S1-20_DSECLZC415 : Data Mining

Lecture #11 – Cluster Analysis

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

- *The slides presented here are obtained from the authors of the books and from various other contributors. I hereby acknowledge all the contributors for their material and inputs.*
- *I have added and modified a few slides to suit the requirements of the course.*

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

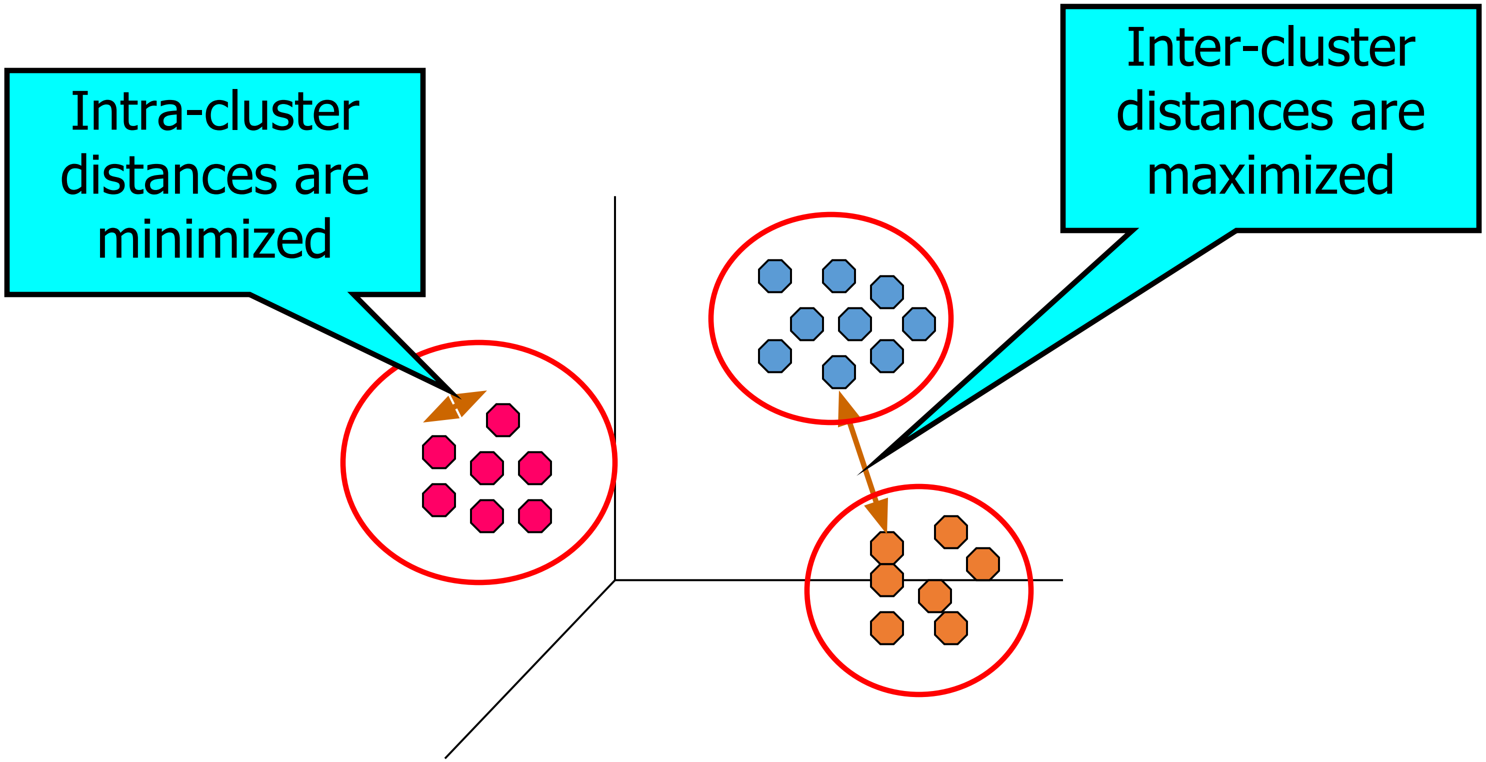
Data Mining

Cluster Analysis

What is Cluster Analysis?

Review

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Quality: What Is Good Clustering?

Review

A good clustering method will produce high quality clusters with

- high intra-class similarity
- low inter-class similarity

The quality of a clustering result depends on both the similarity measure used by the method and its implementation

The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

Types of Clusterings

Review

A clustering is a set of clusters

An important distinction among types of clusterings : *hierarchical* and *partitional* sets of clusters

Partitional Clustering
- A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

Hierarchical clustering
- A set of nested clusters organized as a hierarchical tree

Partitional Clustering Method *K-Means*

Review

Strength: *Efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$.

- Compare to PAM (Partitioning around Medoids) : $O(k(n-k)^2)$, CLARA (Clustering LARge Applications) : $O(ks^2 + k(n-k))$

Comment: Often terminates at a *local optimal*.

Weakness

- Applicable only to objects in a continuous n-dimensional space
  - Using the k-modes method for categorical data
  - In comparison, k-medoids can be applied to a wide range of data
- Need to specify *k,* the *number* of clusters, in advance
- Sensitive to noisy data and *outliers*
- Not suitable to discover clusters with *non-convex shapes*

innovate achieve lead

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

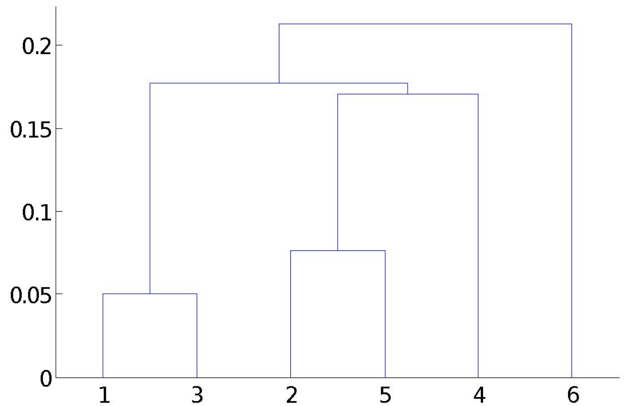
Hierarchical Methods

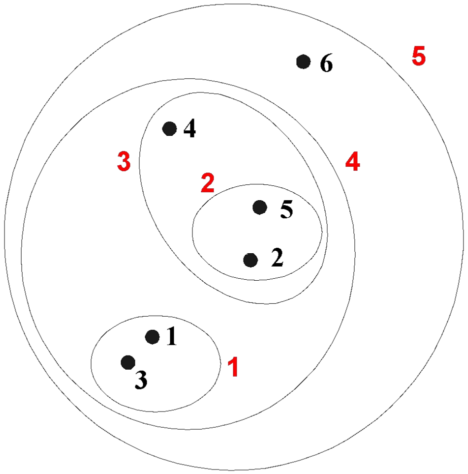
Hierarchical Clustering

Produces a set of nested clusters organized as a hierarchical tree

Can be visualized as a dendrogram

– A tree like diagram that records the sequences of merges or splits





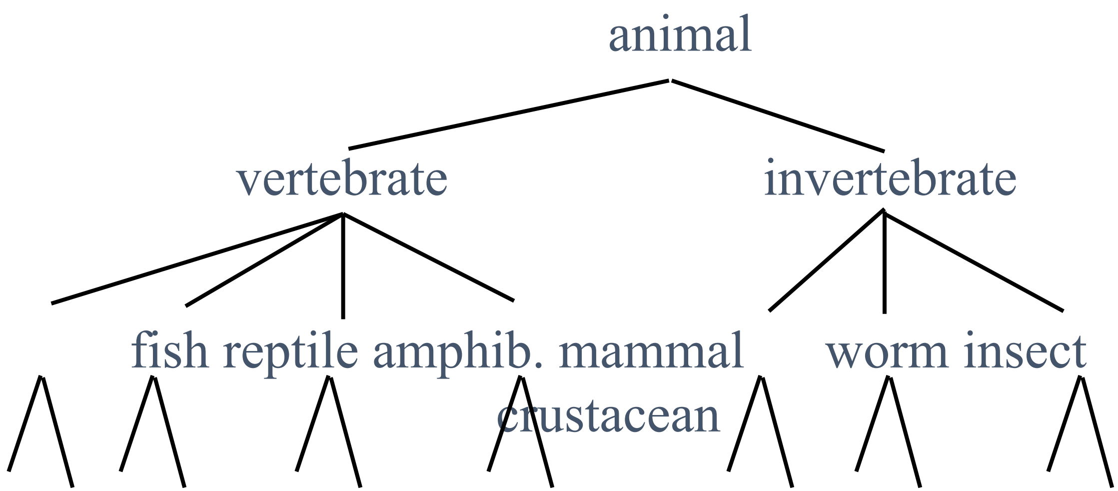
Strengths of Hierarchical Clustering

Do not have to assume any particular number of clusters

- Any desired number of clusters can be obtained by ‘cutting’ the dendogram at the proper level

They may correspond to meaningful taxonomies

- Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)



Hierarchical Clustering

Two main types of hierarchical clustering

- Agglomerative:
  - Start with the points as individual clusters
  - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left

