Data Mining Unit 4 Handwritten

Data Mining And Analytics (SRM Institute of Science and Technology)

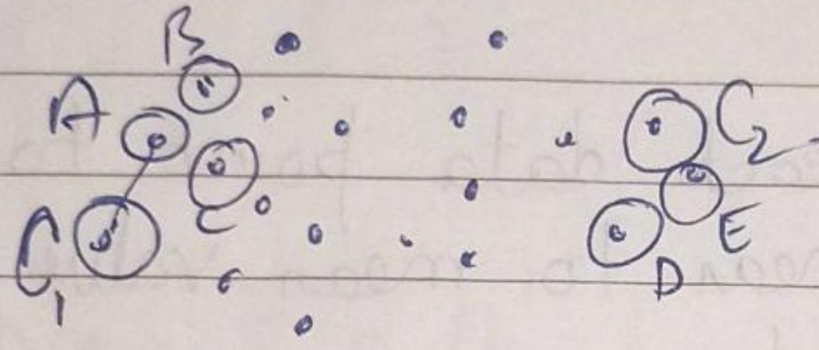
Studocu is not sponsored or endorsed by any college or university
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

Unit - 4

Clustering → Clustering is the task of grouping a set of object in such a way that object in same groups are more similar to each other than those in other group.

① K-means

↳ K - no. of clusters.



C_1	C_2
A	D
B	E
C	⋮
⋮	

Suppose we want to have 2 cluster. $k=2$.

Step 1 → Select 2 random point as centroid C_1, C_2
Now measure distance of all point from C_1 and C_2.

C_1	C_2
A	D
B	E
C	F
$\frac{A+B+C}{3}$	$\frac{D+E+F}{3}$

↳ Mean of above will become Centroid.

* Distance is measured as Euclidean distance.

This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner

Algorithm k-means :-

Input → k - Number of clusters.
Data Set (D)
Output → Set with k clusters.

Algo

① Select k point randomly from the data Set D.
② Repeat
   a. Assign each data point to the cluster that is near to mean value (centroid) of the cluster.
   b. Update the centroid value of the cluster with the mean value of members.
3. Until there is no change in cluster.

Note → Do it 3-4 times in exam

Q. Apply k-means clustering to the following data-points.

	X_1	X_2
A	5	1
B	4	2
C	3	3
D	1	4
E	6	7

where k = 2.

Scanned by TapScanner

Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

Sol → Let A represent C_1 and B represent C_2.

	Distance from C_1	Distance from C_2	Cluster
A	$\sqrt{(5-5)^2+(1-1)^2}=0$	$\sqrt{(5-4)^2+(2-1)^2}=\sqrt{2}$	C_1
B	$\sqrt{(5-4)^2+(2-1)^2}=\sqrt{2}$	$\sqrt{(4-4)^2+(2-2)^2}=0$	C_2
C	$2\sqrt{2}$	$\sqrt{2}$	C_2
D	5	$\sqrt{13}$	C_2
E	$\sqrt{37}$	$\sqrt{29}$	C_2

New, C_1 = A [∵ only 1 (A) has C_1]

C_2 = $\frac{B+C+D+E}{4}$ ← point form (mean of A, B, C, D)

$C_1 = \left(\frac{4+3+1+6}{4}, \frac{2+3+4+7}{4}\right)$

$\Rightarrow C_1 = (3.5, 4)$.

	Distance Square from C_1	Distance Square from C_2.	Cluster
A	$(5-3.5)^2+(1-4)^2=0$	$(5-3.5)^2+(1-4)^2=11.25$	C_1
B	$\sqrt{2}$	4.15	C_1
C	$2\sqrt{2}$	1.25	C_2
D	5	6.25	C_2
E	37	15.25	C_2

This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner

New, $C_1 \rightarrow \left(\frac{A+B}{2}\right)$

$\rightarrow \left[\left(\frac{5+4}{2}\right), \left(\frac{1+2}{2}\right)\right]$

$= (4.5, 1.5)$

New $C_2 \rightarrow \left(\frac{C+D+E}{3}\right)$

$\rightarrow \left[\frac{1+6+3}{3}, \frac{3+4+7}{3}\right]$

$C_2 \rightarrow (3.3, 4.7)$.

	Distance Square from C_1	Distance Square from C_2	Clusters
A	1.25	16.58	C_1
B	0.5	7.78	C_1
C	9.5	2.98	C_2
D	18.5	5.78	C_2
E	32.5	12.58	C_2

Since Now, Cluster doesn't changed.

So, we stop.

