Clustering

张诏月

Programming

Clustering (cluster analysis) is the process of separating a data set into several subsets

It can be applied to a variety of biological study cases such as

- Sequence analysis
- Microarray data analysis
- Phylogenetic analysis

Clustering methods

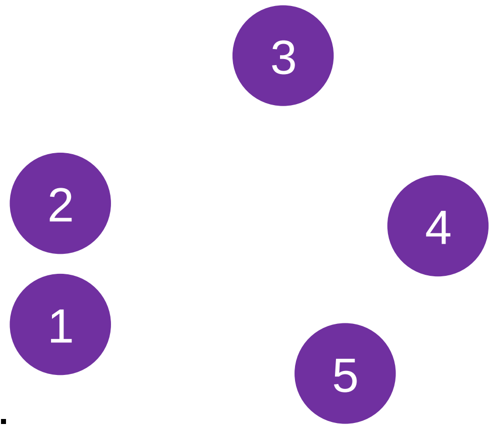
- Hierarchical method (Distance based analysis)
- Partitioning method (K-means clustering)
- Density-based method
- Grid-based method

Hierarchical method

A hierarchical method generates a hierarchical decomposition of the given data set

Agglomerative hierarchical method (bottom-up approach)

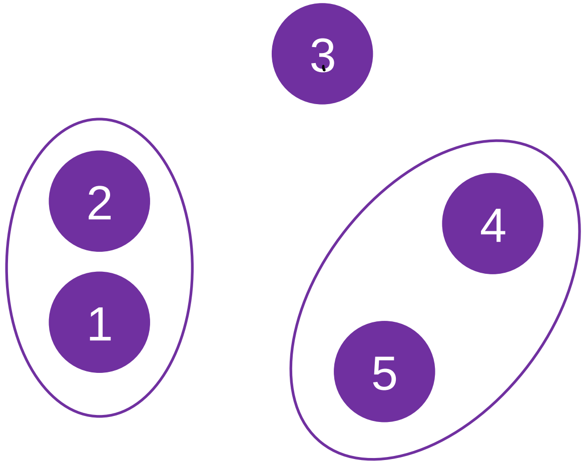
- It starts with each data object composing a separate cluster.
- It successively merges the objects or clusters close to each other.
- Until all clusters are merged into one cluster.



Divisive Hierarchical method (top-down clustering approach)

- It begins with the root, in which all observations are included in a single cluster.
- At each step of the algorithm, the current cluster is split into two clusters that are considered most heterogeneous.
- Until all observations are in their own cluster.

Hierarchical method



The definition of the distance between two clusters

– Single linkage/connectedness/minimum/nearest neighbour

$$D(A,B) = \min_{a_i \in A, b_j \in B} d(a_i, b_j)$$

– Complete linkage/diameter/maximum/furthest neightbour

$$D(A,B) = \max_{a_i \in A, b_j \in B} d(a_i, b_j)$$

– Average linkage/Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

$$D(A,B) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} d(a_i, b_j)$$

– Ward’s linkage/error sum of squares criterion

$$D(A,B) = \mathrm{Var}(A \cup B) - (\mathrm{Var}(A) + \mathrm{Var}(B))$$

– Centroid linkage/Unweighted Pair-Group Method using Centroid approach (UPGMC)

– Median linkage/Weighted Pair-Group Method using Centroid approach (WPGMC)

Distance

Euclidean Distance $$\sqrt{\sum_{i=1}^{n}(X_i^{(a)}-X_i^{(b)})^2}$$

Manhattan Distance $$\sum_{i=1}^{n}|X_i^{(a)}-X_i^{(b)}|$$

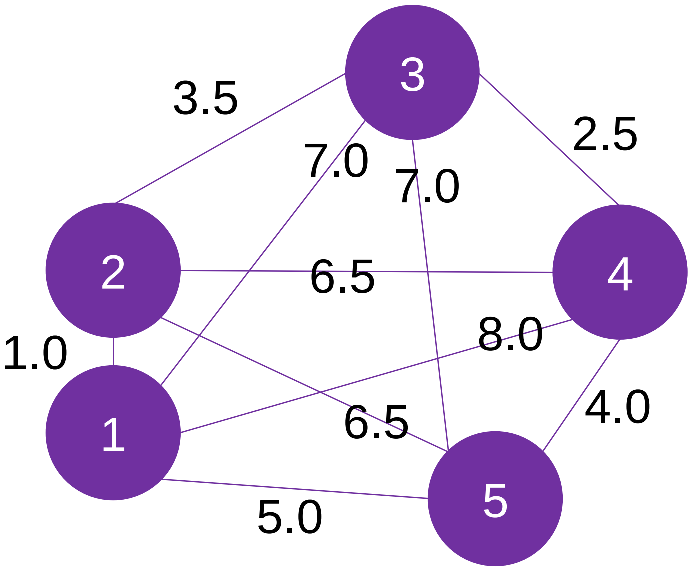
Minkowski Distance $$(\sum_{i=1}^{n}|X_i^{(a)}-X_i^{(b)}|^p)^{\frac{1}{p}}$$