- Divisive:
  - Start with one, all-inclusive cluster
  - At each step, split a cluster until each cluster contains a point (or there are k clusters)

Traditional hierarchical algorithms use a similarity or distance matrix

- Merge or split one cluster at a time

Agglomerative Clustering Algorithm

More popular hierarchical clustering technique

Basic algorithm is straightforward

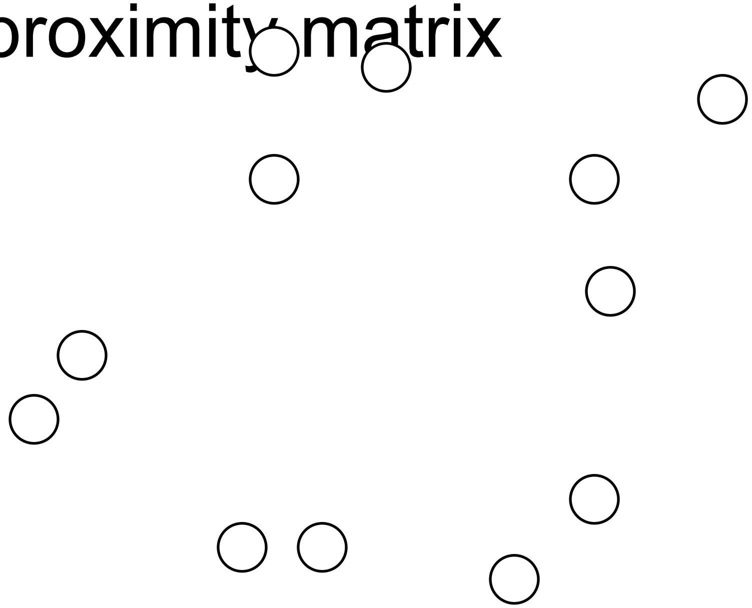
1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

Key operation is the computation of the proximity of two clusters

- Different approaches to defining the distance between clusters distinguish the different algorithms

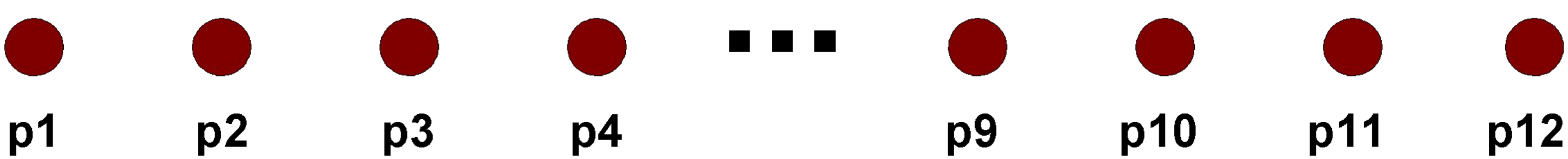
Starting Situation

- Start with clusters of individual points and a proximity matrix



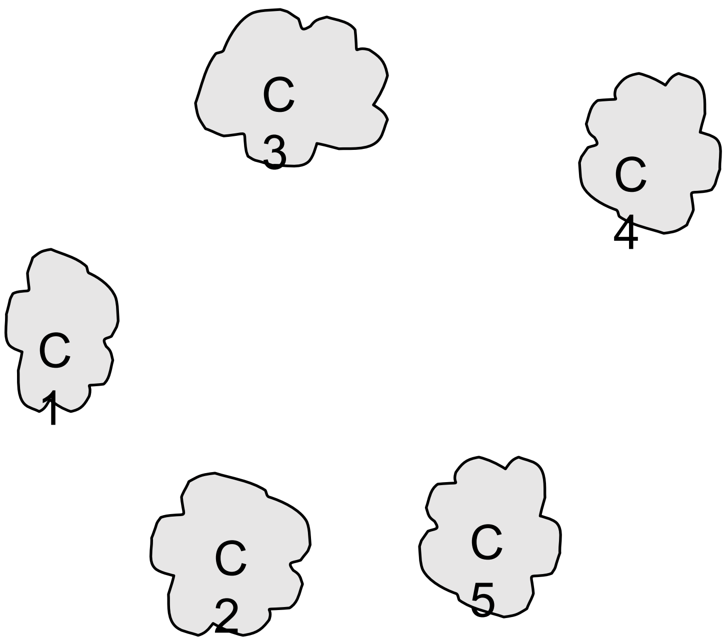
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



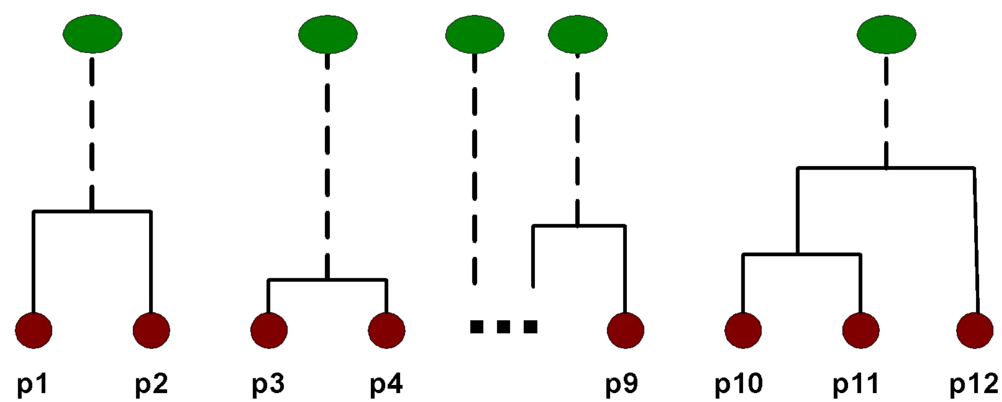
Intermediate Situation

- After some merging steps, we have some clusters



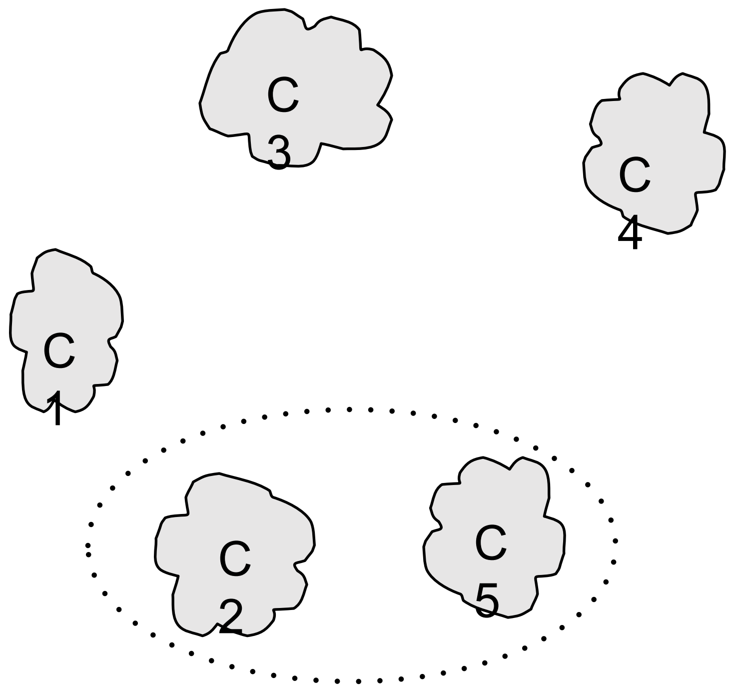
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

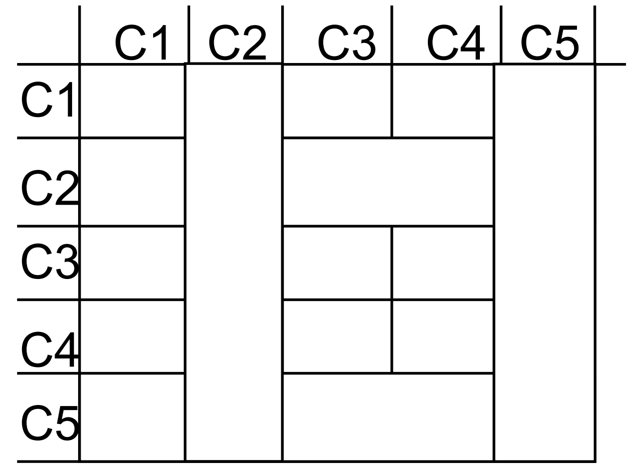
Proximity Matrix



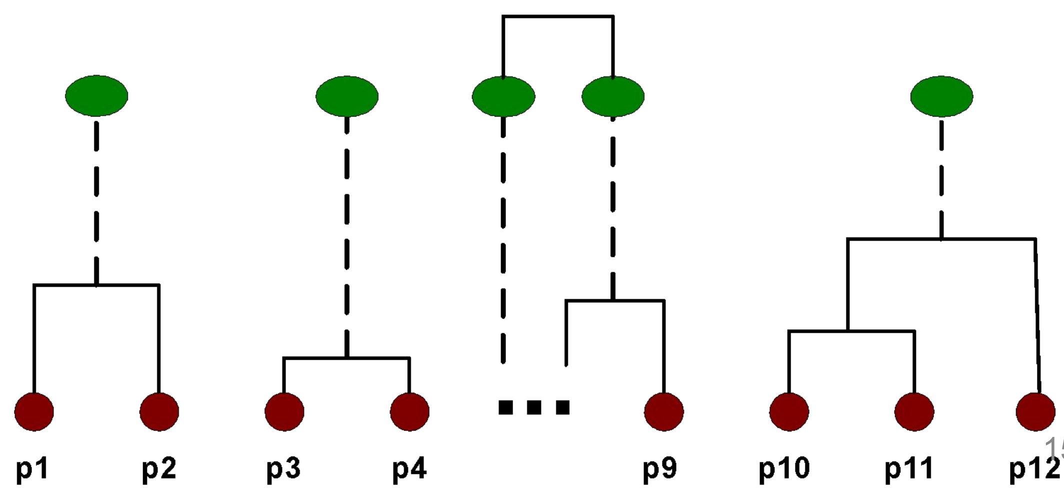
Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.

