Agglomerative Heirarchical Clustering

→ Consider the following set of 6 one-dimensional data points:

18, 22, 25, 42, 27, 43.

→ Apply the agglomerative heirarchical clustering algorithm to build the heirarchical clustering dendrogram.

→ Merge the clusters using Min distance & update the proximity matrix accordingly.

→ clearly show the proximity matrix corresponding to each iteration of the algorithm.

Calculate Distance Matrix:-

Step 1 →

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

As '1' is the minimum distance bet. 42 & 43, merge 42 & 43 to make a cluster.

43 is merged (2) with 42

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	(1) ✓
43	25	21	18	16	(1) ✓	0

minimum distance

(42, 43)

Step 2 :- After merging matrix will be,

	18	22	25	27	42,43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42,43	24	20	17	15	0

Next '2' is the min. distance bet.) 25 & 27, so merge 25 & 27 to make another cluster.

	18	22	25	27	42,43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	(2) ✓	17
27	9	5	(2) ✓	0	15
42,43	24	20	17	15	0

(42, 43) (25, 27)

step 3 → After merging, matrix will be,

	18	22	25,27	42,43
18	0	4	7	24
22	4	0	3	20
25,27	7	3	0	17
42,43	24	20	17	0

Next min. value is 3.

So, merge 22 & (25, 27)

	18	22	25,27	42,43
18	0	4	7	24
22	4	0	3 ✓	20
25,27	7	3 ✓	0	17
42,43	24	20	17	0

(42, 43) ((25, 27), 22)

step 4 :

	18	22,25,27	42,43
18	0	4	24
22,25,27	4	0	20
42,43	24	20	0

Next, min. distance is 4 ④

So, merge 18 & (22, 25, 27).

	18	22,25,27	42,43
18	0	④ ✓	24
22,25,27	④ ✓	0	20
42,43	24	20	0

(42, 43), (((25, 27), 22), 18)

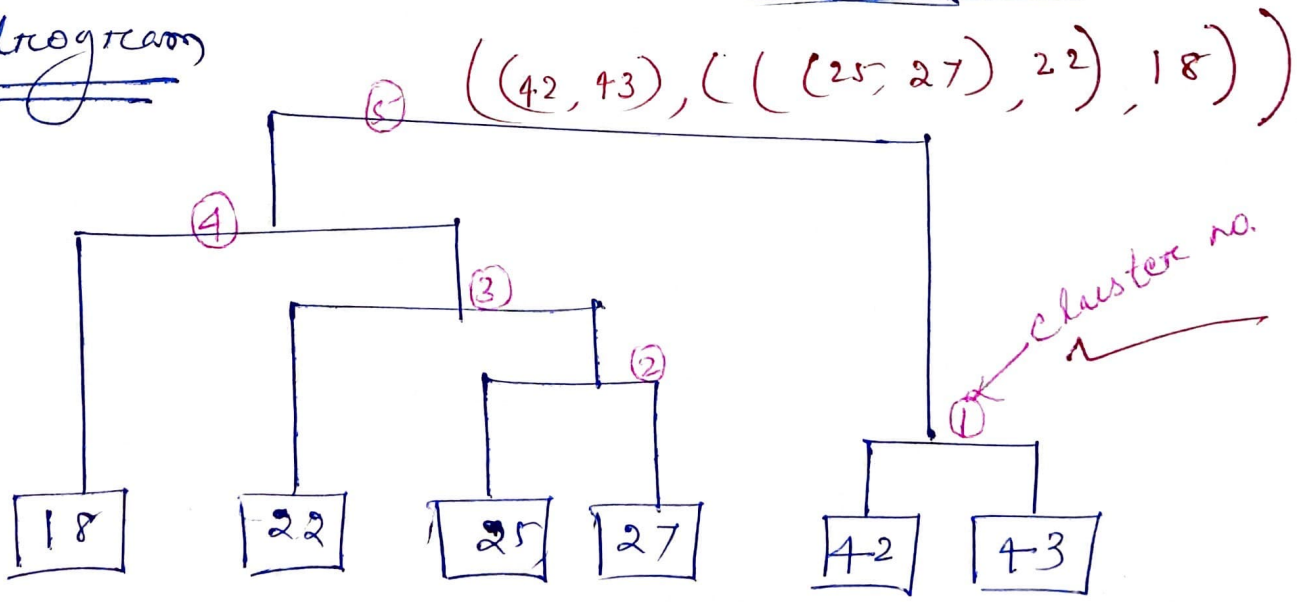
Steps → Matrix will be

	18, 22, 25, 27	42, 43
18, 22, 25, 27	0	24
42, 43	24	0

Step6 →

	18, 22, 25, 27, 42, 43
18, 22, 25, 27, 42, 43	0

Dendrogram



ID3 Algorithm Example

Refer the previous dataset, which consists of 4 attributes & 1 target-output.

First we need to find out which attribute is having maximum information gain, then that attribute is considered as the root node. So, we will calculate IG for each & every attribute.

Attribute : Outlook.

Values (Outlook) = Sunny, Overcast, Rain.

[NOTE: to calculate IG for each attribute, Entropy for the whole dataset & Entropy for the individual attribute value is needed]

So, $S = [9+, 5-]$ (for whole dataset)

$$\text{Entropy}(S) = -P_{yes}\log_2 P_{yes} - P_{NO}\log_2 P_{NO}$$

$$= -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= -\frac{9}{14}\frac{\log_2 9}{\log_2 14} - \frac{5}{14}\frac{\log_2 5}{\log 14}$$

$$= 0.94.$$

$S_{sunny} \leftarrow [2+, 3-]$

$Entropy(S_{sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$= 0.971$

$S_{Overcast} \leftarrow [4+, 0-]$

$Entropy(S_{overcast}) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)$

$= 0$

$S_{Rain} \leftarrow [3+, 2-]$

$Entropy(S_{Rain}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$

$= 0.971$

Now,

$Gain(S, Outlook) = Entropy(S) - \sum_{v \in (sunny, Overcast, Rain)} \frac{|S_v|}{|S|} Entropy(S_v)$

$\therefore Gain(S, Outlook)$

$= Entropy(S) - \frac{5}{14} Entropy(S_{sunny}) - \frac{4}{14}$

$Entropy(S_{overcast}) - \frac{5}{14} Entropy(S_{Rain})$

Gain (S, Outlook) =

$= 0.94 - \frac{5}{14} \times 0.971 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.971$

$= 0.2464$

Attribute : Temp.

