Classification

Binary | Multi-Class | Ordinal

Types of Logistic Regression:-

- Binomial - Binomial LR, there can be only two possible types of dependent variable.
- Multinomial LR - There can be three or more possible unordered types of the dependent variable.
- Ordinal LR - There can be three or more possible ordered types of dependent variable. E.g.- movie rating, [low, medium & high].

Sigmoid function (Logistic func)

The value of logistic Regression must be b/w 0 and 1, which cannot go beyond this limit.

DECISION TREE

Entropy & Information Gain -

Entropy can be thought of as how much variance the data has. For example, a dataset has only blue balls, the entropy is low or zero whereas if a dataset has different coloured balls, the entropy will be high.

The formula for calculating entropy is given as:

$$E = -\sum_{i}^{c} p_i \log_2 p_i$$

Now, that we have calculated entropy for both sections, we can determine the quality of split by weighing the entropy of each

branch by how many elements it has.
Since, there are 4 pts in left partition, the weight will be 0.4 and 6 pts in right partition, the weight will be 0.6.

$$E_{split} = W_{left} E_{left\ split} + W_{right} E_{right\ split}$$

Information Gain is equals to how much entropy, we have removed.
Higher information gain is equals to more entropy removed.

Entropy can be defined as how unpredictable the dataset is and information gain is calculated for ~~gain~~ a split by subtracting the weighted entropies of each branch from the original entropy.
When training a decision tree using these metrics, the best split is chosen by maximumizing information gain.

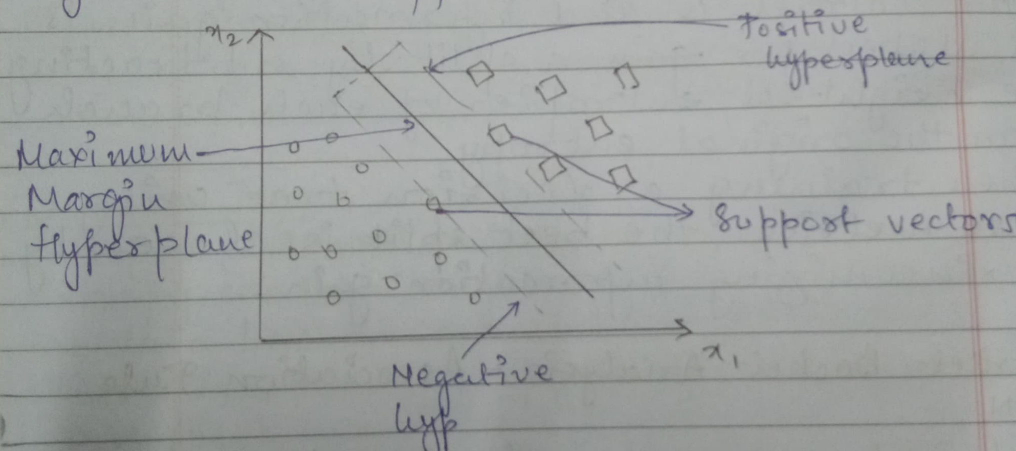
→ Market Basket Analysis - Association Rule
→ ANN
→ Biological neurons has three main components:-
1) Dendrites
2) Soma
3) Axon

Dendrites is an input receiving neuron which receives signal

Soma sums the incoming signals when sufficient input is received.
Axon transfers the signal to other cell

SUPPORT VECTOR MACHINE (SVM)

SVM is a popular supervised classification as well as regression algorithm. Now, the goal of SVM is to create the best line or decision boundary that can segregate two classes which can put the data in correct category class. SVM chooses the extreme point that help in creating hyperplane. The extreme cases are called as support vectors and hence, the algorithm is support vector machine.



There are two types of SVM-: Linear and Non-Linear SVM
Linear SVM is used linearly separable data which means the dataset which can be clearly classified into two classes by using a straight line

Non-linear SVM is used for non-linearly seperable data, which means dataset cannot be classified into two classes by a straight line.

UNSUPERVISED MACHINE LEARNING ALGORITHM

The unsupervised ML algorithm can be classified into two types :- Clustering and Association.