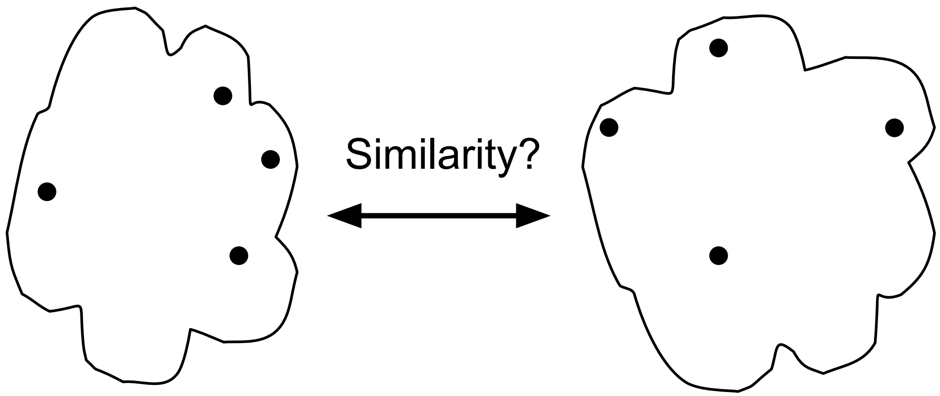




Proximity Matrix



How to Define Inter-Cluster Similarity

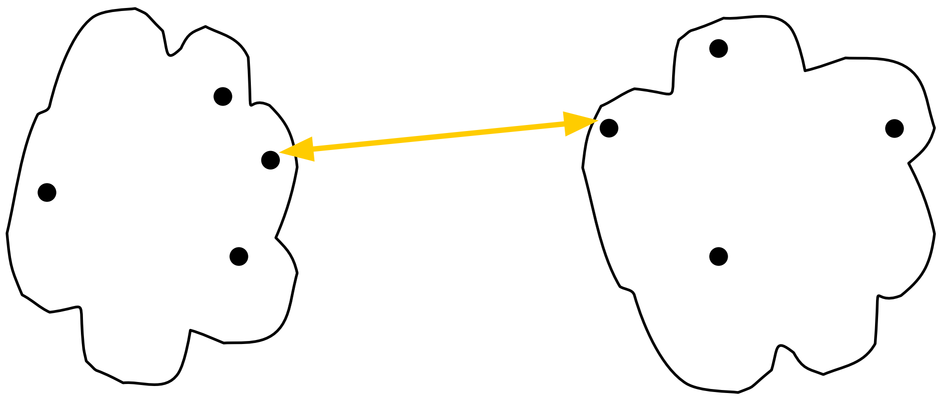


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function

How to Define Inter-Cluster Similarity

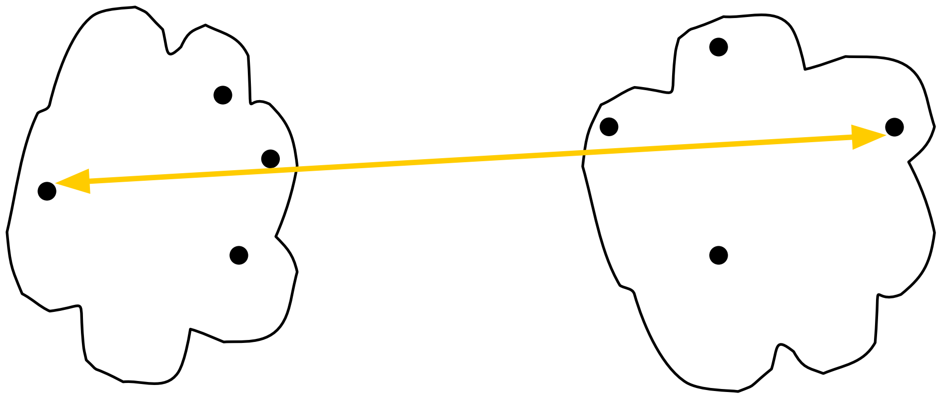


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function

How to Define Inter-Cluster Similarity

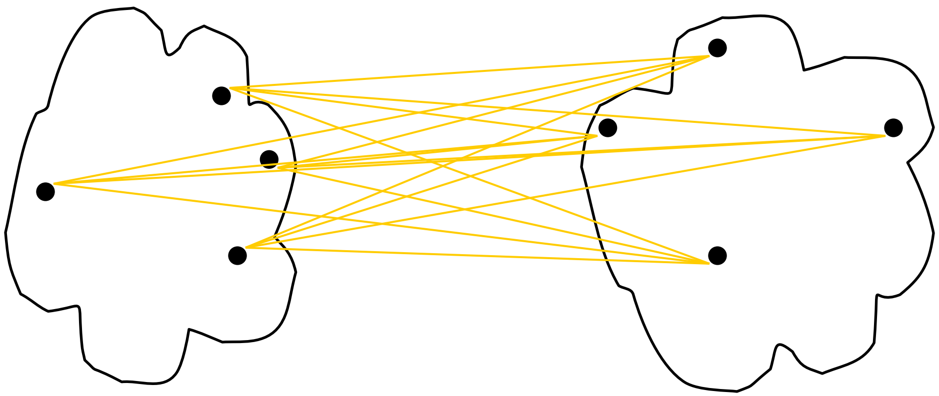


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function

How to Define Inter-Cluster Similarity

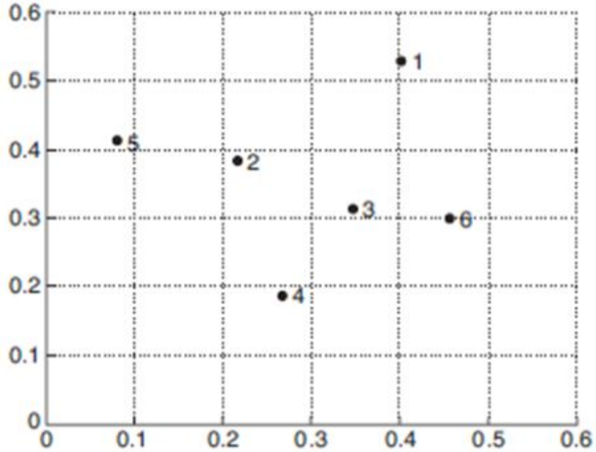


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function

Clustering Example



Set of 6 two-dimensional points.

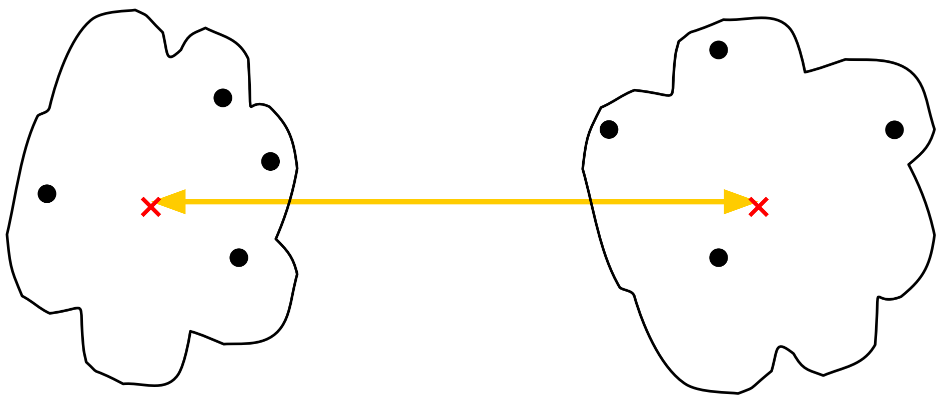
Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

xy coordinates of 6 points.

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Euclidean distance matrix for 6 points.

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Euclidean distance matrix for 6 points.

