

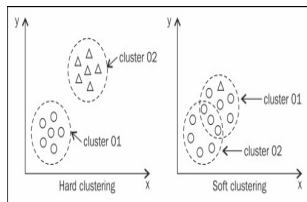
Clustering

ML Instruction Team, Fall 2022

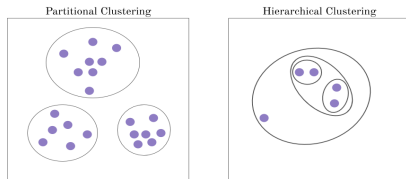
CE Department
Sharif University of Technology

Categories of Clustering

- **Hard:** Where each point is assigned to exactly one cluster.
- **Soft:** Where each point can be assigned to several clusters with certain probabilities that add up to 1.
- **Partitional:** Where all clusters are on the same level.
- **Hierarchical** Where the clustering is done from fine to coarse by merging points successively to larger and larger clusters (agglomerative hierarchical clustering).



(a) Hard vs Soft Clustering, [Source](#)



(b) Partitional vs Hierarchical Clustering, [Source](#)

Figure: Different clustering techniques

K-Means

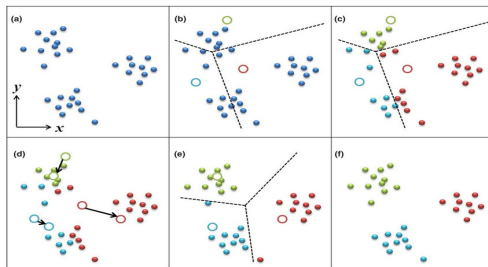


Figure: K-means in practice, [Source](#)

- The idea is to represent each cluster k by a center point c_k and assign each data point x_n to one of the clusters k .
- The center points and the assignment are then chosen such that the mean squared distance between data points and center points

$$J = \sum_{n=1}^N \sum_{n \in c_k} \|x_n - c_k\|^2$$

is minimized.

K-Means

- If the assignment is fixed, it is easy to show that the optimal choice of the center positions is given by:

$$c_k = \frac{1}{N_k} \sum_{n \in c_k} x_n$$

- If the center points are fixed, it is obvious that each point should be assigned to the nearest center position.
- The K -means algorithm now consists of applying these two optimizations in turn until convergence.
- The initial center locations?
- Convergence Guaranty?
- Drawbacks:
 - ▶ A paramount drawback of this and many other clustering algorithms is that the number of clusters is not determined.
 - ▶ The result of the algorithm is not necessarily a global optimum of the error function.
- Solutions
 - ▶ Problem 1?
 - ▶ Problem 2?

Davies-Bouldin (DB) Index

- To evaluate the quality of a clustering a plethora of validity indices have been proposed, one of which is Davies-Bouldin (DB) index.
- **Cluster Dispersion:** Which can be interpreted as a generalized standard deviation.

$$\delta_k := \sqrt{\frac{1}{N_k} \sum_{n \in c_k} \|x_n - c_k\|^2}$$

- **Cluster Similarity:** Is defined such that two clusters are considered similar if they have large dispersion relative to their distance.

$$S_{kl} := \frac{\delta_k + \delta_l}{\|c_k - c_l\|}$$

- Considering aforementioned definitions, an overall validation of the clustering can be done by the DB index:

$$V_{DB} := \frac{1}{k} \sum_{k=1}^K \max_{l \neq k} S_{kl}$$

<https://replicate.com/nateraw/stable-diffusion-videos>

- The DB index does not systematically depend on K and is therefore suitable to find the best optimal number of clusters.

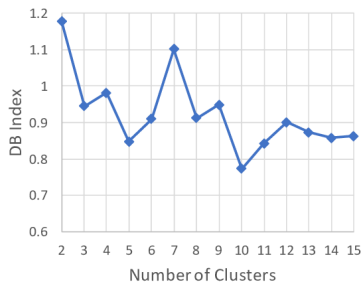


Figure: DB Index, [Source](#)

Thank You!

Any Question?