Read:

Creating a file: Using file.create() function, a new file can be created from console or truncates if already exist.

Syntax:
file.create(" ")

Example: # create a file
The file created can be seen
In your working directory.

Reading a file: using read.table() function R, files can be read and output is shown as dataframe.

Syntax: read.table(file)

Ex: # Reading txt file
new.iris <- read.table(file= "GFG.txt")

# print
print(new.iris)

writing into a csv file
R can create csv file form existing data frame. The write.csv() function is used to create the csv file. This file gets created in the working directory.

```
# create a data frame
data <- read.csv("input.csv")
retval <- subset(data, as.Data(start_date) > as.Date(2004-10-01"))
# write filtered data into a new file.
write.csv(retval, "output.csv")
newdata <- read.csv("output.csv")
print(newdata)
```

Data Manipulation

round(x,n) # round the values of x to n decimal places

ceiling(x) # vector x of smallest integers > x.

floor(x) # vector x of largest integer < x.

as.integer # truncates real x to integers (compare to round (x,0)).

Statistics

min() → Lowest value from given data

mean() → Average value.

median() → Middle value Q_1, Q_2, Q_3

sum() → Total

var() → # produces the variance covariance matrix.

sd() → # standard deviation.

Transformation

fivenum() → # Turkey five numbers min, lowerhinge, median, upper hinge, max.

table() → # frequency counts of entries, ideally the entries are factors (although it works with integers or even reals).

Scale(data, scale = T) → # centers around the mean and scales by the sd)

input and display

read table(filename, header= True) →
read files with tables in first row
read a tab or space delimited file.

read.table(filename, header = True, sep = ",")
# read csv files

x = c(1:10) → # create a data vector with elements 1-10.

vect = c(x,y) → # combine them into vector of length 2n.

mat = cbind(x,y) → # combine them into a n x 2 matrix.

Day-2

Write down a list of 6 numbers where the mean is 0, the mode is 1 and the range is 3.

Write down 8 numbers, 1 to 8, and demonstrate how to calculate the interquartile range

Here are ten numbers.

7 6 8 4 5 9 7 3 6 7

a) Range
b) Mode
c) Median
d) Mean
e) Interquartile range

7.Find the interquartile range of the following data.: 12, 17, 34, 56, 32, 23, 83, 31, 9, 36

8.The points scored by a football team during each game from last season are: 14, 21, 18, 21, 14, 35, 42, 42, 21, 14 Calculate the mean, median, and mode: Mean: Median: Mode:

9. Use the data provided to answer the questions below: 14, 9, 3, 11, 13, 16, 4, 11, 19, 10, 8, 20, 13, 7, 12

Find the median of the data set. This value is known as Quartile 2 or Q2 for short

Find the median of the first half of the data set. This value is known as Quartile 1 or Q1.

. Find the median of the second half of the data set. This value is known as Quartile 3 or Q3.

Subtract the Q3 value from Q1. This is the Interquartile Range or IQR.

Unit -II

1.9 people take a test. Their scores out of 100100 are:

56, 79, 77, 48, 90, 68, 79, 92, 7156,79,77,48,90,68,79,92,71

Calculating the Range

2.Find the range of 12, 8, 4, 16, 15, 15, 5, 15, 10, 812,8,4,16,15,15,5,15,10,8

3.There were 55 members of a basketball team who had a mean points score of 1212 points each per game.

4.One of the team members left, causing the mean point score to reduce to 1010 points each per game.What was the mean score of the player that left?: Find the mode and range of the list of numbers below: 280, 350, 320, 400, 350,280,350,320,400,350, 490, 590, 470, 280, 410, 350490,590,470,280,410,350

5.Dani recorded the heights of members of her extended family, to the nearest cm. Find the median of their heights (listed below):

163, 185, 164, 170, 188, 154, 168, 179163,185,164,170,188,154,168,179

6.

For the following list of numbers, $1, 2, -2, 0, 1, 8, 3, -3, 2, 4, -2, 2$ find the:

a) Range
b) Mode
c) Median
d) Mean
e) Interquartile range

Central Tendency:- cental around statistics.

Mean:- The average of the number a calculated.

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Total no of data values}(n)}$$

Mean (census$ Total·Males)

Median :- Middle value (sorted)

Median (census $ Total. Males)

Mode:- Most frequently occuring values

Mode (census $ Total. Males)

Measures of dispersion

Inter Quartile Range | Quartiles | Mid Range | Outlier.