	P1	P2	P3, P6	P4	P5
P1	0	.24	.22	.37	.34
P2		0	.15	.2	.14
P3, P6			0	.155	.28
P4				0	.29
P5					0

	P1	P2,P5	P3,P6	P4
P1	0	.24	.22	.37
P2,P5		0	.15	.2
P3,P6			0	.155
P4				0

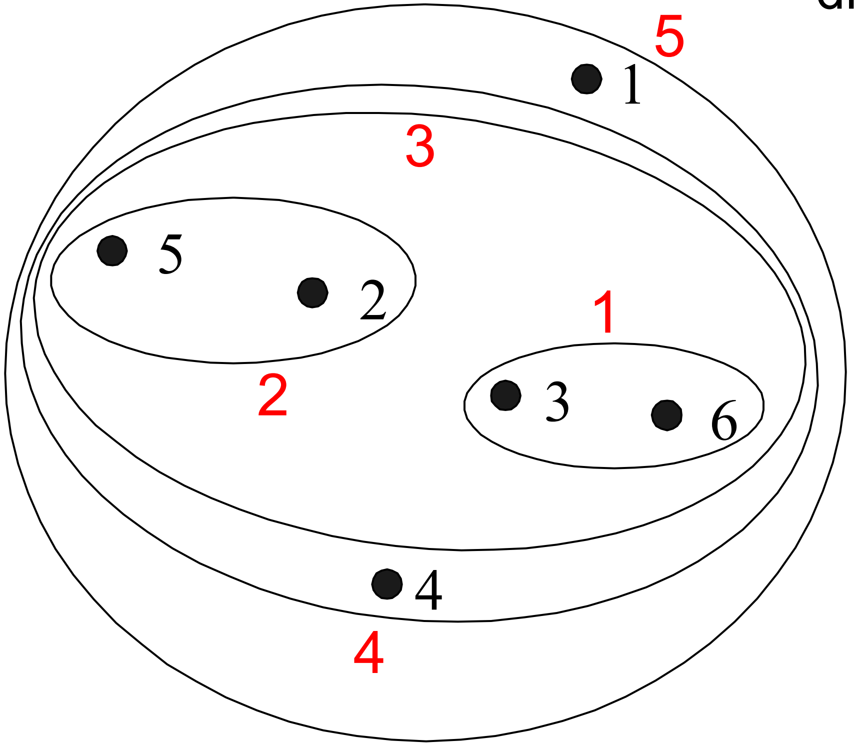
	P1	P2,P5, P3,P6	P4
P1	0	.22	.37
P2,P5, P3,P6		0	.155
P4			0

Hierarchical Clustering example - min

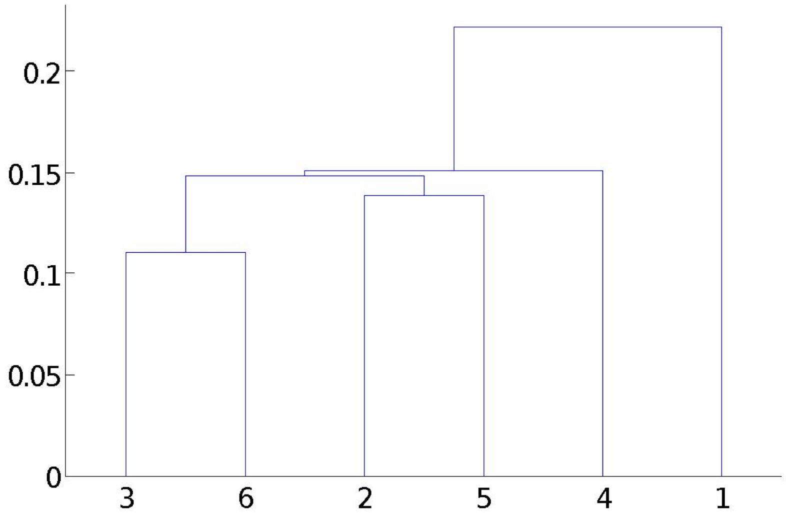
	P1	P2,P5,P3, P6,P4
P1	0	0.22
P2,P5,P3, P6,P4		0

Hierarchical Clustering: MIN

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points, i.e., by one link in the proximity graph.

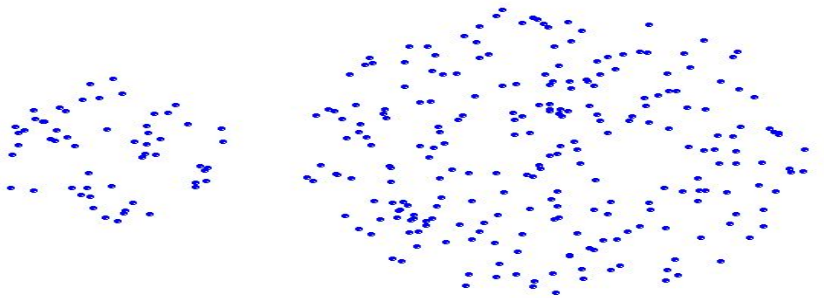


Nested Clusters

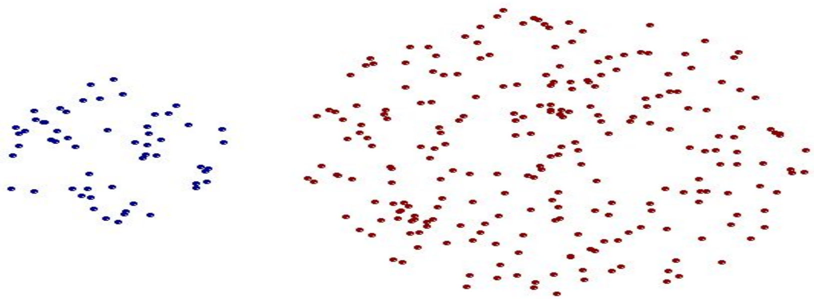


Dendrogram

Strength of MIN



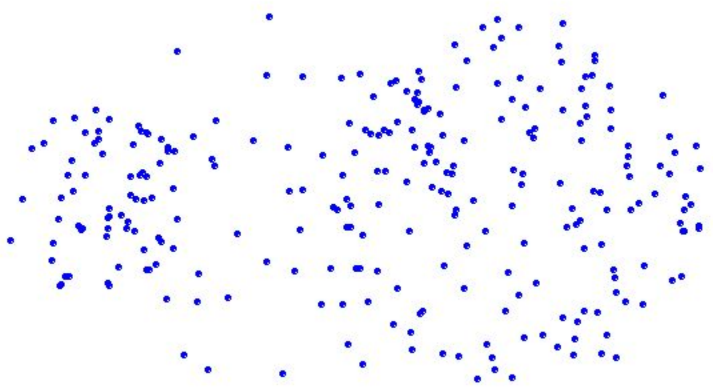
Original Points



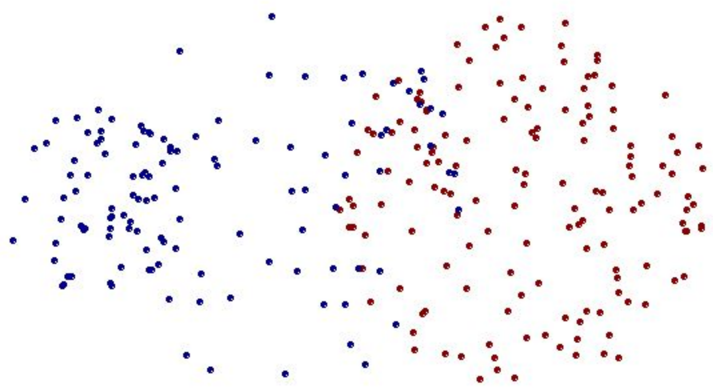
Two Clusters

• Can handle non-elliptical shapes

Limitations of MIN



Original Points



Two Clusters

- Sensitive to noise and outliers

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Euclidean distance matrix for 6 points.

	P1	P2	P3, P6	P4	P5
P1	0	.24	.23	.37	.34
P2		0	.25	.2	.14
P3, P6			0	.22	.39
P4				0	.29
P5					0

	P1	P2,P5	P3,P6	P4
P1	0	.34	.23	.37
P2,P5		0	.39	.29
P3,P6			0	.22
P4				0

	P1	P2,P5	P3,P6, P4
P1	0	.34	.37
P2,P5		0	.39
P3,P6, P4			0

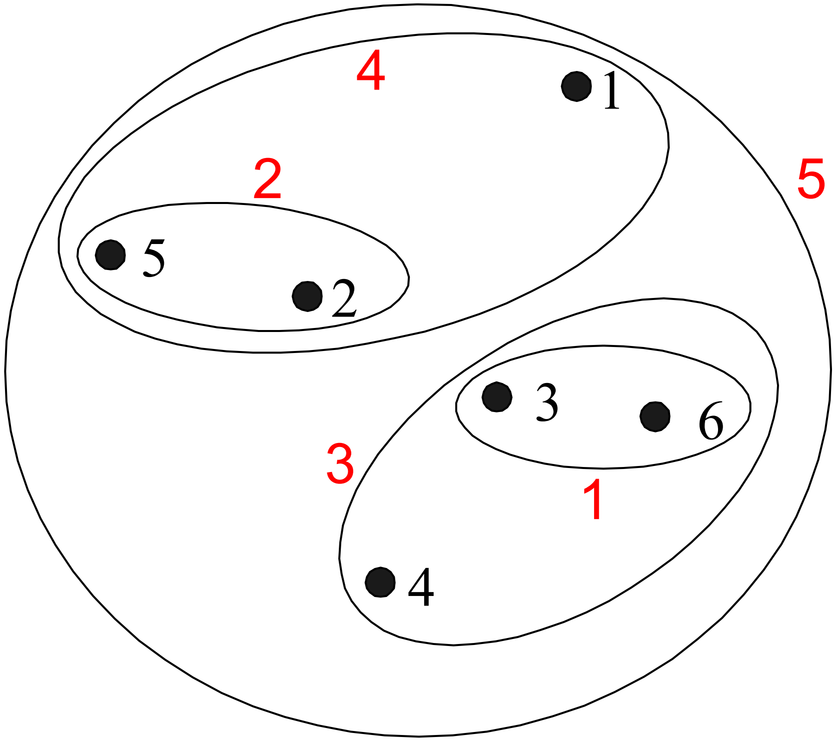
	P1, P2,P5	P3, P6,P4
P1, P2,P5	0	0.39
P3, P6,P4		0

