Clustering (part 4)

Mohammed Brahimi & Samy Belkacem

Outline

- ■ Clustering evaluation
    - ❑ Why cluster evaluation ?
    - ❑ Types of cluster evaluation measures
- ❑ Unsupervised evaluation
    - ❑ Cohesion vs Separation
    - ❑ Silhouette Coefficient
- ❑ Supervised evaluation
    - ❑ Entropy
    - ❑ Precision, Recall, F-measure

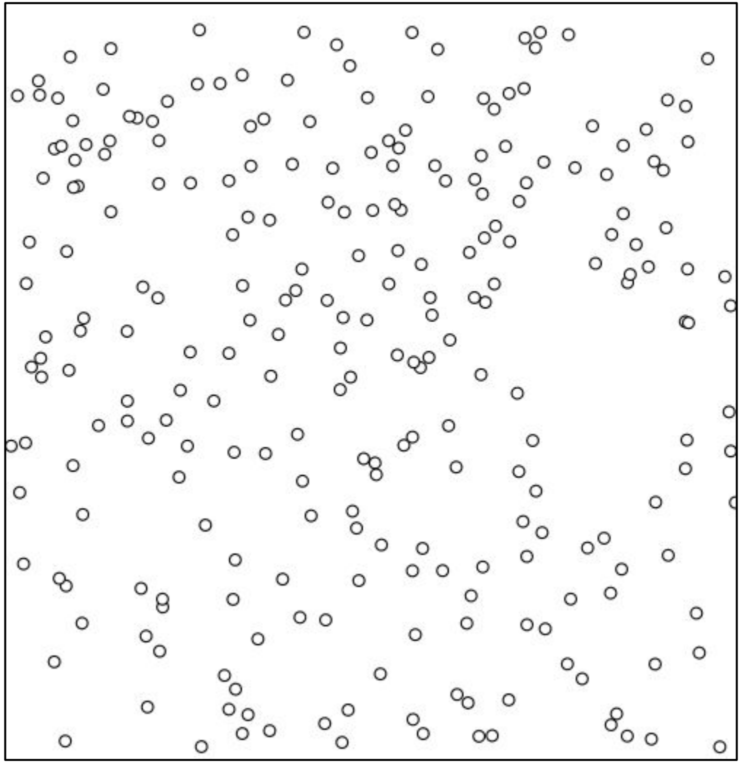
Why cluster evaluation ?

- Generate a random data points.
- Data without any structure

Question:

What is the result of applying K-Means with K=3?

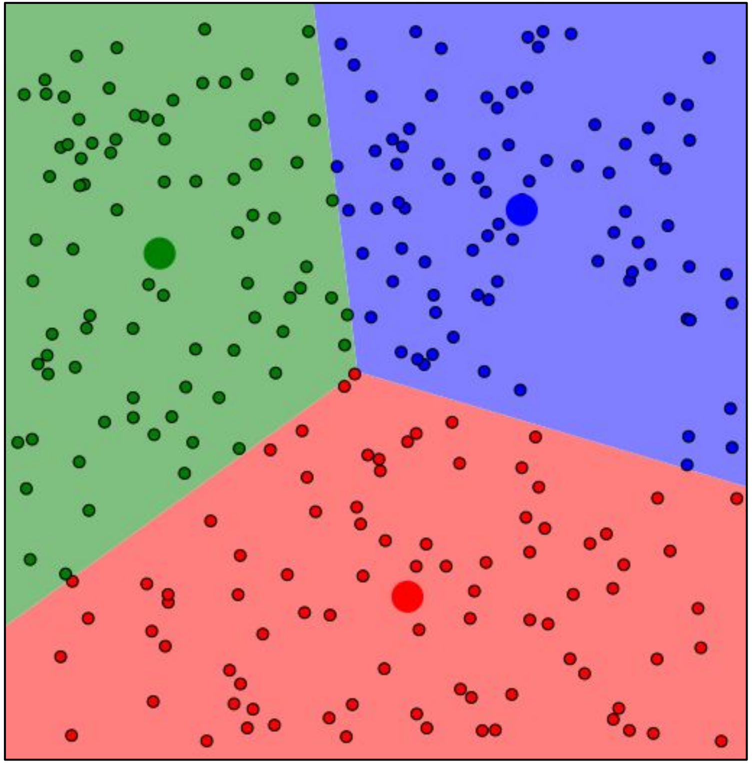
The following link can be used: [K-Means Animation]



Why cluster evaluation ?