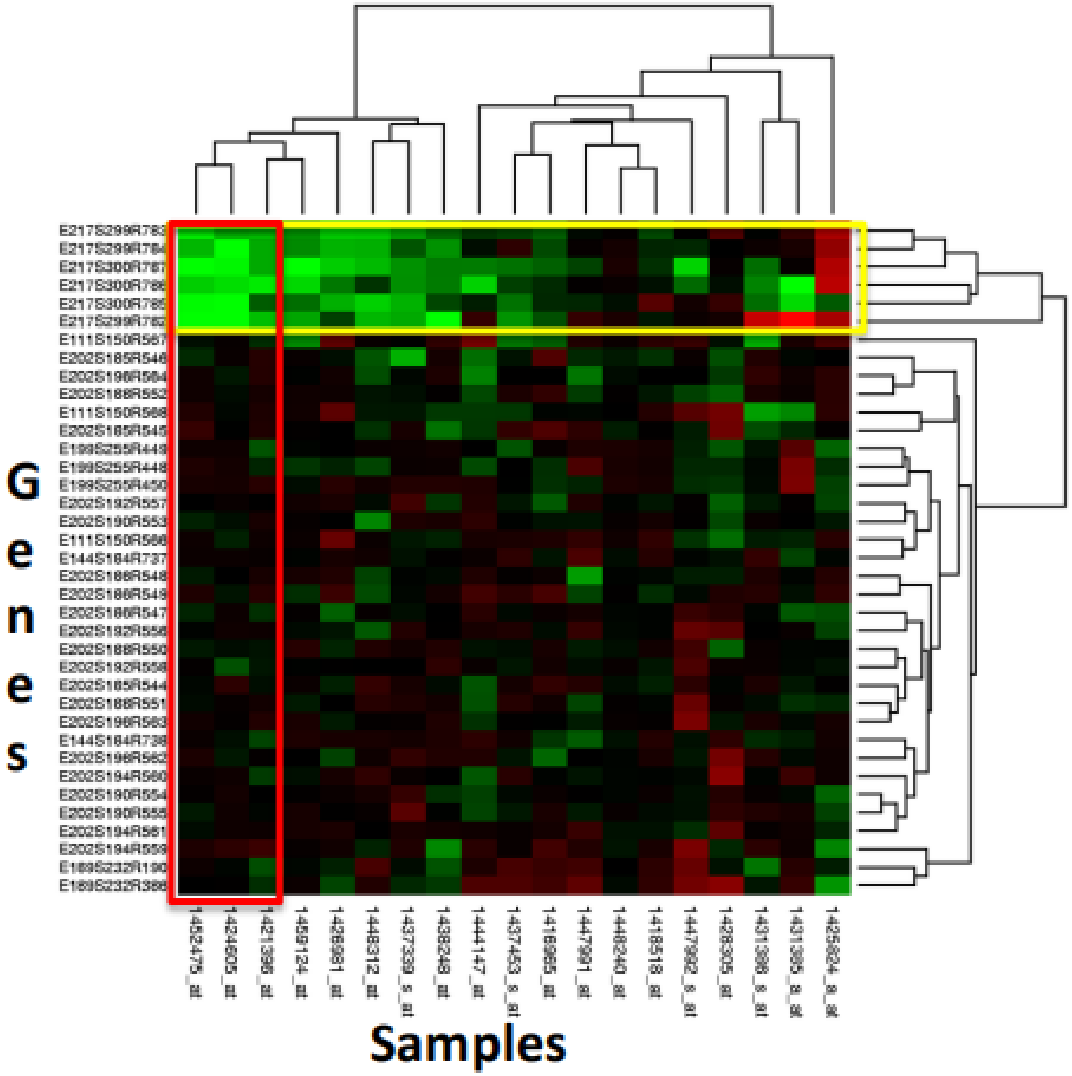
1 - Cosine Similarity $$c(X,Y)=\frac{X\cdot Y}{|X|\,|Y|}=\frac{\sum_{i=1}^{n}X_iY_i}{\sqrt{\sum_{i=1}^{n}X_i^2}\sqrt{\sum_{i=1}^{n}Y_i^2}}$$

1 - Pearson Correlation Coefficient $$\rho(X,Y)=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i-\mu_Y)^2}}$$

1 - Jaccard Coefficient $$\mathrm{J}(\mathrm{A},\mathrm{B})=\frac{|A\cap B|}{|A\cup B|}$$

