Class 8

Data Mining

Decision tree Algo

age select করা হয় Info. gain, Gini, Gain Ratio ইত্যাদি,

Info. Gain

Class same না হলে bias হয় output হয় entropy

বের করা হয়,

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

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

$$= 0.9402$$

∴ Perfectly imbalance হলে '0'

(যদি সব same class হয়)

$$IG(D, A) =$$

↓

Attribute

$$IG(S, Outlook) = H(S) - \sum_{v \in \{S,O,R\}} \frac{|S_v|}{|S|} \cdot H(S_v)$$

Cardinality কতগুলো Data আছে

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

$$-\left(\frac{5}{14} \times 0.97\right) + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97$$

Sunny | Overcast | Rain

$$= 0.247$$

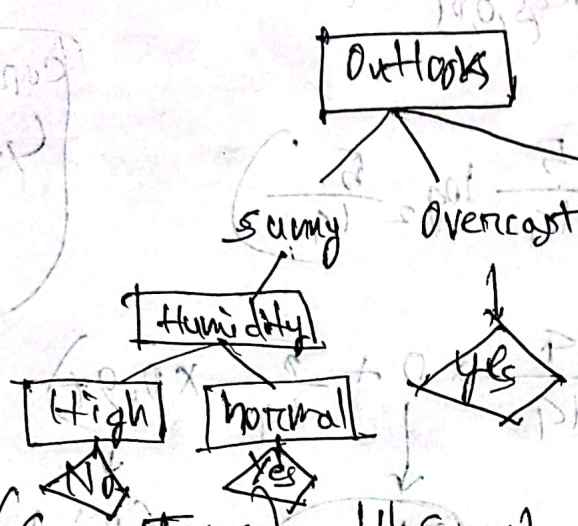
$$IG(S, Humidity) = 0.9402 - \left\{\left(\frac{7}{14} \times 0.985\right) + \left(\frac{7}{14} \times 0.916\right)\right\}$$

$$= 0.1519$$

$$IG(S, Temp) = 0.9402 - \left\{\frac{4}{14} \times \left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{6}{14}\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) + \frac{4}{14}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)\right\}$$

$$=$$

① Outlook has highest IG of root node.



$$IG(S_{sunny}, Temp) = H(Sunny) - \sum_{v \in \{H,N\}} \frac{|S_v|}{|S|} H(S_{sunny_v})$$

$$= \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) - \left(\frac{2}{5}\times 0 + \frac{2}{5}\times 1 + \frac{1}{5}\times 0\right)$$

$$= 0.97 - 0.4$$

$$= 0.57$$

high ↓ normal ↓

$$IG(Sunny, Humidity) = 0.97 - \left(\frac{3}{5}\times 0 + \frac{2}{5}\times 0\right)$$

$$= 0.97$$

$$IG(Sunny, Wind) = 0.97 - \left(\frac{2}{5}\times 1 + \frac{3}{5}\times 0.92\right)$$

$$= 0.97 - 0.952$$

$$= 0.018$$

∴ IG(Sunny, Humidity) is accepted, because it is high.

Class 3

Data mining

Decision tree

Gini এর ক্ষেত্রে যত কম তত ভাল,

Gini Index:

$$G(S) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$G(\text{Outlook}) = 0.48 \times \frac{5}{14} + 0 + 0.48 \times \frac{5}{14} = 0.342$$

$$G(\text{Outlook} \to \text{Sunny}) = 1 - \left\{\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right\} = 0.48$$

$$G(\text{Outlook} \to \text{Overcast}) = 1 - \left\{\left(\frac{4}{4}\right)^2\right\} = 0$$

$$G(\text{Outlook} \to \text{Rain}) = 1 - \left\{\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right\} = 0.48$$

Wind

$$G(\text{Wind}) = 0.375 \times \frac{8}{14} + 0.5 \times \frac{6}{14} = 0.43$$

$$G(\text{Wind} \to \text{weak}) = 1 - \left\{\left(\frac{6}{8}\right)^2 + \left(\frac{2}{8}\right)^2\right\} = 0.375$$

$$G(\text{wind} \to \text{strong}) = 1 - \left\{\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right\} = 0.5$$

$\therefore G(\text{Humidity}) = 0.36$

$$G(T \to Hot) = 1 - \left\{\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right\} = 0.5$$

$$G(T \to Mild) = 1 - \left\{\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right\} = 0.44$$