- Generate a random data points.
- Data without any structure

Clusters found in Random Data !!



The following link can be used: **K-Means Animation**

Why cluster evaluation ?

- **To avoid Detecting clusters in random Structure**
  - Uncovering whether non-random structure exists in the data.
- **To evaluate Clustering Results**
  - Assessing how well the clustering aligns with the data without external reference.
- **To compare with external known patterns**
  - Comparing clustering results to externally known information, e.g., class labels.
- **To compare different Clusterings and algorithms**
  - Evaluating and comparing different sets of clusters for quality.

“*The validation of clustering structures is the* ***most difficult and frustrating part of cluster analysis****.*

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

Algorithms for Clustering Data, Jain and Dubes

Types of cluster evaluation measures

- **Unsupervised (Internal)**: measure the goodness of a clustering structure without respect to external information.
    - The ground truth is not available.
    - **Examples:** Cohesion, separation, SSE, Silhouette Coefficient.

- **Supervised (External) :** measure the extent to which cluster labels match externally supplied class labels.
    - The ground truth is available.
    - **Examples:** Entropy, Precision, Recall, F-measure.

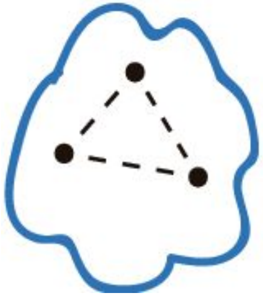
Outline

- ❑ Clustering evaluation
    - ❑ Why cluster evaluation ?
    - ❑ Types of cluster evaluation measures
- ■ Unsupervised evaluation
    - ❑ Cohesion vs Separation
    - ❑ Silhouette Coefficient
- ❑ Supervised evaluation
    - ❑ Entropy
    - ❑ Precision, Recall, F-measure

Cohesion vs Separation

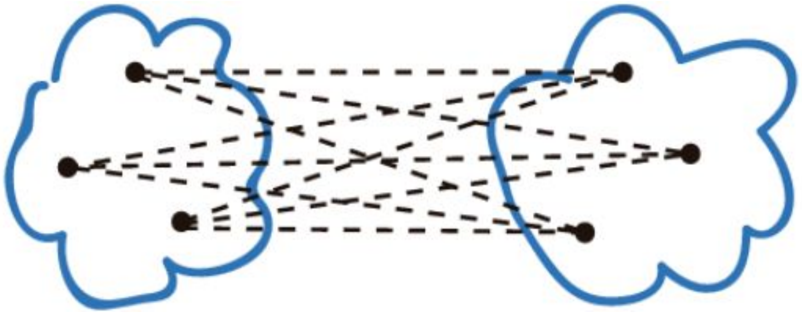
Cluster cohesion (Compactness)

- Measure how closely related object in a cluster.



Cluster Separation

- Measure how distinct or well- separated a cluster is from other clusters.



Graph-Based View

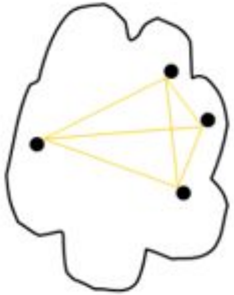
Weighted graph where the weights are the distances between data points.

- **Cohesion:** Sum of proximities in a cluster.

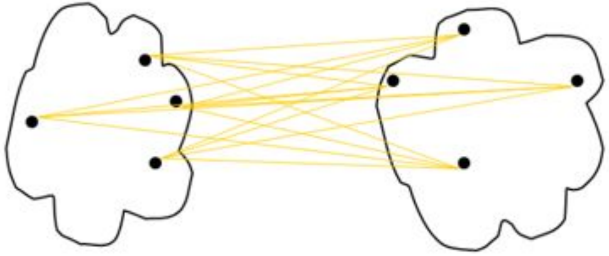
$$cohesion(C_i) = \sum_{\substack{\mathbf{x}\in C_i \\ \mathbf{y}\in C_i}} proximity(\mathbf{x}, \mathbf{y})$$

- **Separation:** Sum of proximities between two clusters.

$$separation(C_i, C_j) = \sum_{\substack{\mathbf{x}\in C_i \\ \mathbf{y}\in C_j}} proximity(\mathbf{x}, \mathbf{y})$$



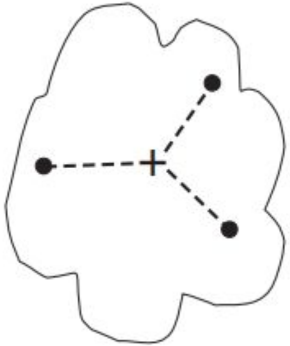
cohesion



separation

Prototype-Based View

Represent a clusters using their centroids.



