to solve classification problems. Categorization is done by using some techniques such as Gini, Chi-square, entropy etc.

Let’s consider a population of people and use decision tree algorithm to identify who like to have a credit card. For example, consider the age and marital status the properties of the population. If age>30 or a person is married, people tend to prefer credit cards much and less otherwise.



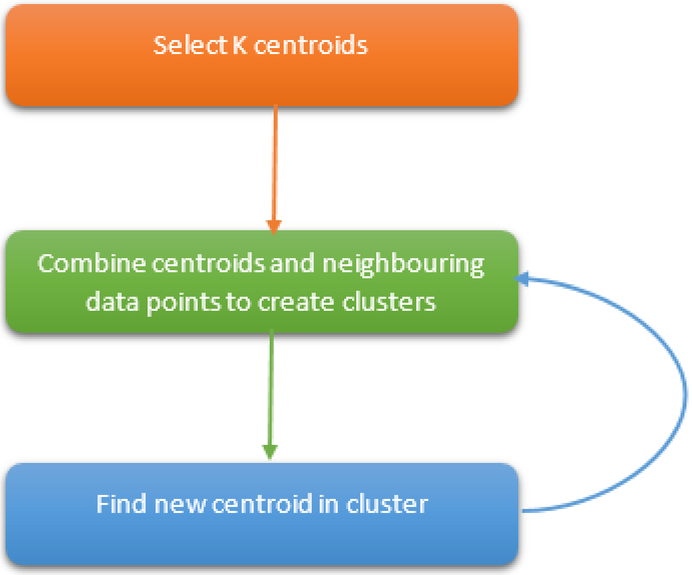
Simple Decision Tree

This decision tree can be further extended by identifying suitable properties to define more categories. In this example, if a person is married and he is over 30, they are more likely to have credit cards (100% preference). Testing data is used to generate this decision tree.

6. K-Means

This is an unsupervised algorithm which provides a solution for clustering problem. The algorithm follows a procedure to form clusters which contain homogeneous data points.

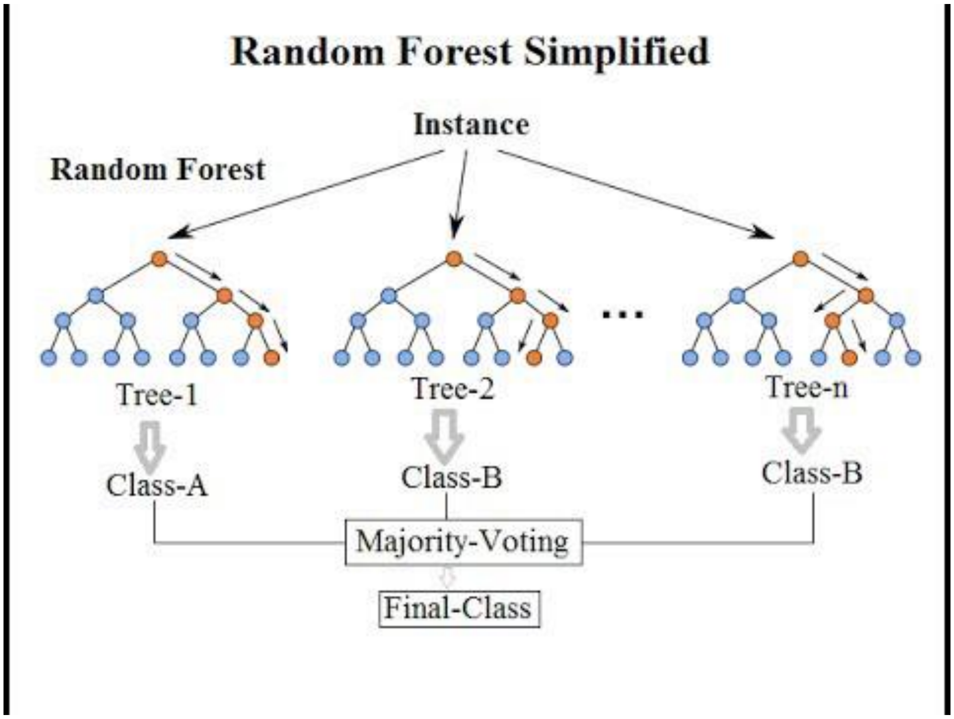
The value of k is an input for the algorithm. Based on that, algorithm selects k number of centroids. Then the neighboring data points to a centroid combines with its centroid and creates a cluster. Later a new centroid is created within each cluster. Then data points near to new centroid will combine again to expand the cluster. This process is continued until centroids do not change.


Select K centroids
Combine centroids and neighbouring data points to create clusters
Find new centroid in cluster

Cluster forming process

7. Random Forest

Random forest can be identified as a collection of decision trees as its name says. Each tree tries to estimate a classification and this is called as a “vote”. Ideally, we consider each vote from every tree and chose the most voted classification.



https://www.kdnuggets.com/wp-content/uploads/rand-forest-1.jpg

8. Naive Bayes

This algorithm is based on the “Bayes’ Theorem” in probability. Due to that Naive Bayes can be applied only if the features are independent of each other since it is a requirement in Bayes’ Theorem. If we try to predict a flower type by its petal length and width, we can use Naive Bayes approach since both those features are independent.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Bayes Equation

Naive Bayes algorithm also falls into classification type. This algorithm is mostly used when many classes exist in the problem.

9. Dimensional Reduction Algorithms

Some datasets may contain many variables that may cause very hard to handle. Especially nowadays data collecting in systems occur at very detailed level due to the existence of more than enough resources. In such cases, the data sets may contain thousands of variables and most of them can be unnecessary as well.

In this case, it is almost impossible to identify the variables which have the most impact on our prediction. Dimensional Reduction Algorithms are used in this kind of situations. It utilizes other algorithms like Random Forest, Decision Tree to identify the most important variables.

10. Gradient Boosting Algorithms

Gradient Boosting Algorithm uses multiple weak algorithms to create a more powerful accurate algorithm. Instead of using a single estimator, having multiple will create a more stable and robust algorithm.

There are several Gradient Boosting Algorithms.

- XGBoost—uses liner and tree algorithms
- LightGBM—uses only tree-based algorithms

The specialty of Gradient Boosting Algorithms is their higher accuracy. Further, algorithms like LightGBM has incredible high performance as well.

Thanks for reading.