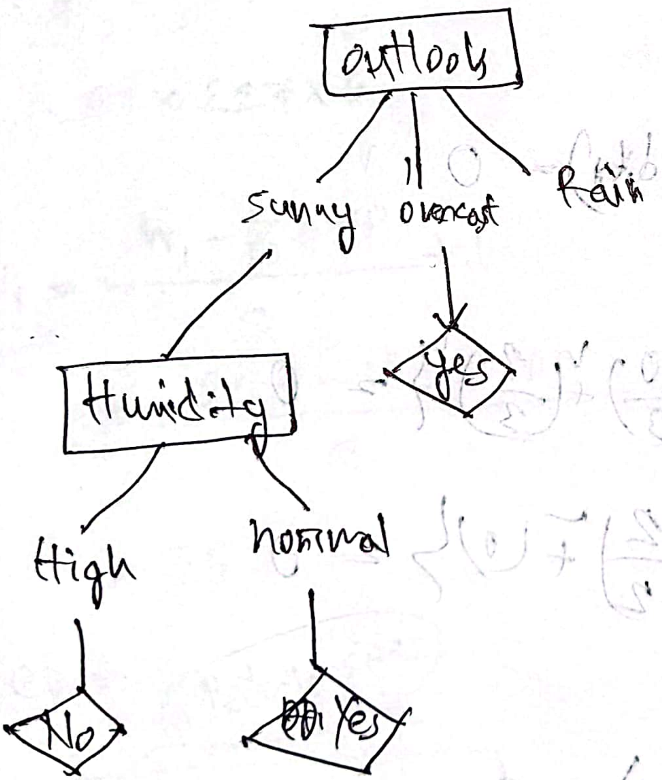
$$G(T \to cool) = 1 - \left\{\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right\} = 0.375$$

$$G(Temp) = 0.5 \times \frac{4}{14} + 0.44 \times \frac{6}{14} + \frac{4}{14} \times 0.375$$

$$= 0.44$$

We take outlook.

Overcast 'yes' fixed থাকত কারণ Overcast এ target class 'yes' হত।



$G(\text{Outlook} = \text{Sunny} \rightarrow \text{Temp}) = 0.2 \times \frac{2}{5} + 0.4 \times \frac{2}{5} + 0.4 \times \frac{1}{5}$

$\Rightarrow 0.2$

$G(O = S \rightarrow T = H) = 1 - \left\{\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right\} = 0$

$G(O = S \rightarrow T = M) = 1 - \left\{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right\} = 0.5$

$G(O = S \rightarrow T = C) = 1 - \left(\frac{1}{1}\right)^2 = 0$

Humidity

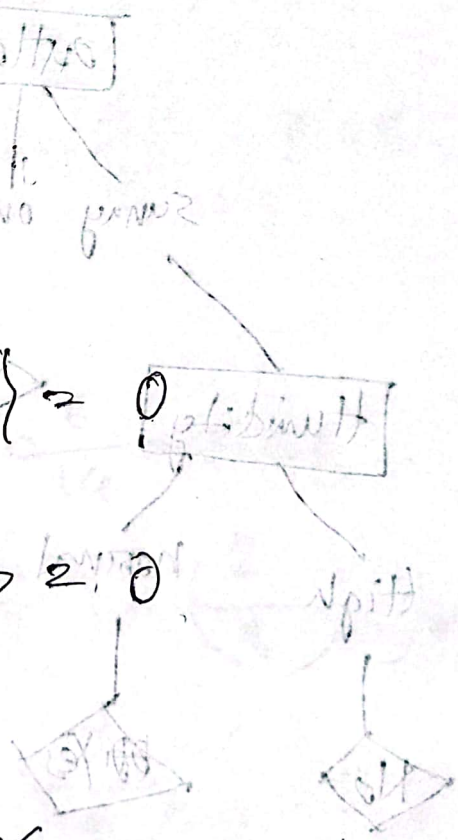
$G(\text{Outlook} = \text{sunny} \rightarrow \text{Humidity}) = 0$

$G(O = S \rightarrow H = H) = 1 - \left\{\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right\} = 0$

$G(O = S \rightarrow H = N) = 1 - \left\{\left(\frac{2}{2}\right)^2 + 0\right\} = 0$

Wind

$G(\text{Outlook} = \text{sunny} \rightarrow \text{wind}) = 0.4 \times 0.5 + 0.4 \times 0.5$