So, Cluster 1 → A, B

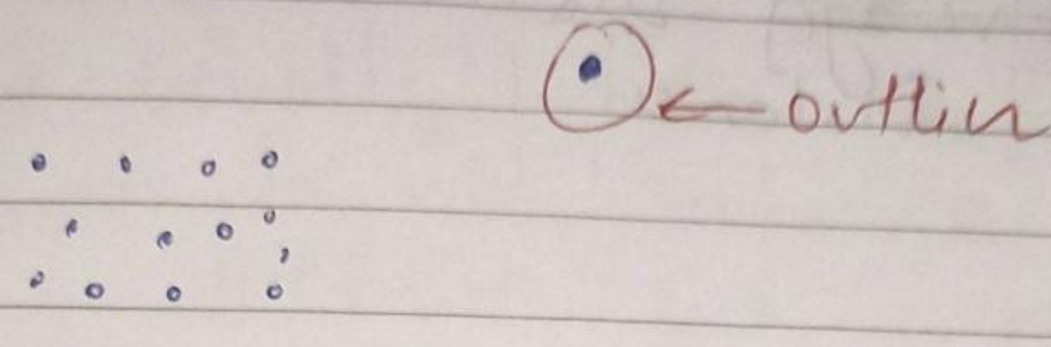
Cluster 2 → C, D, E

Scanned by TapScanner

Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

Disadvantage of K-mean Clustering :-

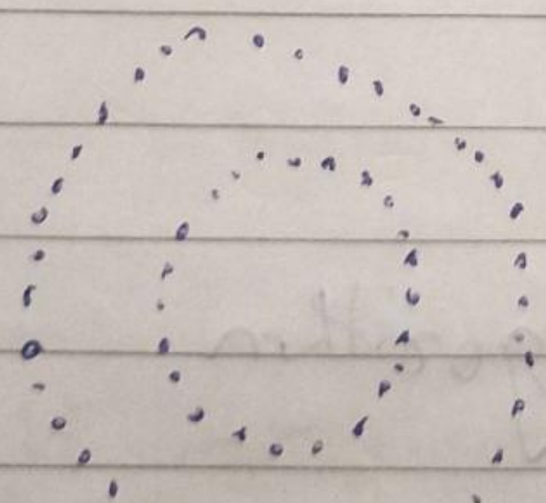
1. The K-means algorithm is sensitive to outliers!



Cluster is imparted by outlier in this.

2. In this when we find new ~~cluster~~ centroid point, we might get a decimal point which may not be any actual point in cluster.
So to overcome this we used new algo which is based on median instead of means.

3.



for any cluster in any shape K-mean is not able to identify shape or divide cluster in form of that shape. For this we used Density Based cluster.

This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner

K-Medoids :-

	X	Y	Distance from C_1 (4, 5)	Distance from (5, 8)
1.	8	7	6	2
2.	3	7	3	7
3.	4	9	4	8
4.	9	6	6	2
5.	8	5	4	6
6.	5	8 $\rightarrow C_2$	—	—
7.	7	3	5	3
8.	8	4	5	1
9.	7	5	3	1
10.	4	5 $\rightarrow C_1$	—	—

$C_1 \rightarrow (2, 3, 5)$, $C_2 \rightarrow (1, 4, 7, 8, 9)$.

Cost = $(3+4+4) + (2+2+3+1+1)$

= 20.

Now, in next step, we choose any other point than C_2 as C_2'. and find cost.

Let say (7, 5) as C_2'.

Then Cost = 22

As 22 > 20

So, undo swap.

& follow this step till end.

Scanned by TapScanner

Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

Algorithm K-medoids :-

Input :- K :- no. of clusters.
D :- data set containing n object
Output :- A set of k-cluster.

1. Randomly chose K object in D as initial representative object or seeds :-
2. Repeat
3. assign each remaining object to the cluster with nearest representative object;
4. randomly select a non-representative object O_{random}.
5. Compute the total cost, S, of swapping representative object O_j with O_{random}.
6. if $S < 0$, then swap O_{random} with O_j to form new set of K representative object
7. Until no change.

Algo ↑

* K-medoid is more robust than k-mean in presence of noise or outliers because a medoid is less influenced by outlier or other extreme values than mean. However processing of K-medoid is more costly than k-mean.

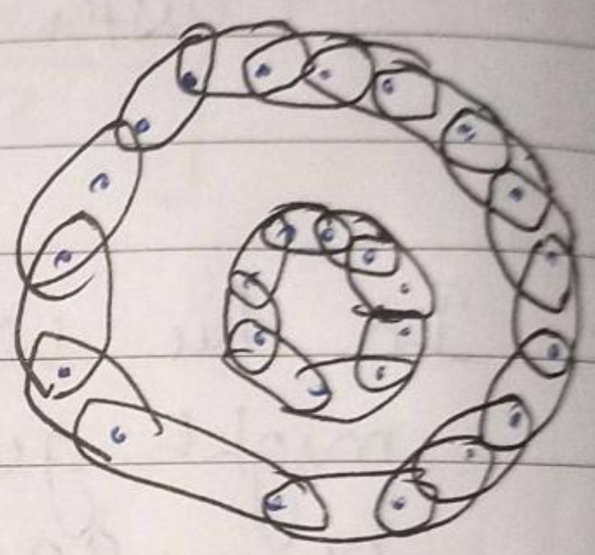
This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner

Distances in clustering Algorithm :-

1) Minimum Distance
2) Maximum Distance
3) Mean ~~(Average)~~ Distance
4) Average Distance

Nearest Neighbour :-

2-D ~~circle~~ (circle)
3-D ~~circle~~ (sphere).



