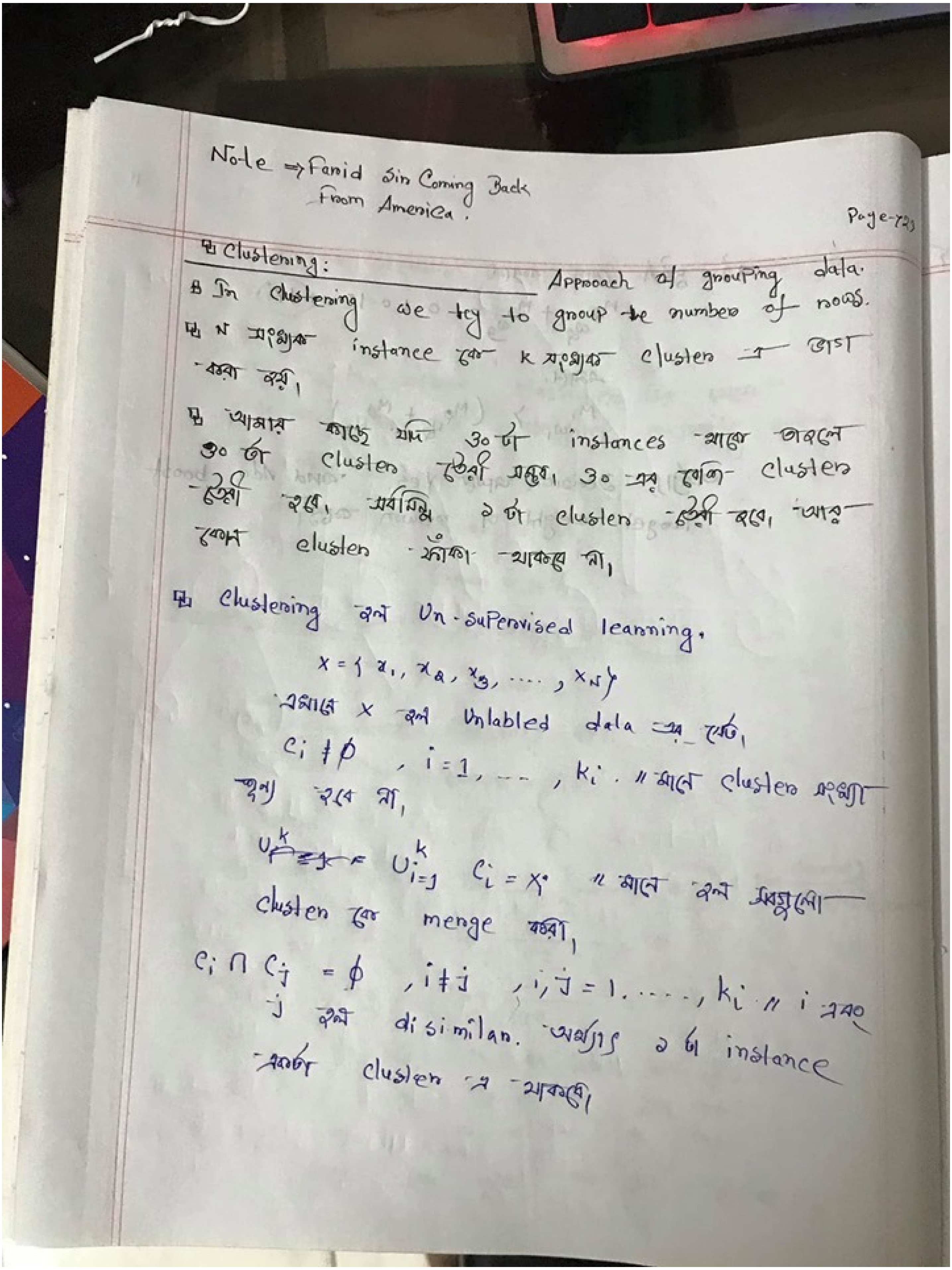


Note ⇒ Farid Sir Coming Back
From America.

Clustering:

Approach of grouping data.

