Study on Machine Learning Algorithms

B.Sandhiya[#1], R.P.S.Manikandan[*2], G.Anitha[#3,] V.Prasath kumar[#4]
[1]Ass. Prof., Department of IT, Sri Shakthi Institute of engineering and technology,Coimbatore,India
[2]Asso. Prof, Department of IT, Sri Shakthi Institute of engineering and technology,Coimbatore,India
[3]Ass. Prof., Department of IT, Sri Shakthi Institute of engineering and technology,Coimbatore,India
[4]Ass. Prof., Department of IT, Sri Shakthi Institute of engineering and technology,Coimbatore,India

Abstract

Machine learning is the field evolved from Artificial Intelligence, goal is to mimic intelligent abilities of human by machines. Here come this paper gives the clear idea about classification algorithm. Classification is a supervised learning where the computer programs learn from the data given to it and the classify the data based on the observation. Classification algorithms are KNN, Decision tree, Random Tree, Support vector machine, Logistic Regression. Classification can be performed on both structured and unstructured data. The goal of classification problem is to find the category to which the new data fall. Data will falls under classification only when the desired output is discrete. Two types of classification, namely Binary classification and Multi label classification.Some examples are speech recognition, handwriting recognition, bio metric identification, document classification etc. In this paper, a novel learning about various Classification algorithms with example and types of classification

Keywords - *Classification, SVM, Binary classification, Multi label classification.*

I. INTRODUCTION

Machine learning (ML), a subfield of artificial intelligence, has evolved out of the need to teach computers how to automatically learn a solution to a problem. Machine learning focuses on development of computer programs to access data and use it, learn for themselves. The process begins with data or observations examples like direct experience or instructions ,in order to look for patterns in data and make better decisions in future with the data that we provide. The ultimate aim is to allow computers to learn automatically with any human intervention and perform actions accordingly.

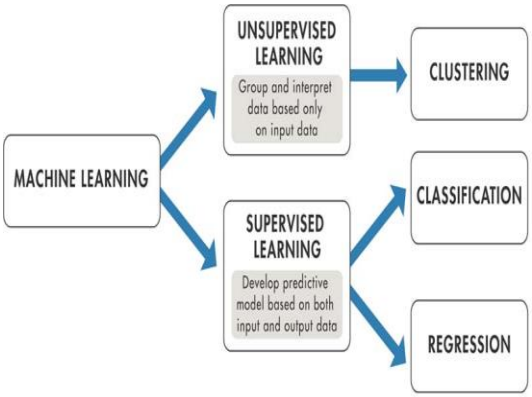
Supervised learning, where one can have input variable(x) and outputvariable(y), with the use of an algorithm to learn the mapping function from the input and output.

$$Y=f(x)$$

The goal is to approximate mapping function so well with the new dataset to predict the output variables.Supervised learning can be grouped into Classification and Regression.A Classification problem is when the output variable is a category like red,blue,green.

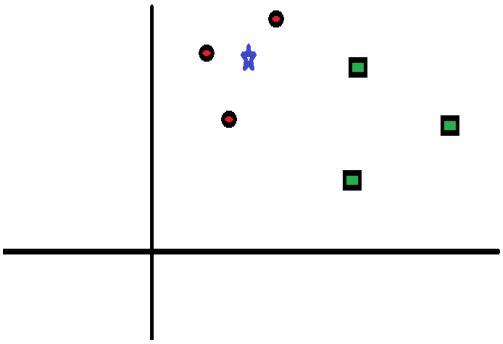
Regression problem is when the output variable is a real value, such as dollars or weight.

Unsupervised learning is where one can have only the input variables(x) and no output variables. Algorithms are left to their own devises to discover and present their interesting structure in the data. Unsupervised learning can be grouped into 2 clustering and Association.



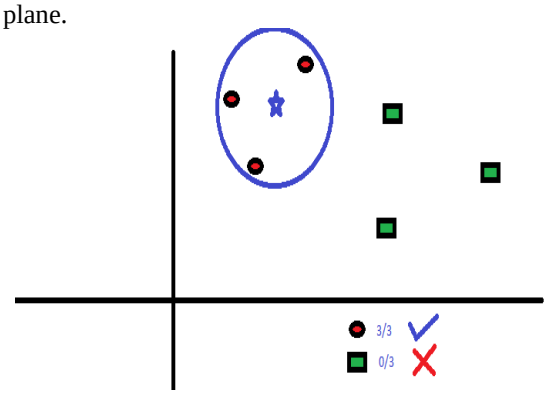
A. Knn algorithm

Knn algorithm is non-parametric method used for classification.It contains closest training examples in the feature space.Let us consider an example to understand this example,Consider a spread a red circles and green squares.



