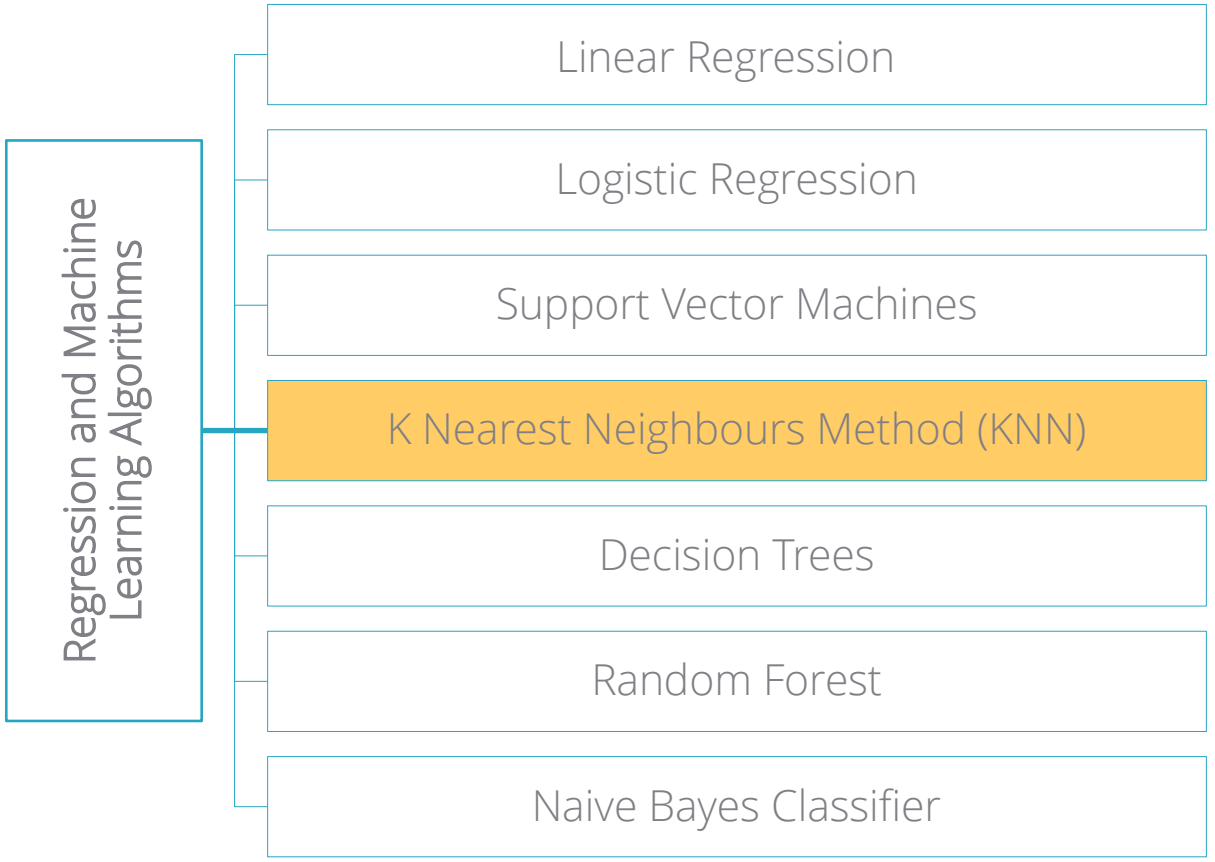
K Nearest Neighbors (KNN) Method
ML ALGORITHM

Machine Learning Methods

Machine Learning is a branch of data science based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

MACHINE LEARNING METHODS

There is lot of overlapping between statistical modeling and machine learning. The Regression Models are used extensively in ML applications.



DATA SCIENCE INSTITUTE

Classical Statistical Models vs. Machine Learning Methods

- There is lot of overlapping between statistical modeling and machine learning. The Regression Models are used extensively in ML applications.