Values (Temp) = Hot, Mild, Cool.

$S = [9+, 5-]$ (for whole dataset)

$Entropy(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$

$= 0.94.$

$S_{Hot} \leftarrow [2+, 2-]$

$Entropy(S_{Hot}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$

$= 1.0$

$S_{Mild} \leftarrow [4+, 2-]$

$Entropy(S_{Mild}) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)$

$= 0.9183.$

$S_{Cool} \leftarrow [3+, 1-]$

$Entropy(S_{Cool}) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

$= 0.8113.$

Hence,

$$Gain(S, Temp) = 0.94 - \frac{4}{14} \times 1.0 - \frac{6}{14} \times 0.9183 - \frac{4}{14} \times 0.8113$$

$$= 0.0289$$

Attribute : Humidity

Values (Humidity) = High, Normal

$S = [9+, 5-]$ (for the whole dataset)

$$Entropy(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.94$$

$S_{High} \leftarrow [3+, 4-]$

$$Entropy(S_{High}) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$$

$$= 0.9852$$

$S_{Normal} \leftarrow [6+, 1-]$

$$Entropy(S_{Normal}) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)$$

$$= 0.5916$$

Now,

$$Gain(S, Humidity) = Entropy(S) - \sum_{v \in (High, Normal)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$\therefore$ Gain (S, Humidity)

$= \text{Entropy}(S) - \frac{7}{14}\text{Entropy}(S_{High}) - \frac{7}{14}\text{Entropy}(S_{Normal})$

$= 0.94 - \frac{7}{14} \times 0.9852 - \frac{7}{14} \times 0.5916$

$= 0.1516.$

Attribute : Wind

Values (Wind) = Strong, Weak.

~~S=[9+, 5-]~~

$S = [9+, 5-]$ (for the whole dataset)

$\text{Entropy}(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$

$= 0.94$

$S_{strong} \leftarrow [3+, 3-]$

$\text{Entropy}(S_{strong}) = 1.0$

$S_{weak} \leftarrow [6+, 2-]$

$\text{Entropy}(S_{weak}) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$

$= 0.8113.$

Gain (S, Wend)

$$= \text{Entropy}(S) - \sum_{v \in (\text{strong, weak})} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$= \text{Entropy}(S) - \frac{6}{14} \text{Entropy}(S_{\text{strong}}) - \frac{8}{14} \text{Entropy}(S_{\text{weak}})$$

