2451-18-733-166

V. Raja Rajeshwari

ASSIGNMENT-II

DSR .

1. What is Predictive modeling? Discuss about evaluation of Predictive models?

Ans Predictive modelling is the process of using known results to create, process and validate a model that can be used to forecast future outcome using statistics

=) It is a tool used in predictive Analytics, a data mining technique that is concerned with forecasting probabilities and trends.

=) examples of specific types of forecasting that benefit business and demand forecasting

headcount planning

clean Analysis.

competition Analysis

finacial risks

Types of predictive modeling

* Regression

* clustering

* Neural networks

* classification
* Time series model
* forecasting

2. What is linear regrisson & list out the critical assumptions of linear regression?

Ans Linear regression is the supervised machine learning model in which the model finds the bestfit linear line between the independent and dependent variables

Critical Assumptions

We can use to understand the relationship b/w twovariables x and y

$y = ax + b$ ①

① Linear relationship

There exists a linear relationship b/w the independent variable x and the dependent variable y.

② Independence

③ Homoscedasticity

④ Normality.

③ Why logistic regression is used for the classification Explain model building stratgies for logistic regression

Ans logistic regression is used as a classification technique. It uses logistic function to model the dependent vasiable

=> The dependent variable is dischotomours in nature

=> As a result, this technique is used while dealing with binary data

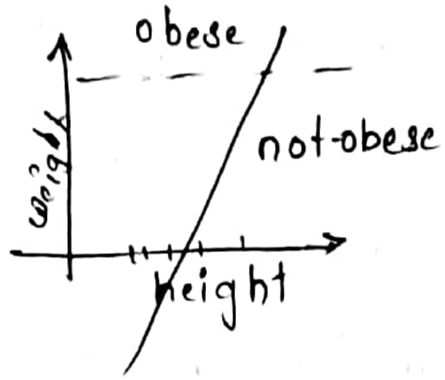
Eg:- customers choosen
spam email /website

=> purposeful solution of variables includes the following steps:

1) univariable analysis
2) multivariable model comparison
3) Linearity Assumption
4) Interactions Among covariate.

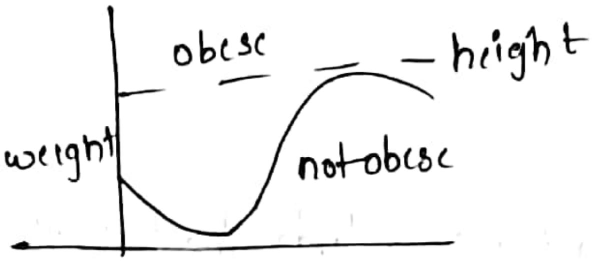
4) ~~E~~ Write in detail about linear and logistic regression?

Ans Linear regression is used to predict the continuous dependent variable using a given set of Independent variables



logistic Regression is used topredict the categofical dependent variables using a given set of independent variables

$$\log\left(\frac{y}{1-y}\right) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$



5) List out the various control structures supported by R programming language?

Ans control statements are expressions used to control the execution and flow of the program based on the condition provided in the sentement

In R programming

* If-condition
* If-else
* for-loop
* nested-loop

6. Breitly describe the datastructures in R-programming language?

Ans A datastructure is a particular way of organizing data in a computer so that it can be used effectividely

=) The most essential datastructure used in 'R' includes:

- Vectors
- Lists
- Dataframes
- Matrics
- Arrays
- Factors

* Vectors

A vector is ordered collection of basic type of given length.

```
x=c(1,3,5,7,8)
  print(x)
olp: [1] 1 3 5 7 8
```

Lists

A list is a generic object consisting of an ordered collection of objects

```
cid = c(1,2,3,4)
```

* while loop
* repeat and break statement
* Return statement
* next statment
* If-condition

Syntax

```
  if(expression)
   {
   statements;
   }
```

* for loop

Syntax

```
 for (value in vectors)
  {
  statements
  }
```

* nested loop

```
 m<-matrix(2:15,2)
   for(r in seq(nrow(m)))
    {
    for(c in seq(ncol(m)))
      {
      print(m[r,c])
    }}
```

```
[1] 2      [1] 8
[1] 4      [1] 10
[1] 6      [1] 12
```

7. Define Object. List themethods for measuring Distance b/w objects?

Ans **objects**

object are the instance of the class. Also everything and to known more look at datatypes in R.