- Statistical Model has model equation, set of regression coefficients and hypothesis testing is performed to test the significance of coefficients.

- Machine Learning Algorithms do not follow above approach but end output is comparable- Either predicted probabilities and predicted values

Introduction

- KNN is a a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.

- In this method, K neighbors of a new case are obtained from training data. The classification is performed using 'majority vote' concept.

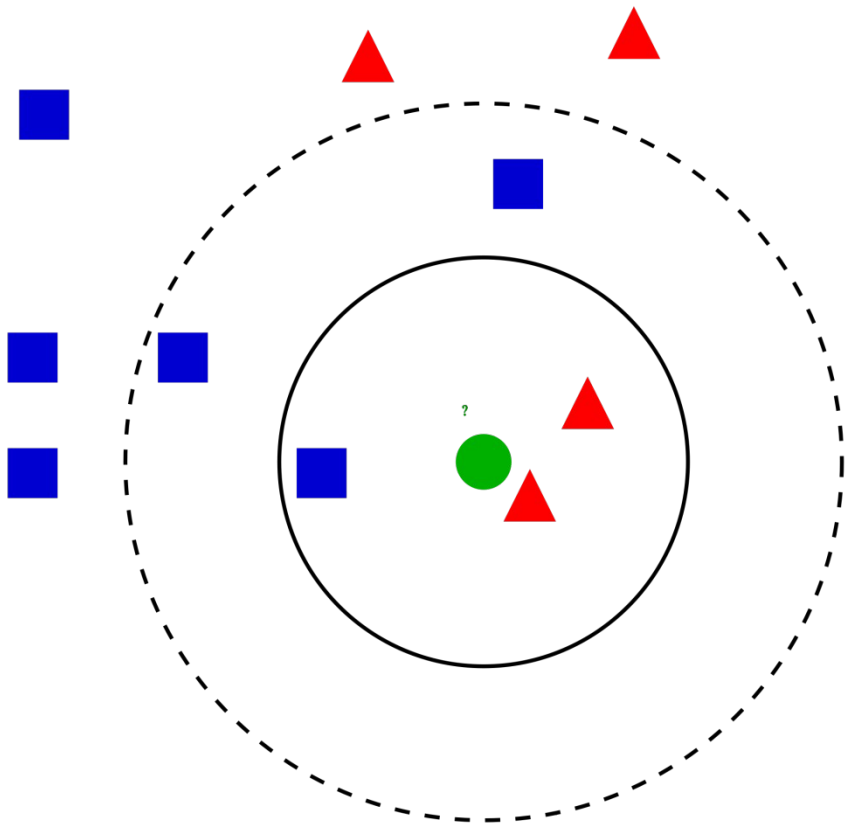
  There is no model equation like logistic regression.

- KNN is conceptually simple, yet able to solve complex problems . The *k*-NN algorithm is among the simplest of all machine learning algorithms.

- KNN is one of the lazy learning algorithm.

- Memory and CPU cost is high.

KNN Classification



There are 2 types of objects:
Y=0

Y=1

What is the “class” of new object (green circle)?

Analyze neighbors
Class of new object

KNN Classification – Distance Concept to Find Neighbor

Age	Current Debt	Default	Distance
25	40,000	0	102000
35	60,000	0	82000
45	80,000	0	62000
20	20,000	0	122000
35	120,000	**0**	22000
52	18,000	0	124000
23	95,000	1	47000
40	62,000	1	80000
60	100,000	**1**	42000
48	220,000	1	78000
33	150,000	**1**	8000
48	142,000	?	