Hierarchical Clustering example - max

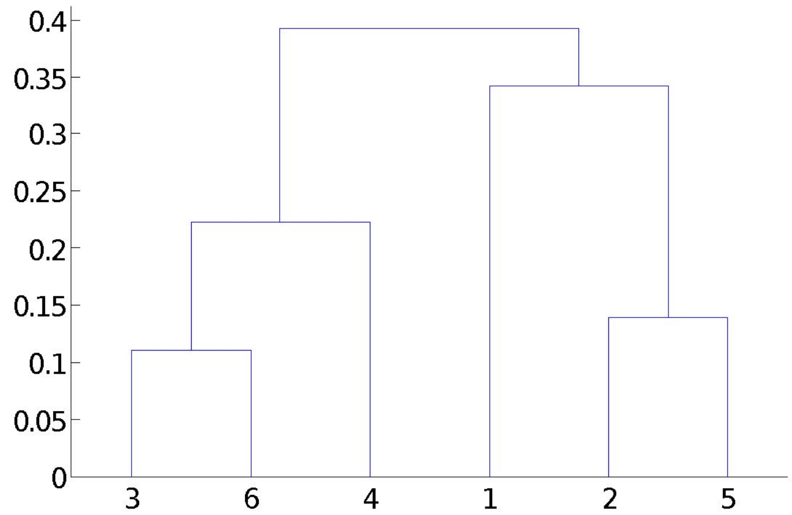
Hierarchical Clustering: MAX

Similarity of two clusters is based on the two least similar (most distant) points in the different clusters