Segmentation
Image Compression
Morphological

Class 4

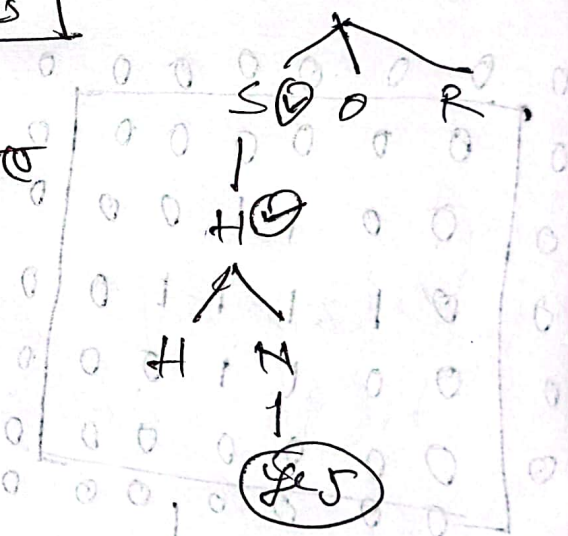
Data Mining

Decision tree

Sunny, Mild, Strong, Normal, Yes

Unknown data ~~দাও~~ Predict করতে

Probability
Conditional Probability



P(A|B)

B event হলে A এর Possibility

P(A|B) → Feature / attribute

↓

class

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \longrightarrow P(A|B)P(B) = P(A \cap B)$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \longrightarrow P(B|A)\,P(A) = P(B \cap A)$$

$$\boxed{P(A|B) = \frac{P(A) * P(B|A)}{P(B)}}$$

↳ Baysian formula.

Naive Baysian :

$P(A|B)$

C_1 C_2

Class

$A = \{yes, No\}$

$B = \{B_1, B_2, \ldots, B_n\}$ → value of feature

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

$$P(yes|B) = \frac{P(yes) * P(B|yes)}{P(B)} = P(yes) * P(B_1|No) * P(B_2|$$

$$P(No|B) = \frac{P(No) * P(B|No)}{P(B)}$$

$$\therefore P(yes|B) = \frac{P(yes) * P(B_1|yes) \cdots * P(B_n|yes)}{P(B_1) \cdot P(B_2) \cdots P(B_n)}$$

$$\therefore P(No|B) = \frac{P(No) \cdot P(B_1|No) \cdot P(B_2|No) \cdots P(B_n|No)}{P(B_1) \cdot P(B_2) \cdots P(B_n)}$$

∴ 2 ta class এর মধ্যে 10 টি class এর same হবে,

$B = \{32, High, No, excellent\}$

$P(yes|32, High, no, excellent) = P(yes) | P(32|yes) \cdot P(High|yes) * P(no|yes) \cdot P(excellent|yes)$

$= \left(\frac{9}{14}\right) * \left(\frac{4}{9}\right) * \left(\frac{2}{9}\right) * \left(\frac{3}{9}\right) * \left(\frac{3}{9}\right)$

$= 0.00705$

$P(No|32, High, no, excellent) = P(No) | P(32|No) \cdot P(High|No) \cdot P(no|no) \cdot P(excellent|no)$

$= \left(\frac{5}{14}\right) \cdot \left(\frac{0}{5}\right) \cdot \left(\frac{2}{5}\right) \left(\frac{1}{5}\right) \left(\frac{2}{5}\right)$

$= 0$

এটা তো not accepted

Class 4

Data Mining

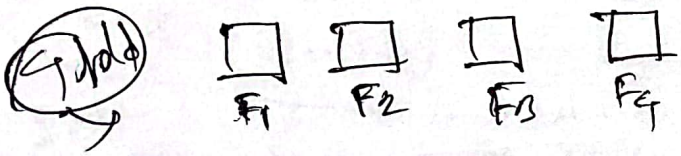
Naive bayes → adv / disadv.

- k-fold, cross validation, hold out method, Bootstrap...