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

"Blue case is new case. Red cases are 3 neighbours of new case based on Distances. Two out of 3 are defaulters. So new case is predicted as defaulter"

DATA SCIENCE INSTITUTE

KNN Classification
Distance Based On Standardized Variables

Age	Current Debt	Default	Distance
0.125	0.11	0	0.7652
0.375	0.21	0	0.5200
0.625	0.31	0	0.3160
0	0.01	0	0.9245
0.375	0.50	0	0.3428
0.8	0.00	0	0.6220
0.075	0.38	1	0.6669
0.5	0.22	1	0.4437
1	0.41	1	0.3650
0.7	1.00	1	0.3861
0.325	0.65	1	0.3771
	0.7		

$$X_s = \frac{X - Min}{Max - Min}$$

Can also use usual (X-mean)/SD
Note that now out of 3 nearest neighbors 2 are non defaulters.

Classification Using KNN Method

- If K=1 then the case is classified using nearest neighbor.
- Usually K is greater than 1. The case is classified to most frequent class of K neighbors. A simple approach to select k is set K=sqrt(n) where n is number of observations in training data.
- KNN can be used to solve regression problem. The estimated value for the case is average of K neighbors.
- Imagine data of millions of customers. Memory and CPU cost is high since distances are to be stored and sorted.

Snapshot of the Data

AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	DEFAULTER
3	17	12	9.30	11.36	5.01	1
1	10	6	17.30	1.36	4.00	0
2	15	14	5.50	0.86	2.17	0
3	15	14	2.90	2.66	0.82	0
1	2	0	17.30	1.79	3.06	1
3	5	5	10.20	0.39	2.16	0
2	20	9	30.60	3.83	16.67	0
3	12	11	3.60	0.13	1.24	0
1	3	4	24.40	1.36	3.28	1
2	0	13	19.70	2.78	2.15	0
1	0	1	1.70	0.18	0.09	0
1	4	0	5.20	0.25	0.94	0

Age groups: 1 (<28 years), 2(28-40 years), 3 (>40 years)

Training and Testing Data Sets

```
#Import csv data BankloanKNN
bankloan<-read.csv(file.choose(),header=T)
library(caret)
index<-createDataPartition(bankloan$SN,p=0.7,list=FALSE)

#Remove variables AGE,SN and DEFAULTER from the data.
#Note that KNN syntax in R requires data without dependent variable.
bankloan2<-subset(bankloan,select=c(-AGE,-SN,-DEFAULTER))
bankloan3<-scale(bankloan2)

traindata<-bankloan3[ind
testdata<-bankloan3[-ind

#create class vectors
Ytrain<-bankloan$DEFAULTER[index]
Ytest<-bankloan$DEFAULTER[-index]
```

Now we have training and testing data sets ready. In addition, we need vector of dependent variable for training and testing data.

DATA SCIENCE INSTITUTE

Running KNN Algorithm in R

```
library(class)
Ytest_pred<-knn(traindata,testdata,k=17,cl=Ytrain) #model stores
    classification
table(Ytest,Ytest_pred)
```

	Ytest_pred	
Ytest	0	1
0	52	11
1	16	37

Knn function requires training data set, testing data set and class vector of training data set.
K=17 specifies number of neighbors.

For each observation in the test data,17 neighbors are obtained. The classification of the observation is based on majority votes.

Sensitivity=37/53=0.70
Specificity=52/63=0.83

Note: Can use confusionMatrix function in caret
confusionMatrix(as.factor(Ytest_pred),as.factor(Ytest),positive ="1")

KNN Method: Majority Voting

- A drawback of the basic "majority voting" classification occurs when the class distribution is skewed. (Imbalanced data problem)

- That is, examples of a more frequent class tend to dominate the prediction of the new example, because they tend to be common among the *k* nearest neighbors due to their large number.

- One option to remove this drawback is to create training data with equal class frequency. This is possible only if data is sufficiently large.

k-Nearest Neighbour Regression

- KNN algorithm can also be used for regression problems.
- The estimated value for the case is average of K neighbors.
- Use knnreg function in caret package to perform k-nearest neighbor regression.

THANK YOU!!

DATA SCIENCE
INSTITUTE