- **Cohesion:** Sum of proximities to the cluster centroid.

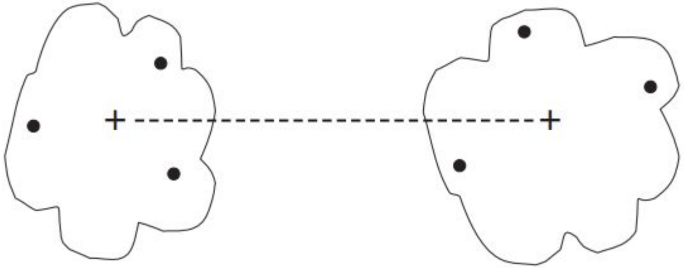
$$cohesion(C_i) = \sum_{\mathbf{x} \in C_i} proximity(\mathbf{x}, \mathbf{c}_i)$$

- **Separation:** Sum of proximites between centroids.
  - Between two centroids

$$separation(C_i, C_j) = proximity(\mathbf{c}_i, \mathbf{c}_j)$$

  - Between a cluster centroid and the global centroid

$$separation(C_i) = proximity(\mathbf{c}_i, \mathbf{c})$$

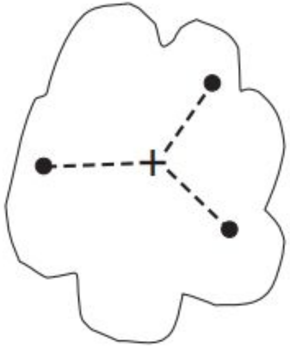


Prototype-Based View

Represent a clusters using their centroids.

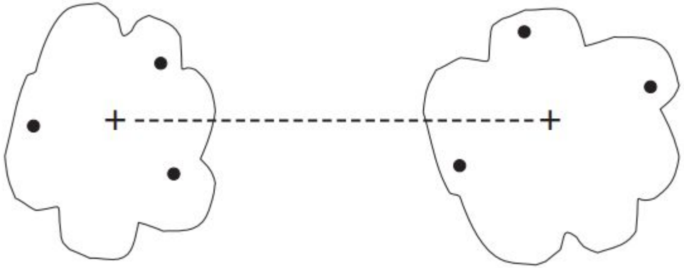
- **Cohesion:** Sum of proximities to the cluster centroid.

$$cohesion(C_i) = \sum_{\mathbf{x} \in C_i} proximity(\mathbf{x}, \mathbf{c}_i)$$



- ***SSE is the sum of prototype based cohesion of all clusters.***
    - Between a cluster centroid and the global centroid

$$separation(C_i) = proximity(\mathbf{c}_i, \mathbf{c})$$

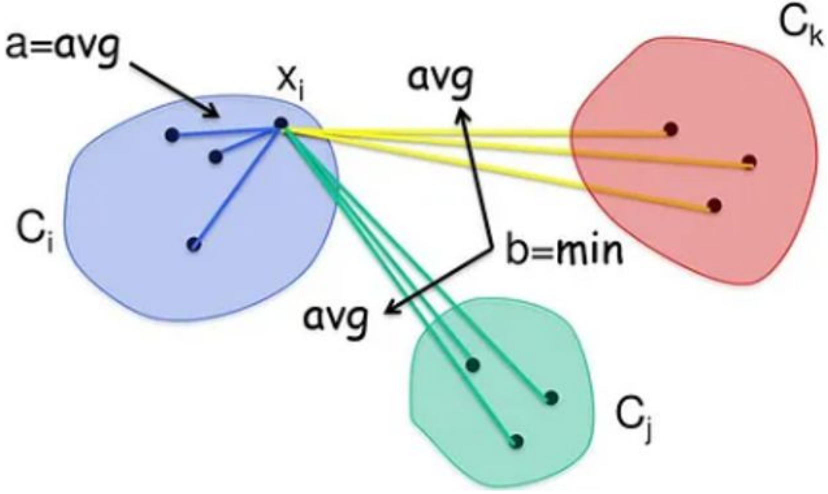


Silhouette Coefficient

- Silhouette coefficient combines **cohesion** and **separation.**
- For an individual point ***i***
  - a = average distance of ***i*** to the points in its cluster
  - b = min (average distance of ***i*** to points in another cluster)
- The silhouette coefficient for a point is

  s = (b – a) / max(a, b)

- Value can vary between -1 and 1.
- The closer to 1 the better.



