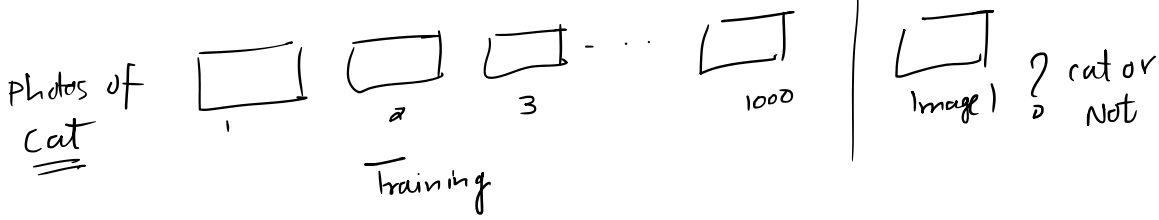
Module 5

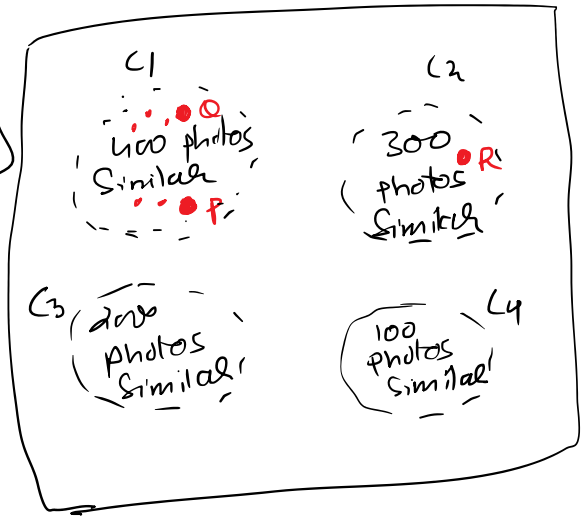
Clustering →

In earlier case → data was labelled.



Supervised learning

Now Consider → 1000 photos (not labelled)



Clustering is the task of dividing the population or data points into no of groups such that the data points in the same group are more similar to other data points in same group than those in other group.

* Aim is to Segregate groups with Similar traits/features and assign them into clusters.

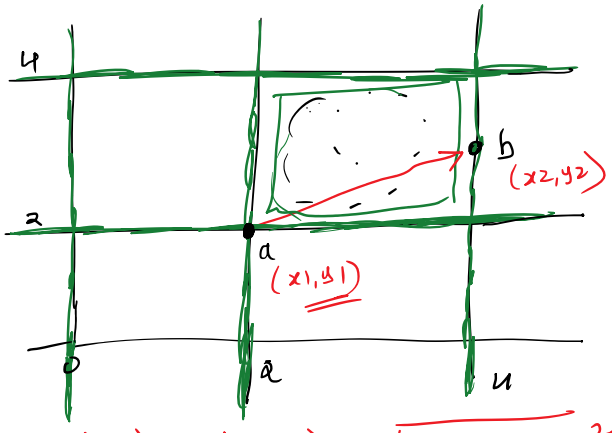
Distance Metric : It is a parameter that tells how close two points are

Given element a, b, c in a set, a distance metric is defined as a function with following properties:

1. Non-negativity → $d(a,b) \geq 0$
2. Symmetry → $d(a,b) == d(b,a)$
3. Triangle Inequality $d(a,c) \leq d(a,b) + d(b,c)$. $\forall a,b,c \in S$

Different Distance Metric

1. **Euclidean Distance** → Most commonly used.
   → When data is dense or continous, this is the best proximity measure.



