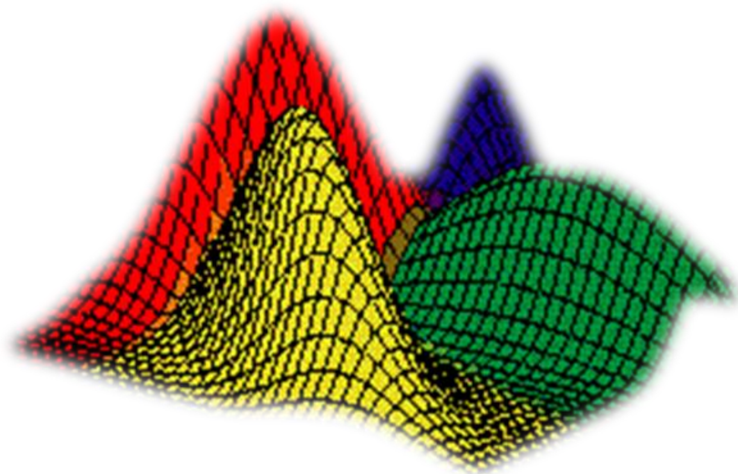


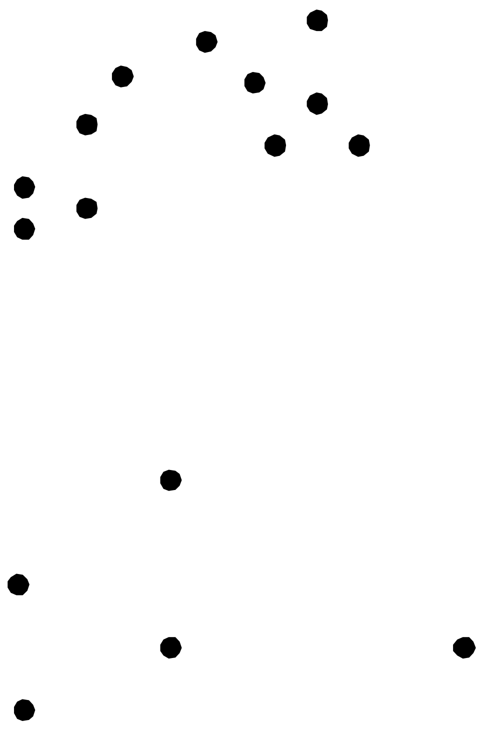
CSE 473: Pattern Recognition

Unsupervised Learning:
Clustering

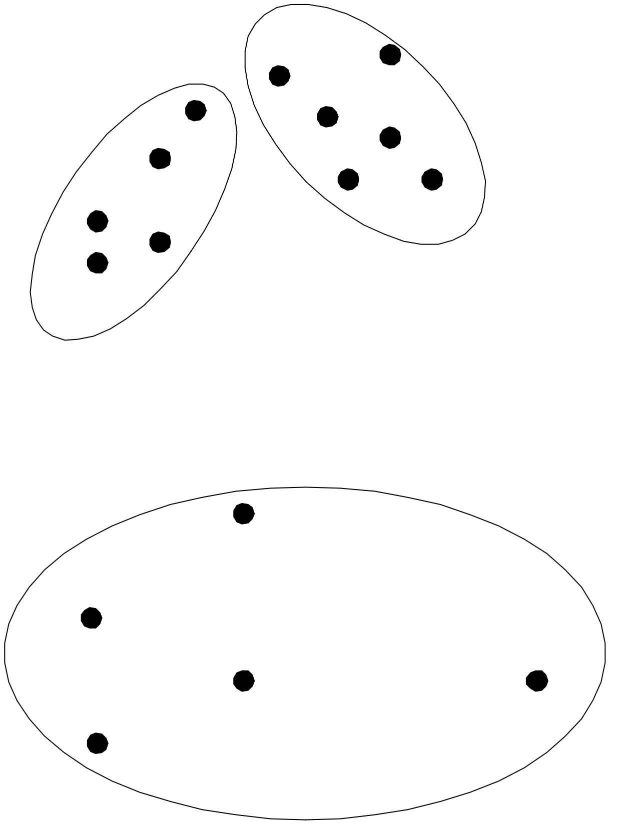
Clustering Algorithms

- 2 Important distinctions: hierarchical and partitional
- 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