To find out the class of Blue star(BS).BS can either be RC or GS and nothing else. The "K" is KNN algorithm is the nearest neighbors we wish to take vote from. Let's say K

= 3. Hence, we will now make a circle with BS as center just as big as to enclose only three data points on the plane.



The closest point to BS is RC.From this we can predict that BS belongs to the class RC. The choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm. So to conclude Knn Algorithm,following factores should be considered for the best K.

The distance is calculated from the Euclidean distance

$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

B. Decision Tree

Decision Trees will comes under the type of Supervised Machine Learning where the data is continuously split according to a specific elements. The tree can be named as,decision nodes and leaves. The leaves are the final outcomes. And the decision nodes are the splited data

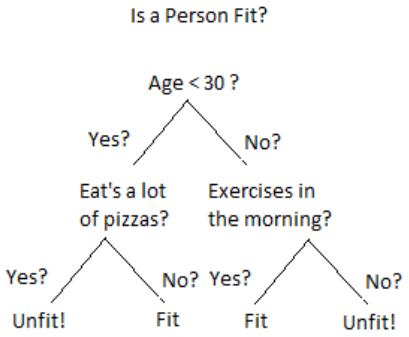


Fig 1.Decision Tree

The above binary tree explains decision tree. With the given information like age, eating habit, and physical activity, etc. to predict whether a person is fit. The decision nodes here are questions like 'What's the age?', 'Does he exercise?', 'Does he eat a lot of pizzas'? And the leaves, which are outcomes like either 'fit', or 'unfit', so it will comes under Binary classification problem.

Entropy: To measure the uncertainty or randomness in data

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

Example, consider the probability of coin toss where the probability of heads is 0.5 and tails is 05.since it has no way of predicting the outcome, the entropy is highest possible. By choice, consider a coin with both sides as head, so the entropy for such an event can be predicted perfectly since we know it always be heads. This event has no randomness, its entropy is zero.

Information gain: Information gain is denoted by IG(S,A) for a set S is the effective change in entropy after deciding on a particular attribute A. It measures the relative change in entropy with respect to the independent variables.

$$IG(S, A) = H(S) - H(S, A)$$

Alternatively,

$$IG(S, A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

Where IG(S, A)is an information gain y applying feature A.

H(S)-Entropy of the entire set.

P(x)*H(x)-calculates the entropy after applying the feature of A.

P(x)-Probability of X.

C. Random Forest

It is a supervised learning algorithm. As the name suggests it forms forest with number of trees. In random forest algorithm,the process of finding information gain or gini index for calculating the root node, the process of finding the root node and splitting the feature nodes will happen randomly.Pseudo code for Random forest is:

1. Randomly select "**K**" features from total "**m**" features where **k << m**
2. Among the "**K**" features, calculate the node "**d**" using the best split point
3. Split the node into **daughter nodes** using the **best split**
4. Repeat the **a to c** steps until "l" number of nodes has been reached
5. Build forest by repeating steps **a to d** for "n" number times to create **"n" number of trees.**

Its main application is banking,medicine,e commerce.The main advantage of random forest algorithm is,it avoids overfitting problem.It is applicable to both regression and classification problems.

Bias variance trade off:There are two sources of error in machine learning, *Bias* is an error that occurs when an algorithm makes too many simplifying assumptions, causing it to predict values that differ from the actual values.

Advantages: A similar Random forest algorithm or the random forest classifier can use for both arrangement and the relapse undertaking.

- Random forest classifier will deal with the missing qualities.
- When we have more trees in the foret, arbitrary forest classifier won't overfit the model.
- Can show the random forest classifier for all out qualities too.

Basic understanding of Random forest algorithm with real time example:

Assume Mady somehow got 2 weeks leave from his office. He needs to put in his 2 weeks by venturing out to the better place. He likewise needs to go to the place he may like.