- In Clustering we try to group the number of rows.
- N সংখ্যক instance কে k সংখ্যক cluster এ ভাগ করা হয়।
- আমার কাছে যদি ৩০ টা instances থাকে তাহলে ৩০ টা cluster তৈরী সম্ভব। ৩০ এর বেশি cluster তৈরী হবে। সর্বনিম্ন ২ টা cluster তৈরী হবে। আর কোন cluster ফাঁকা থাকবে না।
- Clustering হল Un-supervised learning.

$$X = \{x_1, x_2, x_3, \dots, x_N\}$$

এখানে X হল Unlabled data এর সেট।

$c_i \neq \phi$, $i = 1, \dots, k$. // মানে cluster সংখ্যা শূন্য হবে না।

$\cup_{i=1}^{k} c_i = X$. // মানে হল সবগুলো cluster কে merge করা।

$c_i \cap c_j = \phi$, $i \neq j$, $i, j = 1, \dots, k$ // i এবং j হল disimilar. অর্থাৎ ২ টা instance একটা cluster এ থাকবে।

□ What are the requirements of Clustering

Ans:

The requirements of clustering are

a. Dealing with large datasets having different attributes.

b. Find arbitarary shape

c. Noise handling

d. More features.

e. Constraint based clustering.

⇒ K Means Clustering Partitioning Method follow করে।

⇒ K Means iterative relocating technique use করে।

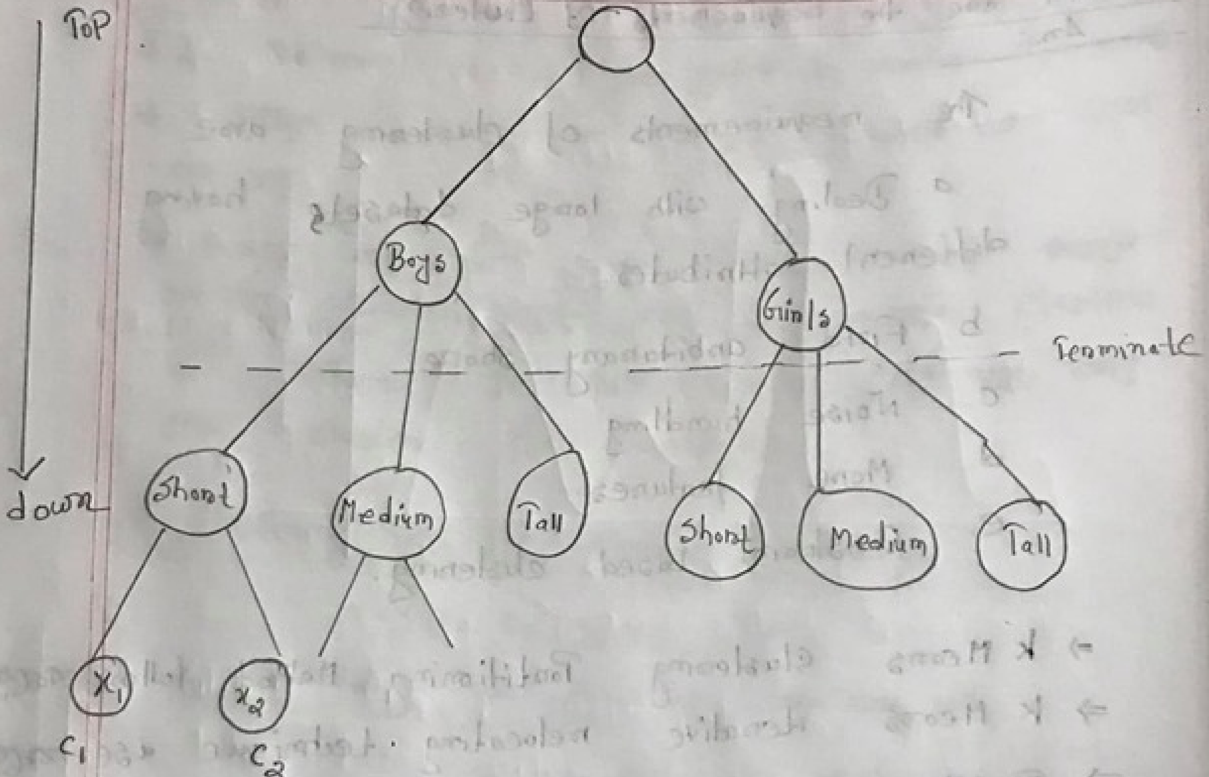
⇒ Partition based clustering এর বড় সমস্যা হল এটা বড় ডাইমেনশন এর ডাটা handle করতে পারে না।