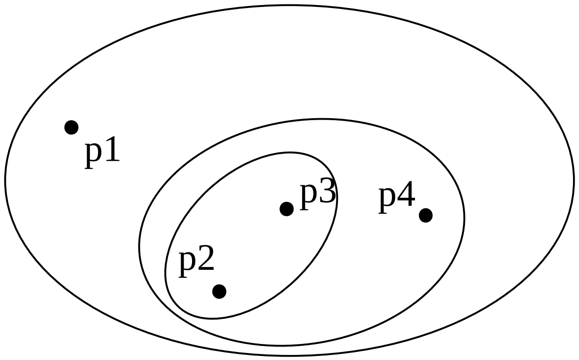


Original Points

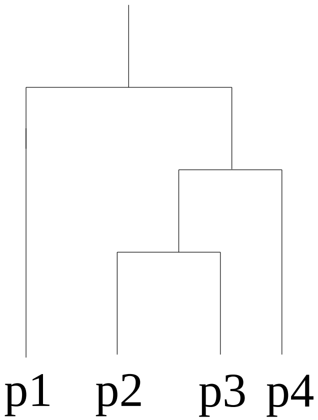


A Partitional Clustering

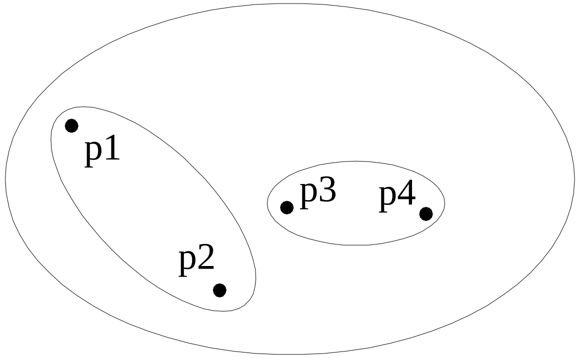
Hierarchical Clustering



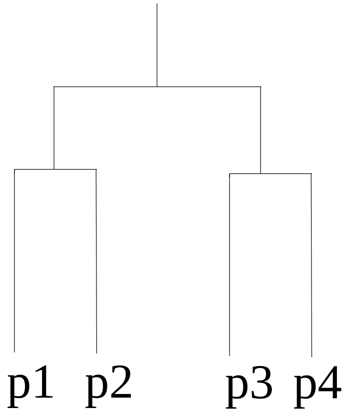
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Hierarchical Clustering

- Hierarchical: A sequence of (nested) clusterings is produced.
  - Agglomerative
  - Divisive
  - Combinations of the above (e.g., the Chameleon algorithm.)

OTHER MAJOR CATEGORIES OF CLUSTERING ALGORITHMS

- Sequential: A single clustering is produced. One or few sequential passes on the data.

- Cost function optimization. For most of the cases a *single* clustering is obtained.
  - Hard clustering (each point belongs exclusively to a single cluster):
    - Basic hard clustering algorithms (e.g., k-means)
    - Branch and bound
    - Boundary detection
    - Genetic clustering algorithms
    - …
    - ..
    - .
  - Fuzzy clustering (each point belongs to more than one clusters simultaneously).
  - Probabilistic clustering (it is based on the *probability* of a point to belong to a cluster).

- Other schemes:
  - Exclusive versus non-exclusive
    - In non-exclusive clusterings, points may belong to multiple clusters.
    - Can represent multiple classes or ‘border’ points
  - Partial versus complete
    - In some cases, we only want to cluster some of the data
  - Heterogeneous versus homogeneous
    - Cluster of widely different sizes, shapes, and densities

- Other schemes:
  - Algorithms based on graph theory (e.g., Minimum Spanning Tree, regions of influence, directed trees).
  - Density based clustering algorithms.
  - Kernel based methods
  - . . . .
  - . . .
  - . .