Clustering refers to formation of groups and Association refers to the different items or product associated with each other.

CLUSTERING

There are three categories of clustering algorithm :-

1) Partitioning based clusters
2) Hierarchical clustering
3) Density-based clustering

Partitioning-based clustering divides the data into non-overlapping groups

Hierarchical clustering - The cluster assignment is done by using a hierarchy.

- Agglomerative - Bottom-up approach
- Divisive - Top down approach

Density-based clustering - The cluster assignment is depended upon density of data points. E.g. - DBSCAN (Density Based Special Clustering Algorithm)

Partitioning Based Clustering -

K-Means Clustering

1) Specify the no. of k clusters to assign.
2) Randomly initialize k centroid
3) Repeat
4) Expectation - Assign each point to its closest centroid.
5) Maximization - Compute the new centroid (mean) of each cluster
6) Until, these centroid positions do not change.

KNN

1) Determine the value of k = nearest neighbour
2) Calculate the distance between query instance and training data.
3) Sort the distance and determine nearest neighbour based on k^{th} minimum distance.
4) Collect the category of nearest neighbours and
5) Use majority voting for prediction.

Acid durability	Strength	Classification	distance
x_1	x_2		$(x1)^2 + (x2)^2$
7	7	Bad ✓	$(7-3)^2 + (7-7)^2 = 16$ (3)
7	4	Bad	$(7-3)^2 + (4-7)^2 = 25$ (4)
3	4	Good ✓	$(3-3)^2 + (4-7)^2 = 9$ (1)
1	4	Good ✓	$(1-3)^2 + (4-7)^2 = 13$ (2)

Query - $(x1 = 3, x2 = 7)$, Class^n = ?

$k = 3$

We are going to select 1, 2 & 3 record.
Prediction ⇒ Good

Hierachical Clustering

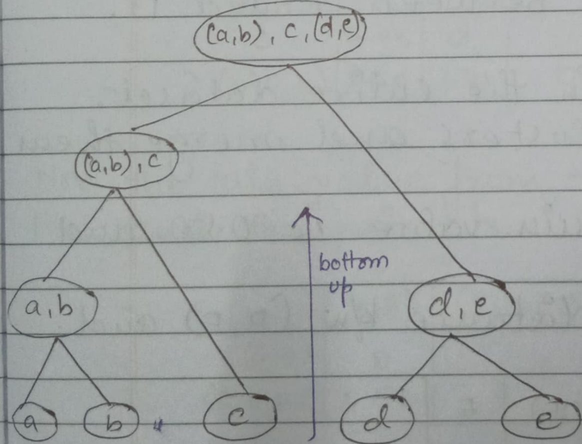
There are two types of hierarchical based clustering :-

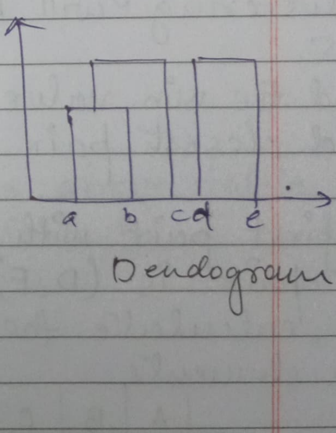
1) Agglomerative - follows bottom up approach
2) Divisive - follows Top down approach

Agglomerative ⟶ AGNES
Agglomerative Nesting

Divisive ⟶ DIANA
Divisive Analysis

Bottom-up approach:





Dendogram

Q for a given set of values, plot a dendogram using the concept of Agglomerative Clustering Technique (i) Use single linkage function (Min)
(ii) Complete linkage (Max)
(iii) Average linkage (Average)
(iv) Centroid (Center)

	A	B	C	D	E	F
A	0					
B	0.71	0				
C	5.66	4.95	0			
D	3.61	2.92	2.24	0		
E	4.24	3.54	1.41	1.00	0	
F	3.20	2.50	2.50	0.50	1.12	0

(i) Single linkage → Minimum distance
The clustering will be between F and D.
Steps:
(i) Find the min. value in the entire dataset.
(ii) Find closest pair clusters and merge them

The first pair with min. value is 0.50 and the pair is (D, F).
Now, calculate the distance b/w (D, F) and other elements.