$$= 0.94 - \frac{6}{14} \times 1.0 - \frac{8}{14} \times 0.8113$$

$$= 0.0478.$$

Now,

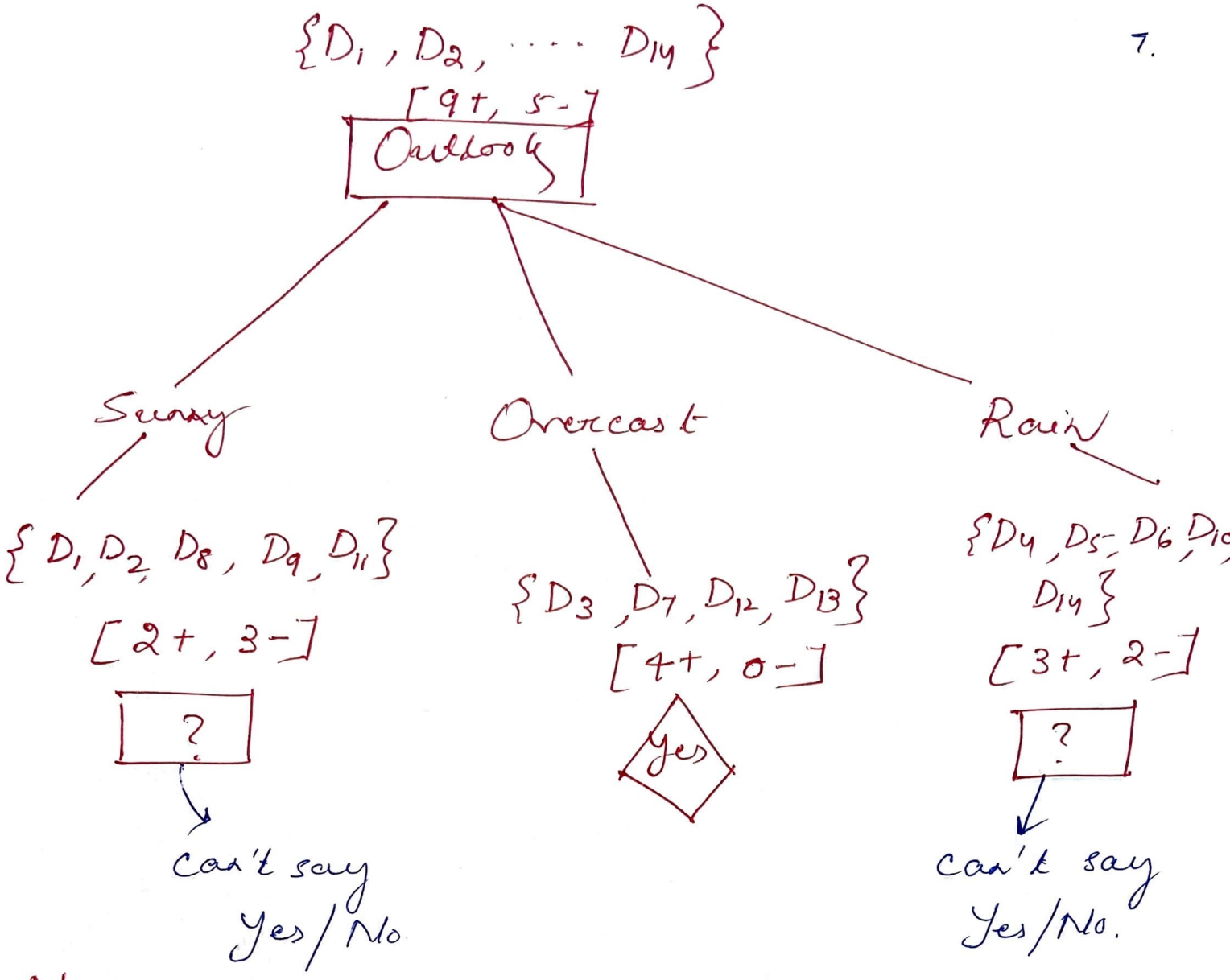
Gain (S, Outlook) = 0.2464 ✓

Gain (S, Temp) = 0.0289

Gain (S, Humidity) = 0.1516

Gain (S, Wend) = 0.0478

Here, Outlook attribute is having maximum IG. So Outlook is considered as the rootnode.



Now,
Consider the Sunny attribute value only;

Day	Temp	Humidity	Wind	Play Tennis
D_1	Hot	High	Weak	No
D_2	Hot	High	Strong	No
D_8	Mild	High	Weak	No
D_9	Cool	Normal	Weak	Yes
D_{11}	Mild	Normal	Strong	Yes

Attribute : Temp

Values (Temp) = Hot, Mild, Cool

$S_{sunny} = [2+, 3-]$

$Entropy(S_{sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$= 0.97$

$S_{Hot} \leftarrow [0+, 2-]$

$Entropy(S_{Hot}) = 0.0$

$S_{Mild} \leftarrow [1+, 1-]$

$Entropy(S_{Mild}) = 1.0$

$S_{cool} \leftarrow [1+, 1-]$

$Entropy(S_{cool}) = 0.0$

$\therefore Gain(S_{sunny}, Temp)$

$= Entropy(S) - \sum_{v \in (Hot, Mild, cool)} \frac{|S_v|}{|S|} Entropy(S_v)$

$= Entropy(S) - \frac{2}{5} Entropy(S_{Hot}) - \frac{2}{5} \cdot Entropy(S_{Mild}) - \frac{1}{5} Entropy(S_{cool})$

$= 0.97 - \frac{2}{5} * 0.0 - \frac{2}{5} * 1.0 - \frac{1}{5} * 0.0$

$= 0.570.$

Attribute: Humidity

Values (Humidity) = High, Normal.

$S_{sunny} = [2+, 3-]$

$Entropy = \frac{2}{5} \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \log_2\left(\frac{3}{5}\right) = 0.97$

$S_{High} \leftarrow [0+, 3-]$, $Entropy(S_{High}) = 0.0$

$S_{Normal} \leftarrow [2+, 0-]$, $Entropy(S_{Normal}) = 0.0$

$\therefore Gain(S_{sunny}, Humidity)$

$= Entropy(S) - \sum_{V \in (High, Normal)} \frac{|S_V|}{|S|} Entropy(S_V)$

$= Entropy(S) - \frac{3}{5} * Entropy(S_{High}) - \frac{2}{5} * Entropy(S_{Normal})$

$= 0.97 - \frac{3}{5} * 0.0 - \frac{2}{5} * 0.0$

$= 0.97$

Attribute : Wind

Values (Wind) = {Strong, Weak}

$S_{sunny} = [2+, 3-]$.

$Entropy(S_{sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$= 0.97$.

$S_{strong} \leftarrow [1+, 1-]$, $Entropy(S_{strong}) = 1.0$

$S_{weak} \leftarrow [1+, 2-]$,

$Entropy(S_{weak}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= 0.9183$.

