A Comparative Study of Classification Algorithms: Social Network Ads

Ashwin Ramesh P  
*Department of Computer Science and Engineering Sapthagiri College of Engineering, Bengaluru-57, Karnataka, India*Ashwini S  
*Department of Computer Science and Engineering Sapthagiri College of Engineering, Bengaluru-57, Karnataka, India*

***Abstract*— The Social media advertising is a market that did not exist just one and a half decade ago, but the rapid increase in popularity of social networking platforms is projected to generate a $37 billion in revenue by 2020.Social advertisements offer the opportunity to build awareness, target based on geographic, demographic and behavioural data. Also, many of the ads can be contextually placed adjacent to relevant topics. Many of the platforms also offer remarketing opportunities for visitors who abandoned the site or shopping cart and returned to the social network. This paper analyses the categorical dataset from kaggle to determine whether a user purchased a particular product from a Social Network ad.**

Keywords—………………………………//\*…………………

# Introduction

Classification is a supervised machine learning approach to a datasets, where we categorize the data into a desired and distinct number of classes where we can assign

labels to each class. There are numerous applications for classification in multiple domains like medical diagnosis, email spam prediction, credit card approval, etc.

There are two types of learners in classification as lazy learners and eager learners.

Lazy learners-They simply store the training data and wait for a testing data to appear. When it does, the classification is conducted based on the most related data within the stored training data. Compared to eager learners, lazy learners have less training time, but take more time in predicting.Ex. k-nearest neighbour.

Eager learners-They construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and lesser time to predict.Ex. Decision Tree.

There are a lot of classifications and algorithms available now, but it is not possible to conclude which one is superior to other. It depends on the application and nature of the data set on which it is applied. Hence, we are performing this comparative study of selected algorithms on the Social Network Ad dataset.

# K-NEAREST NEIGHBOUR ALGORITHM (k-nn)

"K closest neighbors"is a basic calculation that stores every accessible case and orders new cases dependent on a likeness measure (e.g., separation capacities). KNN has been utilized in factual estimation and example acknowledgment as of now in the start of 1970's as a non-parametric method.

Algorithm**—**

1.Load the data set.

2.Initialize K to your picked number of neighbors.

3. For every model in the dataset,

3.1 Calculate the separation between the inquiry model and the present model from the dataset.

3.2 Add the separation and the list of the guide to an arranged accumulation

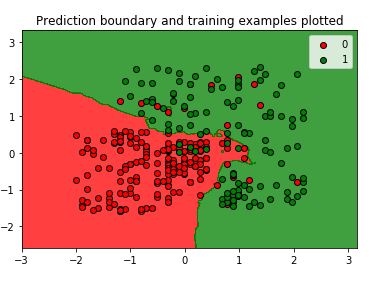
4. Sort the arranged accumulation of separations and files from littlest to biggest (in climbing request) by the separations

5. Pick the primary K sections from the arranged gathering

6. Get the marks of the chose K sections

7. In the event that relapse, return the mean of the K marks

Graph—



# support vector machine (SVM)

"Support Vector Machine" (SVM) is a directed Machine Learning calculation which can be utilized for both arrangement or relapse difficulties. In any case, it is for the most part utilized in arrangement issues. In this calculation, we plot every data item thing as a point in n-dimensional space (where n is number of highlights you have) with the estimation of each element being the estimation of a specific organize. At that point, we perform order by finding the hyper-plane that separate the two classes very well.

Algorithm**—**

1.start

2.initialize the parameter c,€,ƴ randomly.

3. preparing the svm utilizing traning dataset with the instated parameters.

4.test the created svm model utilizing approval data.

5. assess wellness work .

6.compare item wellness assessment with item’s individual best .

7.compare new item terms and update the position of the item as need to be .

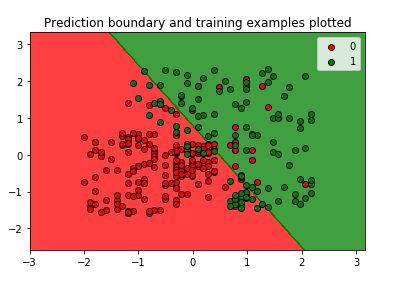
8.stopping rule met?

If no goto step 3 .

9. If yes ideal C,€,ƴ

10.Stop.

Graph—



# 4.NAIVE BAYES ALGORITHM

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Algorithm—

1.Read the training dataset T:

2.calculate the mean and standard deviation of the

dataset

3.repeat

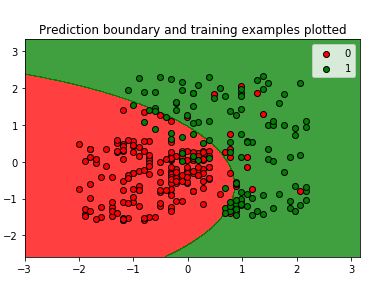
Figure the likelihood of fi utilizing the guass density equation in each class:

Until the proability of all indicator factors (f1….f2… .f3… … .fn) has been determined.