↓

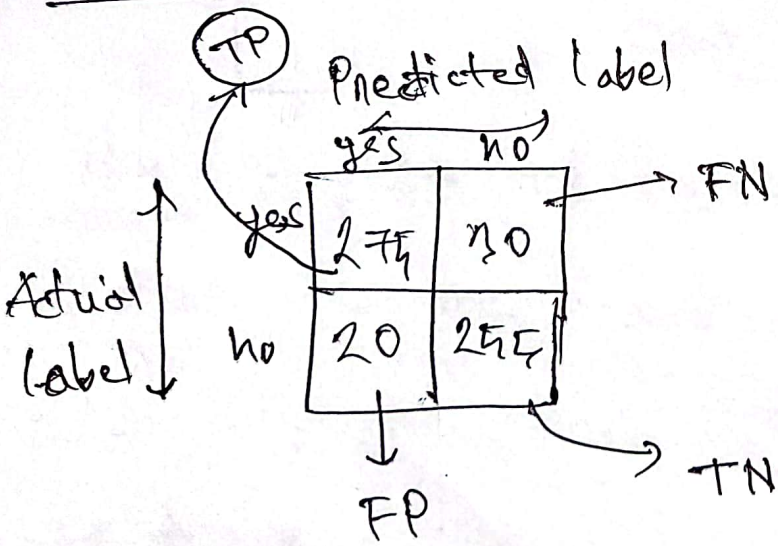
Define.

k-fold



Train	Test
F_1, F_2, F_3	F_4
F_1, F_2, F_4	F_3
F_1, F_3, F_4	F_2
F_1, F_2, F_4	F_3

Confusion



$$Acc = \frac{TP + TN}{TP + TN + FN + FP}$$

$$= 0.91$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Predict করছে Positive

$$\text{Recall / Sensitivity} = \frac{TP}{TP + FN}$$

আসলে যদি কোনো হয়, আমাদের মডেল FN কম হবে যা অনেক ভালো।

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\therefore F_1 \text{ score} = \frac{2 \times \text{Prec} * \text{Recall}}{\text{Prec} + \text{Recall}}$$

4 class

Class no 0

0 কে Negative predict করছে,

Predicted (columns), Actual (rows)

Actual \ Predicted	0	1	2	3
0	TP_0	FN_0	FN_0	FN_0
1	FP_0	TP_1	TN_0	TN_0
2	FP_0	TN_0	TP_2	TN_0
3	FP_0	TN_0	TN_0	TP_3

0 → Prediction = 0 → TP

0 → Prediction = 1 → FN

Class no 1

	0	1	2	3
0		FP_1		
1	FN_1	TP_1	FN_1	FN_1
2		FP_1		
3		FP_1		

Class no 2

	0	1	2	3
0			FP_2	
1			FP_2	
2	FN_2	FN_2	TP	FN
3			FP_2	

Class 6

Data Mining

clustering → (Unsupervised)

soft cluster → 1 টি data multiple cluster এ থাকতে পারে।

Hard → 1 টি data 1 টি cluster এ থাকে।

Partition based → K-means

K=2 মানে 2 ধরণের Data থাকবে।

Problem → k value আগে থেকে select করতে হয়, যা Problem

তৈরি করে।

সব Data include করা যায় না; কিছু noise

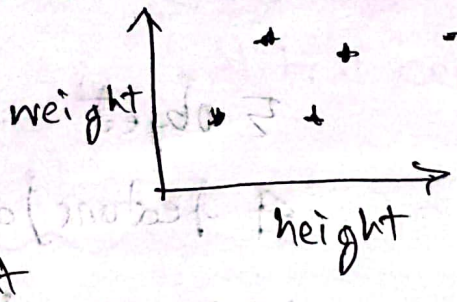
create করে,

Solu → Elbow method.