K-Nearest Neighbour (KNN)

↳ Supervised Algorithm

	X_1	X_2	Y.	Distance
A	1	2	C_1	1
B	2	3	C_1	1
C	1	4	C_2	1
D	1	5	C_2	4

(1, 3) → ? → to which cluster it will go?

Scanned by TapScanner

Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

Density Mase Clustering :- Continue growing the cluster as density of neighbourhood exceeds threshold.

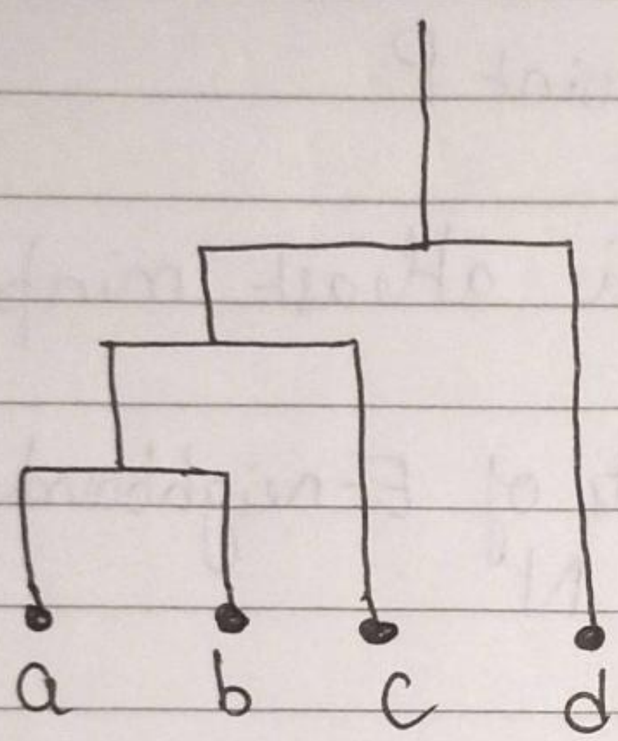
DBSCAN Algorithm :-

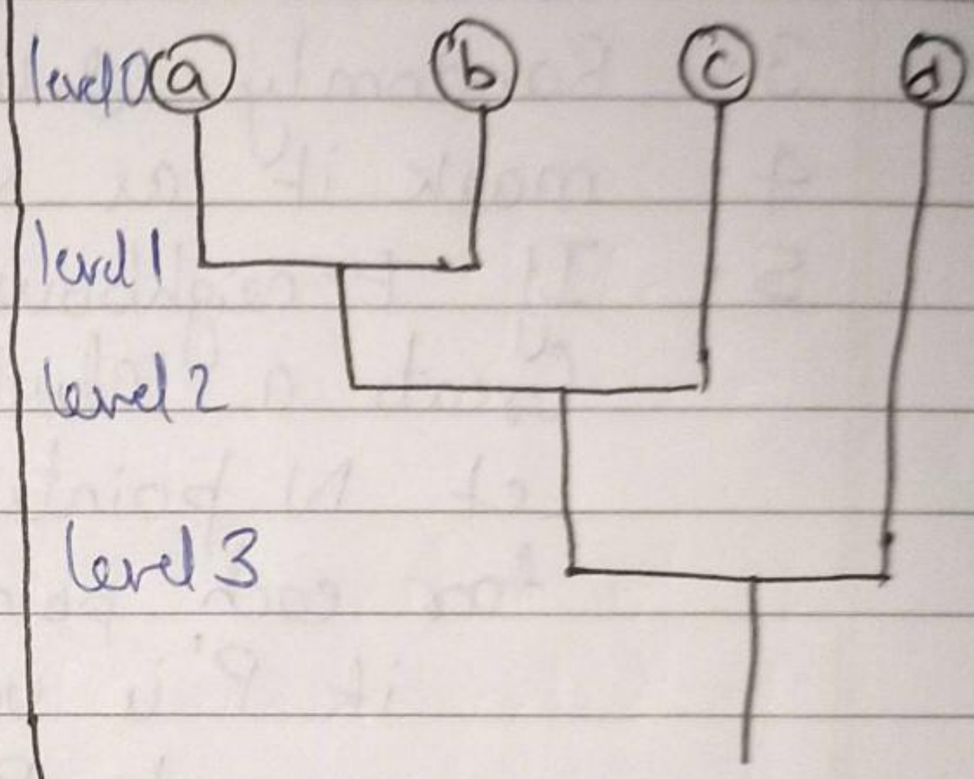
1. Mark ALL OBJECTS as UNVISITED.
2. do
3. Randomly select some point P
4. mark it as visited
5. If E-neighbourhood of P is atleast minpts.
   Creat a cluster C
   Let N points are objects of E-neighbourhood of P
   for each point P' from N
      if P' is unvisited
      mark P' as visited
      if E-neighbourhood of P' has atleast minpts.
      add these points to N.
      if P' is not a part of cluster then add it to cluster C.

end for
Output C
else mark P as noise
until all points are visited.