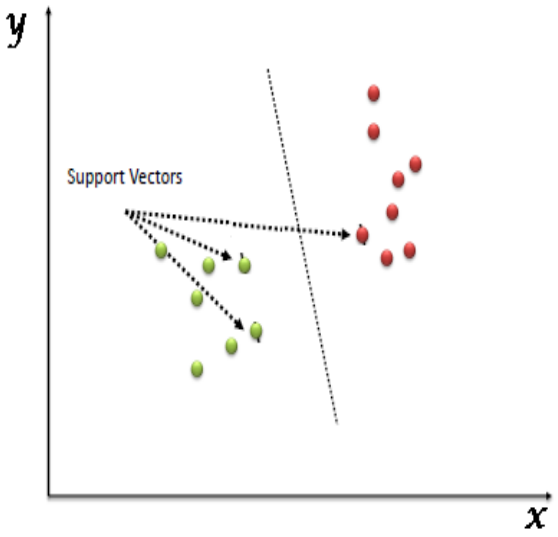
So he chose to get some information about the spots he may like. At that point his companion begun getting some information about his past outings. It's much the same as his closest companion will ask, You have been visited the X put did you like it?

In view of the appropriate responses which are given by Mady, his best begin suggesting the place Mady may like. Here his best shaped the decision tree with the appropriate response given by Mady. As his closest companion may prescribe his best place to Mady as a companion. The model will be biased with the closeness of their kinship. So he chose to request that couple of more companions prescribe the best place he may like. Presently his companions made some irregular inquiries and every one prescribed one place to Mady. Presently Mady considered the place which is high votes from his companions as the last place to visit.

In the above Mady trip arranging, two primary intriguing calculations decision tree algorithm and random tree algorithm utilized. I trust you discover it as of now. At any rate, I might want to feature it once more.

4. Support vector machine(SVM)

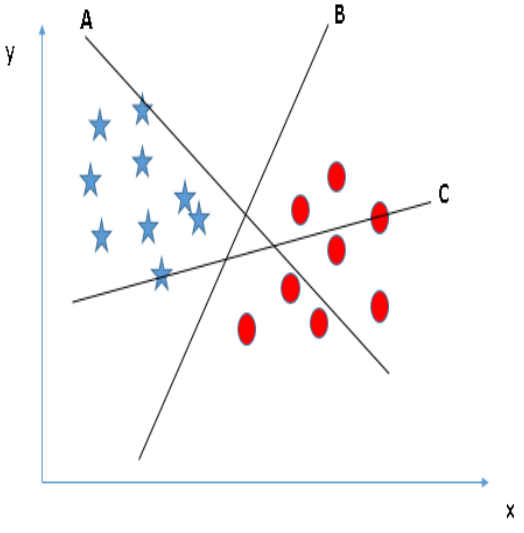
Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be utilized for both classification or regression challenges. Notwithstanding, it is generally utilized in classification issues. In this algorithm, we plot every datum item as a point in n-dimensional space (where n is number of features you have) with the estimation of each component being the estimation of a specific arrange. At that point, we perform grouping by finding the hyper-plane that separate the two classes extremely well.



Support Vectors are basically the co-ordinates of individual perception. Bolster Vector Machine is a boondocks which best isolates the two classes (hyper-plane/line).

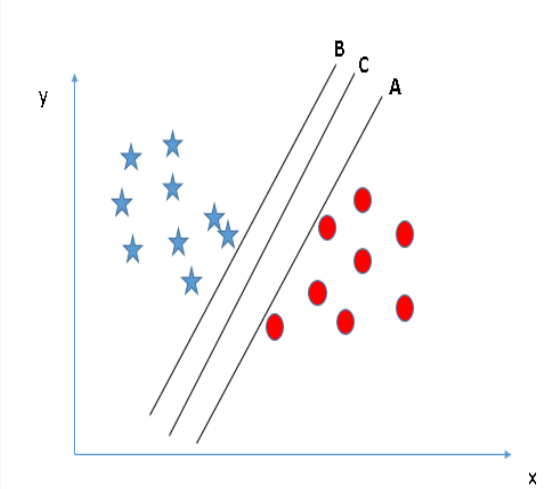
How does it work?

Identifying the right hyper plane(scenario 1): Here, we have three hyper-planes (A, B and C). Presently, distinguish the privilege hyper-plane to characterize star and circle.

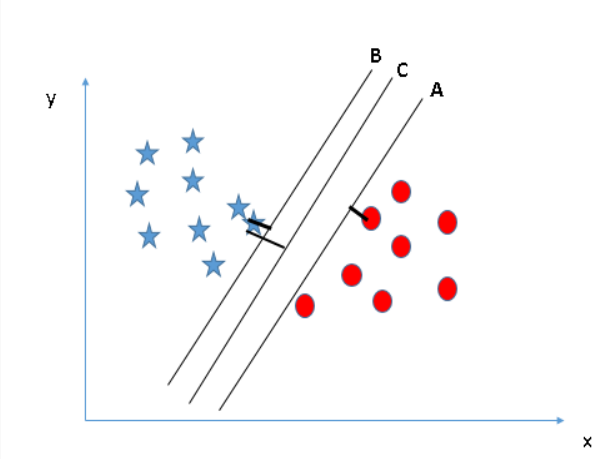


Thumb rule to identify the right hyper plane, "Select the hyper-plane which isolates the two classes better". In this situation, hyper-plane "B" has astoundingly played out this activity.