Euclidean Distance (vector) $= d(a,b) = \sqrt{\sum_i (a_i - b_i)^2}$

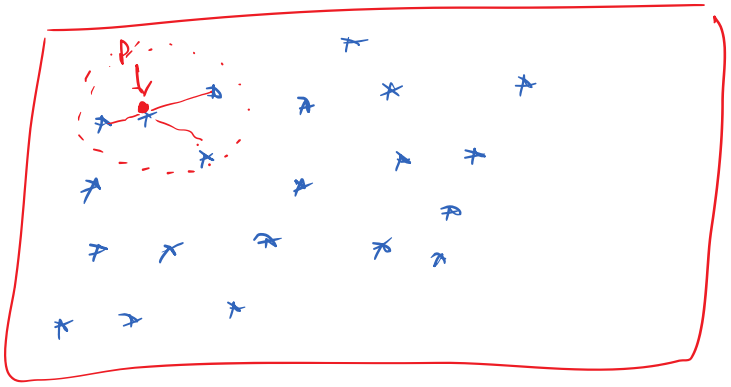
$= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

If we a point $p = (p_1, p_2 \ldots p_n)$ and $q = (q_1, q_2 \cdots q_n)$.

$d(p,q) = d(q,p) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$

$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$

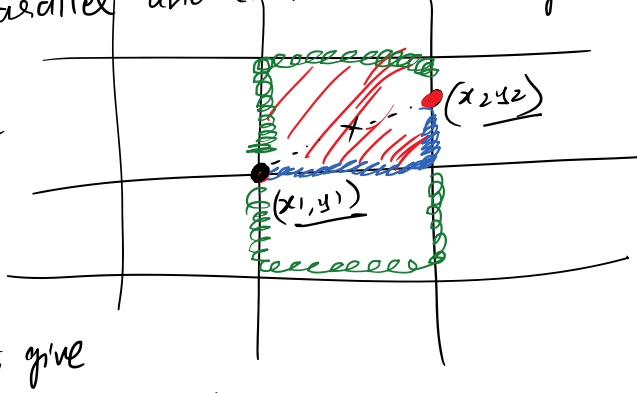
$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$



Manhattan Distance

* If we are trying to find the distance bet^n two buildings that are separated by several blocks. We can only walk through sidewalk that are parallel and cannot walk diagonally through buildings

In such case we cannot use Euclidean distance. Here Euclidean distance will not give realistic estimate for the distance



Manhattan distance.

$= |x_1 - x_2| + |y_1 - y_2|$

* Manhattan Distance is a metric in which the distance between two points is the sum of absolute difference of their Cartesian coordinates.

* Total sum of difference bet^n the x coordinates & y coordinates.

* Manhattan distance metric is also known as Manhattan length, Rectilinear distance, L_1 distance, L_1 Norm, City block distance taxi cab metric.,

* If the data points has n features.

Manhattan Distance $d(a,b) = |a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3) \cdots + |a_n - b_n|$

$$= \sum_{i=1}^{n} |a_i - b_i|$$

Syllabus
- Graph Based Clustering Using Minimum Spanning Tree
- Estimate Maximization Algo for Clustering
- DBSCAN

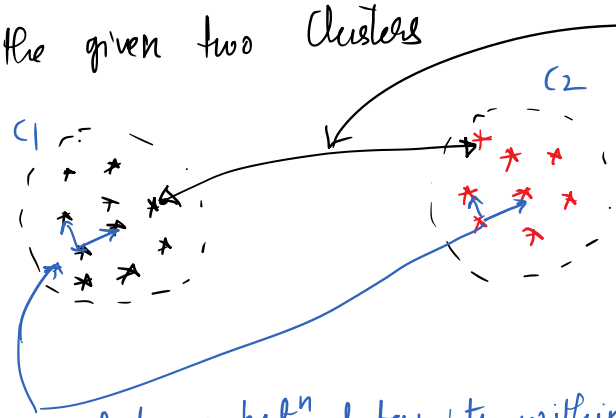
(1) Graph Based Clustering →

(1) First Consider How to Construct Graph of given data.

→ k Neighbourhood graph. ✓

Epsilon Neighbourhood graph ✓

Consider the given two Clusters

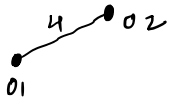


Distance betⁿ data points betⁿ diff clusters are maximum.