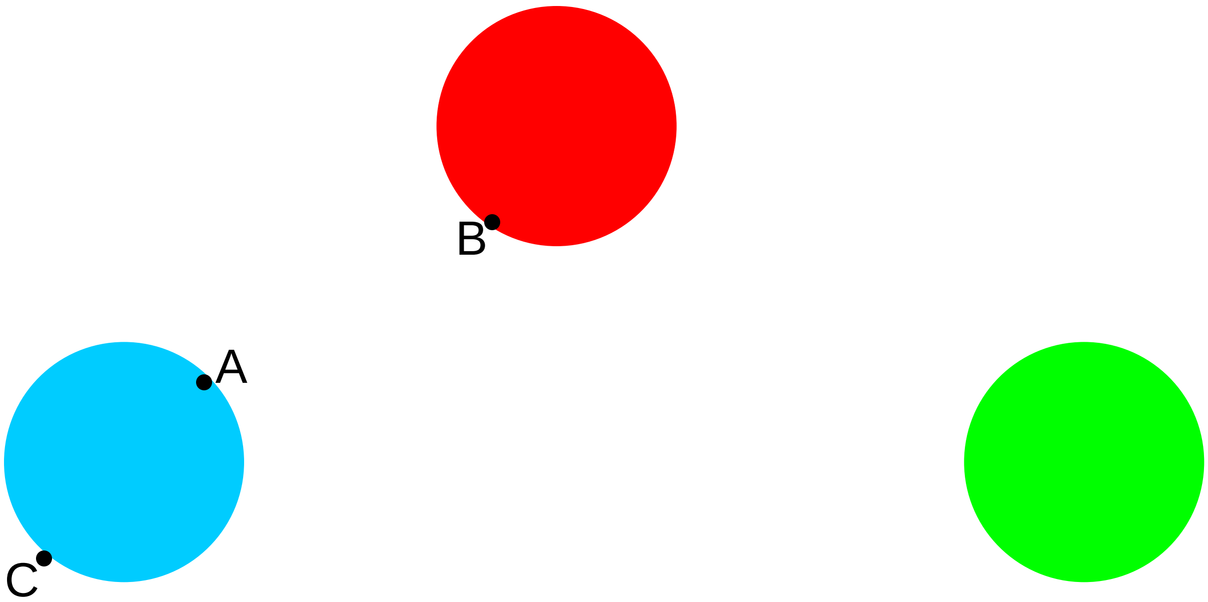
Clustering and Clusters :Difference?

Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

Types of Clusters: Well-Separated

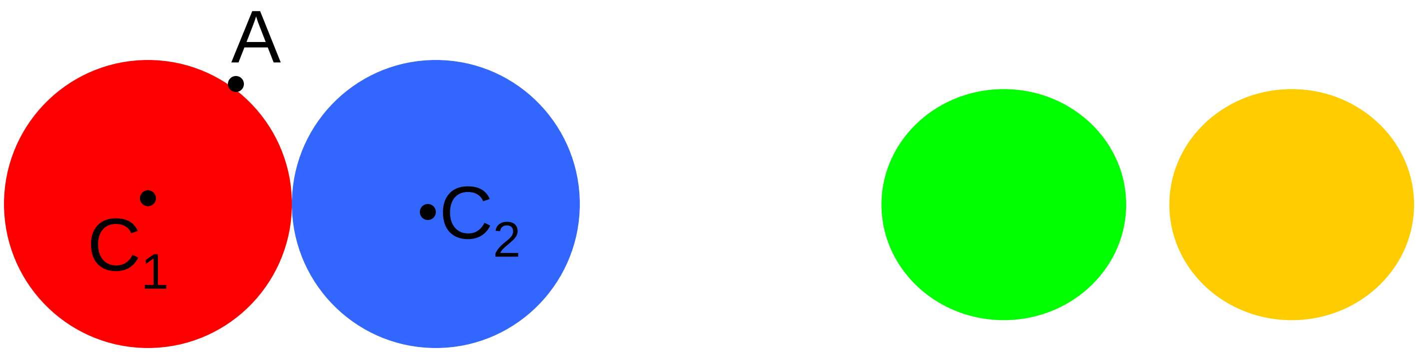
- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