Identifting the right Hyper plane(Scenario 2): Here, we have three hyper-planes (A, B and C) and all are isolating the classes well. Presently, How would we be able to distinguish the privilege hyper-plane?

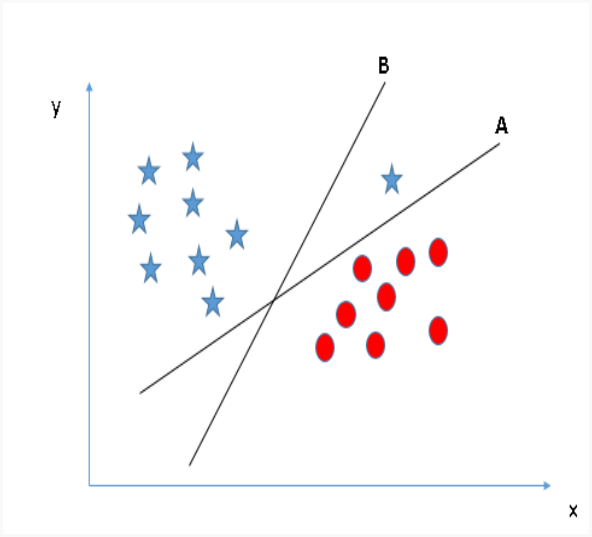


Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**. Let's look at the below snapshot:



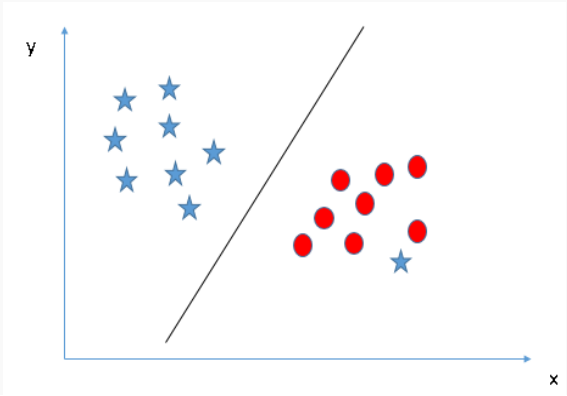
Above, you can see that the edge for hyper-plane C is high when contrasted with both An and B. Consequently, we name the correct hyper-plane as C. Another lightning purpose behind choosing the hyper-plane with higher edge is robustness. In the event that we select a hyper-plane having low edge, there is high possibility of miss-classification.

Identifying the right hyper-plane(scenario 3):HINT:By using the rules defined in previous section identify the hyper plane



Some of you may have chosen the hyper-plane B as it has higher edge contrasted with A. Be that as it may, here is the trick, SVM chooses the hyper-plane which arranges the classes precisely preceding amplifying edge. Here, hyper-plane B has a grouping mistake and A has arranged all accurately. In this way, the privilege hyper-plane is A.

***Can we classify two classes*:** Two classes utilizing a straight line, as one of star lies in the region of other(circle) class as an outlier.



one star at opposite end resembles an exception for star class. SVM has a component to overlook exceptions and discover the hyper-plane that has most extreme edge. Subsequently, we can state, SVM is strong to outlier.

CONCLUSION

In this Paper we have given the explanation of about Knn algorithm,Decision tree,Random forest,Support vector machine algorithms of machine learning with some explanations and graph.This would helps the beginners to learn basics about algorithms.

REFERENCES

[1] Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers . In Proceedings of the Fifth Annual Workshop on Computational Learning Theory.
[2] www.analyticsvidhya.com
[3] V.Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995
[4] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. "A Practical Guide to Support Vector Classification" . Deptt of Computer Sci.National Taiwan Uni, Taipei, 106, Taiwan http://www.csie.ntu.edu.tw/~cjlin 2007
[5] RSES 2.2 User's Guide Warsaw University http://logic.mimuw.edu.pl/»rses ,January 19, 2005
[6] http://logic.mimuw.edu.pl/»rses