Building dendrogram



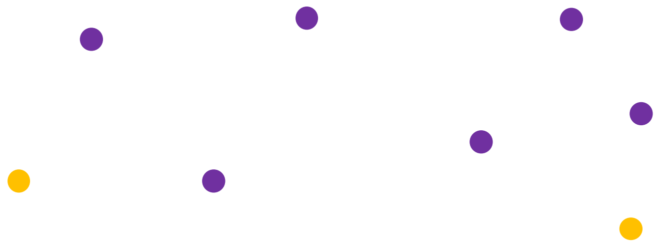


Partitioning method (K-means clustering)

- Randomly assign K objects from the dataset(D) as cluster centres(C)
- (Re) Assign each object to which object is most similar based upon mean values.
- Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.

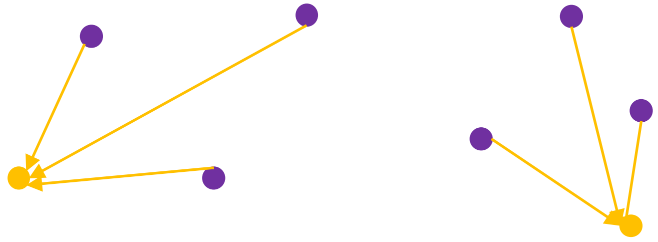
Partitioning method (K-means clustering)

- Randomly assign K objects from the dataset(D) as cluster centres(C)
- (Re) Assign each object to which object is most similar based upon mean values.
- Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.



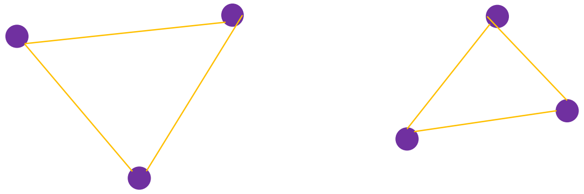
Partitioning method (K-means clustering)

- Randomly assign K objects from the dataset(D) as cluster centres(C)
- (Re) Assign each object to which object is most similar based upon mean values.
- Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.



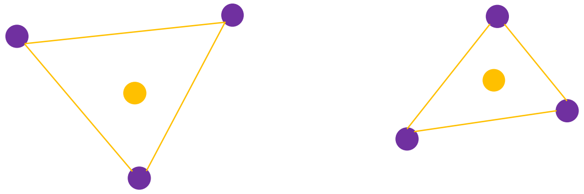
Partitioning method (K-means clustering)

- Randomly assign K objects as cluster centres(C)
- (Re) Assign each object to which object is most similar based upon mean values.
- Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.



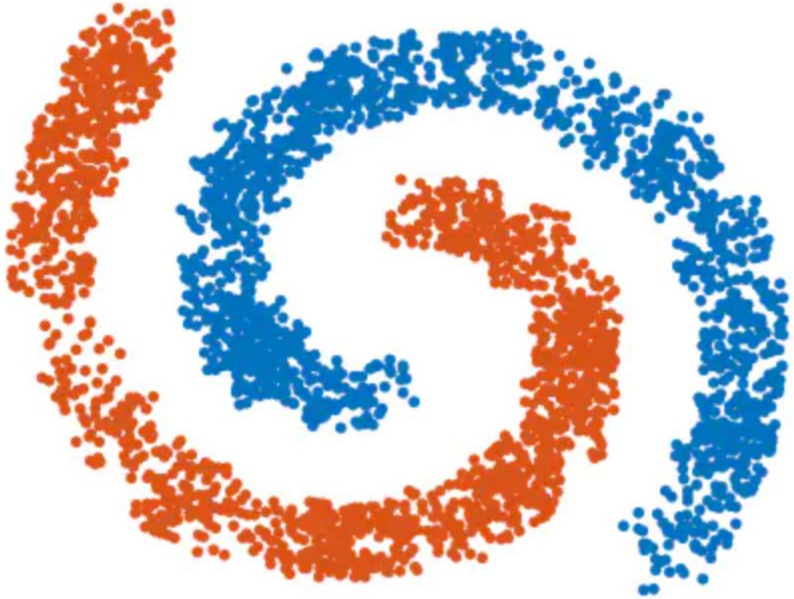
Partitioning method (K-means clustering)

- Randomly assign K objects as cluster centres(C)
- (Re) Assign each object to which object is most similar based upon mean values.
- Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.