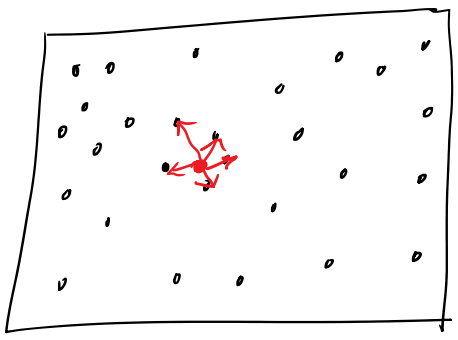
Here the distance betⁿ datapoints within the cluster is Minimum.

* In Graph Based Clustering the objects are represented as nodes in graph.

* The weight of edge / branch betⁿ the objects defines the distance betⁿ the objects



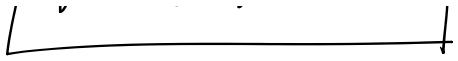
(1) k Neighbourhood Approach.



* To find distance of a point with all other points will be computationally challenging.
* we will consider nearest k points

let k=5.

So we will calculate distance of nearest 5 data points and will be



n = 1000

So we will calculate distance ... nearest 5 data points are will be labelled as Nearest Neighbour of node

Challenge → Choosing K value is challenging task in graph Construction.

(2) Epsilon Neighbourhood Graph → Objects within a radius of epsilon from a given object are considered nearest neighbour.

Here P_1 & P_2 are neighbours of P.

Challenge :→ Selecting a proper epsilon value.

* Graph Based Clustering Using Minimum Spanning Tree →

Algo

1. Determine MST of given graph. ✓
2. Delete branch iteratively. ✓
3. Each Connected Component is a cluster ✓

→ Different Stralgies to Delete Branches.

(1) Delete Branch with max weight →. ✓
(2) Delete Inconsistent Branch → ✓
✓ (3) Delete by Analysis of weight.

(1) Deletion of Branch with max weight

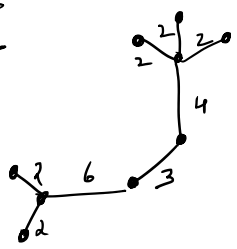
... two clusters by deleting the branch with

(1) Deletion of Branch with [illegible]

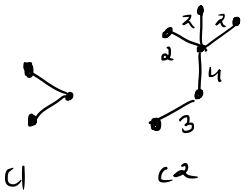
1. In each step create two clusters by deleting the branch with max weight
2. Repeat until the desired no of clusters is reached

no of cluster desired is 3

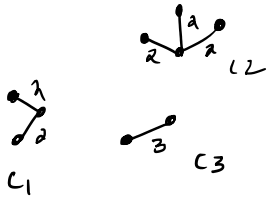
Consider MST



Step 1) delete edge with weight 6



Step 2) delete edge with weight 4.

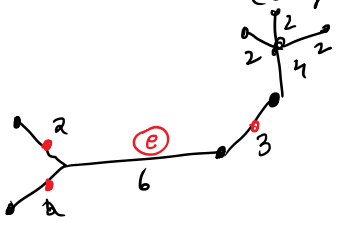


3 clusters

(2) Delete Inconsistent Branches.

* A branch is inconsistent if the corresponding weight (de) is much larger than the reference value $\overline{de}$
* The reference value $\overline{de}$ can be defined by the average weight of all the branches (edges) adjacent to edge e

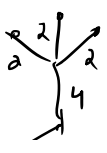
Consider MST:

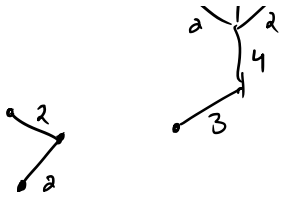


Here $\overline{de} = \frac{2+2+3}{3} = \frac{7}{3} = 2.3$

Here $de = 6 > \overline{de}$

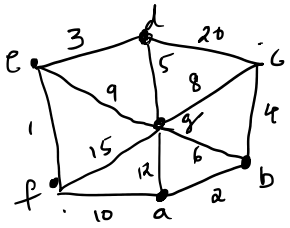
So delete the edge e.