$\therefore Gain(S_{sunny}, Wind)$

$= Entropy(S) - \sum_{v \in (Strong, Weak)} \frac{|S_v|}{|S|} Entropy(S_v)$

$= Entropy(S) - \frac{2}{5} Entropy(S_{strong}) - \frac{3}{5} Entropy(S_{weak})$

$= 0.97 - \frac{2}{5} * 1.0 - \frac{3}{5} * 0.918$

$= 0.0192$

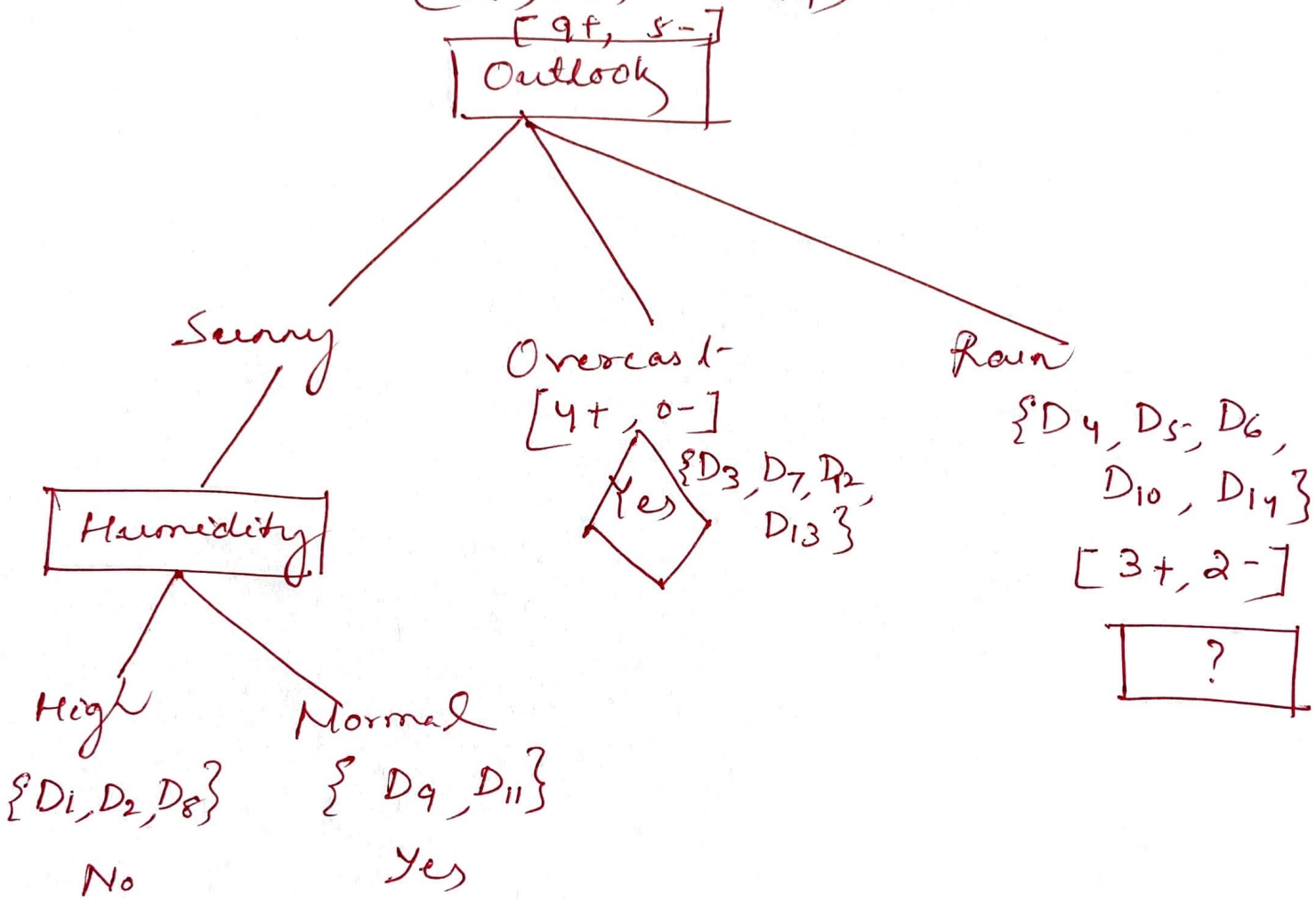
~~Hence~~, As,

Gain(Ssunny, Temp) = 0.570

Gain(Ssunny, Humidity) = 0.97 ← Having max Gain

Gain(Ssunny, Wind) = 0.0192

Now, the tree will be,



Next, consider Rain attribute.

Dataset will be,

Day,	Temp	Humidity	Wind	Play Tennis
D_4	Mild	High		Yes
D_5	Cool	Normal		Yes
D_6	Cool	Normal		No
D_{10}	Mild	Normal		Yes
D_{14}	Mild	High		No.

Attribute : Temp.

Values (Temp) = Hot, Mild, Cool.

$S_{Rain} = [3+, 2-]$.

Entropy $(S_{Rain}) = 0.97$.

$S_{Hot} \leftarrow [0+, 0-]$

Entropy $(S_{HOT}) = 0.0$

$S_{Mild} \leftarrow [2+, 1-]$.

Entropy $(S_{Mild}) = 0.9183$

$S_{cool} \leftarrow [1+, 1-]$

Entropy $(S_{cool}) = 1.0$.

Gain $(S_{Rain}, Temp) = 0.0192$

Attribute : Humidity

Values (Humidity) = High, Normal.

$S_{Rain} \leftarrow [3+, 2-]$

Entropy $(S) = 0.97$.

$S_{High} \leftarrow [1+, 1-]$, Entropy $(S_{High}) = 1.0$

$S_{Normal} \leftarrow [2+, 1-]$, Entropy $(S_{Normal}) = 0.918$

$\therefore$ Gain $(S_{Rain}, Humidity) = 0.0192$

Attribute : Wind

Values (Wind) = Strong, Weak.

$S_{Rain} \leftarrow [3+, 2-]$, Entropy $(S_{Rain}) = 0.97$.

$S_{strong} \leftarrow [0+, 2-]$, Entropy $(S_{strong}) = 0.0$

$S_{weak} \leftarrow [3+, 0-]$, Entropy $(S_{weak}) = 0.0$

Gain $(S_{Rain}, Wind) = 0.97$

Now,

Gain $(S_{Rain}, Temp) = 0.0192$ } Same.