	A	B	C	(D,F)	E
A	0				
B	0.71	0			
C	5.66	4.95	0		
D,F	3.20	2.50	2.24	0	
E	4.24	3.54	1.41	1.00	0

Calculate the Distance b/w (D, F) and A by using ~~Eucle~~ single linkage, i.e. minimum distance

$\min[D_{DA}, D_{FA}]$

Now, the min. value from the above table is 0.71 and the pair is (A, B).

	(A, B)	C	(D, F)	E
(A, B)	0			
C	4.95	0		
(D, F)	2.50	2.24	0	
E	3.59	1.41	1.00	0

Now, the min. value from the above table is 1.00 with pair (D, F, E).

	(A, B)	C	(D, F, E)
(A, B)	0		
C	4.95	0	
(D, F, E)	2.50	1.41	0

Now, the min. value from the above table is 1.41 and the pair is (C, D, E, F)

	(A, B)	(C, D, E, F)
(A, B)	0	
(C, D, E, F)	2.50	0

Q For a given dataset, partition the data into two clusters ~~for~~ by using Euclidian distance problem

	h	w
1	185	72
2	190	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	188	67
12	177	76

$N = 12$

$C = 2$

$D = ED$

$ED = \sqrt{(x_h - h_1)^2 + (x_w - W_1)^2}$

h_1 ↓ centroid value

x_h ↓ Observed value

$C_1 = \{1, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

$C_2 = \{2, 3\}$

For 3,

$ED = \sqrt{(168-185)^2 + (60-72)^2}$ from C1

$= \sqrt{289 + 144} = 20.81$

$ED = \sqrt{(190-168)^2 + (60-56)^2}$ from C2

$= \sqrt{484 + 16} = \cancel{2}2 \cdot 36$

$= 4.4$

For 4,

Centroid will be (185, 72) for C1 and (169, 58)

$ED = \sqrt{(179-185)^2 + (68-72)^2}$ w.r.t C1

$= \sqrt{36 + 16} = 7.21$

$ED = \sqrt{(179-169)^2 + (68-58)^2}$ w.r.t C2

$= 14.14$

for 5,
Centroid is (182, 70) for C1 and (169, 58) for C2.

$ED = \sqrt{(182-182)^2 + (72-70)^2}$ w.r.t C1

$= 2$

$ED = \sqrt{(182-169)^2 + (72-58)^2}$ w.r.t C2

$= \sqrt{169 + 196} = 19.10$

For 6,
Centroid is (182, 70.67) for C1 and (169, 58) for C2

$ED = \sqrt{(188-182)^2 + (77-70.67)^2}$ w.r.t C1

$= \sqrt{36 + 40.07} = 8.721$

$ED = \sqrt{(188-169)^2 + (77-58)^2}$ w.r.t C2

$= \sqrt{361 + 361} = 26.87$

for 7,
Centroid is (183.5, 72.25) for C1 and (169, 58) for C2.

$ED = \sqrt{(180-183.5)^2 + (71-72.25)^2}$ w.r.t C1

$= \sqrt{12.25 + 1.5625} = 3.716$

$ED = \sqrt{(180-169)^2 + (\cancel{58}\ 71-58)^2}$ w.r.t C2

$= \sqrt{121 + 169} = 17.029$

Q find clusters,

d	x	y
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5
6	4.5	5
7	3.5	4.5

$N = 7$
$C = 2$

$C1 = \{1\}$

$C2 = \{2, 3, 4, 5, 6, 7\}$

for 3,

$ED = \sqrt{(3-1)^2 + (4-1)^2}$ w.r.t C1

$= \sqrt{4+9} = 3.6$

$ED = \sqrt{(3-1.5)^2 + (4-2)^2}$ w.r.t C2

$= \sqrt{2.25 + 4} = 2.5$

for 4,
centroid is (1, 1) for C1 and (2.25, 3) for C2.

$ED = \sqrt{(5-1)^2 + (7-1)^2}$ w.r.t C1

$= \sqrt{16 + 36} = 7.2$

$ED = \sqrt{(5-2.25)^2 + (7-3)^2}$ w.r.t. C2

$=$

$= \sqrt{7.5625 + 16} = 4.85$

for 5,
Centroid is (1, 1) for C1 and (3.625, 5) for C2.