Determined by all pairs of points in the two clusters

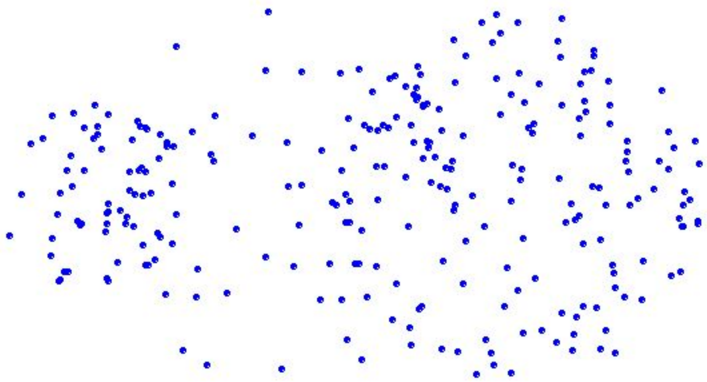


Nested Clusters

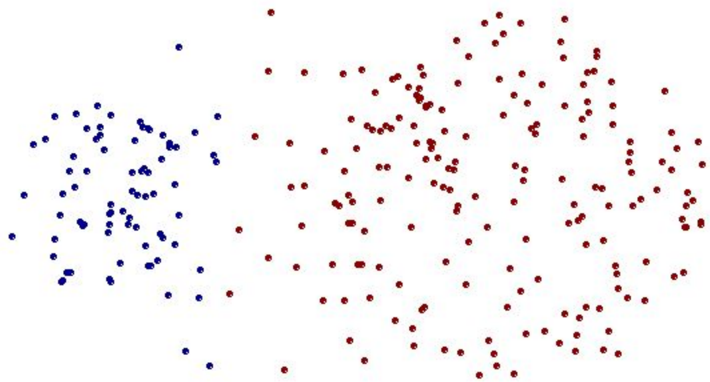


Dendrogram

Strength of MAX



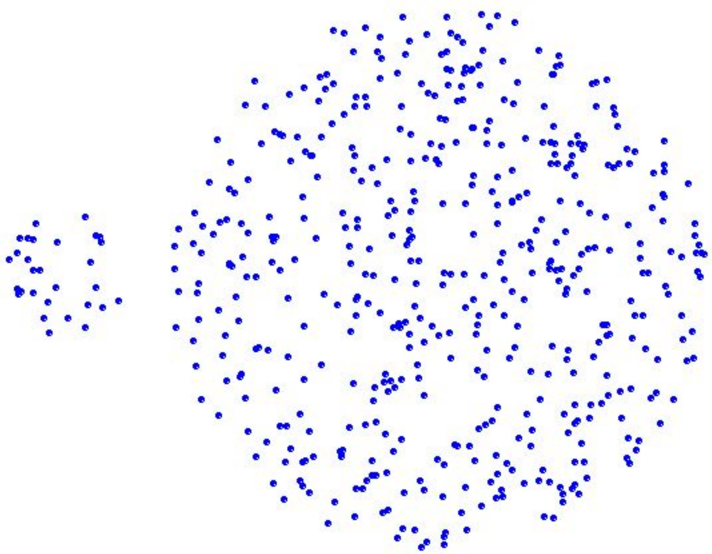
Original Points



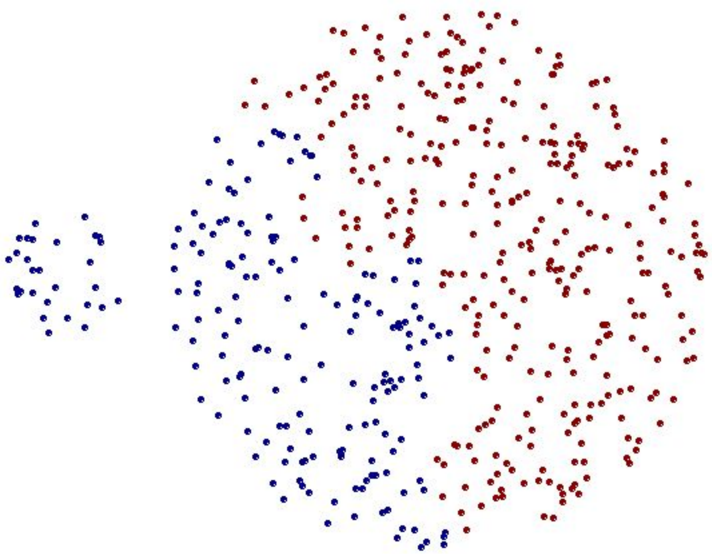
Two Clusters

- Less susceptible to noise and outliers

Limitations of MAX



Original Points

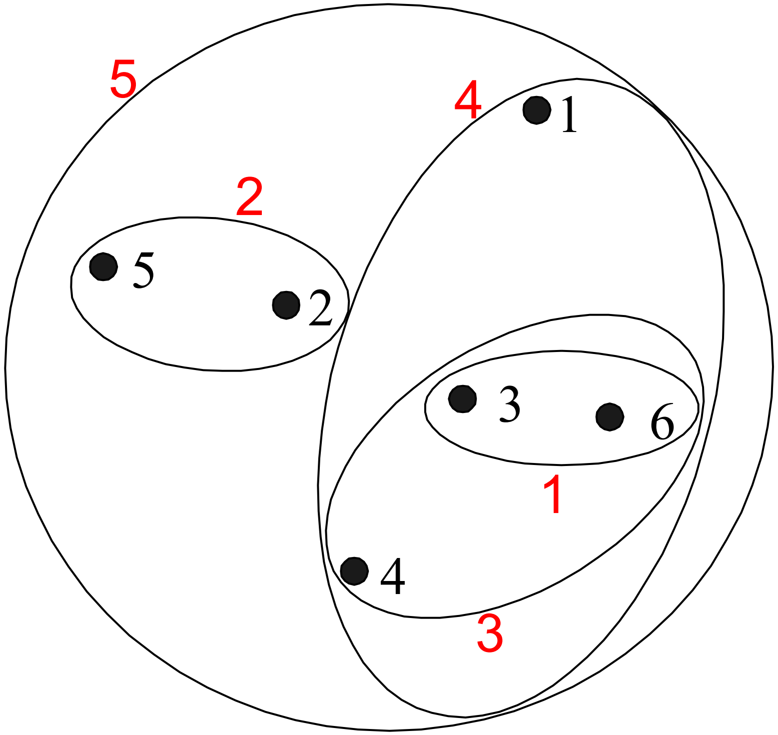


Two Clusters

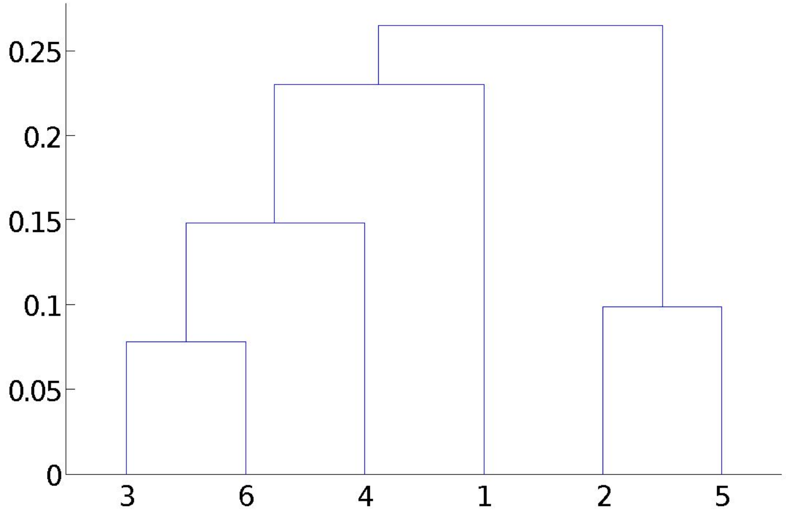
- Tends to break large clusters
- Biased towards globular clusters

Hierarchical Clustering: Group Average

Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

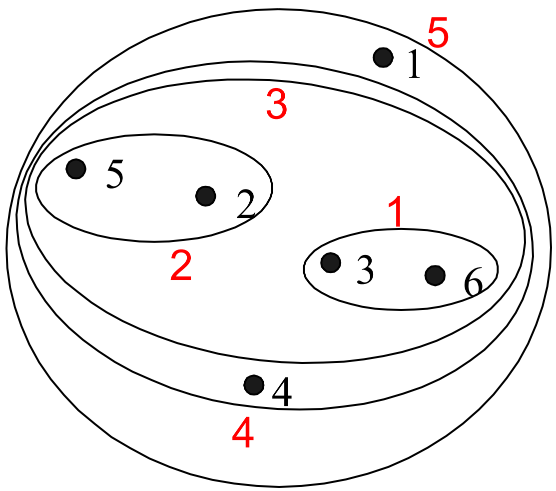


Nested Clusters



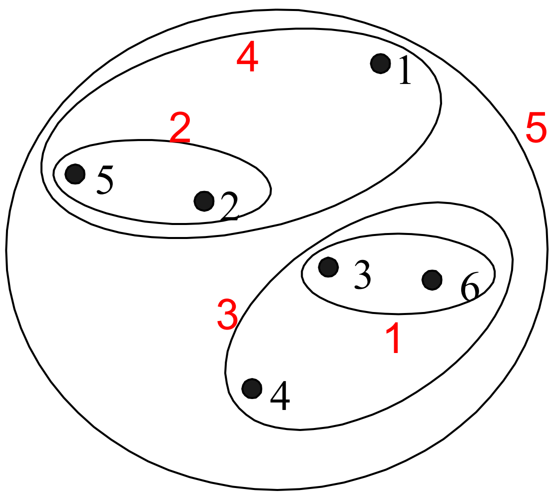
Dendrogram

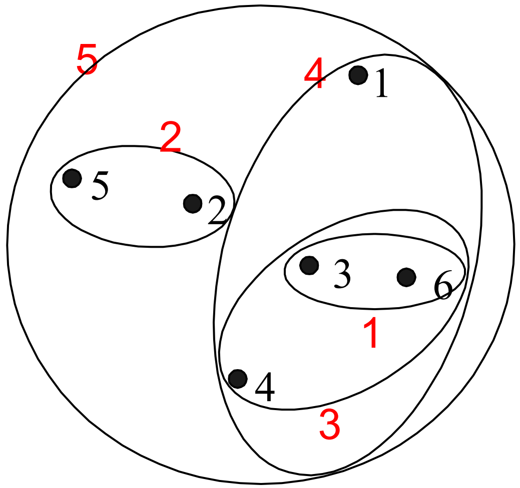
Hierarchical Clustering: Comparison



MIN

MAX





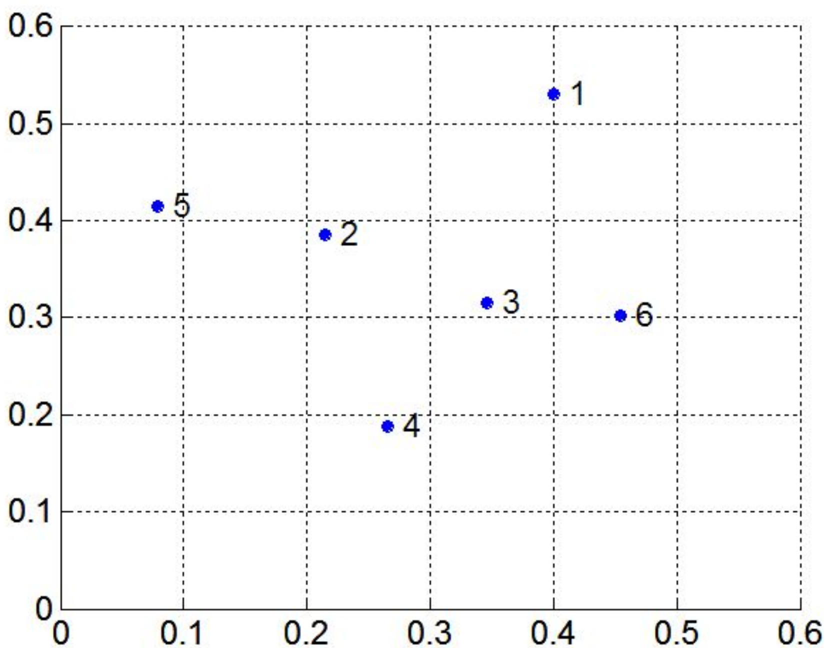
Group Average

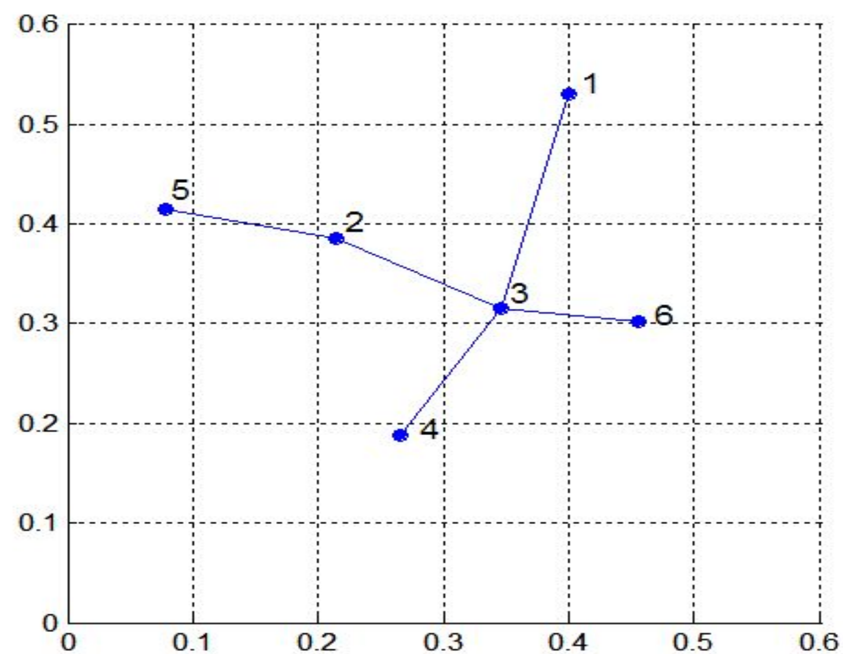
MST: Divisive Hierarchical Clustering

- Build MST (Minimum Spanning Tree)
  - Start with a tree that consists of any point
  - In successive steps, look for the closest pair of points (p, q) such that one point (p) is in the current tree but the other (q) is not
  - Add q to the tree and put an edge between p and q

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Euclidean distance matrix for 6 points.





MST: Divisive Hierarchical Clustering

- Use MST for constructing hierarchy of clusters

Algorithm 7.5 MST Divisive Hierarchical Clustering Algorithm

1: Compute a minimum spanning tree for the proximity graph.
2: **repeat**
3:     Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).
4: **until** Only singleton clusters remain

Hierarchical Clustering: Time and Space requirements

O(N^2) space since it uses the proximity matrix.

- N is the number of points.