→ method for measuring distance b/w objects

Euclideandistance

The most common distance is euclidean distance

Edist(x,y) ← sqrt ((x[1]-y[1])^2 + (x[2]-y[2]]^2 + ...).

* **Manhattandistance**

Manhattan distance measure distance in the no. of horizontal and vertical units it takes to get from onepoint to the other

mdist ← sum (abs(x[1]-y[1]) + abs((x[2]-y[2] + ...)

* **Cosine Similarility**

cosine similarility is a common similarity metric in text Analysis

dot (x,y) ← sum(x[1]*y[1]+x[2]*y[2]+ ...)

cosin(x,y) ← dot(x,y) / sqr(dot (x,x) or dot (y.y))

$$\cos(\theta) = \frac{A \cdot B}{|A||B|}$$

Dataframes

Dataframes are generic data objects of R which are used to store tabular form

=) Dataframes are the foremost popular data objects in R programming

=) Each item in a single column must be of same datatype

```
Name = c("sweety", "Madhu", "Ramani")
lang = c("R", "python", "c++)
df = data.frame(Name, lang)
print(df)
Name    lang
sweety   R
Madhu   python
Ramani    c++.
```

Matrices

A matrix is a rectangular arrangement of number in rows and columns

```
A = matrix(c(1,2,3,4,5,6,7,8,9))
   nrow=3, ncol=3, byrow=TRUE)
print(A)
         [,1]  [,2]  [,3]
[1,]      1     2     3
[2,]      4     5     6
[3,]      7     8     9
```

8 . Define list and datafiame in R and explain various opaations on lists and datafames with suitable example.

Ans A list is generated using list() function.

Geating lists

```
Geek-list<-list("Geek", TRUE 127)
print(Geek-list)
```

* Naming the elements of a list

Accessing elements of a list l

Adding, deleting and updating elements of a list

Meging elements of a list t

converting a list to vector

Datafaames

Datafaames are generic data objects of R which are used to stoge the tabular data

creating a datafame using vector

Accessing rows and columns

selecting of the subset of the dataframe

Editing datafames

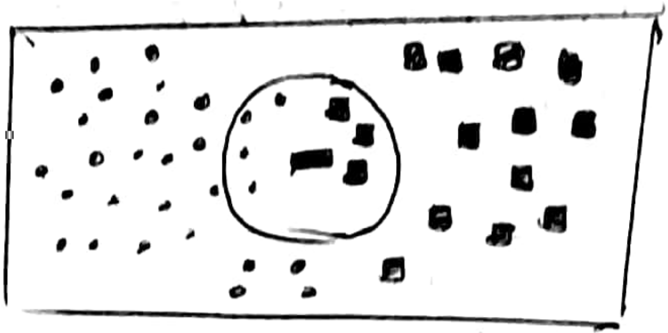
5 Explain k-Nearest Neighbours Algorithm and its implementation in R programming language?

Ans K-Nearest Neighbors(KNN)
KNN can be defined as a supervised machine learning algorithm that can classify new data point into target class based on neighbouring data points features

For example: consider a KNN algorithm for a machine that differentiates between apples and mangoes. To perform this a dataset of apples and mangoes are must be provided as input and then model must be trained in such a way that it detects that fruit based on certain characteristics.

For examples, features of an apple would be red in colour, round in shape etc similarly features of mango would be yellow in colour, oval in shape etc An image is provided to the model as input measures suchas Eucledian and mahantann distance are used to classify the closeness b/w data points When k=3 the neduest neighbours are 2 squares and 1 one dot When k=3 It will assigned to classB

But ^ k=7 the neighbours would be 4 dots and 3 squares



When it is classified the new data point will be assigned to class A.

