Clustering is a hard problem

Many possibilities: What is best clustering?

2 clusters: easy
Clustering is a hard problem

Many possibilities: What is best clustering?

Clustering

- Hierarchical clustering
- K-means clustering
- How many clusters?
Hierarchical Clustering

To cluster a set of data \( D = \{P_1, P_2, \ldots, P_N\} \), hierarchical clustering proceeds through a series of partitions that runs from a single cluster containing all data points, to \( N \) clusters, each containing 1 data point.

Two forms of hierarchical clustering:

- **Agglomerative**: Starts with \( N \) independent clusters: \( \{P_1\}, \{P_2\}, \ldots, \{P_N\} \)
- **Divisive**: Begins with one big cluster and splits it into two clusters

Methods differ in their definition of inter-cluster distance (or similarity)

1) Single linkage clustering

Distance between closest pairs of points:

\[
d(A, B) = \min \{d(P_i, P_j) \mid P_i \in A, P_j \in B\}
\]

2) Complete linkage clustering

Distance between farthest pairs of points:

\[
d(A, B) = \max \{d(P_i, P_j) \mid P_i \in A, P_j \in B\}
\]
Agglomerative hierarchical clustering techniques

3) Average linkage clustering
Mean distance of all mixed pairs of points:
\[ d(A, B) = \frac{\sum_{i \in A} \sum_{j \in B} d(P_i, P_j)}{N_A N_B} \]

4) Average group linkage clustering
Mean distance of all pairs of points:
\[ d(A, B) = \frac{\sum_{i \in A} \sum_{j \in B} d(P_i, P_j)}{N_A^2} \]

Clustering

- Hierarchical clustering
- K-means clustering
- How many clusters?

K-means clustering

The k-means algorithm partitions the data into k mutually exclusive clusters

K = 3
K-means clustering

Algorithm description
- Choose the number of clusters, K
- Randomly choose initial positions of K centroids

K = 3

(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

Algorithm description
- Choose the number of clusters - K
- Randomly choose initial positions of K centroids
- Assign each of the points to the "nearest centroid" (depends on distance measure)

K = 3

(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

Algorithm description
- Choose the number of clusters - K
- Randomly choose initial positions of K centroids
- Assign each of the points to the "nearest centroid" (depends on distance measure)
- Re-compute centroid positions
- If solution converges ⇒ Stop!

K = 3

(http://www.weizmann.ac.il/midrasha/courses/)
Clustering is hard: it is an unsupervised learning technique. Once a Clustering has been obtained, it is important to assess its validity!

The questions to answer:

- Did we choose the right number of clusters?
- Are the clusters compact?
- Are the clusters well separated?

To answer these questions, we need a quantitative measure of the cluster sizes:

- intra-cluster size
- inter-cluster distances
Inter cluster size

Several options:
- Single linkage
- Complete linkage
- Average linkage
- Average group linkage

Intra cluster size

Several options:
- Complete diameter:
  \[ \Delta(S) = \max_{x,y \in S} d(x,y) \]
- Average diameter:
  \[ \Delta(S) = \frac{1}{N(N-1)} \sum_{x \neq y} d(x,y) \]
- Centroid diameter:
  \[ \Delta(S) = \frac{2}{N} \sum_{x \in S} d(x,C) \]

Cluster Quality

For a clustering with K clusters:

1) Dunn's index
   \[ D = \min_{1 \leq i \neq j \leq K} \left( \min_{x \in S_i, y \in S_j} \frac{\delta(S_i,S_j)}{\max(\Delta(S_i),\Delta(S_j))} \right) \]
   Large values of D correspond to good clusters

2) Davies-Bouldin's index
   \[ DB = \max_{1 \leq i \leq K} \left( \frac{\Delta(S_i) + \Delta(S_j)}{\delta(S_i,S_j)} \right) \]
   Low values of DB correspond to good clusters
Cluster Quality: Silhouette index

Define a quality index for each point in the original dataset:

- For the ith object, calculate its average distance to all other objects in its cluster. Call this value ai.
- For the ith object and any cluster not containing the object, calculate the object’s average distance to all the objects in the given cluster. Find the minimum such value with respect to all clusters; call this value bi.
- For the ith object, the silhouette coefficient is

\[ s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \]

Note that:

\[-1 \leq s(i) \leq 1\]

- \(s(i) = 1\), i is likely to be well classified
- \(s(i) = -1\), i is likely to be incorrectly classified
- \(s(i) = 0\), indifferent

Cluster Quality: Silhouette index

Cluster silhouette index:

\[ S(X_i) = \frac{1}{N} \sum_{j=1}^{N} s(j) \]

Global silhouette index:

\[ GS = \frac{1}{K} \sum_{k=1}^{K} S(X_k) \]

Large values of GS correspond to good clusters