CEFET/RJ

Clustering

Eduardo Ogasawara
eduardo.ogasawara@cefet-rj.br
https://eic.cefet-rj.br/~eogasawara

What is Cluster Analysis?

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning
  - no predefined classes (i.e., learning by observations vs. learning by examples: supervised)
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Applications of Cluster Analysis

- Data reduction
  - Summarization: Preprocessing for regression, PCA, classification, and association analysis
- Prediction based on groups
  - Cluster & find characteristics/patterns for each group
- Finding K-nearest Neighbors
  - Localizing search to one or a small number of clusters
- Outlier detection: Outliers are often viewed as those “far away” from any cluster

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Clustering: Application Examples

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Climate: understanding Earth climate, find patterns of atmospheric and ocean

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Basic Steps to Develop a Clustering Task

- Feature selection
  - Select info concerning the task of interest
  - Minimal information redundancy
- Proximity measure
  - Similarity of two feature vectors
- Clustering criterion
  - Expressed via a cost function or some rules
- Clustering algorithms
  - Choice of algorithms
- Validation of the results
  - Validation test (also, clustering tendency test)
- Interpretation of the results
  - Integration with applications

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters
  - high intra-class similarity: cohesive within clusters
  - low inter-class similarity: distinctive between clusters
- The quality of a clustering method depends on
  - the similarity measure used by the method
  - its implementation, and
  - Its ability to discover some or all the hidden patterns

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Measure the Quality of Clustering

- Dissimilarity/Similarity metric
  - Similarity is expressed in terms of a distance function, typically metric: d(i, j)
  - The definitions of distance functions are usually rather different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables
  - Weights should be associated with different variables based on applications and data semantics
- Quality of clustering:
  - There is usually a separate "quality" function that measures the "goodness" of a cluster.
  - It is hard to define "similar enough" or "good enough"
    - The answer is typically highly subjective

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Considerations for Cluster Analysis

- Partitioning criteria
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)
- Separation of clusters
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- Similarity measure
  - Distance-based (e.g., Euclidian, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- Clustering space
  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Requirements and Challenges

- Scalability
  - Clustering all the data instead of only on samples
- Ability to deal with different types of attributes
  - Numerical, binary, categorical, ordinal, linked, and mixture of these
- Constraint-based clustering
    - User may give inputs on constraints
    - Use domain knowledge to determine input parameters
- Interpretability and usability
- Others
  - Discovery of clusters with arbitrary shape
  - Ability to deal with noisy data
  - Incremental clustering and insensitivity to input order
  - High dimensionality

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- Dissimilarity (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Data Matrix and Dissimilarity Matrix

- Data matrix
  - n data points with p attributes
  - Two modes: objects and attributes

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
  - n data points, but registers only the distance
  - A triangular matrix
  - Single mode: distances

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Distance on Numeric Data: Minkowski Distance

- Minkowski distance: A popular distance measure
  - $d(i,j) = \sqrt[h]{\left|x_{i_1} - x_{j_1}\right|^h + \left|x_{i_2} - x_{j_2}\right|^h + \cdots + \left|x_{i_p} - x_{j_p}\right|^h}$
  - where i = (x_{i_1}, x_{i_2}, …, x_{i_p}) and j = (x_{j_1}, x_{j_2}, …, x_{j_p}) are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm)
- Properties
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positive definiteness)
  - d(i, j) = d(j, i) (Symmetry)
  - d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)
- A distance that satisfies these properties is a metric