Example of Central Tendency.

Ex:- 30, 30, 32, 38, 60

$$\text{Mean} = \frac{30+30+32+38+60}{5} = \frac{190}{5} = 38$$

Median = 32

Mode = 30.

Inter Quartile Range :- Measures of variability, based on dividing a data set into quartiles.

Example:-

→ 2, 3, 5, 7, 11, 13, 17, 19, 23, 29

Ten prime numbers are (Increasing order)

Q_1 → Lower Q·P (Quartile part)
Q_2 → Median
Q_3 → Upper Q·P

No of values = 10 → Even number

∴ Median is Mean of 11 & 13

$$Q_2 = \frac{11+13}{2} = \frac{24}{2} = 12$$

Q_1 part; 2, 3, 5, 7, 11 → 5 → odd no.

Central value → 5 $\boxed{Q_1 = 5}$

Q_3 part:- 13, 17, 19, 23, 29 → 5 C.V → 19.

The Subtraction $\boxed{Q_1 \& Q_3 \rightarrow 19 - 5 = 11}$

11 is inter quartile Range Value.

Quartile (census $ Total. Males, 0.25)

Quartiles:- Compare set of observations

Ex:- 4, 6, 7, 8, 10, 23, 24 n=7

Lower :- $Q_1 = \frac{(n+1)}{4} - \frac{7+1}{4} = 2^{nd}$ item → 6

Median:- $Q_2 = \frac{n+1}{2} = \frac{7+1}{2} = 4^{th}$ item → 8

Upper:- $Q_3 = \frac{3(n+1)}{4} = 6^{th}$ item - 23

Mid Range:- The Arithmetic mean of largest & Smallest Value.

range (census$ Total. Males)

$$\text{Mid range} = \frac{(\text{Maximum Value} + \text{Minimum value})}{2}$$

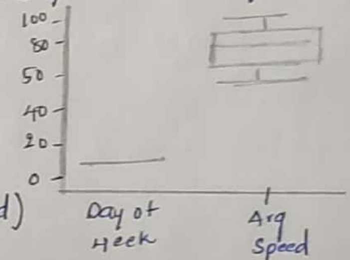
Ex:- 2, 5, 6, 7, 8, 9, 4 → $\frac{9+2}{2} = \frac{11}{2}$

PLOTTING GRAPHS USING R-TOOL

Bar plot:- Syntax:- barplot (data, xlab, ylab)

Ex:- travels-travel times [which (travel $ Day of Week == 'Friday')]

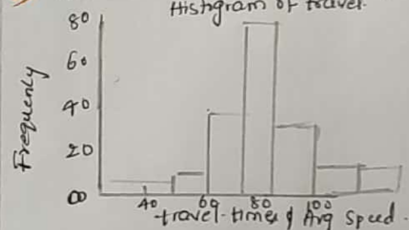
names (travel-times) <- in 1:c ("Day of Week", Avg Speed")



HISTOGRAM

→ Which displays of statistical information.

→ Hist (travel·times $ Avg. Speed)

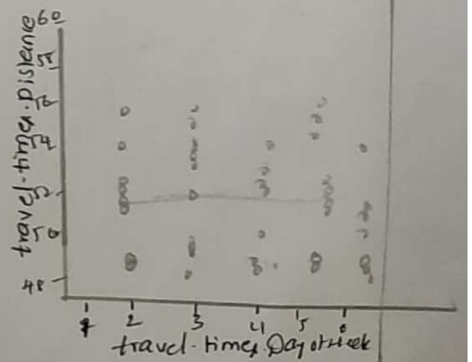


SCATTERPLOT:- Extent of

→ Correlation.

→

Scatter·smooth (travel·times $ Day of Week, travel $ Distance)



PERFORM CORRELATION ANALYSIS AND NORMALIZATION USING R-TOOL

Correlation Analysis:

steps involved :-

1. Create a new table with required dataframes
2. After that apply the formula or query for the chi-square test

Quries:-

* diabetes1 <- table(diabetes$Age · diabetes$Insulin)
* diabetes1
* chi sq·test (diabetes1)

Min Max Normalization Formula

Marks
8
10
15
20

Min: The minimum value of the given attribute

Here Min is 8.

Max: The maximum value of given attribute.

Here Max is 20.