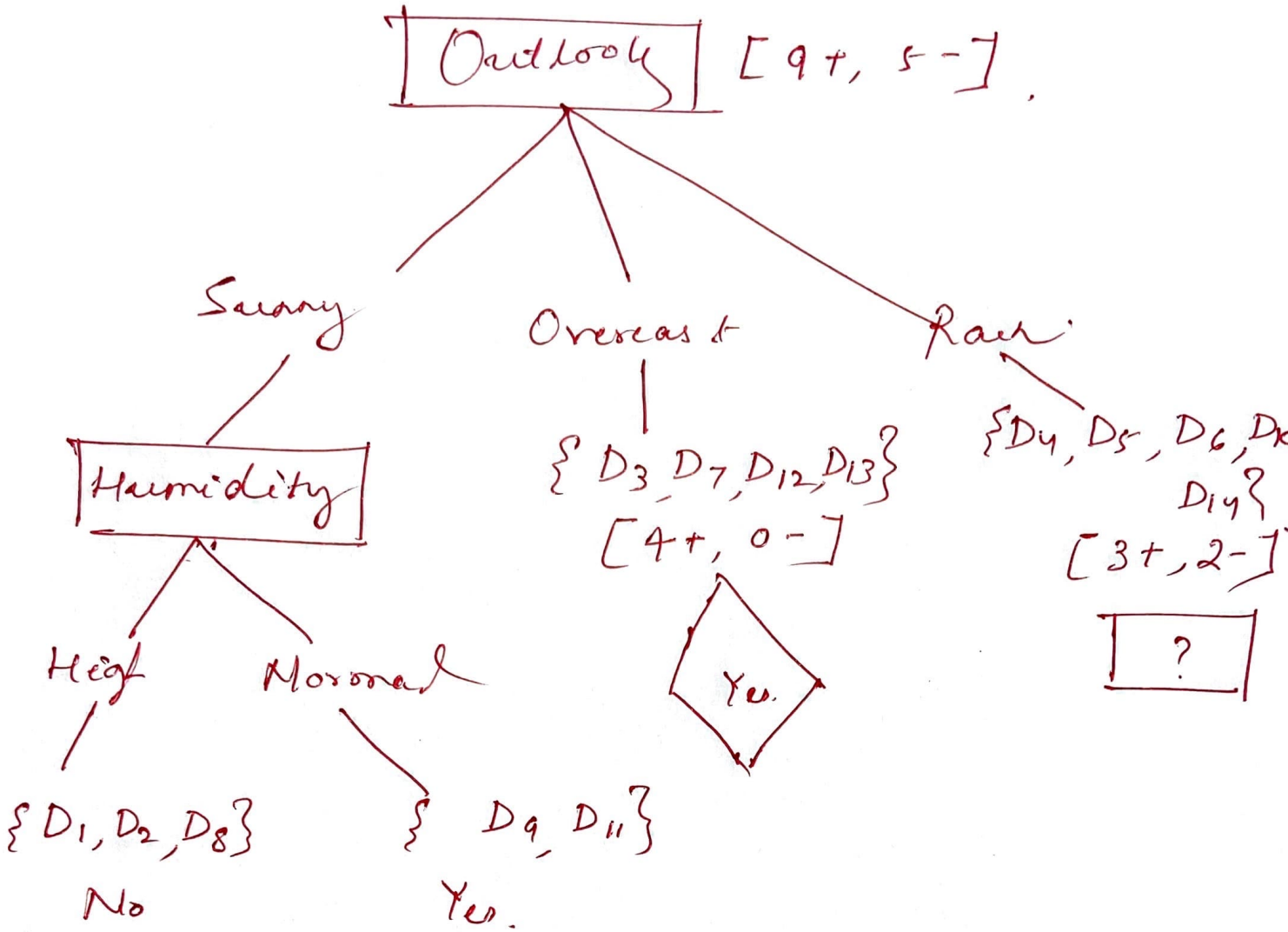
Gain $(S_{Rain}, Humidity) = 0.0192$ }

Gain $(S_{Rain}, Wind) = 0.97$

So, Not desirable

Here, the tree will be 14.

$\{D_1, D_2, \ldots D_{14}\}$



❖ K-MEANS CLUSTERING ALGORITHM

Step-01:

- Choose the number of clusters K.

Step-02:

- Randomly select any K data points as cluster centres.
- Select cluster centres in such a way that they are as farther as possible from each other.

Step-03:

- Calculate the distance between each data point and each cluster center.
- The distance may be calculated either by using given distance function or by using Euclidean distance formula.

Step-04:

- Assign each data point to some cluster.
- A data point is assigned to that cluster whose centre is nearest to that data point.

Step-05:

- Re-compute the centre of newly formed clusters.
- The centre of a cluster is computed by taking mean of all the data points contained in that cluster.

Step-06:

Keep repeating the procedure from Step-03 to Step-05 until any of the following stopping criteria is met-

- Centre of newly formed clusters do not change
- Data points remain present in the same cluster
- Maximum number of iterations are reached

Problem on K-means Clustering:

Cluster the following eight points (with (x, y) representing locations) into three clusters:

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

$$P(a, b) = |x2 – x1| + |y2 – y1|$$

Use K-Means Algorithm to find the three cluster centres after the second iteration.

Solution-

We follow the above discussed K-Means Clustering Algorithm-

Iteration-01:

We calculate the distance of each point from each of the centre of the three clusters.

- The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the centre of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

P(A1, C1)

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |10 – 10|

= 0

Calculating Distance Between A1(2, 10) and C2(5, 8)-

P(A1, C2)

= |x2 – x1| + |y2 – y1|

= |5 – 2| + |8 – 10|

= 3 + 2

= 5

Calculating Distance Between A1(2, 10) and C3(1, 2)-

P(A1, C3)

$= |x2 – x1| + |y2 – y1|$

$= |1 – 2| + |2 – 10|$

$= 1 + 8$

$= 9$

In the similar manner, we calculate the distance of other points from each of the centre of the three clusters.

Next,

- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose centre is nearest to it.

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (5, 8) of Cluster-02	Distance from center (1, 2) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

From here, new clusters are-

Cluster-01:

First cluster contains points-

- A1(2, 10)

Cluster-02:

Second cluster contains points-

- A3(8, 4)
- A4(5, 8)
- A5(7, 5)
- A6(6, 4)
- A8(4, 9)

Cluster-03:

Third cluster contains points-

- A2(2, 5)
- A7(1, 2)

Now,

- We re-compute the new cluster centres.
- The new cluster centre is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

We have only one point A1(2, 10) in Cluster-01.

- So, cluster center remains the same.

For Cluster-02:

Centre of Cluster-02

= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)

= (6, 6)

For Cluster-03:

Centre of Cluster-03

= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5)

This is completion of Iteration-01.

Iteration-02:

- We calculate the distance of each point from each of the centre of the three clusters.
- The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the centre of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

P(A1, C1)

$= |x2 – x1| + |y2 – y1|$

$= |2 – 2| + |10 – 10|$

$= 0$

Calculating Distance Between A1(2, 10) and C2(6, 6)-

P(A1, C2)

$= |x2 – x1| + |y2 – y1|$

$= |6 – 2| + |6 – 10|$

$= 4 + 4$

$= 8$

Calculating Distance Between A1(2, 10) and C3(1.5, 3.5)-

P(A1, C3)

$= |x2 – x1| + |y2 – y1|$

$= |1.5 – 2| + |3.5 – 10|$

$= 0.5 + 6.5$

$= 7$

In the similar manner, we calculate the distance of other points from each of the centre of the three clusters.