O(N^3) time in many cases

- There are N steps and at each step the size, N^2, proximity matrix must be updated and searched
- Complexity can be reduced to O(N^2 log(N)) time for some approaches

Hierarchical Clustering: Problems and Limitations

Once a decision is made to combine two clusters, it cannot be undone

No objective function is directly minimized

Different schemes have problems with one or more of the following:

- Sensitivity to noise and outliers
- Difficulty handling different sized clusters and convex shapes
- Breaking large clusters

innovate achieve lead

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Density based Cluster Analysis

Density-Based Clustering Methods

Clustering based on density (local cluster criterion), such as density-connected points

Major features:

- Discover clusters of arbitrary shape
- Handle noise
- One scan
- Need density parameters as termination condition

Several interesting studies:

- DBSCAN: Ester, et al. (KDD'96)
- OPTICS: Ankerst, et al (SIGMOD'99).
- DENCLUE: Hinneburg & D. Keim (KDD'98)
- CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

Density-Based Clustering: Basic Concepts

Two parameters:

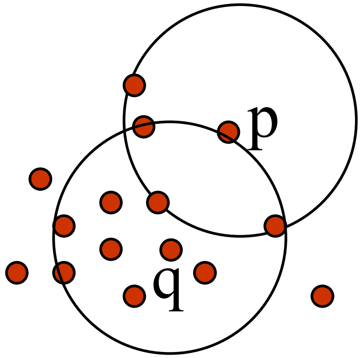
- *Eps*: Maximum radius of the neighbourhood
- *MinPts*: Minimum number of points in an Eps-neighbourhood of that point

$N_{Eps}(p)$: {q belongs to D | dist(p,q) ≤ Eps}

Directly density-reachable: A point *p* is directly density-reachable from a point *q* w.r.t. *Eps*, *MinPts* if

- *p* belongs to $N_{Eps}(q)$
- core point condition:

$$|N_{Eps}(q)| \geq MinPts$$



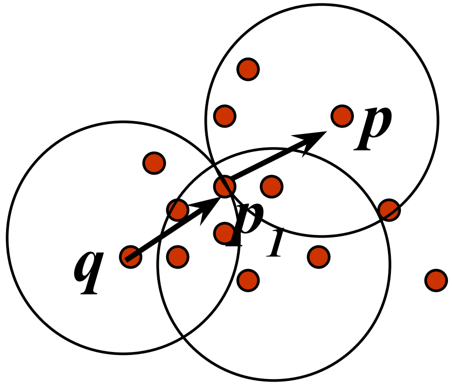
MinPts = 5

Eps = 1 cm

Density-Reachable and Density-Connected

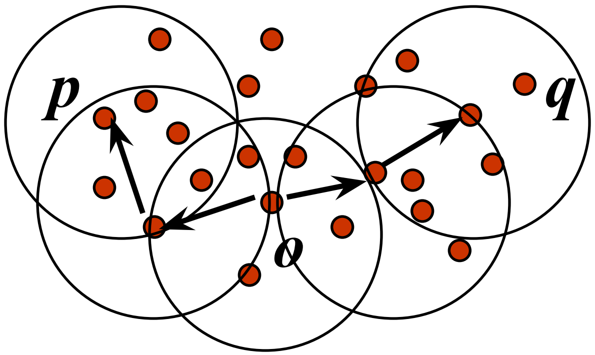
Density-reachable:

- A point p is density-reachable from a point q w.r.t. *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i



Density-connected

- A point p is density-connected to a point q w.r.t. *Eps*, *MinPts* if there is a point o such that both, p and q are density-reachable from o w.r.t. *Eps* and *MinPts*



DBSCAN

DBSCAN is a density-based algorithm.

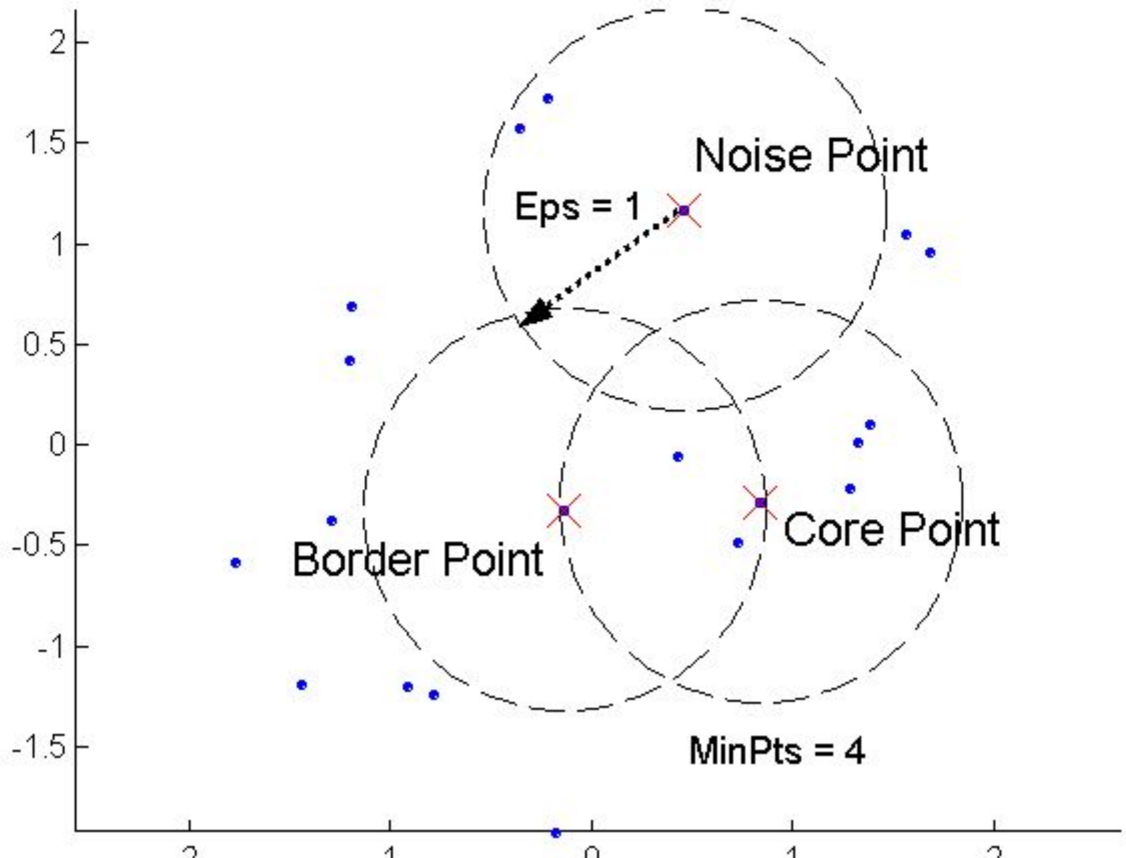
- Density = number of points within a specified radius (Eps)

- A point is a **core point** if it has more than a specified number of points (MinPts) within Eps
  - These are points that are at the interior of a cluster

- A **border point** has fewer than MinPts within Eps, but is in the neighborhood of a core point

- A **noise point** is any point that is not a core point or a border point.

DBSCAN: Core, Border, and Noise Points



DBSCAN Algorithm

- Eliminate noise points
- Perform clustering on the remaining points

$current\_cluster\_label \leftarrow 1$

for all core points **do**

  if the core point has no cluster label **then**

    $current\_cluster\_label \leftarrow current\_cluster\_label + 1$

    Label the current core point with cluster label $current\_cluster\_label$

  end if

  for all points in the Eps-neighborhood, except i^{th} the point itself **do**

    if the point does not have a cluster label **then**

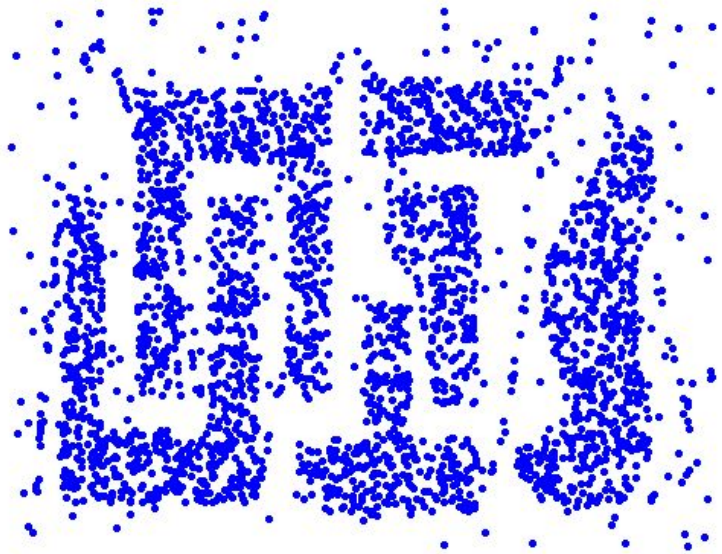
      Label the point with cluster label $current\_cluster\_label$