This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner

Agglomerative Approach	Devisive Approach
→ In this object are merged to form a cluster	→ In this clusters are split into object.
→ It use bottom up approach.	→ It use top down approach.
→ Represented using Dendogram	→ Represented using Dendogram.
→ Uses closed group object to merge so euledian distance is calculated between various point to decide the merging point	





*Note → On viewing from top of the devise approach diagram, it look like Agglomerative clustering and from bottom it look like devisive approach.

Scanned by TapScanner
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

BIRCH (Balanced Iterative Reduced Clustering Using Hierarchy) :-

K-means usually have time complexity of $O(n^2)$. So to improvise then a new method BIRCH was introduced which has the time complexity of $O(n)$. It provides scalability.

* BIRCH is designed for clustering a large amount of numerical data by integration of hierarchal clustering and other clustering methods such as iterative partitioning.

It overcome two difficulties of agglomeric clustering methods :- i) Scalability
ii) the inability to undo what was done in previous step.

* BIRCH introduces two concept :-
i) Clustering factor
ii) clustering factor tree

* Clustering factor → is a vector (3-D) summerising information about clusters. $< n, LS, SS >$ ← CF

- n → no. of points in cluster
- LS → Linear Sum of n points
- SS → Square Sum of data points

This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner

Q. Given three points (2,5), (3,2), & (9,3). Compute the clustering factor.

Sol →

$$n = 3.$$

$$\text{Linear Sum} = \sum_{x_i=1}^{n} x_i$$

$$= (2+3+4), (5+2+3)$$

$$= (9, 10).$$

$$\text{Square Sum} = \sum_{x=1}^{n} x_i^2.$$

$$= 4+9+16 \quad 25+9+4$$

$$= (29, 38).$$

$$CF_1 = \langle 3, (9,10), (29,38) \rangle \quad \text{Ans}$$

How to merge 2 - clustering factor ?