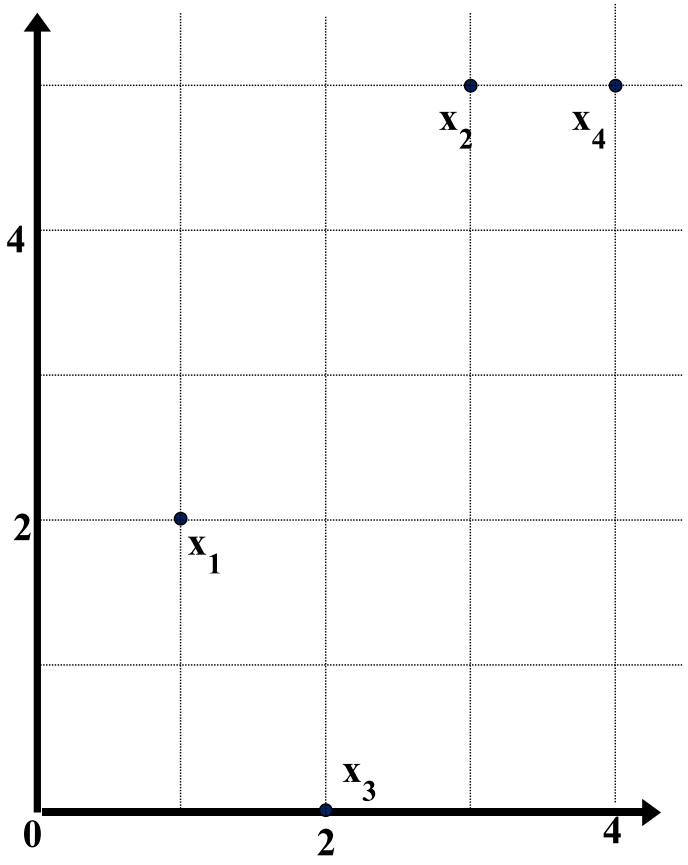
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Special Cases of Minkowski Distance

- h = 1: Manhattan (city block, L1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors
  - $d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \cdots + |x_{i_p} - x_{j_p}|$
- h = 2: (L2 norm) Euclidean distance
  - $d(i,j) = \sqrt{|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \cdots + |x_{i_p} - x_{j_p}|^2}$
- h $\rightarrow \infty$: "supremum" (Lmax norm, L∞ norm) distance
  - This is the maximum difference between any component (attribute) of the vectors
  - $d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{i_f} - x_{j_f}|^h \right)^{\frac{1}{h}}$

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Example:
Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix

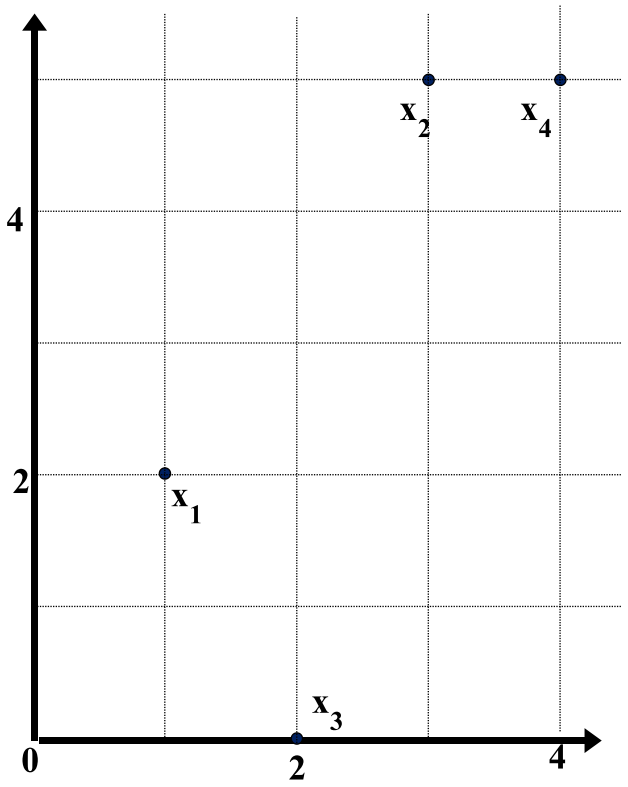
(with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Proximity Measure for Binary Attributes

- A contingency table for binary data
- Distance measure for symmetric binary variables
  - $d(i,j) = \frac{r+s}{q+r+s+t}$
- Distance measure for asymmetric binary variables
  - $d(i,j) = \frac{r+s}{q+r+s}$
- Jaccard coefficient (similarity measure for asymmetric binary variables):
  - $d(i,j) = \frac{q}{q+r+s}$

		Object j		
		1	0	sum
Object i	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p

Note: Jaccard coefficient is the same as "coherence"

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Dissimilarity between Binary Variables