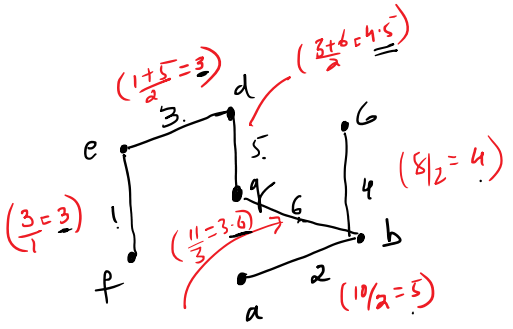


2 clusters

Q) Construct MST and provide Clustering of graph Using Inconsistent branches.



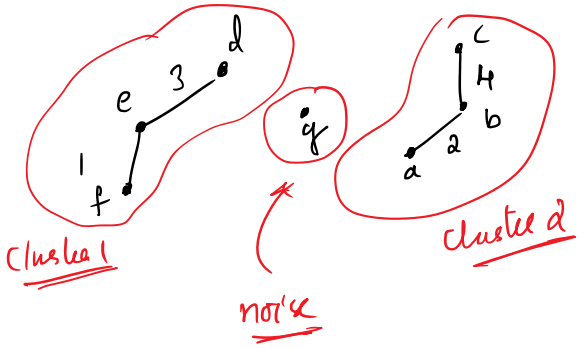
Step 1) MST of the Graph.



Step 2) For dilation we are using Inconsistent branch approach.

Here edge with 5 > reference value (4.5) so delete edge 5

Also edge with 6 > reference value (3.6) so delete edge 6.



(2) Estimation Maximization Algo

→ This Algo is for Maximum Likelihood Estimation in presence of hatent variable (missing)

→ The most discussed application of EM Algorithm is used for Clustering with Mixture Model. (where data points have missing value)

* Estimation (E-step) → Estimate are expectations for machine & then classify the data into some classes.

* Maximization (M-step) → Whatever we estimated should now be maximized.

* Iteratively Repeat E-step and M step. until the values converges (no difference)

Imagine there are 2 coins A & B (Biased) Here One is likely to get Head more no of times and other is likely to get tail more no of tail

* You pick one at random and toss it Which one was it

Sol: Repeat this 5 time.

{ * pick a coin randomly (known whether A or B)
* Toss 10 times
* Read No of Head & Tails
}

Consider the Result

	Coin A	Coin B
First Round		5H, 5T
Second Round	9H, 1T	

Consider the

First Round		5H, 5T
Second Round	9H, 1T	
Third Round	8H, 2T	
Fourth Round		4H, 6T
Fifth Round	7H, 3T	
Total	24H, 6T	9H, 11T

Probability of getting Head with Coin A $= \frac{24}{30} = 0.80$

" " " " " Coin B $= \frac{9}{20} = 0.45$

Coin A yields head 80% of time and Coin B 45% of time.

* What if only results are given.
& (we need to guess the percentage of heads that each coin yields &) Part I
Also (need to guess which coin was picked at each round of toss.) Part B

First Round		5H, 5T	?
Second Round	9H, 1T		? ✓
Third Round	8H, 2T		? ✓
Fourth Round		4H, 6T	? ✓
Fifth Round	7H, 3T		? ✓

Assume -> the probability of Head for Coin A = 60% $\Rightarrow \theta_A = 0.6$, Coin B = 50% $\Rightarrow \theta_B = 0.5$ } Randomly.

Initial

First Round $\Rightarrow$ 5H, 5T.