    end if

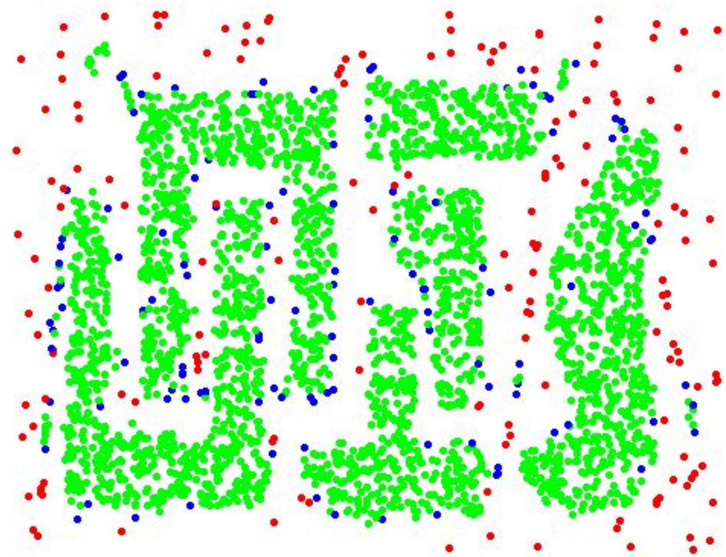
  end for

end for

DBSCAN: Core, Border and Noise Points



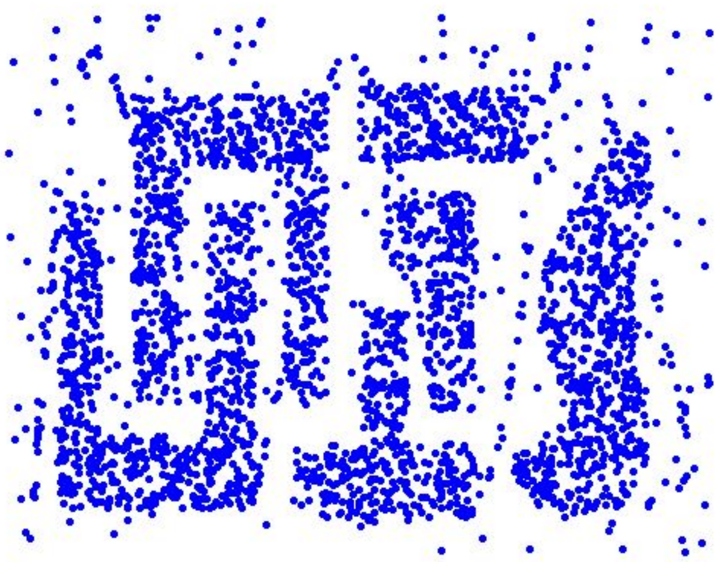
Original Points



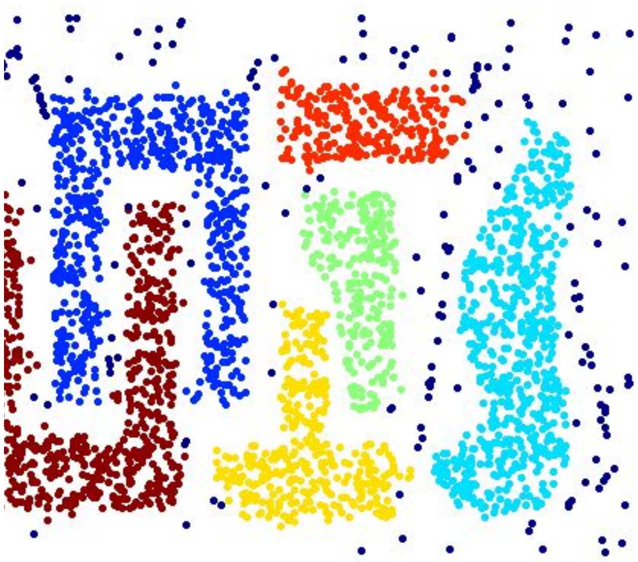
Point types: core, border and noise

Eps = 10, MinPts = 4

When DBSCAN Works Well



Original Points

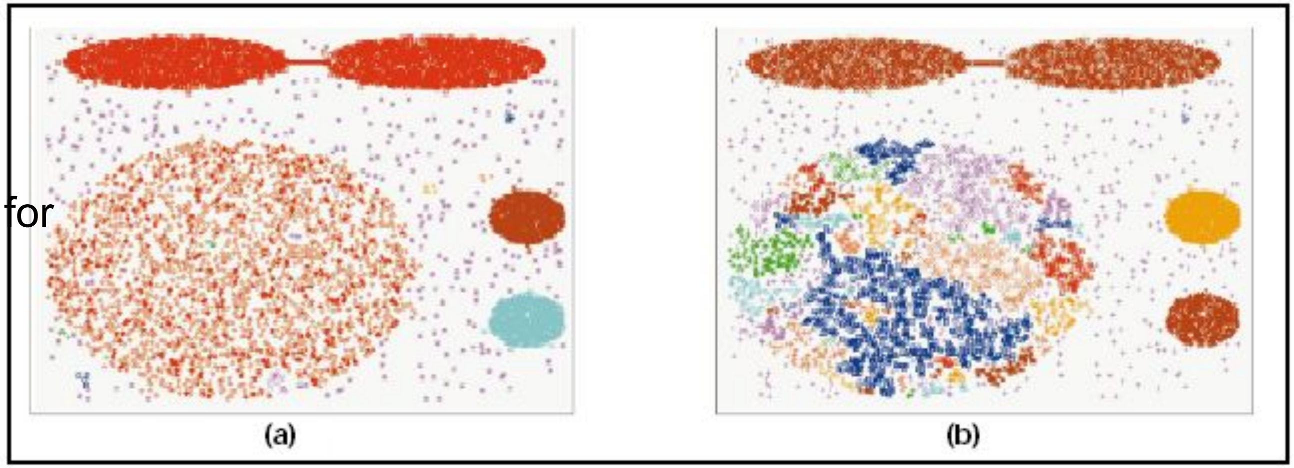


Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

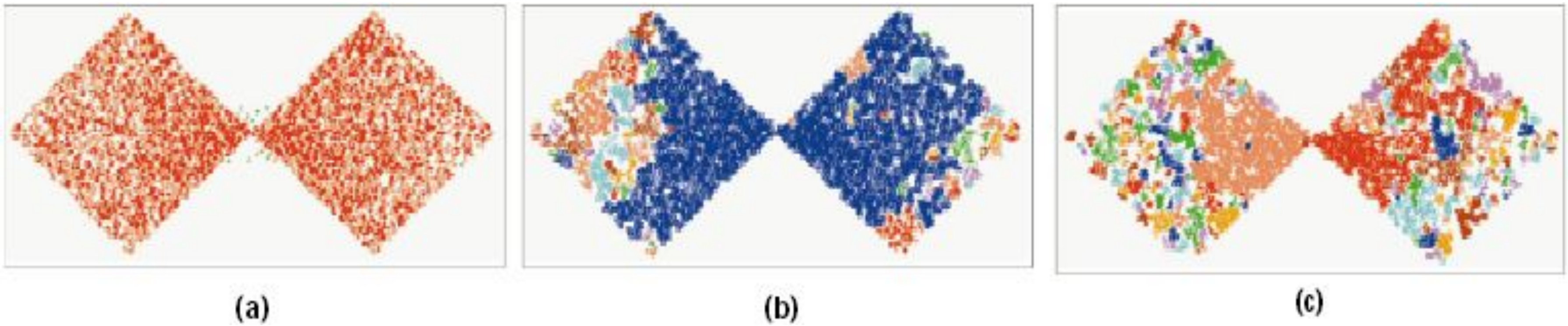
DBSCAN: Sensitive to Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.



DBSCAN does not work well for
- Varying densities
- High-dimensional data

Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.

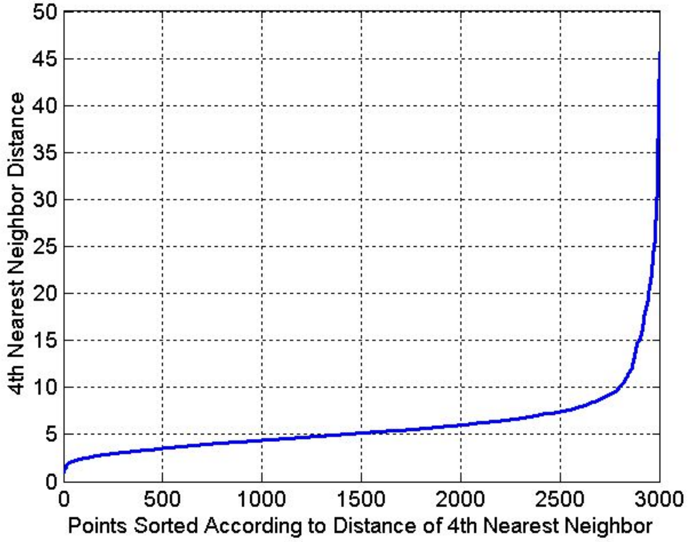


DBSCAN: Determining EPS and MinPts

Idea is that for points in a cluster, their k^{th} nearest neighbors are at roughly the same distance

Noise points have the k^{th} nearest neighbor at farther distance

So, plot sorted distance of every point to its k^{th} nearest neighbor



Prescribed Text Books

	Author(s), Title, Edition, Publishing House
T1	Tan P. N., Steinbach M & Kumar V. “Introduction to Data Mining” Pearson Education
T2	Data Mining: Concepts and Techniques, Third Edition by Jiawei Han, Micheline Kamber and Jian Pei Morgan Kaufmann Publishers
R1	Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner by Vijay Kotu and Bala Deshpande Morgan Kaufmann Publishers