Types of Clusters: Center-Based

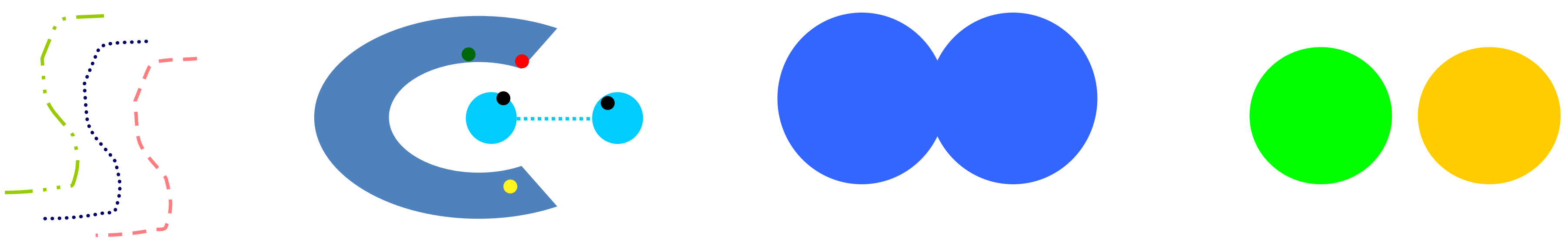
- Center-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster



4 center-based clusters

Types of Clusters: Contiguity-Based

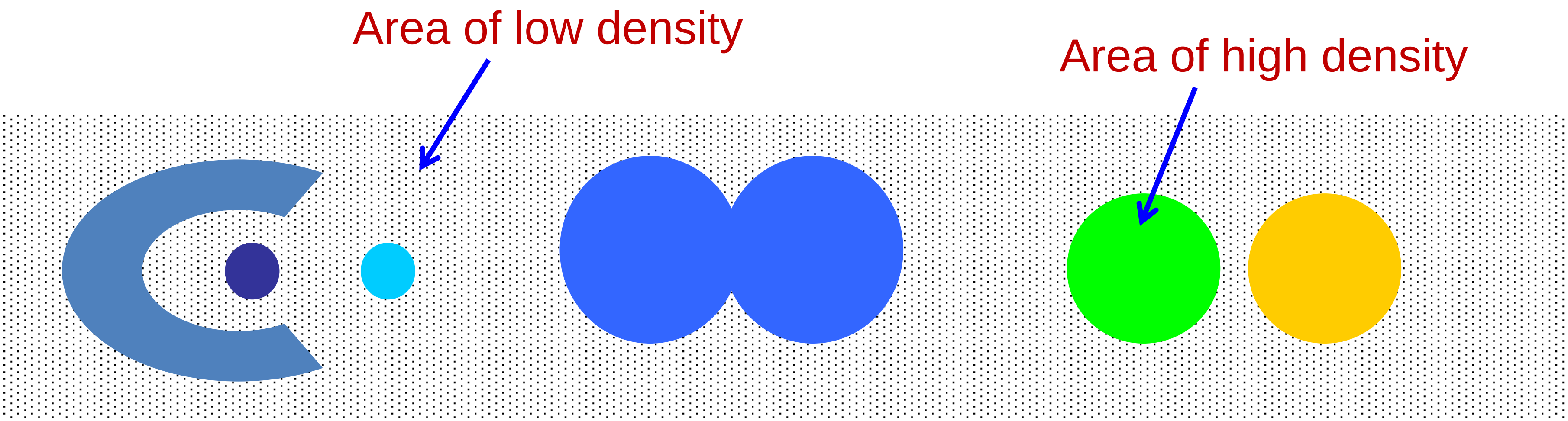
- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



8 contiguous clusters

Types of Clusters: Density-Based

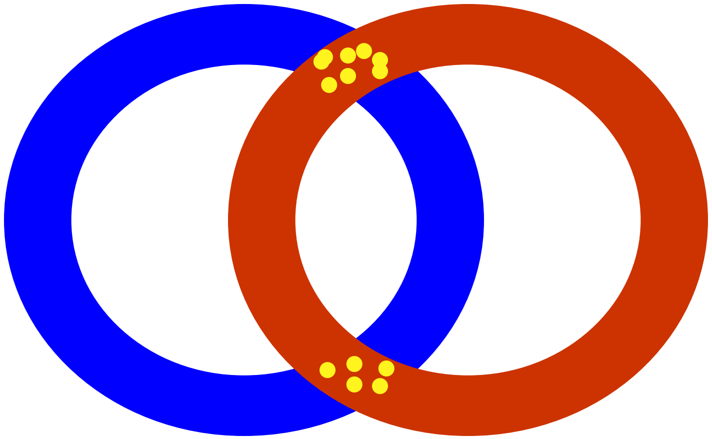
- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

Types of Clusters: Conceptual Clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.