$$Cf_2 = \langle 3, (35,36), (417, 440) \rangle$$

$$Cf_3 = Cf_1 + Cf_2$$

$$= \langle 3, (9,10), (29,38) \rangle + \langle 3, (35,36), (417,440) \rangle$$

$$= \langle 3, (44,46), (446, 478) \rangle. \quad \text{Ans}$$

Scanned by TapScanner

Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

How to find Centroid of C.F :-

* Centroid = $\frac{\sum x_i}{n} = \frac{LS}{n}$

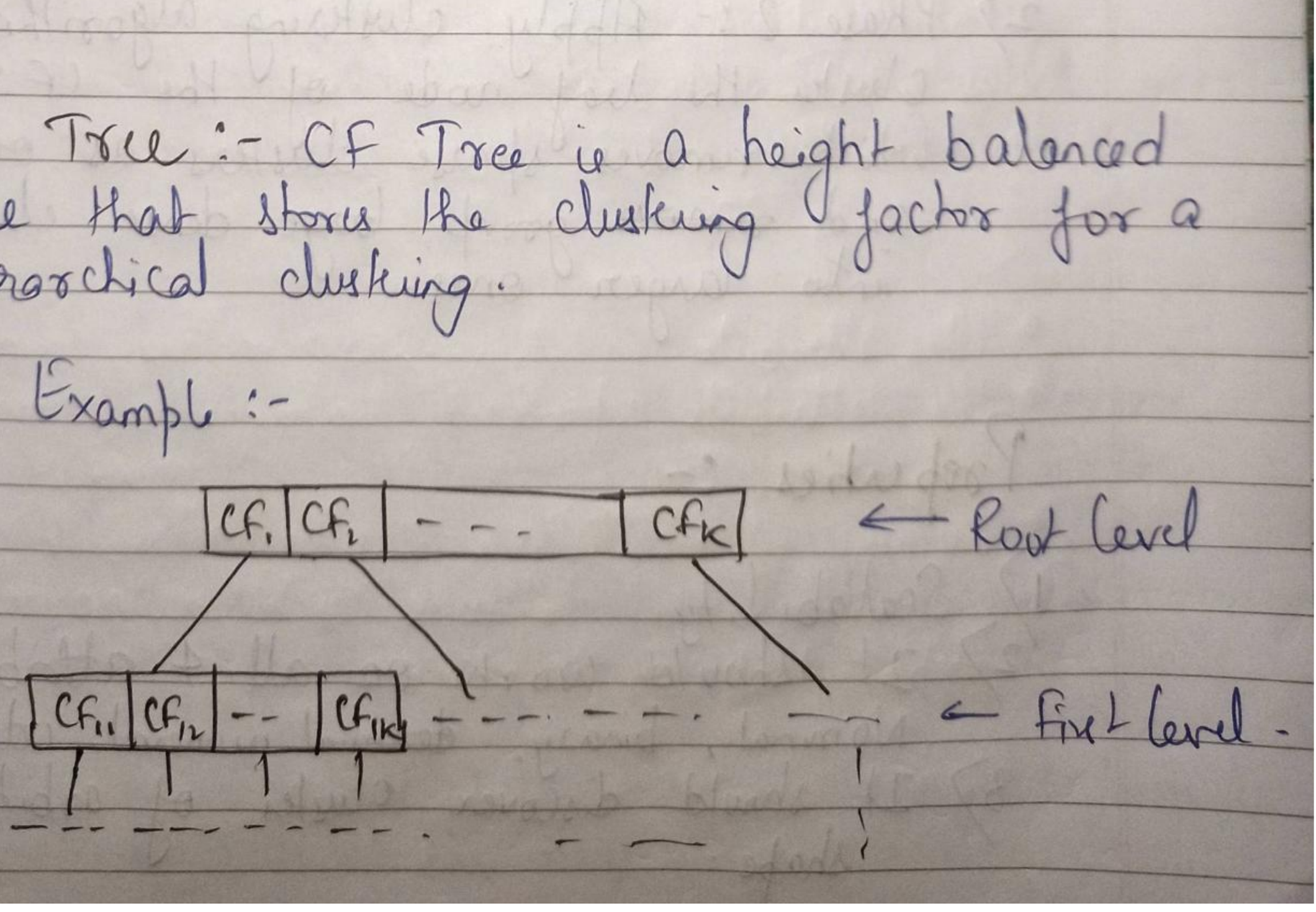
$= \left(\frac{9}{3}, \frac{16}{3}\right)$ in above example.

* Radius = $\sqrt{\frac{\sum (x_i - x_0)^2}{n}} = \sqrt{\frac{n \cdot SS - 2LS^2 + n \cdot LS}{n^2}}$

* Diameter = $\sqrt{\frac{2 \cdot n \cdot SS - 2 \cdot LS^2}{n(n-1)}}$

* CF Tree :- CF Tree is a height balanced tree that stores the clustering factor for a hierarchical clustering.

Example :-



This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner

* A CF Tree has 2 parameters :

i) Balancing Factor :- Specifies maximum no. (Branching Factor) :→ of children per non-leaf node.

ii) Threshold → Specifies maximum diameter of ~~subset~~ subclusters stored at the leaf node of the tree.

These two parameters influence the size of resulting tree.

BIRCH Algorithm works in 2 phases :-

1) Phase 1 :- Scan the database to build an initial in-memory CF Tree with multilevel compression of data.

2) Phase 2 :- Apply clustering algorithm to cluster the leaf node of the CF Tree, which removes sparse clusters as outliers and ~~remove~~ groups ~~large~~ dense clusters into larger ones.

Properties :-

1) Scalability
2) It should work on all 4 attributes (i.e Nominal, binary, ~~decimal~~ numeral, ordinal).
3) It should discover cluster of arbitrary shape.