Implementation of KNN in Rprogramming language.

consider a problem to analyze the bank data and then build a machine learning model that can predict whether loan for applicant can be approved or rejected depending on its socio-economic profile The bank credit dataset ~~consists~~ includes the information about 1000's of applicant's such as name, age, account balance, loan records etc Based upon this data, prediction can be made that whether loan can be approved or rejected. This problem can be solved by KNN algorithm by classifying loan apprequest into two classes

1. Approved
2. Rejected.

steps to solve this problem as follows -

step1: Importing the dataset -

Import DATA the dataset of applicants -

loan ← read.csv("c:/usars/desktop/DATASET/knn/credits-csv")

The structure of data would be

```
str(loan)
'dataframe': 1000obs. of 21variables:-
$ credibility: int 1 1 1 1 1 1 1 1 1 - -
$ Account Balance: int 1 2 1 1 1 1 4 2
$ Duration of credits month: int 18 9 12 12 10 8 6 18 24
$ purpose int 2 0 9 0 0 0 1 2 3
$ credit amount int 1049 2799 841 2122 2171
$ values savings stocks : int 1 1 2 1 1 1 1 3 - -
$ Length of current employment: int 2 3 4 3 3 2 4 2 1 1
$ instalment percent : int 4 2 2 3 4 1 1 2 4 1
$ Marial status : int 4 2 3 2 3 2 2 1 1 1
$ Guarantors int 1 1 1 1 1 1 1 1 - -
$ Duration in current Address : int 4 2 4 2 4 3 4 4 4 - -
$ Age years : int 21 36 23 39 38 40
$ No of dependices : int 1 2 1 1 2 1 2 1 1
$ Telephone : int 1 1 1 1 1 - - -
```

Step: Data cleaning

The structure of dataset consists of 21 predictable variables that are used to decide whether loan for applicant can be approved or rejected.

Some of the variables might not be useful in predicting the loan. For example. ~~credit~~ concurrent credits telephones etc. can be removed Because such type of variable leads to complexity of machine learning model

loan.subset ← loan[c('Credibility', 'Age.years', 'Account balance')]

The dataset now would be

str(loan.subset)

'dataframe': 1000 obs of 8 variables

$ credibility int 1 1 1 1 1 1

$ age.years int 29 30 49 . .

int 2 3 11 1 1

$ Marital status int 3929 4232 - -

$ occupation int 1 2 2 1 1 1 4 2 1

$ Account balance int 2 3 4 3 3 2 4 4 1

$ length of current employment int 20 9 0 0 0 0 3 3 - -

$ purpose

Step3: Data Normalization.

Normalization of data set is mandatory to make the output unbiased consider below observations

head(loan.subset)

	Creditbility	Age.years	Maritalstatus	Occupation	Account Balance
2	1	49	4	1	1
3	1	63	3	1	2
4	1	78	2	2	4
5	1	58	1	1	3
6	1	15	1	1	2

Credit Amount	Length of employments	Purpose
2943	2	2
5678	3	0
9629	4	9
1423	3	0
2428	3	6
1928	2	0
		6

Here credit amount variable has value scale in 1000's and the remaining are in single or two digitits. If the data is not normalized outcome is in baised form

```
normalize <- function(x)
 {
 return ( x- min(x)))/(max(x) -min(x))
 }
```

Now store the normalized data set in loan subset and then remove crediability variable because it is

Explain K-Means Algorithm and its implementation in R programming

K-means

The K-means is an iterative clustering algorithm in which objects are moved ~~over~~ among set of, clusters until desired set is achieved. It is most popular and commonly used method. the algorithm is based on the concept of most specified input parameter (K). A set of 'n' objects are divided into 'k' clusters. A high degree of similarity among ~~objects~~. elements in clusters is obtained.

Working of K-means Algorithm

Algorithm's input

the number of desired clusters are denoted by 'k'

A dataset containing 'n' objects denoted by 'D'.