2 Overlapping Circles

SEQUENTIAL CLUSTERING ALGORITHMS

- Characteristics:
  - One or very few passes on the data
  - No a-prior knowledge about the number of clusters
  - The clusters are defined with the aid of
    - An appropriately defined distance $d\ (\underline{x}, C)$ of a point (x) from a cluster (C).
    - A threshold Θ associated with the distance.

➢ Basic Sequential Clustering Algorithm (BSAS)

- $m=1$ \ {number of clusters}\
- $C_m=\{\underline{x}_1\}$
- For $i=2$ to N
  - Find C_k: $d(\underline{x}_i,C_k)=min_{1\leq j\leq m}d(\underline{x}_i,C_j)$
  - If $(d(\underline{x}_i,C_k)>\Theta)$ AND $(m<q)$ then
    - o $m=m+1$
    - o $C_m=\{\underline{x}_i\}$
  - Else
    - o $C_k=C_k\cup\{\underline{x}_i\}$
    - o Where necessary, update representatives (*)
  - End {if}
- End {for}

Basic Sequential Clustering Algorithm (BSAS)

- $m=1$ \{number of clusters}\
- $C_m=\{\underline{x}_1\}$
- For $i=2$ to N
  - Find C_k: $d(\underline{x}_i,C_k)=min_{1\leq j\leq m}d(\underline{x}_i,C_j)$
  - If $(d(\underline{x}_i,C_k)>\Theta)$ AND $(m<q)$ then
    - o $m=m+1$
    - o $C_m=\{\underline{x}_i\}$
  - Else
    - o $C_k=C_k\cup\{\underline{x}_i\}$
    - o Where necessary, update representatives (*)
  - End {if}
- End {for}

(*) When the mean vector $\underline{m}_C$ is used as representative of the cluster C with n_c elements, the updating in the light of a new vector $\underline{x}$ becomes

$$\underline{m}_C^{new}=(n_C\,\underline{m}_C+\underline{x})\,/\,(n_C+1)$$

➢ Remarks:

- The order of presentation of the data in the algorithm plays important role in the clustering results.
  - Different order of presentation may lead to totally different clustering results

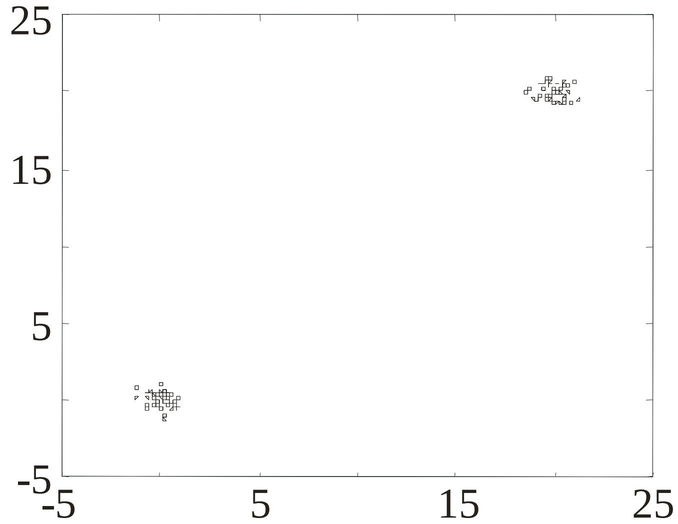
➢ Remarks:

- The order of presentation of the data in the algorithm plays important role in the clustering results.
  - Different order of presentation may lead to totally different clustering results
- In BSAS the decision for a vector $\underline{x}$ is reached prior to the final cluster formation.