- Example
  - Gender is a symmetric attribute
  - The remaining attributes are asymmetric binary
  - Let the values Y and P be 1, and the value N be 0

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Method 1: Simple matching
  - m: # of matches, p: total # of variables
  - $d(i,j) = \frac{p-m}{p}$
- Method 2: Use many binary attributes
  - creating a new binary attribute for each of the M nominal states

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank $r_{if} \in \{1, \dots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing i-th object in the f-th variable by $z_{if} = \frac{r_{if}-1}{M_f-1}\{1, \dots, M_f\}$
  - compute the dissimilarity using methods for interval-scaled variables

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Cosine Similarity

- A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: gene features in micro-arrays, …
- Applications: information retrieval, biologic taxonomy, gene feature mapping, …
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then
  - $\cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{|d_1||d_2|}$,
- where $\cdot$ indicates vector dot product, $|d|$: the length of vector d

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Example: Cosine Similarity

- cos(d1, d2) = (d1 • d2) /||d1|| ||d2|| ,
  - where • indicates vector dot product, ||d|: the length of vector d

- Ex: Find the similarity between documents 1 and 2.

  - d1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
  - d2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

  - d1•d2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25
  - ||d1||= (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)0.5=(42)^0.5 = 6.481
  - ||d2||= (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)0.5=(17)^0.5 = 4.12
  - cos(d1, d2) = 0.94

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Combining Mixed Types

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects
  - $d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{f} d_{ij}^{f}}{\sum_{f=1}^{p} \delta_{ij}^{f}}$
  - δ_{ij}^{f} = 0
    - if (1) either xif or xjf is missing
    - or xif = xjf = 0 and attribute is binary asymmetric
  - δ_{ij}^{f} = 1, otherwise
  - f is binary or nominal: d_{ij}^{f} = 0 if xif = xjf , or d_{ij}^{f} = 1 otherwise
  - f is numeric: use a normalized distance
  - f is ordinal: convert to ranks rif and compute zif

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Major Clustering Approaches (I)

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, CAMELEON
- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSCAN, OPTICS, DenClue
- Grid-based approach:
  - based on a multiple-level granularity structure
  - Typical methods: STING, WaveCluster, CLIQUE

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Major Clustering Approaches (II)

- Model-based:
  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
  - Typical methods: EM, SOM, COBWEB
- Frequent pattern-based:
  - Based on the analysis of frequent patterns
  - Typical methods: p-Cluster
- User-guided or constraint-based:
  - Clustering by considering user-specified or application-specific constraints
  - Typical methods: COD (obstacles), constrained clustering
- Link-based clustering:
  - Objects are often linked together in various ways
  - Massive links can be used to cluster objects: SimRank, LinkClus

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Partitioning Algorithms: Basic Concept

- Partitioning method: Partitioning a database D of n objects into a set of k clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)
- Given k, find a partition of k clusters that optimizes the chosen partitioning criterion
  - $E = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, c_i)^2$
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: k-means and k-medoids algorithms
  - k-means: Each cluster is represented by the center of the cluster
  - k-medoids or PAM (Partition around medoids): Each cluster is represented by one of the objects in the cluster

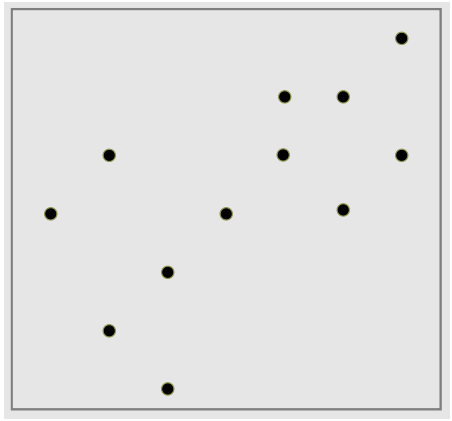
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

The K-Means Clustering Method

- Given k, the k-means algorithm is implemented in four steps:
  - [1] Partition objects into k nonempty subsets
  - [2] Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., mean point, of the cluster)
  - [3] Assign each object to the cluster with the nearest seed point
  - [4] Go back to Step 2, stop when the assignment does not change