Compute the likelihood that it was Coin A & Coin B Using

Binomial Distribution i.e.

$$P(k) = \theta^k (1-\theta)^{n-k}$$

n = no of toss
k = successful (head)

θ => is probability on n trails with k success

Likelihood of A = $P_A(h)^k (1 - P_A(h))^{n-k}$

Here n=10, k=5

$= (0.6)^5 \times (1-0.6)^{10-5} = 0.000796824$

Likelihood of B = $P_B(h)^k (1 - P_B(h))^{n-k}$

n=10 k=5

$= (0.5)^5 \times (1-0.5)^{10-5} = 0.0009765625$

Normalize by Using $\frac{A}{A+B}$ [To Convert Likelihood into Probability]

for coin A = $0.000796824 / (0.000796824 + 0.0009765625)$

$= 0.45$

coin B = $0.0009765625 / (0.000796824 + 0.0009765625)$

$= 0.55$

Estimating Likely no of head & tails for coin A & B for First Round.

for coin A = $0.45 * 5 = 2.2$ H
$= 0.45 * 5 = 2.2$ T

for coin B = $0.55 * 5 = 2.8$ H
$= 0.55 * 5 = 2.8$ T. } First Round

Consider Round 2 (9H, 1T)

Likelihood of A = $0.6^9(1-0.6)^{10-9} = 0.0040310784$

likelihood of B = $0.5^9(1-0.5)^{10-9} = 0.0009765625$

Normalize A & B

$P(A) = 0.80$

$P(B) = 0.20$

Estimate likelihood no of Head & Tail for Coin A & B in Round 2

No of H for coin A = $0.80 \times 9 = 7.2$ H

T " " A = $0.8 \times 1 = 0.8$ T.

No of H for Coin B = $0.2 \times 9 = 1.8$ H

T " B = $0.2 * 1 = 0.2$ T.

After 5 Rounds

	Coin A	Coin B
Round 1	2.2H, 2.2 T	2.8 H, 2.8 T.
Round 2	7.2H, 0.8T	1.8 H, 0.2 T.
Round 3	5.9 H 1.5 T	2.1 H 0.5 T
Round 4	1.4 H 2.1 T	2.6 H 3.9 T.
Round 5	4.5 H 1.9 T	2.5 H 1.1 T
Total	21.3 H 8.6 T	11.7 H 8.4 T.

Probability of getting Head

$Q_A = \frac{21.3}{(21.3+8.6)} = 0.71$

$Q_B = \frac{11.7}{(11.7+8.4)} = 0.58$.

After 1 step of E & M $Q_A = 0.71$, $Q_B = 0.58$ ✓

Round until New Value of Q_A & Q_B is very Similar.

Repeat until New Value of θ_A & θ_B is very similar to Earlier.

In this Example

The Value Converge to

$\theta_A = 0.8$ ✓✓

$\theta_B = 0.52$ ✓

Part I Answered Q Percentage of Head Coin A yields = 80%

B yields = 52%

Now Consider the Result →

$\theta_A = 0.8$

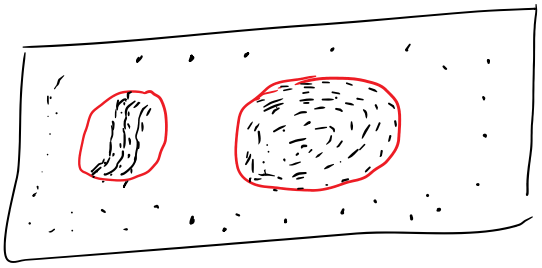
$\theta_B = 0.52$

	H	T	Which coin?
Round 1	5	5	$P(H) = \frac{5}{10} = 0.5$ = Coin B
Round 2	9	1	$P(H) = \frac{9}{10} = 0.9$ = Coin A
Round 3	8	2	$P(H) = \frac{8}{10} = 0.8$ Coin A
Round 4	4	6	$P(H) = \frac{4}{10} = 0.4$ Coin B
Round 5	7	3	$P(H) = \frac{7}{10} = 0.7$ Coin A

// Sol^n to Part 2

DBSCAN (Density Based Spatial Clustering of Application With Noise)

* In Dense Region there is possibility of cluster than in Sparse Region
* Density based approach is better if the cluster is of arbitrary shape.