➢ Remarks:

- The order of presentation of the data in the algorithm plays important role in the clustering results.
  - Different order of presentation may lead to totally different clustering results
- In BSAS the decision for a vector $\underline{x}$ is reached prior to the final cluster formation.
- BSAS perform a single pass on the data. Its complexity is $O(N)$.

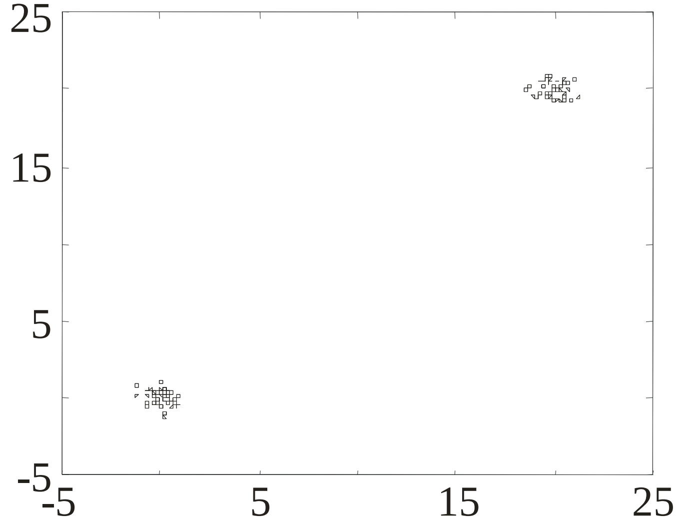
- Estimating the number of clusters in the data set:



Estimating the number of clusters in the data set:

Let $BSAS(\Theta)$ denote the $BSAS$ algorithm when the dissimilarity threshold is Θ.

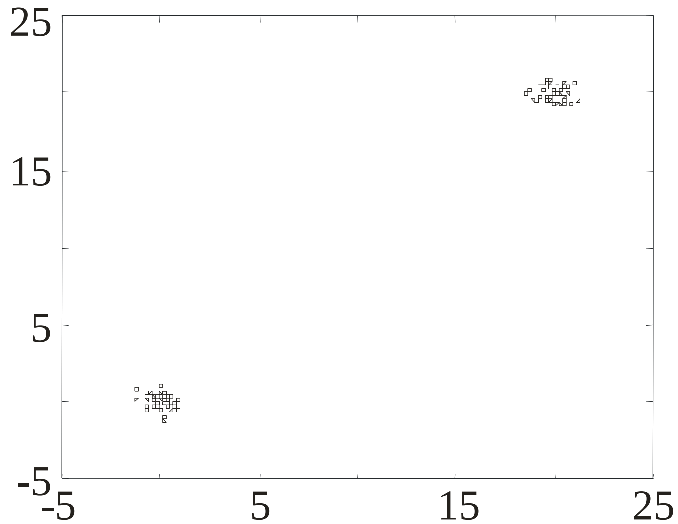
- Find
  - a, the maximum and
  - b, the minimum distances among the points



Estimating the number of clusters in the data set:

Let $BSAS(\Theta)$ denote the $BSAS$ algorithm when the dissimilarity threshold is Θ.

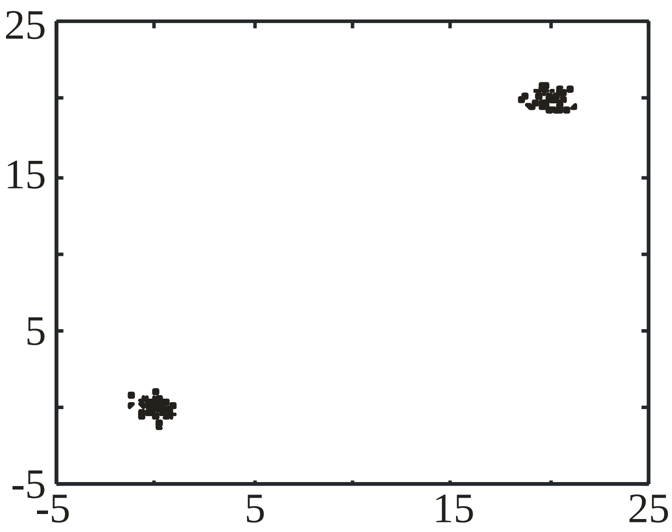
- For $\Theta=a$ to b step c
  - Run s times $BSAS(\Theta)$, each time presenting the data in a different order.
  - Estimate the number of clusters m_{Θ}, as the most frequent number resulting from the s runs of $BSAS(\Theta)$.
- Next Θ