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

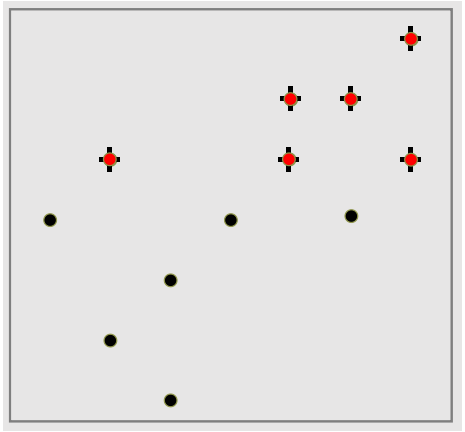
An Example of K-Means Clustering



The initial data set

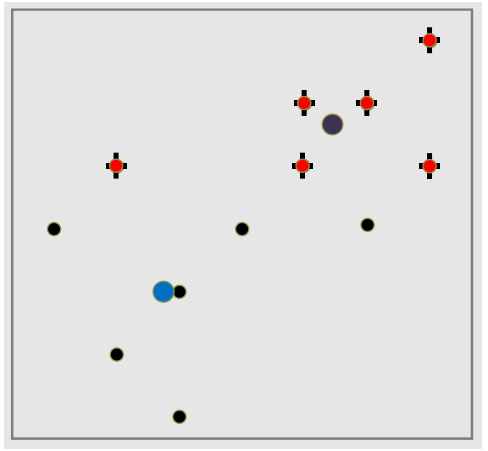
K=2

Arbitrarily partition objects into k groups



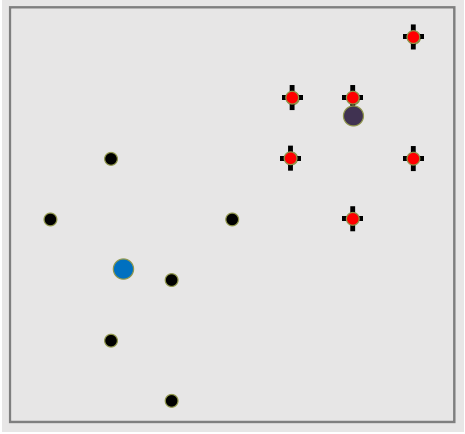
Loop if needed

Update the cluster centroids

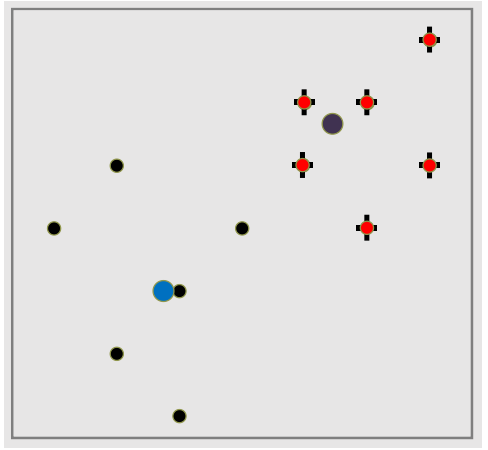


Reassign objects

- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change



Update the cluster centroids



[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

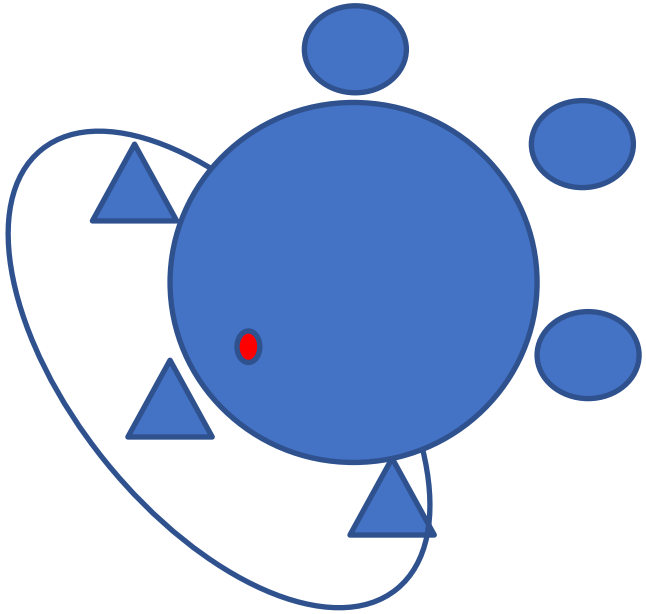
Comments on the K-Means Method

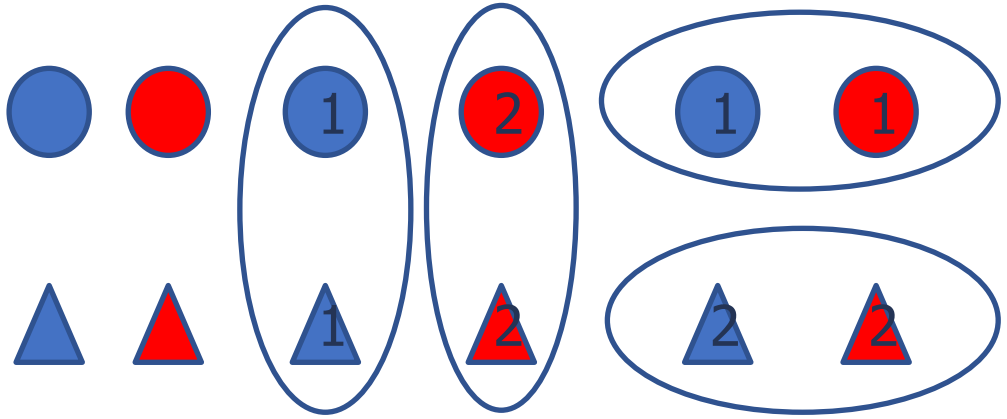
- Strength: Efficient: $O(tkn)$, where n is number objects, k is number of clusters, and t is number of iterations. Normally, k, $t << n$.
- Comment: Often terminates at a local optimal
- Weakness
  - Applicable only to objects in a continuous n-dimensional space
    - Using the k-modes method for categorical data
    - In comparison, k-medoids can be applied to a wide range of data
  - Need to specify k, the number of clusters, in advance
    - there are ways to automatically determine the best k
  - Sensitive to noisy data and outliers
  - Not suitable to discover clusters with non-convex shapes

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

What Is the Problem of the K-Means Method?

- The mean of a cluster might be meaningless
- Local minimum should be achieved due to initial configuration