Scanned by TapScanner

Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

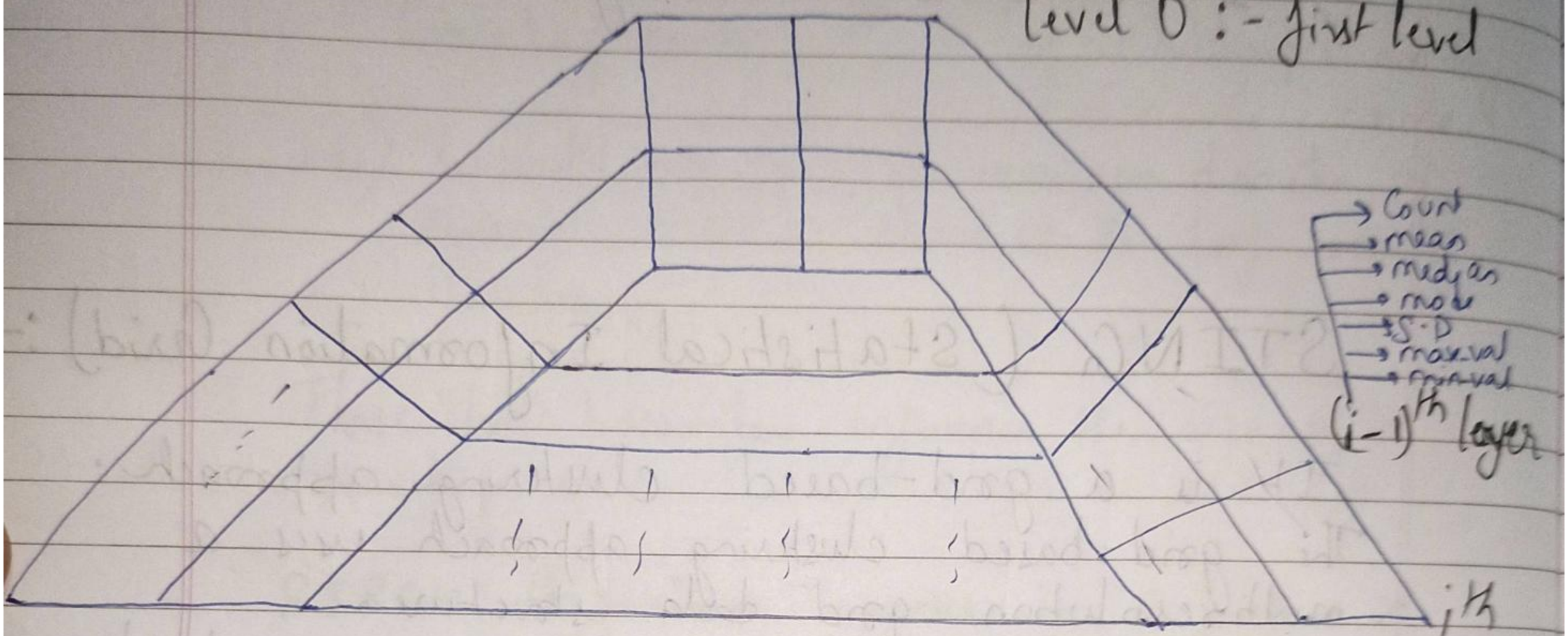
4) It should be incremental.
5) It should work on high dimensional data.

STING (Statistical Information Grid):-

It is a grid-based clustering approach.
The grid-based clustering approach uses a multiresolution grid data structure.
It quantizes the object space into a finit number of cells that forms a grid structure on which all of the operation for clustering are performed.
The main advantage of the approach is it fast processing time, which is typically independent of the number of data objects, yet dependent on only the number of cells in each dimension in the quantized space.

STING:- Sting is a grid based ~~multiresolut~~ multiresolution clustering technique in which the spatial area is divided into rectangular cells.
There are usually several level of each rectangular cells corresponding to different levels of resolution, and these cells form a hierarchical structure, each cell at higher level is partioned to form a number of cells at the next lower level.

This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner



A hierarchical structure of STING clustering

Statistical parameters of higher level cells can be easily computed from parameters of lower-level cells. These parameters include the following :-

- the attribute independent parameter, (count)
- the attribute dependent parameter, (mean, S.D, min, max)
- the type of distribution that the attribute value in the cell follows, such as normal, uniform, exponential or none.

Scanned by TapScanner

Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

Advantages of STING :-

Sting offers several advantages :-

i) the grid-based computation is query independent because the statistical information stored in each cell represent the summary information of data in grid cell, independent of query.

ii) the grid structure facilitates parallel processing and incremental updating.

iii) the ~~med~~ method's efficiency is a major advantage. time complexity of generating clusters is O(n).