V: V is the representative value of attribute

For ex: $V_1 = 8$, $V_2 = 10$, $V_3 = 15$ & $V_4 = 20$

new Max: 1

new Min: 0

$$\boxed{v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A}$$

For marks as 8:

$$MinMax = \frac{V - Min\ marks}{Max\ marks - Min\ marks}(newMax - newMin) + newMin$$

$$MinMax = \frac{(8-8)}{20-8} * (1-0) + 0$$

$$MinMax = \frac{0}{12} * 1$$

$$MinMax = 0$$

For marks as 10:

$$MinMax = \frac{(10-8)}{20-8} * (1-0) + 0$$

$$MinMax = \frac{(2)}{12} * 1$$

$$MinMax = 0.16$$

Z-Score

$$\text{standard deviation} = \sqrt{\frac{\Sigma\left(\text{every individual value of marks} - \text{mean of marks}\right)^2}{n}}$$

$$\boxed{S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}}$$

mean of marks = $8 + 10 + 15 + 20 / 4 = 13.25$

$$S = \sqrt{\frac{(8-13.25)^2 + (10-13.25)^2 + (15-13.25)^2 + (20-13.25)^2}{4}}$$

$$= \sqrt{\frac{27.56 + 10.56 + 3.06 + 45.56}{4}} = \sqrt{\frac{86.74}{4}} = \sqrt{21.6} = \underline{\underline{4.6}}$$

Mean $(\mu) = 13.25$

standard deviation $(\sigma) = 4.6$

$$Zscore = \frac{x - \mu}{\sigma} = \frac{8 - 13.25}{4.6} = -1.14$$

$$Zscore = \frac{x - \mu}{\sigma} = \frac{10 - 13.25}{4.6} = -0.7$$

$$Zscore = \frac{x - \mu}{\sigma} = \frac{15 - 13.25}{4.6} = 0.3$$

$$Zscore = \frac{x - \mu}{\sigma} = \frac{20 - 13.25}{4.6} = 1.4$$

marks	marks after z-score normalization
8	-1.14
10	-0.7
15	0.3
20	1.4

z-score Normalization:

* A <- c (diabetes $ Age)
* Mean <- mean(A)
* std <- sd(A)
* zscore <- (A-mean)|std
* Zscore

Decimal Scaling Formula

A value of attribute A is can be normalized by the formula

$$\boxed{\text{normalized value of attribute} = \frac{v'}{10'}}$$

Salary Bonus	Formula	CGPA Normalized after Decimal scaling
400	400/1000	0.4
310	310/1000	0.31

$$\text{Decimal scaling} = \frac{A}{100}$$

Day-4

APRIORI ALGORITHM

Transaction ID	ITEMS
T1	Hot Dogs, Buns, Ketch up
T2	Hot Dogs, Buns
T3	Hot dogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchep
T6	HotDogs, Coke, chips

Find the "Frequent Item Sets" & generate "Association Rules" on this.

Assume, Min Support threshold

$\boxed{S = 33.33\%}$

Min Confident Threshold

$\boxed{C = 60\%}$

Solution:-

Min Support Count $= \frac{33.33}{100} \times 6$

Min Support Count = 2

Item Set	Sup count
Hot Dogs	4
Buns	2
Ketchup	2
Coke	3
Chips	4

⬇

Item Set	Sup.count
Hot Dogs, Buns	2
Hot Dogs, Ketchup	1
Hot Dogs, Coke	2
Hot Dogs, chips	2
Buns, Ketchup	1
Buns, Coke	0
Buns, chips	0
Ketchup coke	0
Ketchup chips	1

⬇

Item Set	Sup.count
Hot Dogs, Buns	2
Hot Dogs, Coke	2
Hot Dogs, chips	2
Coke, Chips	3

⬇

Item Set	Sup.count
Hot Dogs, Buns, coke	0
Hot Dogs, Buns, chips	0
Hot Dogs, coke, chips	2

⬇

Item Set	Sup.count
Hot Dogs, coke, chips	2

* There is only one itemset with min support 2. So, only one itemset is frequent.