Hierarchical Clustering:

এটা ২ Approach এ কাজ করে।

i. Top-down

ii. Bottom-Up



- Bottom Up Approach কে Similarity Approach বলা হয়। Bottom Up এ Similar cluster গুলোর Merge হয়। শেষে গিয়ে ১টা Cluster এ পরিণত হয়।

- Hierarchical Clustering কে যেকোন ১টা level পর্যন্ত terminate করানো যায়।

- Density Based Clustering: কোন একটা Space এ dataPoints কোথায় আছে, সেটা density based clustering দিয়ে measure করা হয়।

প্র What is Similarity Measure? Write the Similarity rules.

Ans:

Based on Column value ; Similarity is measured.

The rules are

a. Each cluster must contain at least one instance.

b. Each instance must belong to exactly one cluster.

প্র Normal k Means randomly Center Pick করে, এর ফলে Global Optima তে algorithm stuck হয়ে যায়।

প্র k Means ++ algorithm একটা মেকানিজম follow করে Center Pick করে। Min, Max, Average Value গুলো Center হিসেবে Pick করে।

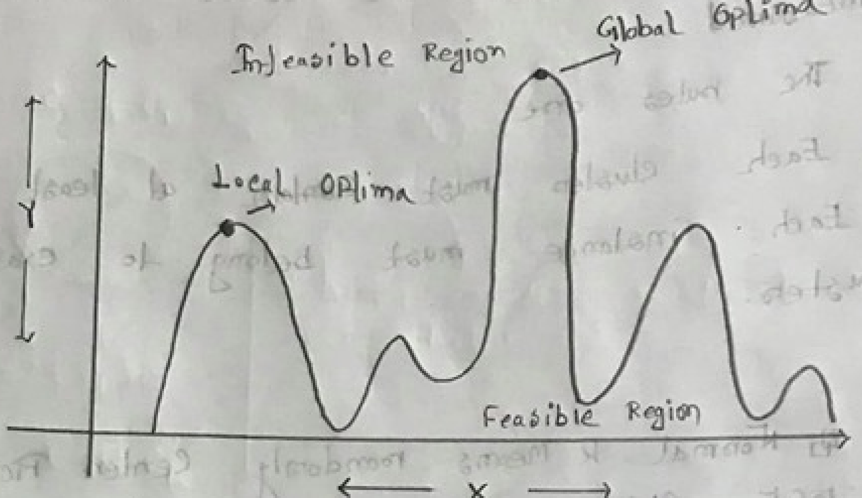
প্র Define Iterative Relocation

Ans:

It's a technique to improve the quality of the initial clusters. It implies that cluster membership is changed to find the local optima.

প্র Define Global Optima Problem Graph

Ans:



Figune: Global Optima Graph

Global Optima represents a very best solution. Local optima is better than it's immediate neighbors.

প্র Define Multimodal

Ans

local optima is also known as multimodal

প্র k-means clustering numerical সংখ্যা নিয়ে কাজ করে।

প্র similarity based clustering categorical value নিয়ে কাজ করে।

প্র sub-space clustering ; ensemble clustering এর একটা Part.

- k-Means Clustering এ Cluster কয়টা হবে, সেই সংখ্যাটা বলে দিতে হয়। Similarity based clustering এ cluster এর সংখ্যা বলা লাগে না। Similarity based clustering automatically cluster তৈরী করে ফেলে।

- Similarity Based Clustering threshold-value use করে। Threshold-value বলে আমরা কতগুলো feature এর উপর base করে clustering করতে চাচ্ছি। অর্থাৎ আমার কাছে 4 টা feature / Column থাকলে আমি 1/4 ; 2/4 ; 3/4 ; 4/4 threshold আকারে Clustering করতে পারি। 4/4 threshold হল tight ; Coupling situation.