Next,

- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose centre is nearest to it.

Given Points	Distance from centre (2, 10) of Cluster-01	Distance from centre (6, 6) of Cluster-02	Distance from centre (1.5, 3.5) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	8	7	C1
A2(2, 5)	5	5	2	C3
A3(8, 4)	12	4	7	C2
A4(5, 8)	5	3	8	C2
A5(7, 5)	10	2	7	C2
A6(6, 4)	10	2	5	C2
A7(1, 2)	9	9	2	C3
A8(4, 9)	3	5	8	C1

From here, new clusters are-

Cluster-01:

First cluster contains points-

- A1(2, 10)
- A8(4, 9)

Cluster-02:

Second cluster contains points-

- A3(8, 4)
- A4(5, 8)
- A5(7, 5)
- A6(6, 4)

Cluster-03:

Third cluster contains points-

- A2(2, 5)
- A7(1, 2)

Now,

- We re-compute the new cluster clusters.
- The new cluster centre is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

Centre of Cluster-01

= ((2 + 4)/2, (10 + 9)/2)

= (3, 9.5)

For Cluster-02:

Centre of Cluster-02

= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)

= (6.5, 5.25)

For Cluster-03:

Centre of Cluster-03

= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5)

This is completion of Iteration-02.

After second iteration, the centre of the three clusters is-

- C1(3, 9.5)
- C2(6.5, 5.25)
- C3(1.5, 3.5)

Calculating probability in Logistic Regression

The dataset of pass/fail in an exam for 5 students is given in the table below. If we use Logistic regression as the classifier & assume the model suggested by the optimizer will become the following for Odds of passing a course.

$$\log(\text{Odds}) = -64 + 2 \times \text{hours}.$$

Hours Studies	Result (1 = Pass, 0 = Fail)
29	0
15	0
33	1
28	1
39	1

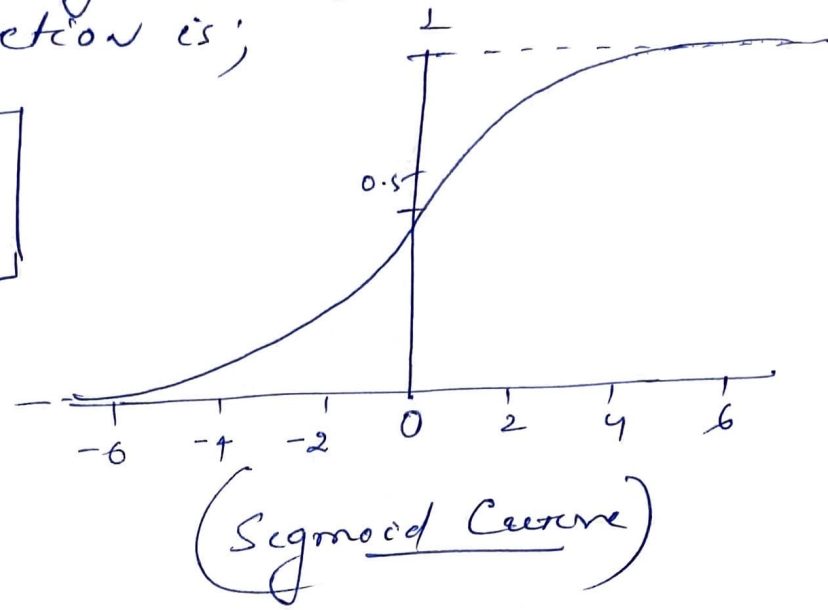
① How to calculate the probability of pass for student who studied 33 hours?

② At least how many hours the student should study that makes sure will pass the course with the probability of more than 95%?

Solution :

As we know ea logistic regression, the sigmoid function is;

$$S(x) = \frac{1}{1+e^{-x}}$$



(Sigmoid Curve)

Q ①

Probability (P)

$$= \frac{1}{1+e^{-z}}$$

(where $z = -64 + 2 \times$ hours) [It is given]

Here hours = 33

$\therefore z = -64 + 2*33$

$= -64 + 66$

$= 2.$

$$P = \frac{1}{1+e^{-2}} \quad (\because z = 2)$$

$= 0.88$

Therefore, a student who studies for 33 hours has a 88% chance of probability of passing the course.

Q(2)

Probability is given

$$P = 0.95 \quad (\text{i.e, } 95\%)$$

Now, already we know,

$$P = \frac{1}{1+e^{-z}}$$

$$\Rightarrow 0.95 = \frac{1}{1+e^{-z}} \quad (\text{# solve for } z)$$

$$\Rightarrow 0.95 \times (1+e^{-z}) = \frac{1}{1+e^{-z}} \times (1+e^{-z})$$

$$\Rightarrow 0.95 \times (1+e^{-z}) = 1$$

$$\Rightarrow 0.95 + 0.95\, e^{-z} = 1$$

$$\Rightarrow 0.95\, e^{-z} = 1 - 0.95$$

$$\Rightarrow 0.95\, e^{-z} = 0.05$$

$$\Rightarrow \frac{0.95\, e^{-z}}{0.95} = \frac{0.05}{0.95} \quad (\because \text{divide } 0.95 \text{ on both sides})$$

$$\Rightarrow e^{-z} = 0.05263$$

$$\Rightarrow \ln|e^{-z}| = \ln|0.05263|$$

$$\Rightarrow -z = -2.94$$

$$\Rightarrow \frac{-z}{-1} = \frac{-2.94}{-1} \quad (\because \text{divide by } -1)$$

$$\Rightarrow \boxed{z = 2.94}$$

$[\because \ln[e^{-z}] = -z]$

Refere back to the log(odds) equation,

$\log(\text{odds}) = -64 + 2 \times \text{hours}$.

i.e, $z = -64 + 2 \times \text{hours}$.