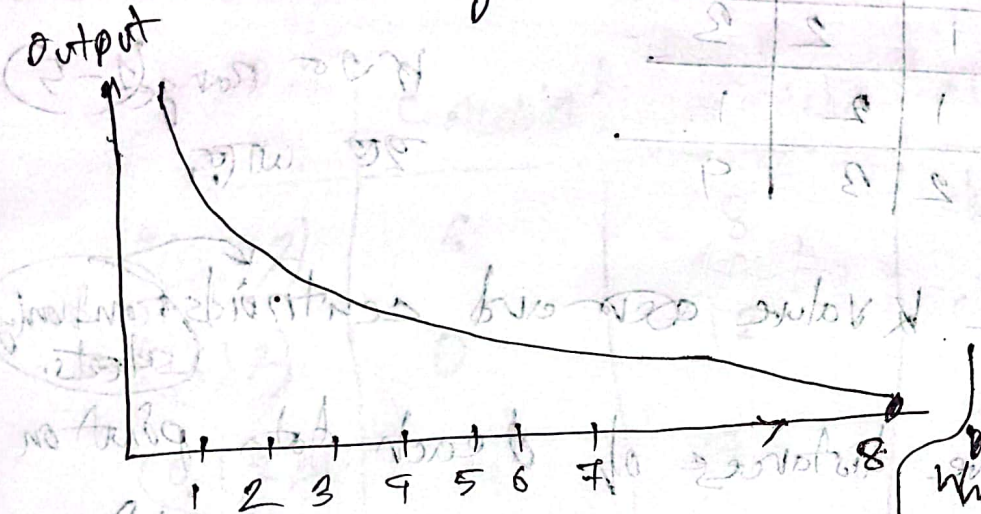
→ Silhouette

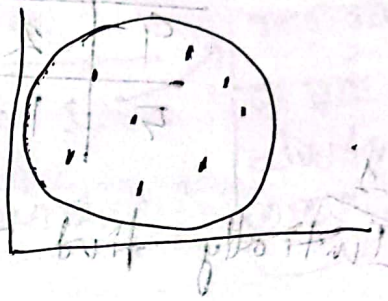
Partitioning | Hierarchical | Density

Elbow



→ k এর মান 1 থেকে 9 এর মধ্যে 2টি ভাগে,





$W_{max} = d_1^2 + d_2^2 + d_3^2 + \dots + d_8^2$

For k=1,

For k=2 হলে,

$W_{ess_1} + W_{ess_2} = W_{ess}$

2 group

সমান আসতে slowly decrease হবে তাই হলো elbow point.

K-means

ID	A_1	A_2	A_3	A_4
1	5	9	3	2
2	7	6	1	2
3	0	1	2	3
4	2	1	2	1
5	1	2	3	4

5 object
4 feature / attribute
k=? Range (1-5)
2 cluster

step-1 Initially find the k value and centroids randomly selects.

step-2 calculate the distances of each data point on objects from all k number of centroids (all centroids)

step-3 → Assign the points come in to their corresponding clusters. (data obj)

step-4: Repeat step (2 & 3) until the clusters are stable.

(k value = centroids)

Centroids (2, 4), $K = 2$

Lets, the initial centroids be (7, 6, 1, 2) & (2, 1, 2, 1)

For Iteration 1:

Data obj	Centroid1	centroid2	cluster
(5, 4, 3, 2)	6	8	cluster 1
(7, 6, 1, 2)	0	12	cluster 1
(0, 1, 2, 3)	14	4	cluster 2
(2, 1, 3, 1)	12	0	cluster 2
(1, 2, 3, 4)	14	6	cluster 2

New centroid,

$$\text{new-centroid 1} = \left(\frac{7+5}{2}, \frac{6+4}{2}, \frac{1+3}{2}, \frac{2+2}{2}\right)$$

$$= (6, 5, 2, 2)$$

$$\text{new-centroid 2} = \left(\frac{0+2+1}{3}, \frac{1+1+2}{3}, \frac{2+2+3}{3}, \frac{3+1+4}{3}\right)$$

$$= (1, 1.33, 2.33, 2.66)$$

complete the iteration, যদি iteration match হয় তাহলে যদি change না হয় তাহলে Answer হবে।

Training করলে → (Draw and loss function)

— X —

K-mean
- Limitation
- Advantage
- Application

Class 7

Data Mining

Density based clustering (DBSCAN)

Density এর উপর clustering করা হয়।

→ Location wise clustering হয়।

∴ two parameter → epsilon and minPoints.

circle এর Radius

ε = 1
minA = 3 → 3 টি point আছে

1. core point, Border point, and Noise.

- core point → যদি 3 এর বেশি point থাকে তাহলে center point হবে / core point
- Border point → যাকে ঘিরে core point হতে But minA = 3 এর থেকে কম
- Noise → যদি শুধু একটি থাকে

2. Trial and error method ব্যবহার করে epsilon (ε) বের করা হয়।

minPoint → Dimention + 1
(Dimention → feature/attribute)

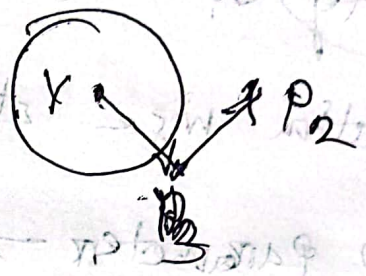
Reachability and Connectivity

X is direct density-reachable from point Y.