~~Ans~~ Write Similarity Based Clustering Algorithm

Ans

Input: $X = \{x_1, x_2, x_3, \dots, x_N\}$ // A set of Unlabelled instances

Output: A set of clusters,

$C = \{C_1, C_2, \dots, C_k\}$

Method:

```
$C = \phi$;
$k = 1$
$C_k = \{x_1\}$;
$C = C \cup C_k$
```

Dataset-এর প্রথম row/instance টা নিয়ে Cluster তৈরী করি,

```
for $i = 2$ to $N$ do
    for $l = 1$ to $k$ do
        find the lth cluster center $x_l \in C_l$ to
        maximize the similarity measure $sim(x_i, x_l)$;
    end for
    if $sim(x_i, x_l) \geq$ threshold_value then
        $C_l = C_l \cup X_i$
    else
        $k = k+1$
        $C_k = \{x_i\}$;
        $C = C \cup C_{ki}$
    end if
end for
```

Answer

Let

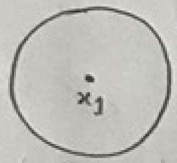
$t = 75\% = \frac{75}{100} = 0.75$ // Which means at least 3 features should be similar.

Step 1:

C_1 is the cluster x_1 would be the center.

~~Sim~~

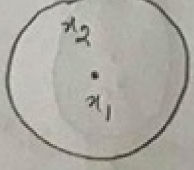
$x_1 \rightarrow C_1$



C_1

Step 2:

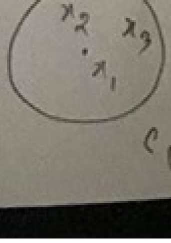
$sim(\underset{\text{Center of Cluster } (C_1)}{x_1}, \underset{\text{Instance}}{x_2})$ // $sim(x_1, x_2) \geq t$: true



C_1

Step 3:

$sim(x_1, x_3)$ // $sim(x_1, x_3) \geq$: true

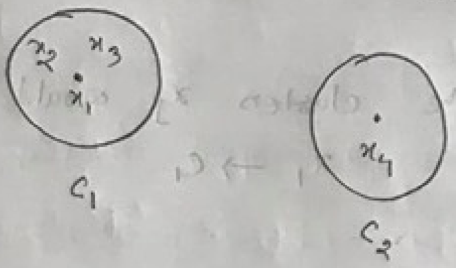


C_1

Step 4: $sim(x_1, x_4)$ // $sim(x_1, x_4) \geq t$: false

So x_4 would become Center of cluster 2

$x_4 \rightarrow C_2$



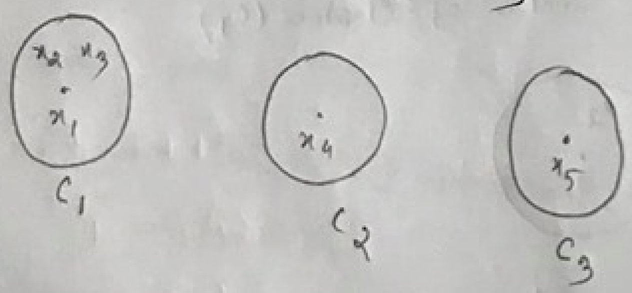
Step 5:

$sim(x_1, x_5)$ // $sim(x_1, x_5) \geq t$: false

$sim(x_4, x_5)$ // $sim(x_4, x_5) \geq t$: false

So, x_5 would become Center of Cluster 3

$x_5 \rightarrow C_3$

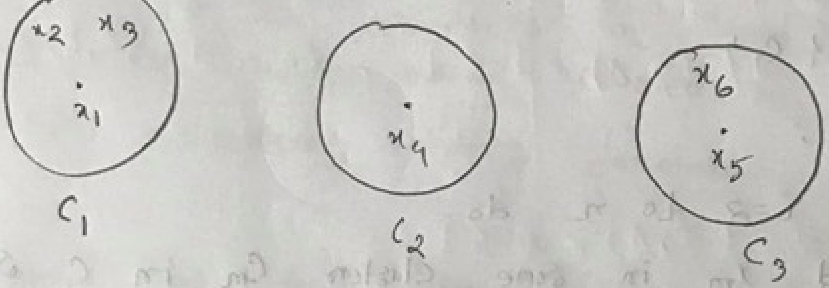


Step 6:

$\text{sim}(x_1, x_6)$ // $\text{sim}(x_1, x_6) \geq t$: false

$\text{sim}(x_4, x_6)$ // $\text{sim}(x_4, x_6) \geq t$: false

$\text{sim}(x_5, x_6)$ // $\text{sim}(x_5, x_6) \geq t$: true



Step 7:

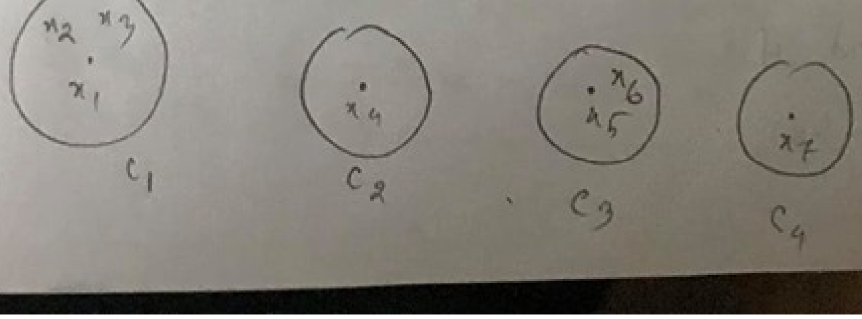
$\text{sim}(x_1, x_7)$ // $\text{sim}(x_1, x_7) \geq t$: false

$\text{sim}(x_4, x_7)$ // $\text{sim}(x_4, ~~x_4~~ x_7) \geq t$: false

$\text{sim}(x_5, x_7)$ // $\text{sim}(x_5, x_7) \geq t$: false

$x_7 \rightarrow C_4$

So, x_7 would become the center of clusters 4.



Q. Write Nearest Neighbor Clustering Algorithm

Ans

Input: $D = \{x_1, x_2, x_3, \ldots, x_n\}$ // A set of instances.

A // Adjacency matrix showing distance between instances

Output: A set of C Clusters.

Method.

```
C_1 = {x_1};
C = {C_1};
k = 1;
for i=2 to n do
    find x_m in some cluster C_m in C so that
    dis(x_i, x_m) is the smallest;
    if dis(x_i, x_m) ≤ t, threshold_value then
        C_m = C_m ∪ X_i
    else
        k = k+1;
        C_k = {x_i};
        C = C ∪ C_k;
    end if
end for
```

⊞ Similarity Based Clustering এ Nested for loop ব্যবহার করা হয়,

⊞ Nearest Neighbor Clustering এ Single for loop ব্যবহার করা হয়।

⊞ Write the advantage of Similarity Based Clustering (SCM)

Ans:

∻ The instances in (SCM) can self-organise local optimal cluster & volumes without using cluster validity functions.

⊞ Look at the given table.

Item	A	B	C	D	E
A	0	1	2	2	3
B		0	2	4	3
C			0	1	5
D				0	3
E					0

Given 5 items with the distance between them. Cluster them using Nearest Neighbor algorithm with a threshold t = 2

Solve. : Given

threshold_value ; $t = 2$

Step 1 :

A : $C_1 = \{A\}$

Step 2 :

B : $dis(B, A) = 1 \le t$; So $C_1 = \{A, B\}$

Step 3 :

C : $dis(C, A) = 2 \le t$

C : $dis(C, B) = 2 \le t$

So, $C_1 = \{A, B, C\}$

Step 4 :

D : $dis(D, A) = 2 \le t$

D : $dis(D, B) = 4 \le t$ it's false

D : $dis(D, C) = 1 \le t$

So, $C_1 = \{A, B, C, D\}$

Step 5:

E: $dis(E,A) = 3 \leq t$ it's false

E: $dis(E,B) = 3 \leq t$ it's false

E: $dis(E,C) = 5 \leq t$ it's false

E: $dis(E,D) = 3 \leq t$ it's false

So, new cluster would be formed. Therefore

$$C_2 = \{E\}$$

Ans..

Write the advantage of Nearest Neighbor Clustering

Ans:

No need to know the number of clusters to discover beforehand. It's different than in k-means & hierarchical.

Nearest Neighbor clustering এর Complexity depend করে dataset এর volume on কতগুলো instances আছে তার উপর,

Nearest Neighbor Clustering এর Time Complexity ~~$O(N^3)$~~ O(ন২)

Space Complexity ~~$O(N^2)$~~ O (ন২)

Q. Write the differences between Classification & Clustering

Ans

Classification	Clustering
a. Known numbers of classes	a. Unknown numbers of classes
b. Based on training set.	b. No Prior knowledge.
c. Used to classify future observations.	c. Used to explore data
d. Supervised Type	d. Unsupervised Type.