CLIQUE (Clustering in Quest) :-

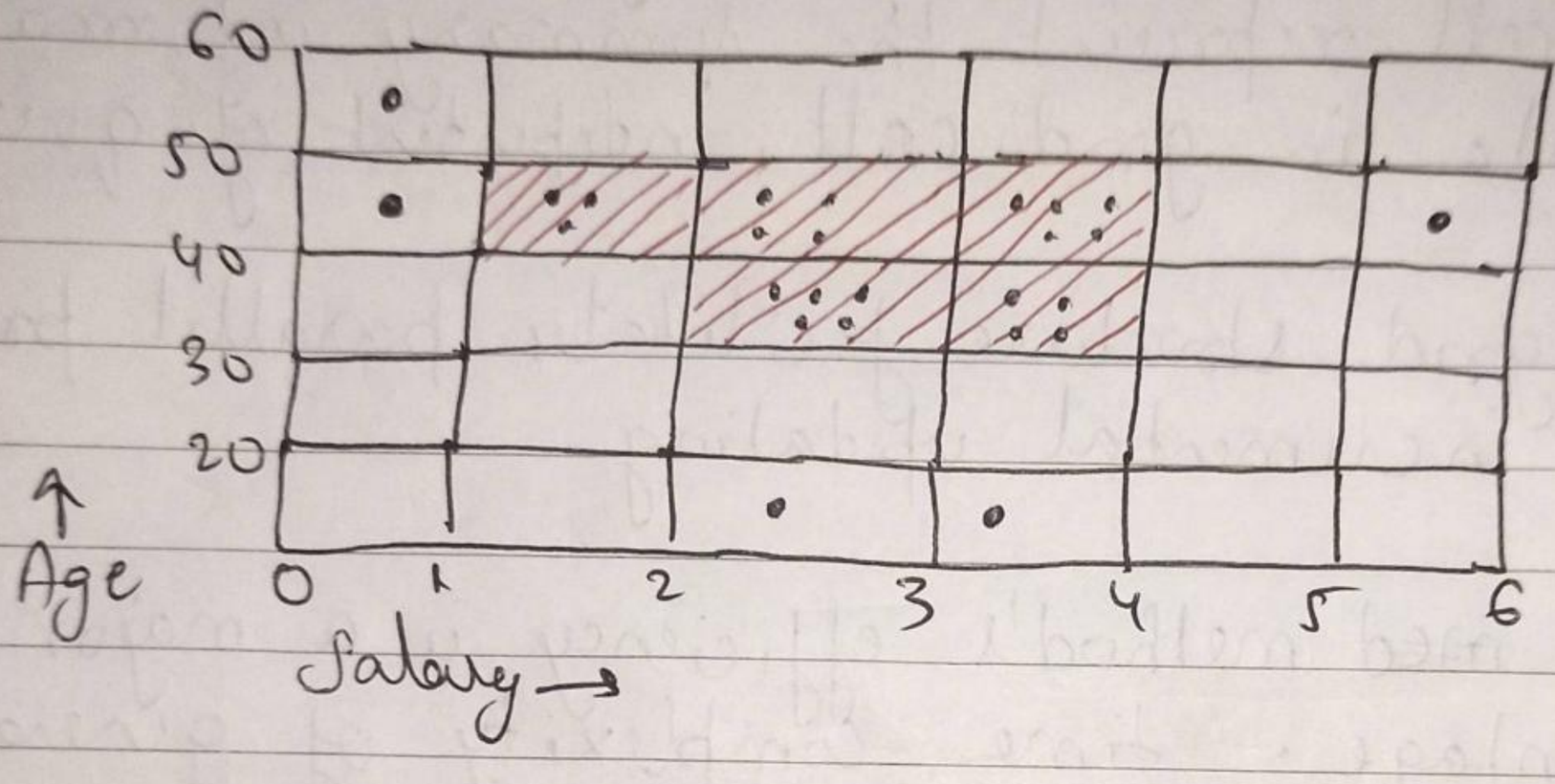
1. It is a grid based clustering algorithm
2. It is density based clustering algorithm
3. It use Apriori Property.

CLIQUE was the first algorithm proposed for dimension-growth subspace clustering in high dimensional space.

This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner

Cells

Since it is based on density Based, there will be a threshold in there?



Threshold value define whether the grid is dense or sparse.

If Threshold = 3.

Then, $T \geq 3$ → dense

$T < 3$ → sparse

Steps :-

1) Identify the subspace with clusking
2) Identify the cluster
3) Generate the minimum descriptor for clusking

Scanned by TapScanner

Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

Advantage :-

1) It is faster than K-means. Moreover, we need not have to find neighbours that we need to find DBSCAN, so it is faster than DBSCAN and K-means both.

2) It can find any shape of cluster
3) It is one of the simplest method.

Disadvantage :-

1) Not suitable for high dimension because
   ⇒ It is hard to visualize when dimension increases

Evaluation of Clustering :-

① Accessing the clustering technique :-

Whether they access the non-random structure or not?

Hopkin's Statics :- used to identify whither data set follow any pattern or not?

⟶ Given a dataset D as a sample of random variable o, determine how far away o is from being uniformly distributed in data space.

This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner

→ Sample n point $p_1, \ldots p_n$ uniformly from D. For each p_i find its nearest neighbour in D.

$$u_i = \min(dis(p_i, v)) \text{ where } v \in D.$$

(Population)

→ Sample n point $q_1 \ldots q_n$ uniformly from D, (Set of neighbours.)
for each q_i finds its nearest neighbour in $D - \{q_i\}$:-

$$y_i = \min\{dist(q_i, v)\} \text{ where } v \in D \text{ \& } v \neq q_i$$

→
$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

→ If Uniformly distributed, $\Sigma x_i = \Sigma y_i$

$$\Rightarrow H = \underline{0.5}$$

↓

Then cluster is uniformly distributed.

→ If D is highly skewed, H is close to 0.