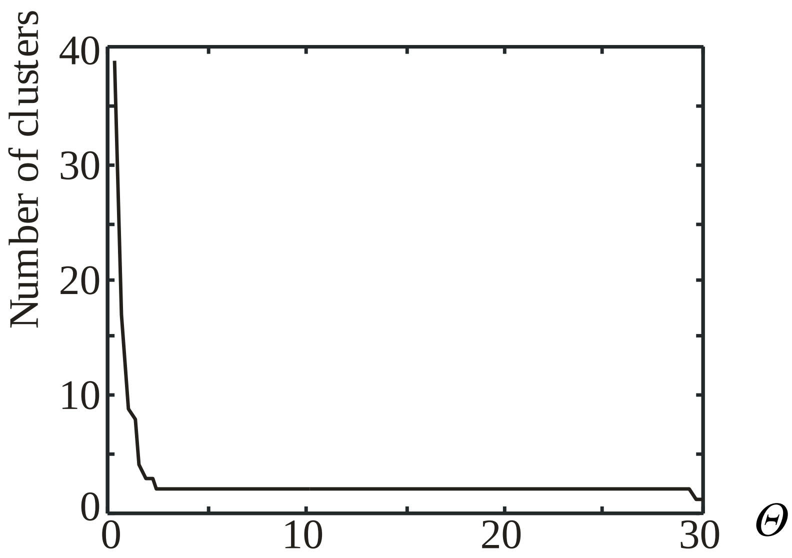


- Estimating the number of clusters in the data set:

Let $BSAS(\Theta)$ denote the $BSAS$ algorithm when the dissimilarity threshold is Θ.

- For $\Theta=a$ to b step c
  - Run s times $BSAS(\Theta)$, each time presenting the data in a different order.
  - Estimate the number of clusters m_{Θ}, as the most frequent number resulting from the s runs of $BSAS(\Theta)$.
- Next Θ
- Plot m_{Θ} versus Θ and identify the number of clusters m as the one corresponding to the widest flat region in the above graph.





➢ MBSAS, a modification of BSAS

In BSAS a decision for a data vector $\underline{x}$ is reached prior to the final cluster formation, which is determined after all vectors have been presented to the algorithm.

➢ MBSAS, a modification of BSAS

In BSAS a decision for a data vector $\underline{x}$ is reached prior to the final cluster formation, which is determined after all vectors have been presented to the algorithm.

- MBSAS deals with the above drawback, at the cost of presenting the data twice to the algorithm.

MBSAS, a modification of BSAS

MBSAS consists of:

- A cluster determination phase (first pass on the data),

  which is the same as BSAS with the exception that no vector is assigned to an already formed cluster. At the end of this phase, each cluster consists of a single element.

- A pattern ~~classification~~ phase (second pass on the data),

  where each one of the unassigned vector is assigned to its closest cluster.

➢ MBSAS, a modification of BSAS

- *Cluster Determination*
  - $m=1$ \ {number of clusters}\
  - $C_m=\{\underline{x}_1\}$
  - For $i=2$ to N
    - Find C_k: $d(\underline{x}_i,C_k)=min_{1\le j\le m}d(\underline{x}_i,C_j)$
    - If $(d(\underline{x}_i,C_k)>\Theta)$ AND $(m<q)$ then
      - o $m=m+1$
      - o $C_m=\{\underline{x}_i\}$
    - End {if}
  - End {for}

➢ MBSAS, a modification of BSAS

- *Pattern ~~Classification~~ Assignment*
  - For i= 1 to N
    - If $\boldsymbol{x}_i$ has not been assigned to a cluster, then
      - Find C_k: $d(\underline{x}_i, C_k) = min_{1 \le j \le m} d(\underline{x}_i, C_j)$
      - $C_k = C_k \cup \{\boldsymbol{x}_i\}$
      - where necessary, update representatives
    - End {if}
  - End {for}

MBSAS, a modification of BSAS

Remarks:

- decision for a vector $\underline{x}$ during the pattern classification phase is reached taking into account all clusters.
- still sensitive to the order of presentation of the vectors.
- requires two passes on the data. complexity is $O(N)$.