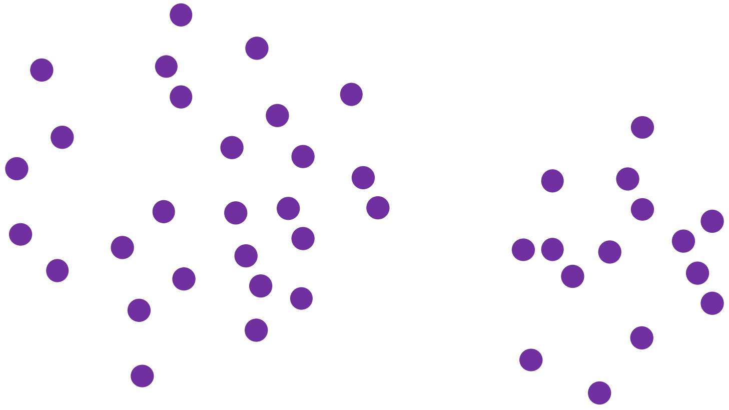
Density-based method

- Randomly assign *p* visited objects.
- Create a circus with radius e, marked objects in the circus visited.
- Until no object is unvisited.



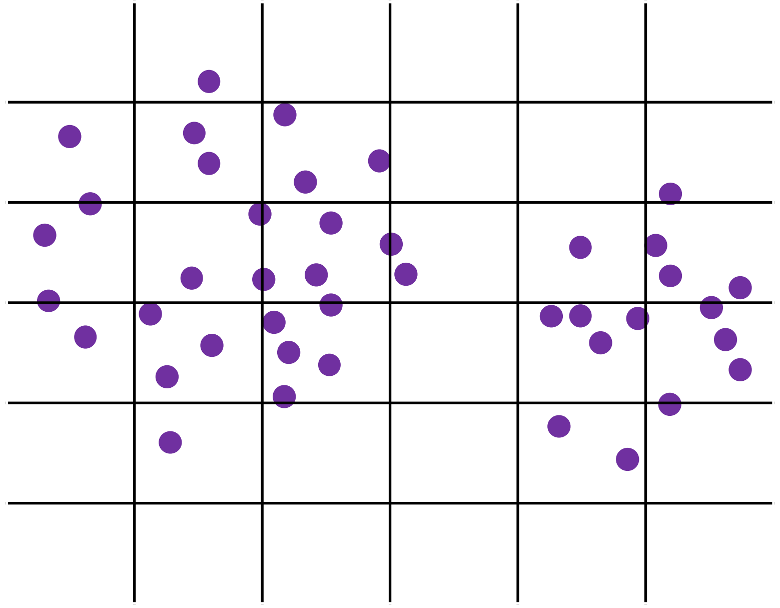
Grid-based method

- Creating the grid structure.
- Calculating the cell density for each cell.
- Sorting of the cells according to their densities and identifying cluster centers.
- Traversal of neighbor cells.



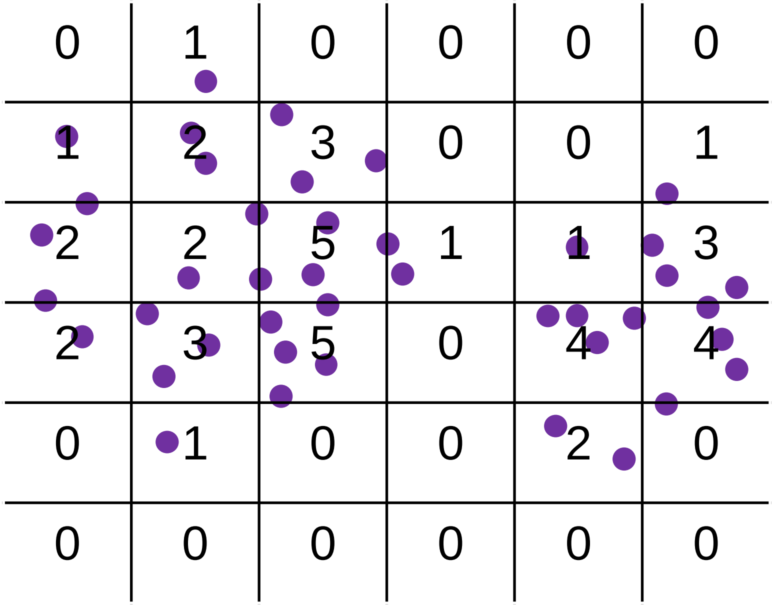
Grid-based method

- Creating the grid structure.
- Calculating the cell density for each cell.
- Sorting of the cells according to their densities and identifying cluster centers.
- Traversal of neighbor cells.



Grid-based method

- Creating the grid structure.
- Calculating the cell density for each cell.
- Sorting of the cells according to their densities and identifying cluster centers.
- Traversal of neighbor cells.



0	1	0	0	0	0
1	2	3	0	0	1
2	2	5	1	1	3
2	3	5	0	4	4
0	1	0	0	2	0
0	0	0	0	0	0