$ED_1 = \sqrt{(3.5-1)^2 + (5-1)^2}$ w.r.t C1

$= \sqrt{6.25 + 16} = 4.72$

$ED_2 = \sqrt{(3.5-3.625)^2 + (5-5)^2}$

$= 0.1.25$

for 6,
Centroid is (1, 1) for C1 and (3.5625, 5) for C2

$ED = \sqrt{(4.5-1)^2 + (5-1)^2}$ w.r.t C1

$= \sqrt{12.25 + 16} = 5.31$

$ED = \sqrt{(4.5-3.5625)^2 + (5-5)^2}$ w.r.t C2

$= 0.9375$

Date:
Page:

for 7;
Centroid is (1, 1) for C1 and (4.03125, 5) for C2.

$$ED = \sqrt{(3.5-1)^2 + (4.5-1)^2} \quad \text{w.r.t } C1$$

$$= \sqrt{6.25 + 12.25} = 4.3$$

$$ED = \sqrt{(3.5-4.03125)^2 + (4.5-5)^2} \quad \text{w.r.t } C2$$

$$= \sqrt{0.5322} = 0.729$$

DBSCAN (Density Based Clustering)
DBSCAN stands for density based spatial clustering for applications with noise.
DBSCAN stands for Density Based Spatial Clustering for Applications with Noise.

DBSCAN requires two technical parameter:-
1. EPS 2. Minimal Point

EPS - It defines the neighbourhood around a data point. If the distance between two points is lower or equal to EPS, they are considered neighbours. If EPS is too small, then large part of the data will be outliers.

Minimal Point - Minimal no. of neighbours within EPS radius is minimal point.

EPS < 1.5 } Core pt.
Min pt = 3
Noise pt

Q For a given dataset, identify the core, boundary and noise points.

	A	B	C	D	E	F
A	0	0.7	5.7	3.6	4.2	3.2
B	0.7	0	4.9	2.9	3.5	2.5
C	5.7	4.9	0	2.2	1.4	2.5
D	3.6	2.9	2.2	0	1	0.5
E	4.2	3.5	1.4	1	0	1.1
F	3.2	2.5	2.5	0.5	1.1	0

For F, 0.5, 1.1 and 0 lies less than 1.5 and is 3 in number so,
F is one of the core pt.
likewise E is also a core pt and D is a core pt. as well.

C is a boundary point because there are 2 minimal pt that has EPS < 1.5
likewise B is also a boundary pt. and A is a boundary pt as well.

APRIORI ASSOCIATION ALGORITHM

Q for the following dataset, generate rules using apriori algorithm. Consider support = 50% and confidence = 75% respectively.

Transaction ID	Items Purchased
1	Bread, cheese, Egg, Juice
2	Bread, cheese, Juice
3	Bread, Milk, Yogurt

$$Support(A) = \frac{\text{No. of transactions in which A occurs}}{\text{Total no. of transactions}}$$

4	Bread, Juice, Milk	
5	Cheese, Juice, Milk	

Sol

Step 1: Create a table of frequent items purchased

Items	Frequency	Support
Bread	4	4/5 = 80 %
Cheese	3	3/5 = 60 %
Egg	1	1/5 = 20 %
Juice	4	4/5 = 80 %
Milk	3	3/5 = 60 %
Yogurt	1	1/5 = 20 %

Choose items that has support greater than 50 % or equal to.

Choosen items = Bread, Cheese, Juice, Milk

Step 2: Make two items candidate set and write down their frequency.