$\Rightarrow 2.94 = -64 + 2 * \text{hours}$.

$\Rightarrow 2\,\text{hours} = 2.94 + 64$

$\Rightarrow \frac{2\,\text{hours}}{2} = \frac{66.94}{2}$ // divide by 2

$\Rightarrow \text{hours} = 33.47$.

$\simeq 33.5$ hours. ✓

[Verify, $z = -64 + 2 * 33.5$

$= -64 + 67$

$= 3$

$P = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-3}} \simeq 0.952$.]

i.e, student needs to study for approximately 33.5 hours. to achive 95%.

—

Problem on multiple linear regression with one or more independent variables

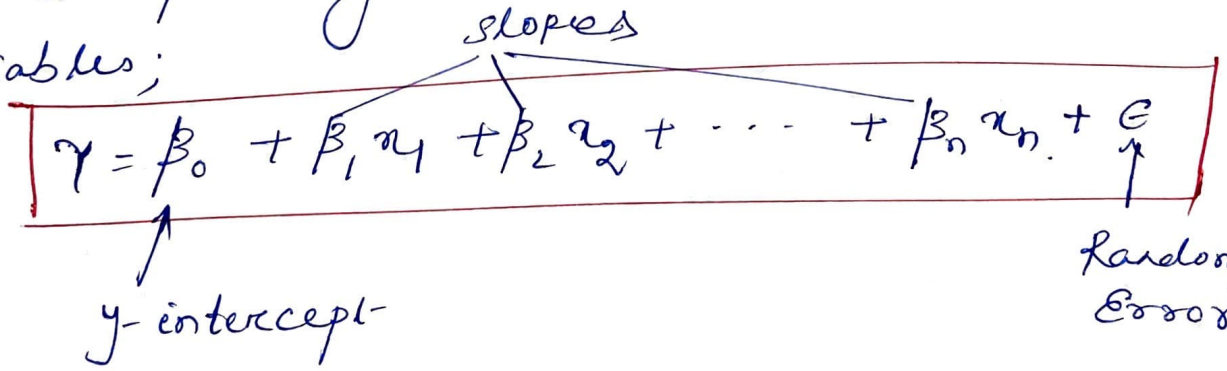
Question

Predict the value of Y given x_1 and x_2

Subject	Y	x_1	x_2
1	-3.7	3	8
2	3.5	4	5
3	2.5	5	7
4	11.5	6	3
5	5.7	2	1
6	?	3	2

Solution :

→ Multiple regression model with k- independent variables;



→ Here, linear regression with 2 independent variables,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Where,

$$\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$$

$$\beta_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$\beta_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

Now,

$$\Sigma x_1^2 = \Sigma x_1 \cdot x_1 - \frac{(\Sigma x_1)(\Sigma x_1)}{N}$$

$$\Sigma x_2^2 = \Sigma x_2 \cdot x_2 - \frac{(\Sigma x_2)(\Sigma x_2)}{N}$$

$$\Sigma x_1 y = \Sigma x_1 \cdot y - \frac{(\Sigma x_1)(\Sigma Y)}{N}$$

$$\Sigma x_2 y = \Sigma x_2 \cdot y - \frac{(\Sigma x_2)(\Sigma y)}{N}$$

$$\Sigma x_1 x_2 = \Sigma x_1 \cdot x_2 - \frac{(\Sigma x_1) \cdot (\Sigma x_2)}{N}$$

Now,

Subject	Y	x_1	x_2	x_1x_1	x_2x_2	x_1x_2	x_1y	x_2Y
1	-3.7	3	8	9	64	24	-11.1	-29.6
2	3.5	4	5	16	25	20	14	17.5
3	2.5	5	7	25	49	35	12.5	17.5
4	11.5	6	3	36	9	18	69	34.5
5	5.7	2	1	4	1	2	11.4	5.7
Σ	19.5	20	24	90	148	99	95.8	45.6

So, $$\beta_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= 2.28$$

$$\beta_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= -1.67$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$$

$$= \frac{19.5}{5} - \frac{2.28 \times 20}{5} - \frac{-1.67 \times 24}{5}$$

$$= 2.796$$

Final regression equation or model is

$$Y = 2.796 + 2.28 x_1 + (-1.67) x_2$$

Now given $x_1 = 3$, $x_2 = 2$, $Y = ?$

$$Y = 2.796 + 2.28 \times 3 - 1.67 \times 2$$
$$= 6.296.$$

PROBLEM ON SIMPLE LINEAR REGRESSION

Problem:1

Find a quadratic regression model for the following data:

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

Solution:

Let the simple linear regression model be

$$y = a + bx$$

Steps to find a and b,

First, find the mean and covariance.

Means of x and y are given by,

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{y} = \frac{1}{n}\sum y_i$$

The variance of x is given by,

$$\operatorname{Var}(x) = \frac{1}{n-1}\sum (x_i - \bar{x}_i)^2$$

The covariance of x and y, denoted by Cov(x, y)is defined as,

$$\operatorname{Cov}(x, y) = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$$

Now the values of a and b can be computed using the following formulas:

$$b = \frac{\operatorname{Cov}(x, y)}{\operatorname{Var}(x)}$$

$$a = \bar{y} - b\bar{x}$$

First, find the mean of x and y,

$$n = 5$$

$$\bar{x} = \frac{1}{5}(1.0 + 2.0 + 3.0 + 4.0 + 5.0)$$

$$= 3.0$$

$$\bar{y} = \frac{1}{5}(1.00 + 2.00 + 1.30 + 3.75 + 2.25)$$

$$= 2.06$$

Next, find the Covariance between x and y,

$$\mathrm{Cov}(x, y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

$$\mathrm{Cov}(x, y) = \frac{1}{4}[(1.0 - 3.0)(1.00 - 2.06) + \cdots + (5.0 - 3.0)(2.25 - 2.06)]$$

$$= 1.0625$$

Now find the variance of x,

$$\mathrm{Var}(x) = \frac{1}{n-1}\sum(x_i - \bar{x}_i)^2$$

$$\begin{aligned}\text{Var}\,(x) &= \frac{1}{4}[(1.0-3.0)^2+\cdots+(5.0-3.0)^2] \\ &= 2.5\end{aligned}$$