Frequent Itemset (I) = {Hot Dogs, coke, chips}

Association Rules:-

* [Hot Dogs ∧ coke] ⇒ [chips]

$$\text{Conf} = \frac{\text{Sup (Hot Dogs} \wedge \text{Coke} \wedge \text{chips)}}{\text{Sup (Hot Dogs} \wedge \text{Coke)}}$$

$= 2/(2\times100) = 100\%$ [Selected]

* [Hot Dogs ∧ chips] ⇒ [coke]

$$\text{Conf} = \frac{\text{Sup(Hot Dogs} \wedge \text{coke} \wedge \text{chips)}}{\text{Sup (Hot Dogs} \wedge \text{chips)}}$$

$= 2/(2\times100) = 100\%$ [Selected]

* [coke ∧ chips] ⇒ [Hot Dogs]

$$\text{Conf} = \frac{\text{Sup (Hot dogs} \wedge \text{coke} \wedge \text{chips)}}{\text{Sup (coke} \wedge \text{chips)}}$$

$= 2/(3\times100) = 66.67\%$ [Selected]

* [Hot Dogs] ⇒ [coke ∧ chips]

$$\text{Conf} = \frac{\text{Sup (Hot Dogs} \wedge \text{coke} \wedge \text{chips)}}{\text{Sup(Hot Dog)}}$$

$= 2/(4\times100) = 50\%$ [Rejected]

* [coke] ⇒ [Hot Dogs ∧ chips]

$$\text{Conf} = \frac{\text{Sup (Hot Dogs} \wedge \text{Coke} \wedge \text{chips)}}{\text{Sup (coke)}}$$

$= 2/[3\times100] = 66.67\%$ [Selected]

* [chips] ⇒ [Hot Dogs ∧ coke]

$$\text{Conf} = \frac{\text{Sup(Hot Dogs} \wedge \text{coke} \wedge \text{chips)}}{\text{Sup(chips)}}$$

$= 2/(4\times100) = 50\%$ [Rejected]

There are 4-Strong results [Min.conf > 60%]

******

FP-Growth Algorithm

* Frequent pattern-Growth Alg

TID	ITEMS BOUGHT	(ORDERED) Freq-Items
100	f, a, c, d, g, i, m, p	f, c, a, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m ✓
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m, n	f, c, a, m, p, v ✓

* Frequent Pattern-Type

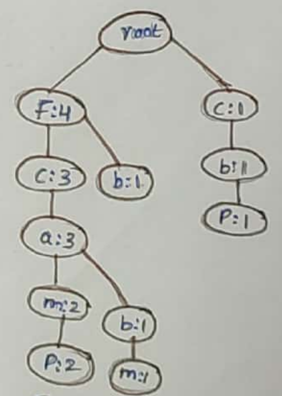


Fig Frequent pattern Tree

* Conditional Pattern Base [CPB]

Item	Conditional Pattern Base
p	{{F,c,a,m:2},{c,b:1}}
m	{{F,c,a:2},{F,c,a,b:1}}
b	{{F,c,a:1},{F:1},{c:1}}
a	{{F,c:3}}
c	{{F:3}}
F	∅

* Conditional-Frequent Pattern Tree

Item	Conditional Pattern Base	{Conditional} FP-Tree
p	{F,c,a,m:2},{c,b:1}	{c:3}
m	{F,c,a:2},{F,c,a,b:1}	{F,c,a:3}
b	{F,c,a:1},{F:1},{c:F}	∅
a	{F,c:3}	{F,c,:3}
c	{{F:3}}	{F:3}
F	∅	∅

Common in all Paths

Frequent Pattern rules

Item	Conditional Pattern Base	Conditional FP-Tree	Freq-Pattern Generated
p	{{F,c,a,m:2},{c,b:1}}	{c:3}	{ <c,p:3> }
m	{{F,c,a:2},{F,c,a,b:1}}	{F,c,a:3}	{ <F,m:3>, <c,m:3> <a,m:3>, <F,c,m:3> <F,a,m:3>, <c,a,m:3>}
b	{{F,c,a:1},{F:1},{c:1}}	∅	{ }
a	{{F,c:3}}	{F, c:3}	{(F,a:3), <c,a:3> <F,c,a:3>}
c	{{F:3}}	{F:3}	{<F,c:3>}
F	∅	∅	{ }

Hence Frequent pattern rules — generated —

297-5

rec	Age	Income	Student	Credit _rating	Buys_computer
r1	<=30	High	No	Fair	No
r2	<=30	High	No	Excellent	No
r3	31..40	High	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31..40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	<=30	Low	Yes	Fair	Yes
r10	>30	Medium	Yes	Fair	Yes
r11	<=30	Medium	Yes	Excellent	Yes
r12	31..40	Medium	No	Excellent	Yes
r13	31..40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No

Day	outlook	temp	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Day-5

DECISION TREE CLASSIFIER

Decision Tree Construction Alg.	Mathematical Measure – Picking test attribute
ID3 – Iterative Dichotomizer	IG – Information Gain
CART – Classification and Regression Tree	Gini Index

Entropy :

Entropy (P) =

$-[P_1 \log(P_1) + P_2 \cdot \log(P_2) + \dots + P_n \log(P_n)]$

* Classes → Tree

n-classes : (T have n-classes)

C1, C2, C3, ... Cn.