Items Pair	Freq	Support	
(Bread, Cheese)	2	2/5 = 40 %	
(Bread, Milk)	2	2/5 = 40%	
(Bread, Juice)	3	3/5 = 60 %	✓
(Cheese, Milk)	1	1/5 = 20 %	
(Cheese, Juice)	3	3/5 = 60%	✓
(Milk, Juice)	2	2/5 = 40 %	✗

Confidence (A, B) = Support of A∪B with reference to A

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

$B \rightarrow J$, $J \rightarrow B$, $C \rightarrow J$, $J \rightarrow C$

$$\text{Confidence}(B \rightarrow J) = \frac{3}{5} \times \frac{5}{4} = \frac{3}{4}$$
$$= 75\%$$

$$\text{Confidence}(J \rightarrow B) = \frac{3}{5} \times \frac{5}{4} = \frac{3}{4}$$
$$= 75\%$$

$$\text{Confidence}(C \rightarrow J) = \frac{3}{5} \times \frac{5}{3} = 100\%$$

$$\text{Confidence}(J \rightarrow C) = \frac{3}{5} \times \frac{5}{4} = 75\%$$

Since, all four combination values are greater than 70% which means all association rules are good

Q Support = 22%, Confidence = 70%

Transaction ID	Items Purchased
1	I_1, I_2, I_5
2	I_2, I_4
3	I_2, I_3
4	I_1, I_2, I_4
5	I_1, I_3
6	I_2, I_3
7	I_1, I_3
8	I_1, I_2, I_3, I_5
9	I_1, I_2, I_3

PROBABILISTIC LEARNING

NAIVE BAYES ALGORITHM

- It is a supervised machine learning algorithm
- based on Bayesian theorem.
- Used for classification.
  E.g. - text classification, spam filtering, Sentiment Analysis
- It is a probabilistic classifier, i.e., it predicts on the basis of probability.

Naive Bayes theorem is comprised of two keywords - Naive & Bayes.
Naive Bayes assumes the occurrence of certain features is independent of the occurrence of other features. For example, fruit classification
Bayes refers to Bayesian theory.

Q Predict outcome for a car. The values of the car is given as:
Car = {Red, Domestic, SUV}

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	~~Sports~~ SUV	Imported	No

Red Sports Imported Yes

formula for Naive Bayes is given as:

$$\underbrace{P(A/B)}_{\text{Posterior}} = \frac{P(B/A) \cdot P(A)}{P(B)}$$

→ Probability of A

Probability of A when B is true

Probability of B

x = [Red, Domestic, SUV]

$P(\text{Red}/\text{Yes}) = \frac{3/5 \times 5/10}{5/10} = \frac{3}{5}$

$P(\text{Domestic}/\text{Yes}) = \frac{2/5 \times 5/10}{5/10} = \frac{2}{5}$

$P(\text{SUV}/\text{Yes}) = \frac{1/\cancel{4} \times \cancel{4}/10}{5/10} = \frac{1}{10} \times \frac{10}{5} = \frac{1}{5}$

$P(\text{Red}/\text{No}) = \frac{2/5 \times 5/10}{5/10} = \frac{2}{5}$

$P(\text{Domestic}/\text{No}) = \frac{3/5 \times 5/10}{5/10} = \frac{3}{5}$

$P(\text{SUV}/\text{No}) = \frac{3/4 \times 4/10}{5/10} = \frac{3}{5}$

Date:
Page:

Probability of 'Yes' $= \frac{3}{5} * \frac{2}{5} * \frac{1}{5}$

$= \frac{6}{125} = 0.024 = 0.048$

Probability of 'No' $= \frac{2}{5} * \frac{3}{5} * \frac{3}{5}$

$= \frac{18}{125} = \cancel{1.6}\ 0.144$

X is not a stolen car.

BAYESIAN BELIEF NETWORK

→ It's a probabilistic graphical model (PGM)
→ Represents a set of variable and conditional dependencies among those variables.
→ suitable for representing probabilistic relationship b/w multiple events.

Uses :-
1. Prediction
2. Optimized Web Search
3. Spam filtering
4. Decision Making under uncertainity
5. Time Series Prediction
6. Disease Diagnosis
7. Reasoning

→ We represents Bayesian B.N. through Directed Cyclic Graph (DCG)

Joint Probability - It is a measure of two events happening at the same time, i.e. p(A and B). The probability of the intersection of A and B, maybe written by P(A∩B).

Conditional Probability - of an event B is the probability that the event will occur given that an event A has already occurred

$$P(B/A) = \frac{P(A \text{ and } B)}{P(A)}$$

If A and B are dependent variables.