4.calculate the probability foreach class:

5.get the best probability.

Graph—



5.DECISION TREE ALGORITHM

Decision Tree algorithm has a place with the group of administered learning calculations. In contrast to other directed learning calculations, decision tree algorithm can be utilized for taking care of relapse and grouping issues as well. The general intention of utilizing Decision Tree is to make a preparation model which can use to foresee class or estimation of objective factors by taking in choice guidelines derived from earlier data(training dataset).

Algorithm—

Input:

T//decision tree

D//Input dataset

Output:

M//Model Prediction

DTProc Alogrithm:

/represent forecast procedure utilizing DT

For every t ɛ D do

n=root hub of T;

while n not leaf hub do

get answer to address on n connected t:

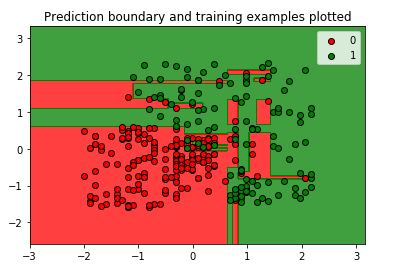
recognize circular segment from t which contains

right answer:

n=node at part of the circular segment:

make expectation for t dependent on marking of n;

Graph—



6.RANDOM FOREST ALGORITHM

Random forest algorithm is a regulated arrangement calculation. As the name recommend, this calculation makes the woodland with various trees. By and large, the more trees in the forest the more vigorous the forest resembles. Similarly in the random forest algorithm, the higher the quantity of trees in the forest gives the high exactness results.

Algorithm—

1. Select haphazardly M highlights from the list of capabilities.

2. For each x in M

a.calculate the information gain

gain(t,x)=E(t)-E(t,x)

E(t)=

E(t,x)=

Where E(t) is the entryopy of the two classes . E(t.x) is the entryopy of feature x.

b.select the hub d which has the most elevated data gain

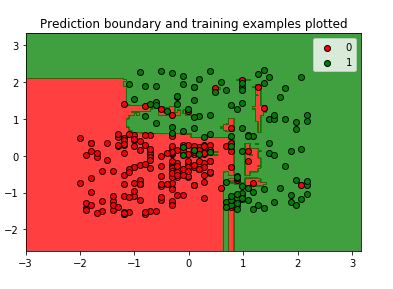
c. parts the hub into sub-hubs

d.repeat steps a,b and c to contruct the tree until arrive at least number of Samples required to part

3. repeat step 1and2 for N times to assemble forest of N

trees.

Graph—

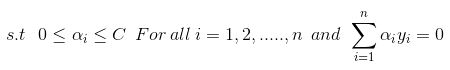


6.KERNEL SVM ALGORITHM

Going to the real piece of the SVM for which it is most

Well known,the bit stunt.The portion is a method for processing the spot result of two vectors x and y in a few (extremely high dimensional) include space, which is the reason piece capacities are at times called "summed up dab item.

Capture16.JPG



Algorithm—

1.start

2.input the dataset

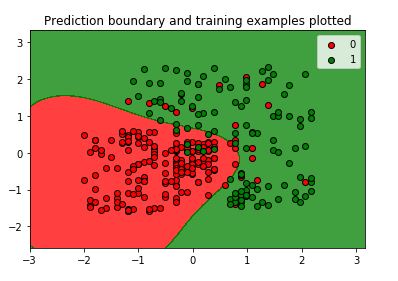
3.group the dataset

4.imply the SVM machine learning with kernel function ie,radial based kernel function(RBF)

5.indicate the hyper-plane

6. whenever got legitimacy and precision isn't worthy,then goto step 4

7.end

Graph— 

## 

|  |  |  |
| --- | --- | --- |
| Classification algorithm | Accuracy | Percentage |
| KNN | 0.93 | 93% |
| SVM | 0.90 | 90% |
| Kernel SVM | 0.93 | 93% |
| Naive Bayes | 0.90 | 90% |
| Decision Tree | 0.91 | 91% |
| Random Forest | 0.92 | 92% |

Conclusion—

In this paper we propose the novel classification approach which is based on classification algorithms. Different classification algorithms were evaluated on the real-world Social network ads and compared with each other. The results shows the KNN and Kernel SVM approach outperform other classification algorithms in terms of both the overall accuracy and performance. Also the visual interpretations of the results generated by different classification algorithms are considered.