Outline

❑ Clustering evaluation

  ❑ Why cluster evaluation ?

  ❑ Types of cluster evaluation measures

❑ Unsupervised evaluation

  ❑ Cohesion vs Separation

  ❑ Silhouette Coefficient

■ Supervised evaluation

  ❑ Entropy

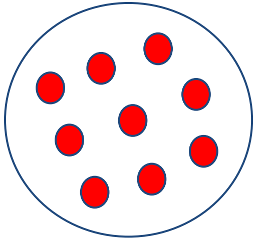
  ❑ Precision, Recall, F-measure

Entropy

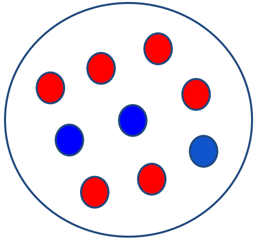
Entropy measures the extent to which the clustering structure matches external class labels.

- Pure cluster is cluster that contain only one class label.
- We measure the purity of a cluster using the entropy.
- How to Use Entropy for Evaluation:
  - Calculate entropy for each cluster.
  - Sum the entropies to get an overall measure.
  - Lower values indicate better alignment with external class labels.

$$s_c = \sum_{1}^{K} \frac{n_k}{N} s_{Lk} : s_{lk} = \sum_{1}^{L} -p_{Lk} \log_2 p_{lk}$$



Pure cluster



Impure cluster

Entropy

k	p_{1k}	p_{2k}	p_{3k}	s_{Lk}
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0

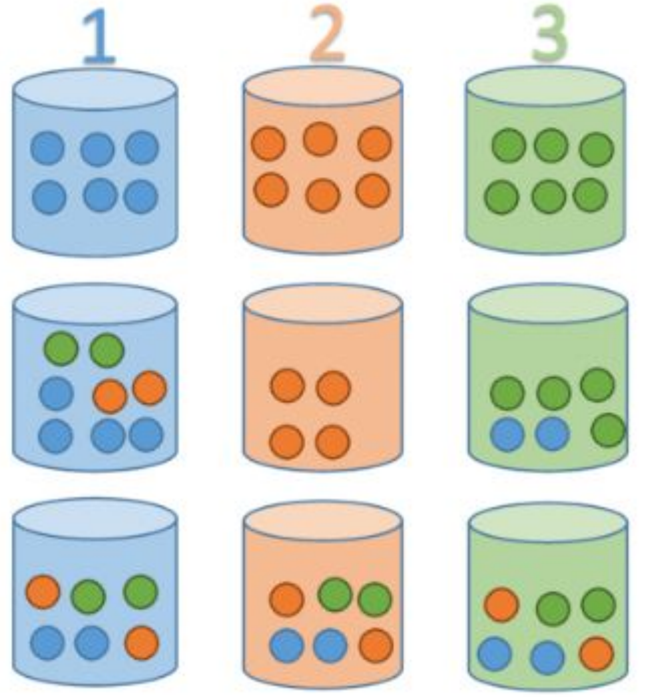
$S_c = 0$

k	p_{1k}	p_{2k}	p_{3k}	s_{Lk}
1	4/8	2/8	2/8	1.5
2	0	1	0	0
3	2/6	0	4/6	0.918

$S_c = 0.971$

k	p_{1k}	p_{2k}	p_{3k}	s_{Lk}
1	2/6	2/6	2/6	1.585
2	2/6	2/6	2/6	1.585
3	2/6	2/6	2/6	1.585

$S_c = 1.585$



$$S_c = \sum_1^K \frac{n_k}{N} s_{Lk} \quad : \quad s_{lk} = \sum_1^L -p_{Lk} \log_2 p_{lk}$$

Precision, Recall, F-measure

- **Precision:** The fraction of a cluster ***i*** that consists of objects of a specified class.

$$\text{Precision}(i, j) = \frac{\text{Number of examples of class } j \text{ in cluster } i}{\text{Size of cluster } i}$$

- **Recall:** The extent to which a cluster contains all objects of a specified class.

$$\text{Recall}(i, j) = \frac{\text{Number of examples of class } j \text{ in cluster } i}{\text{Number of examples of class } j}$$

- **F-measure:** A combination of precision and recall that measures the extent to which a cluster contains only objects of a particular class and all objects of that class.

$$F(i, j) = \frac{2 \times \text{Precision}(i, j) \times \text{Recall}(i, j)}{\text{Precision}(i, j) + \text{Recall}(i, j)}$$