Algorithm's output

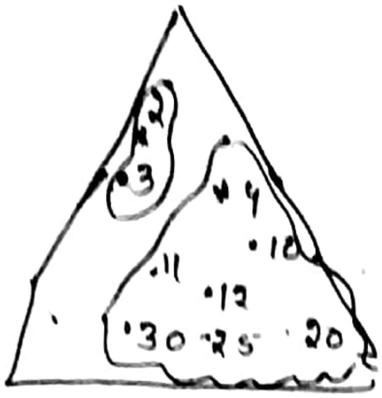
K: A set consisting of 'k' clusters

Procedure:

Step1: Initially select 'k' objects randomly from D, as initial cluster centers

Step2: Depending upon the distance between the object and the cluster mean, each remaining object assinged to cluster to which it is most similar.

Euclidean distance b/w mean and objects to classify objects into two clusters



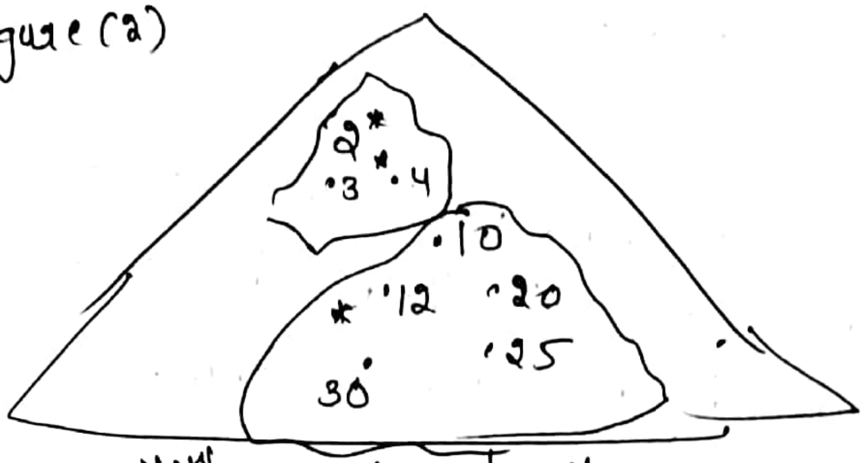
Fig(1) initial partitioning with $m_1 = 2$ and $m_2 = 4$

Now evaluate new mean for the resulting clusters and again partition is done base of euclidean distance as follows

$$m_1 = \frac{2+3}{2} = 2.5$$

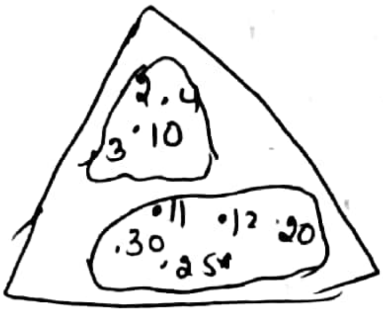
$$m_2 = \frac{4+10+11+12+20+25+30}{7} = 16$$

$m_1 = 2.5$ and $m_2 = 16$ and resulting cluster is shown in figure (2)

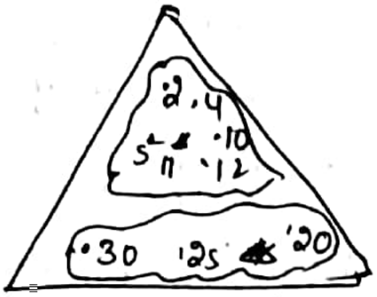


Fig(2) New clusters formed with $m_1 = 2.5$ and $m_2 = 16$.

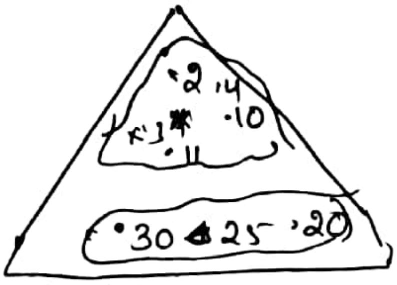
The process iterates for the successive values of the mean to enhance the partition of clusters



clustas = m1=3 m2=18



clustas with m1=4.75 and m2=19.5



clusters with m1=7 and m2=25

Thus finally the correct parition of the numered objects into two clustas is achieved.

K-Means Implementation in R

The K-means algorithm can be implemented by predefining. Here K represents numbu of clusters to be defined. Steps to implement K-means are as followls.