Y → X

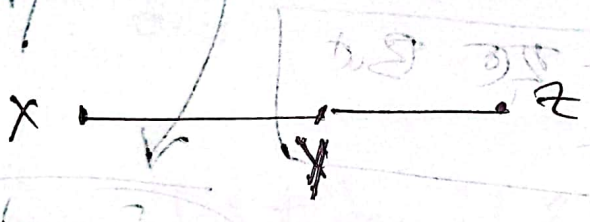
① Y से P2 indirectly density-reachable.

② X से Y Density reachable.



Density connected

X से Y और Y से Z Same distance reachable.



Algorithm steps

①

Example:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	0								
x_2	1	0							
x_3	1.41		0						
x_4				0					
x_5					0				
x_6						0			
x_7							0		
x_8								0	
x_9									0

euclidian → $d_{x_1 x_2} = \sqrt{(0-1)^2 + (0-0)^2} = 1$

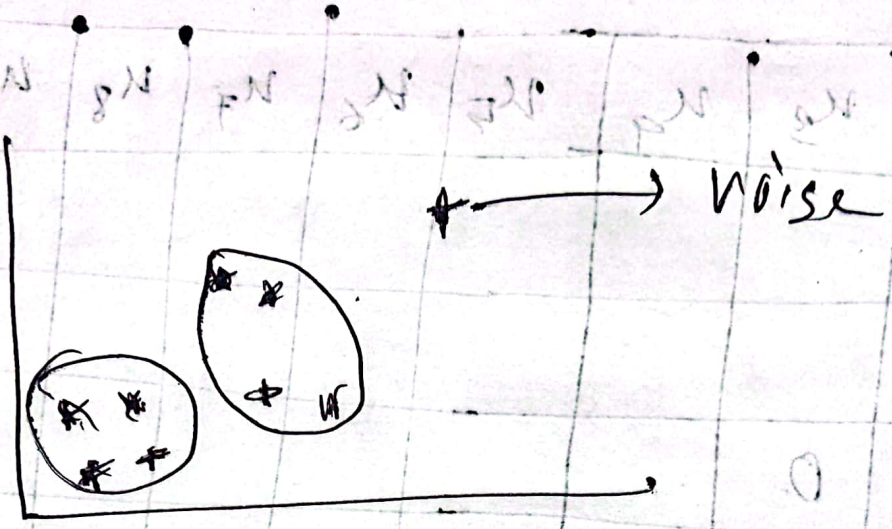
(★) যদি ∈ এর মান ~~অন্যান্য~~ distance থেকে বেশি হয় তবে নিজে neighbourhood তৈরী করবে।

min point = 3

(★) x_4, x_8 → border point, x_9 = noise.

(★) এর মাঝে Directly and density reachable রয়েছে

এটি কাজ clustering করছে দেখা যাচ্ছে,



(৪) ২টি cluster.

YOLO

→ less function. (मतलब का फीचर नहीं)

→ training

transfer

→ ~~Real~~ ≠ step. transfer learning.

Class 8 edureka

Data Mining

Data ~~Mining~~ Warehouse

"Technology use OLAP"

meta data → data ~~or~~ info

- Meta Data → Data Info
- Aggregate → सारांश रूपी
- Raw data → data

OLAP

- Multidimensional data.
- OLAP vs OLTP → 2 dimensional

Types of OLAP

MOLAP vs ROLAP vs HOLAP

OLAP operation

- Roll up
- Drill Down → multidimentional
- Slice → 2D,
- Dice → sub-cube বিশ্লেষণ
- Pivot → Transpose matrix ও concept,
  (Rotated করে)

② Dimensions

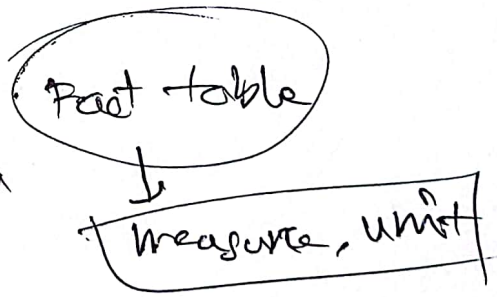
③ Facts and Measure

↳ Dimension key measure.

Schemas

Star, Snowflake,

Star schema → 1 fact table.



Snowflake →

Galaxy schema → more than 1 fact table.

(Figure দিয়ে schema explain করা)