CHAPTER 11 – CLUSTERING ALGORITHMS I

❖ Number of possible clusterings

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?

Answer:

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

➢ Examples:

$$S(15,3)=2\,375\,101$$
$$S(20,4)=45\,232\,115\,901$$
$$S(100,5)=10^{68}\,!!$$

- 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.

SEQUENTIAL CLUSTERING ALGORITHMS

- The common traits shared by these algorithms are:

  - 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.
  - The clusters are defined with the aid of
    - An appropriately defined distance $d\,(\underline{x},C)$ of a point from a cluster.
    - 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}

(*) 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 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.