Key Features

To understand DBSCAN Algo we need to understand.

ϵ Maximum radius of neighbourhood.

MinPts: Minimum No of points in an ϵ neighbourhood of that point.

ϵ neighbourhood of $q = \{4\}$

ϵ —————— $p = \{3\}$

Density of q is high

Density of p is low.

Terminologies → Given ϵ and Minpts then Categorize the objects into 3 group

(1) Core Point → if for a point the neighbourhood has the points $\geq$ Minpoint. then it is Core point.

ie Minpts = 4

No of points in neighbourhood of $q = 4 \rightarrow$ q is Core point

So q is core point

② Border Point → If for a point the neighbourhood has points < Minpt but it should be neighbours of core point

Minpt = 4

No of points in neighbourhood of p = 3 < Minpt

& p is neighbours of q (core point)

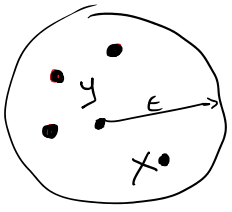
So 'p' is border point

③ Outlier → If a point is neither a core point nor a border point, it is called outlier.

"It is outlier"

* Directly Density Reachable → A point X is directly density reachable from point Y wrt epsilon & Minpt if

1. X belong to neighbourhood of Y (dist(X,Y) <= ϵ)
2. Y is core Point.



Minpt = 4

① as Y a core point →

no of point in neighbourhood of Y = 5 > Minpt

∴ Y is core point

② X belong to neighbourhood of Y.

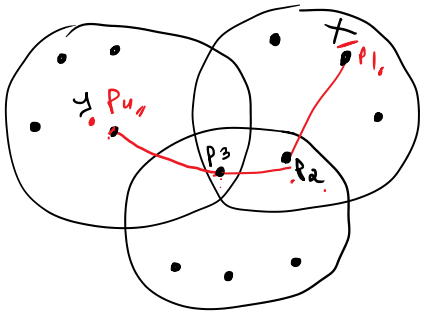
∴ X is Directly Density Reachable from Y.

② Density Reachable →

A point X is density reachable from Point Y wrt ϵ & Minpt

A point X is density reachable from Point Y wrt ϵ & Minpt if there is a chain of points $P_1, P_2 \ldots P_n$ and $P_1 = X$ & $P_n = Y$ such that P_{i+1} is directly density Reachable from P_i

Minpt = 4 Here X & Y are Core points



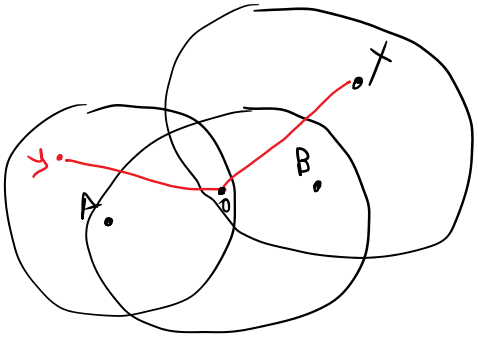
Here P_2 is directly Density Reachable from X

P_3 is directly Density Reachable from P_2

Y is directly Density Reachable from P_3

∴ X is Density Reachable from Y.

Density Connected → A point X is density Connected from point Y wrt ϵ and Minpt if there exists a point O such that both X & Y are density Reachable from O wrt ϵ & Minpt



Y & O are direct Density Reachable

X & O are direct Density Reach

Thus X & Y are density Connected

Direct $\xrightarrow{\text{within } \epsilon}$ Direct Density Reachable.

Via Some chain → " "

2 points are Density Reachable through Common pt → Density Connected

Algo

1. Arbitrary Select a Point P
2. Retrieve all points density Reachable from P wrt ϵ & Min Point
3. If P is Core Point then Cluster is formed.
4. If p is border point then DBSCAN visits next point of Dataset
5. Continue the process untill all the points are processed.