* Identify the class of an element of T is

$P = \frac{|C_1|}{|T_1|}, \frac{|C_2|}{|T_2|} \dots \frac{|C_n|}{|T_n|}$ (P)

where

* Training Database Contains

class attribute & non-class attribute

* Weighted Average of INFO (T_i)

$$INFO(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} INFO(T_i)$$

→ lower the value of INFO(X,T) lower the uncertainty.

→ So, choose the attribute X, lowest value for INFO(X,T)

* Information Gain (IG) due to split on an attribute X is

$Gain(X,T) = INFO(T) - INFO(X,T)$

* Attribute with the Max-Info-Gain is chosen always at the decision-node.

* **Example :**

DAY	OUTLOOK	TEMP	HUM (Humidity)	WIND	CLASS (P.B – Play Ball)
D1	Sunny	Hot	High	Weak	NO
D2	Sunny	Hot	High	Strong	NO
D3	Overcast	Hot	High	Weak	YES
D4	Rain	Mild	High	Weak	YES
D5	Rain	Cool	Normal	Weak	YES
D6	Rain	Cool	Normal	Strong	NO
D7	Overcast	Cool	Normal	Strong	YES
D8	Sunny	Mild	High	Weak	NO
D9	Sunny	Cool	Normal	Weak	YES
D10	Rain	Mild	Normal	Weak	YES
D11	Sunny	Mild	Normal	Strong	YES
D12	Overcast	Mild	High	Strong	YES
D13	Overcast	Hot	Normal	Weak	YES
D14	Rain	Mild	High	Strong	NO

* Build the Decision-Tree

* 2-classes & 4-Non-class attributes

(Yes & No) (NO)

* Determining Entropy

9 – yes, 5 – no, 14 – Total

INFO (T) = ENTROPY (P) =

$Entropy\left(\frac{9}{14}, \frac{5}{14}\right) =$

$-\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14}$

$\boxed{E = 0.94}$

$$INFO(outlook, T) = \sum_{i=1}^{3} \frac{|T_i|}{|T|} INFO(T_i)$$

$= \frac{5}{14} INFO\left(\frac{3}{5}, \frac{2}{5}\right) + \frac{4}{14} INFO(1, 0) + \frac{5}{14} INFO\left(\frac{3}{5}, \frac{2}{5}\right)$

$= 0.694$

$$INFO(humidity, T) = \sum_{i=1}^{2} \frac{|T_i|}{|T|} \times INFO(T_i)$$

$= \frac{7}{14} INFO\left(\frac{3}{7}, \frac{4}{7}\right) + \frac{7}{14} INFO\left(\frac{6}{7}, \frac{1}{7}\right)$

$= 0.78845045$

$$INFO(Temperature, T) = \sum_{i=1}^{3} \frac{|T_i|}{|T|} INFO(T_i)$$

$= \frac{4}{14} INFO\left(\frac{2}{4}, \frac{2}{4}\right) + \frac{6}{14} INFO\left(\frac{4}{6}, \frac{2}{6}\right) + \frac{4}{14} INFO\left(\frac{3}{4}, \frac{1}{4}\right)$

$= 0.9110633$

$$INFO(wind, T) = \sum_{i=1}^{2} \frac{|T_i|}{|T|} \cdot INFO(T_i)$$

$= \frac{6}{14} INFO\left(\frac{3}{6}, \frac{3}{6}\right) + \frac{8}{14} INFO\left(\frac{6}{8}, \frac{2}{8}\right)$

$= 0.89215892$

INFO (Outlook, T) = 0.694

INFO (humidity, T) = 0.78845045

INFO (Temperature, T) = 0.9110633

INFO (wind, T) = 0.89215892

• Gain (X,T) = INFO(T) – INFO(X,T)

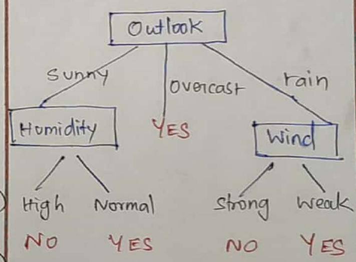
* Gain (outlook, T) = 0.94 – 0.694

* Gain (Humidity, T) = 0.94 – 0.78845

* Gain (Temp, T) = 0.94 – 0.9110633

* Gain (wind, T) = 0.94 – 0.892158

∵ Outlook attribute has the highest gain, it is chosen as a root-node.