➢ The maxmin algorithm, a variant of MBSAS

Let X= the set of all data points

W = the set of all points that have been chosen to form clusters up to the current iteration step.

➢ The maxmin algorithm, a variant of MBSAS

cluster formation:

- For each $\underline{x} \in X\text{-}W$ determine $d_{\boldsymbol{x}} = min_{\underline{\boldsymbol{z}} \in W}\, d(\underline{x}, \underline{z})$
- Determine $\underline{y}$: $d_{\boldsymbol{y}} = max_{\underline{x} \in X\text{-}W} d_{\boldsymbol{x}}$
- If $d_{\boldsymbol{y}}$ is greater than a pre-specified threshold then
  - this vector forms a new cluster
- else
  - the cluster determination phase of the algorithm terminates.
- End {if}

pattern classification:

assign each unassigned vector to its closest cluster

➢Remarks:

- The maxmin algorithm is more computationally demanding than MBSAS.
- However, it is expected to produce better clustering results.

- A two-threshold sequential scheme (TTSAS)
  - The formation of the clusters, as well as the assignment of vectors to clusters, is carried out concurrently (like BSAS and unlike MBSAS)
  - Two thresholds Θ_1 and Θ_2 ($\Theta_1 < \Theta_2$) are employed

- A two-threshold sequential scheme (TTSAS)

  - The general idea is the following:
    - If the distance $d(\underline{x},C)$ of $\underline{x}$ from its closest cluster, C, is greater than Θ_2 then:
      - A new cluster represented by $\underline{x}$ is formed.
    - Else if $d(\underline{x},C)<\Theta_1$ then
      - $\underline{x}$ is assigned to C.
    - Else
      - The decision is postponed to a later stage.
    - End {if}

- A two-threshold sequential scheme (TTSAS)
  - The general idea is the following:
    - If the distance $d(\underline{x},C)$ of $\underline{x}$ from its closest cluster, C, is greater than Θ_2 then:
      - A new cluster represented by $\underline{x}$ is formed.
    - Else if $d(\underline{x},C)<\Theta_1$ then
      - $\underline{x}$ is assigned to C.
    - Else
      - The decision is postponed to a later stage.
    - End {if}

The unassigned vectors are presented iteratively to the algorithm until all of them are classified.

- Remarks:
  - In practice, a few passes (≥ 2) of the data set are required.
  - TTSAS is less sensitive to the order of data presentation, compared to BSAS.

- Refinement stages

Why necessary?

The problem of closeness of clusters: “*In all the above algorithms it may happen that two formed clusters lie very close to each other*”.

- Refinement stages
  - A simple merging procedure
    - (A) Find C_i, C_j ($i<j$) such that $d(C_i,C_j)=min_{k,r=1,\ldots,m,k\neq r}d(C_k,C_r)$
    - If $d(C_i,C_j)\leq M_1$ then \{ M_1 is a user-defined threshold \}
      - Merge C_i, C_j to C_i and eliminate C_j .
      - If necessary, update the cluster representative of C_i .
      - Rename the clusters $C_{j+1},\ldots,C_m$ to $C_j,\ldots,C_{m-1}$, respectively.
      - $m=m-1$
      - Go to (A)
    - Else
      - Stop
    - End {if}

- Reassignment of vectors

Why necessary?

The problem of sensitivity to the order of data presentation:

"A vector $\underline{x}$ may have been assigned to a cluster C_i at the current stage but another cluster C_j may be formed at a later stage that lies closer to $\underline{x}$"

- A simple reassignment procedure
  - For i=1 to N
    - Find C_j such that $d(\underline{x}_i,C_j)=min_{k=1,\ldots,m}d(\underline{x}_i,C_k)$
    - Set $b(i)=j$ \{ $b(i)$ is the index of the cluster that lies closet to $\underline{x}_i$ \}
  - End {for}
  - For j=1 to m
    - Set $C_j=\{\underline{x}_i \in X: b(i)=j\}$
    - If necessary, update representatives
  - End {for}