Now, find the intercept and coefficients,

$$b = \frac{\text{Cov}\,(x,y)}{\text{Var}\,(x)}$$

$$a = \bar{y} - b\bar{x}$$

$$\begin{aligned}b &= \frac{1.0625}{2.5} \\ &= 0.425\end{aligned}$$

$$\begin{aligned}a &= 2.06 - 0.425 \times 3.0 \\ &= 0.785\end{aligned}$$

Therefore, the linear regression model for the data is,

$$y = 0.785 + 0.425x$$

PROBLEM ON SIMPLE LINEAR REGRESSION WITH ONE INDEPENDENT VARIABLE

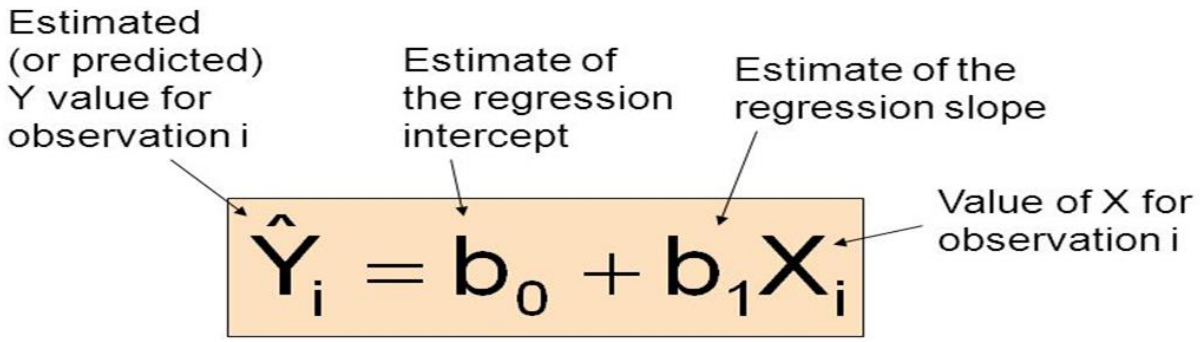
Problem:2

how to use a regression equation to predict the glucose level given the age.

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81
7	55	?

Solution:

However, there can be only one dependent variable (y) in the regression equation.



Here we need to find the value of bo, b1 using the following equation.

$$b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

<u>Step 1</u>: Make a chart of your data, filling in the columns in the same way as you would fill in the chart if you were finding the Pearson's Correlation Coefficient

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

<u>Step 2:</u> Use the following equations to find b0 and b1.

Find b0:

$$b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b_0 = \frac{(486)(11409) - (247)(20485)}{6\,(11409) - (247)^2}$$

$$b_0 = \frac{4848979}{7445} = 65.14$$

Find b1:

$$b_1 = \frac{n\,(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b_1 = \frac{6\,(20485) - (247)(486)}{6(11409) - (247)^2}$$

$$b_1 = \frac{2868}{7445} = 0.385335$$

<u>Step 3</u>: ***Insert the values into the equation.***

$y' = b_0 + b_1 * x$

$y' = 65.14 + (0.385225 * x)$

<u>Step 4:</u> Prediction – the value of y for the given value of x = 55

$y' = 65.14 + (0.385225 * 55)$

$y' = 86.327$

PROBLEM ON MULTIPLE LINEAR REGRESSION WITH ONE OR MORE INDEPENDENT VARIABLES

Problem:3

Predict the value of Y given X_1 and X_2

SUBJECT	Y	X_1	X_2
1	-3.7	3	8
2	3.5	4	5
3	2.5	5	7
4	11.5	6	3
5	5.7	2	1
6	?	3	2

Predict the value of Y given X1 and X2

- Finding the values of *b* (the slopes) is tricky for k>2 independent variables, and you really need matrix algebra to see the computation
- At this point, you should notice that all the terms from the one variable case appear in the two variable case.
- In the two variable case, the other X variable also appears in the equation. For example, X_2 appears in the equation for b_1. s.

The Multiple Regression Model

Idea: Examine the linear relationship between 1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with k Independent Variables:

Y-intercept | Population slopes | Random Error

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Linear Regression 2 independent variable

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Linear Regression 2 independent variable

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N}$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N}$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$

Linear Regression 2 independent variable

SUBJECT	Y	X_1	X_2	X_1X_1	X_2X_2	X_1X_2	X_1Y	X_2Y
1	-3.7	3	8	9	64	24	-11.1	-29.6
2	3.5	4	5	16	25	20	14	17.5
3	2.5	5	7	25	49	35	12.5	17.5
4	11.5	6	3	36	9	18	69	34.5
5	5.7	2	1	4	1	2	11.4	5.7
∑	19.5	20	24	90	148	99	95.8	45.6

Linear Regression 2 independent variable

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{32.8 * 17.8 - 3 * (-48)}{10 * 32.8 - 3 * 3} = 2.28$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{10 * (-48) - 3 * 17.8}{10 * 32.8 - 3 * 3} = -1.67$$

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 = \frac{19.5}{5} - \frac{2.28 * 20}{5} - \frac{-1.67 * 24}{5} = 2.796$$

Linear Regression 2 independent variable

Final Regression equation or Model is:

$$\boldsymbol{Y = 2.796 + 2.28\, x_1 - 1.67 x_2}$$

Now given $x_1 = 3$ *and* $x_2 = 2$ $Y = ?$

$$Y = 2.796 + 2.28 * 3 - 1.67 * 2$$
$$= \mathbf{6.296}$$