Decision Tree Generated.

Note :

• Outlook < Sunny, Overcast, Rain → $\sum_{i=1}^{3}$

• Humidity < High, Normal → $\sum_{i=1}^{2}$

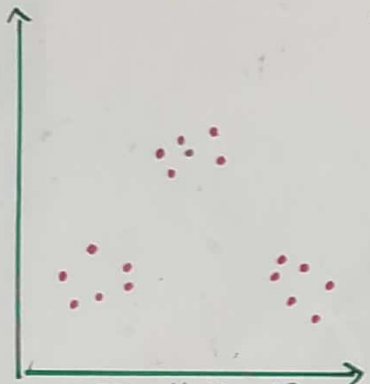
• Temp < Hot, Cool, Mild → $\sum_{i=1}^{3}$

wind < weak, strong → $\sum_{i=1}^{2}$

Day-16

K-Means Clustering:

* K-Means clustering is an unsupervised iterative clustering technique.
* It partitions the given data set into k predefined distinct clusters.
* A cluster is defined as a collection of data point exhibiting certain similarities.


BEFORE K-MEANS

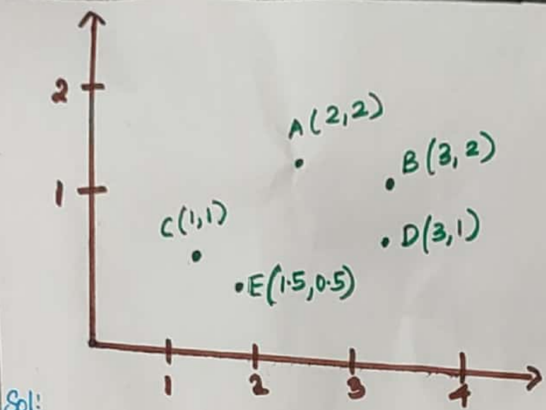

AFTER K-MEANS

It partitions the data set such that:

* Each data point belongs to a cluster with the nearest mean.
* Data points belonging to one cluster have high degree of similarity.
* Data points belonging to different clusters have high degree of dissimilarity.

PROBLEM:

Use K-Means Algorithm to create two clusters.



Sol:

→ Iteration 01:

We calculate the distance of each point from each of the center of the two clusters.

Distance is calculated by using euclidean distance formula.

→ Calculating Distance Between A(2,2) and $C_1(2,2)$:

$P(A,C1)$

$= \text{sqrt}\left[(x_2-x_1)^2+(y_2-y_1)^2\right]$

$= \text{sqrt}\left[(2-2)^2+(2-2)^2\right] = 0$

→ Calculating Distance Between A(2,2) and $C_2(1,1)$:

$P(A, C2)$

$= \text{sqrt}\left[(x_2-x_1)^2+(y_2-y_1)^2\right]$

$= \text{sqrt}\left[(1-2)^2+(1-2)^2\right] = 1.41$

Given points	Distance from Center (2,2)	Distance from center (1,1)	Point belongs to cluster
A(2,2)	0	1.41	C1
B(3,2)	1	2.24	C1
C(1,1)	1.41	0	C2
D(3,1)	1.41	2	C1
E(1.5,0.5)	1.58	0.71	C2

Cluster-01: contains points A(2,2), B(3,2), E(1.5,0.5), D(3,1)

Cluster-02: contains points C(1,1), E(1.5,0.5)

→ Now we re-compute the new cluster clusters.

→ The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

center of Cluster-01

$= ((2+3+3)/3, (2+2+1)/3)$

$= (2.67, 1.67)$

For cluster-02:

center of cluster-02

$= ((1+1.5/2, (1+0.5)/2)$

$= (1.25, 0.75)$

* This is completion of Iteration-01.
* Next, we go to iteration-02, iteration-03 and so on until the centers do not change anymore.