2) Empirical Method :-

We find No. of clusters in this. for n points, No. of clusters

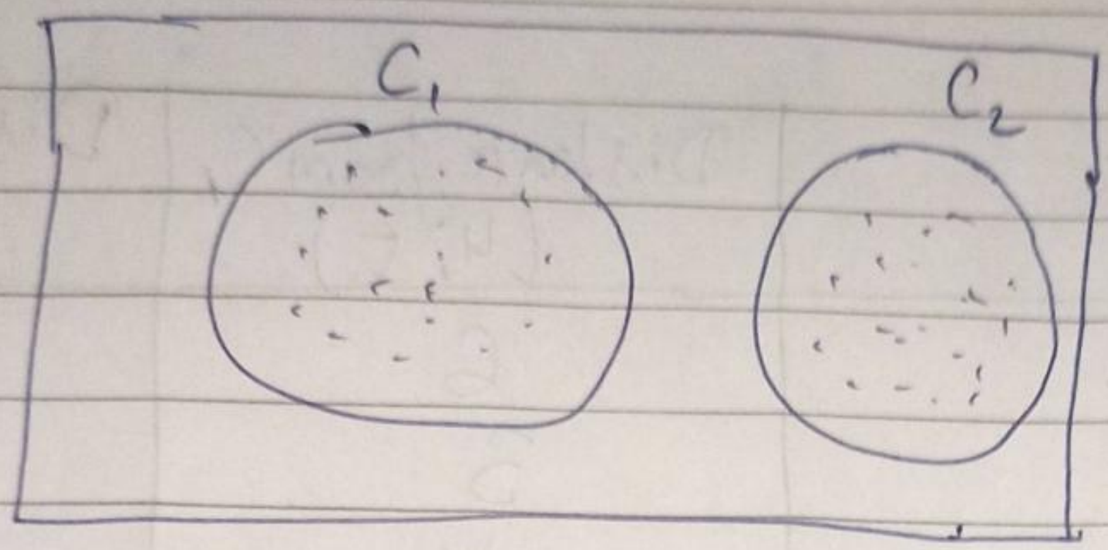
$k = \sqrt{\frac{n}{2}}$ for dataset of n points.

each cluster having $\sqrt{2n}$ points.

Scanned by TapScanner

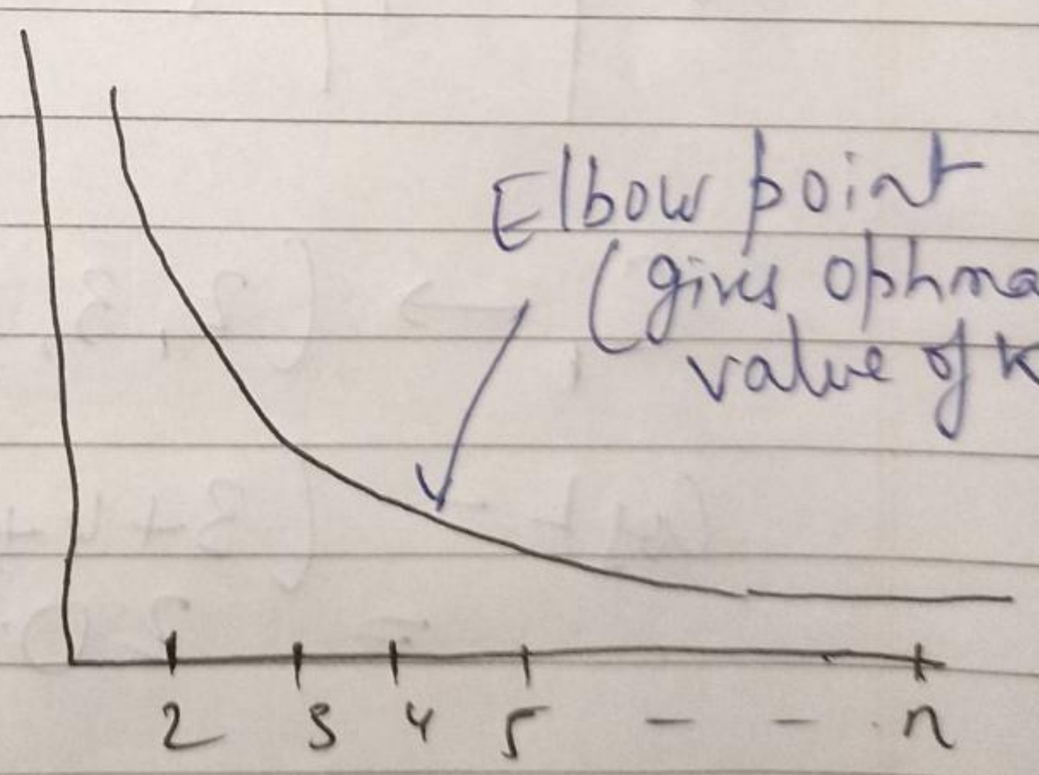
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)

3) Elbow Method :- (intracluster).



→ We the turning point in the curve of sum of within cluster variance with respect to no. of cluster.

for 2 to n :-
find distance & draw it on graph



Can't take k = 1 as data is already given at k = 1.

Silhouette Method / Score :-
depend on inter cluster.

This document is available free of charge on studocu
Downloaded by S SRI KRISHNA (s.sri.krishna.050202@gmail.com)
Scanned by TapScanner