الگوریتم‌های خوشه‌بندی

# CLUSTERING ALGORITHMS

- Number of possible clusterings

  به چند طریق مختلف می‌توان N نقطه را در m گروه مختلف قرار داد.

  Let $X=\{\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_N\}$.

  Question: In how many ways the $N$ points can be assigned into m groups?

  $S(N,1)=1$
  $S(N,N)=1$
  $S(N,m)=0 \quad m>N$

  Answer:

$$S(N,m)=\frac{1}{m!}\sum_{i=0}^{m}(-1)^{m-1}\binom{m}{i}i^N$$

  - Examples:

$$S(15,3)=2\,375101$$
$$S(20,4)=45\,232115901$$
$$S(100,5)=10^{68}!!$$

❖ A way out:

➢ Consider only a small fraction of clusterings of $X$ and select a "sensible" clustering among them

- Question 1: Which fraction of clusterings is considered?
- Question 2: What "sensible" means?

- The answer depends on the specific clustering algorithm and the specific criteria to be adopted.

# MAJOR CATEGORIES OF CLUSTERING ALGORITHMS

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

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

- **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)
    - $k$-medoids algorithms
    - Mixture decomposition
    - Branch and bound
    - Simulated annealing
    - Deterministic annealing
    - Boundary detection
    - Mode seeking
    - Genetic clustering algorithms
  - Fuzzy clustering (each point belongs to more than one clusters simultaneously).
  - Possibilistic clustering (it is based on the *possibility* of a point to belong to a cluster).

- Other schemes:
  - Algorithms based on graph theory (e.g., Minimum Spanning Tree, regions of influence, directed trees).
  - Competitive learning algorithms (basic competitive learning scheme, Kohonen self organizing maps).
  - Subspace clustering algorithms.
  - Binary morphology clustering algorithms.

الگوریتم های خوشه بندی ترتیبی:

# SEQUENTIAL CLUSTERING ALGORITHMS

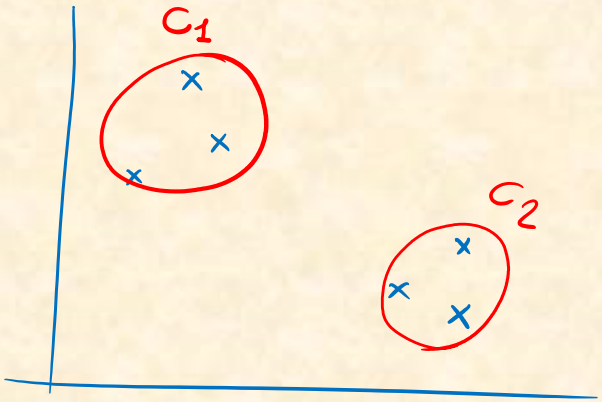
- The common traits shared by these algorithms are:

یک یا چند گام

  - One or very few passes on the data are required.
  - The number of clusters is not known a-priori, except (possibly) an upper bound, $q$. تعداد خوشه ها به صورت پیش فرض داده نمی شود. فقط حداکثر max
  - The clusters are defined with the aid of
    - An appropriately defined distance $d(x,C)$ of a point from a cluster. معیار فاصله از هر خوشه
    - A threshold $\Theta$ associated with the distance. حداکثر مقدار روی معیار d

# Basic Sequential Clustering Algorithm (BSAS)

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

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

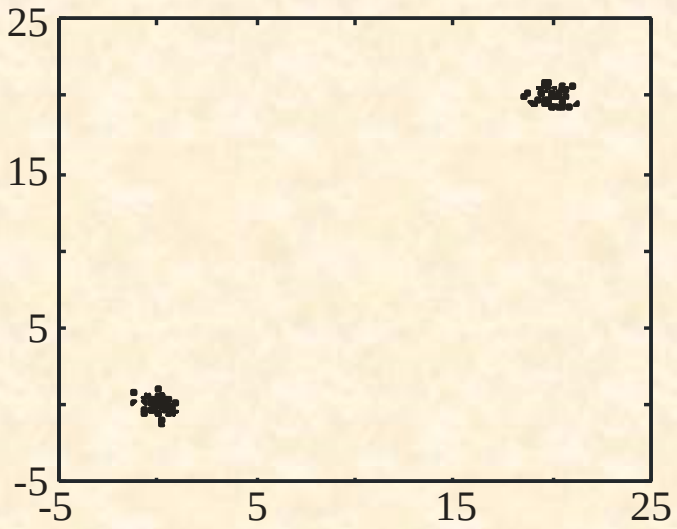
## ➢ Remarks:

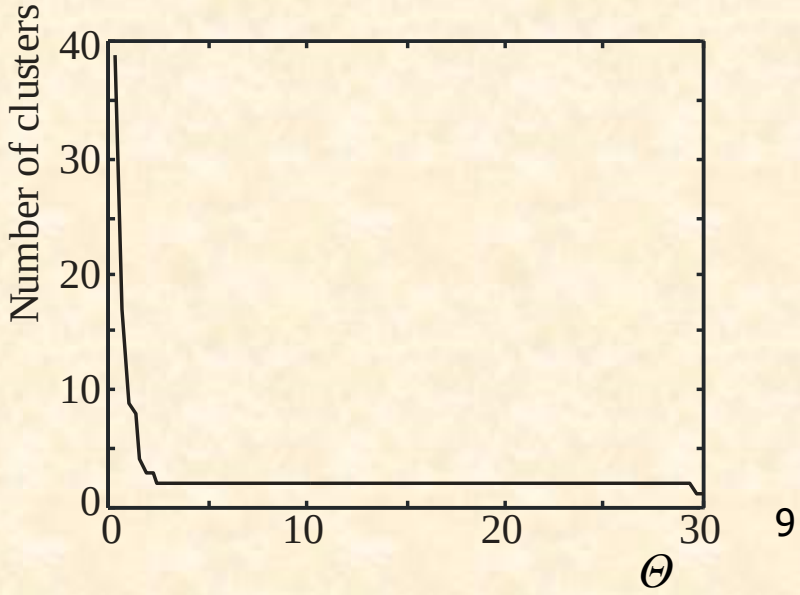
- The order of presentation of the data in the algorithm plays important role in the clustering results. Different order of presentation may lead to totally different clustering results, in terms of the number of clusters as well as the clusters themselves.
- In BSAS the decision for a vector $\underline{x}$ is reached prior to the final cluster formation.
- BSAS perform a single pass on the data. Its complexity is $O(N)$.
- If clusters are represented by point representatives, compact clusters are favored.

## Estimating the number of clusters in the data set:

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

- 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 $m_\Theta$, as the most frequent number resulting from the $s$ runs of $BSAS(\Theta)$.
- Next $\Theta$
- Plot $m_\Theta$ versus $\Theta$ 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 deals with the above drawback, at the cost of presenting the data twice to the algorithm.
- 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.

## ➢ Remarks:

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

## The maxmin algorithm

Let $W$ be the set of all points that have been chosen to form clusters up to the current iteration step. The formation of clusters is carried out as follows:

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

After the formation of the clusters, each unassigned vector is assigned to its closest cluster.

➢Remarks:

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

## ❖ Refinement stages

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

➢ 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}

- 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}

## ❖ 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 $\Theta_1$ and $\Theta_2$ ($\Theta_1<\Theta_2$) are employed
- The general idea is the following:
  - If the distance $d(\underline{x},C)$ of $\underline{x}$ from its closest cluster, $C$, is greater than $\Theta_2$ 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 ($\geq 2$) of the data set are required.
  - TTSAS is less sensitive to the order of data presentation, compared to BSAS.