1. In the first step k centeres are defined and every clusta is assigned to the cluster that has closest centre to it.

2. In the second step the centres are redefined by using the observation of each clusta. The column means are used for defining the centroid

The above two steps are repeated until the centres are converge therefore K-means algorithm said to be iteative.

List out the various performance metrics for classification?

Performance Measures

Performance of a classification model is evaluated once the predictive model is build. It determines whether the model can predict the outcome of new observation test data. that is not used in the training the model. The performance of a model can be accessed by comparing the predicted outcom values against known outcome values.

The commonly used metrics and methods to access the performance of predictive classification models are as follows.

1. Average classification Accuracy

It represents the porportion of correctly classified observation. the complete classification rate corresponds to the part of observations that are classified correctly. The first step of accessing the performance of a model is to determine the raw classification accuracy.

The raw classification error rate can be inversely defined as porportion of observation that are misclassified.

error rate = 1-accuracy.

The raw classification accuracy and error can be determined by comparing the observed classes in test data and predicted classes by model

```
>accuray <- mean(observedclass == predictedclass)
>accuray
[1] 0.08
>error <- mean(observed class != = predictedclass)
>error
[1] 0.192
```

The binary classifiea in this example makes two types of error

=> It can wrongly ~~class~~ assign invidiudal who has diabetes positive to diabetes negative

=> It can wrongly assign individual who has diabetics negative to diabetics positive

2. confusion matrix

The main factor required for performance evaluation of ~~mat~~ classification models includes correct or incorrect predictions of no.of testcases by the model. These test records are given in a tabular form known as confusion matrix.

Actual class	Predicted class	
	classY	classX
classy	fyy	fyx
classX	fxy	fxx

The above specified table outlines the binary classificaton problem and each individual entry is given as fxy Here x represents no of records belongsto class x andy represents no of records belongs to classy ·

fxx·

This entry specifies the no·of records belonging to class'x' There is no mis-prediction in this entry ,since the records of class'x' are predicted as records from class 'x'

fxy·

This entry specifies the no·of records belonging to class'x' There is misprediction in this entry since the records of class are incorrectly predict as records from class'y'.

fyy

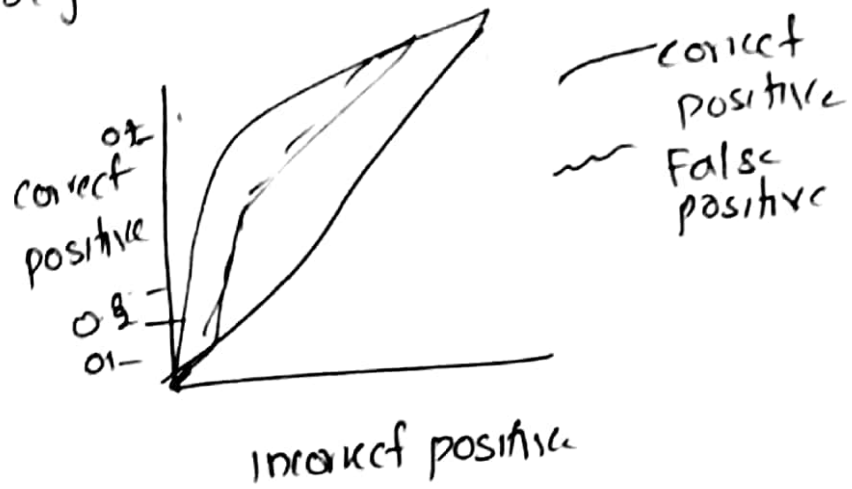
This entry specifies the no·of records belonging to class'y'. There is no misprediction in this entry

4) ROC curve

The graphical representation that exhibits the performance of predicted classification model is known as receiver operating characteritics curve.

Applications of ROC curve

ROC curves depicts the tradeoff b/w porportion of correctly determined positive tuples known as true positive rate and the porportion of negative tuples incorrectly tuples as positive tuples known as negative rate for given model


correct positive
incorrect positive
correct positive
False positive