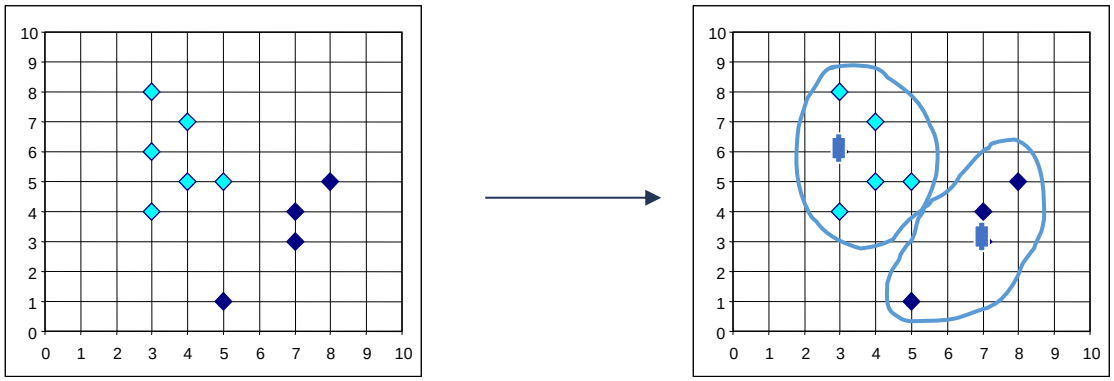




[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

The K-Medoid Clustering Method

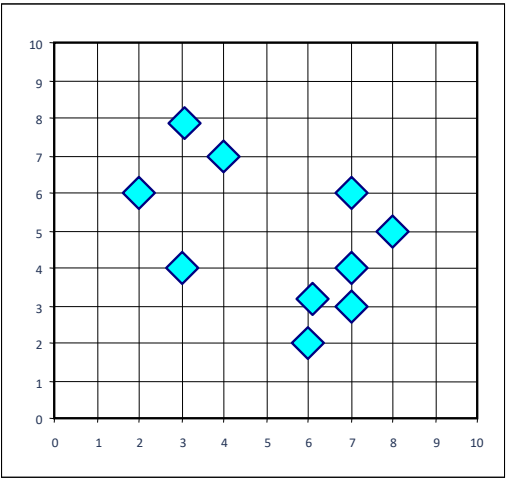
- K-Medoids Clustering
  - Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster
  - Find representative objects (medoids) in clusters
  - Starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
  - PAM works effectively for small data sets, but does not scale well for large data sets (due to the computational complexity)



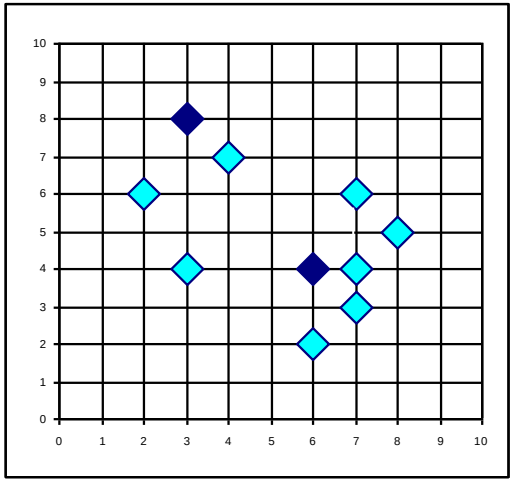
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

PAM: A Typical K-Medoids Algorithm

K=2

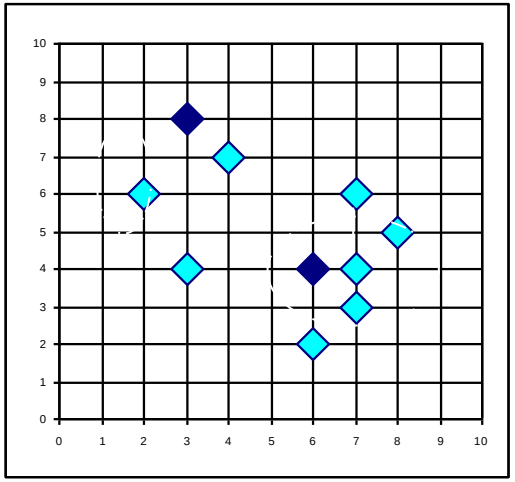


Arbitrary choose k object as initial medoids

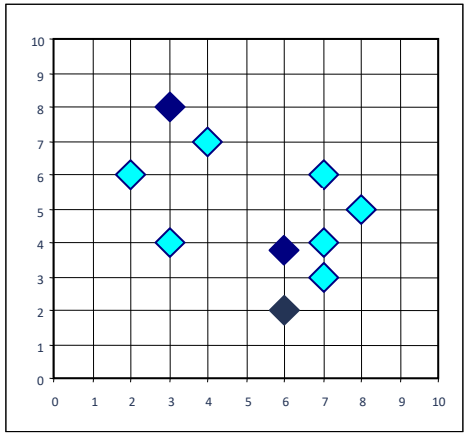


Assign each remaining object to nearest medoids

Total Cost = 20

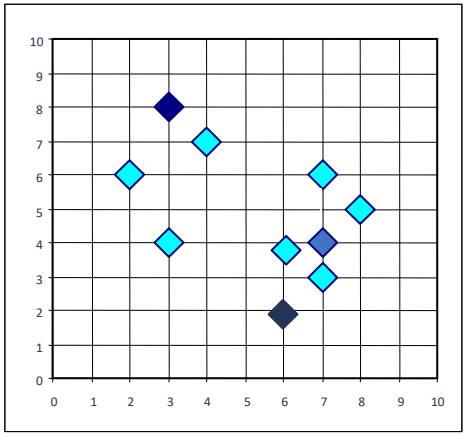


Randomly select a nonmedoid object, O_{ramdom}



Compute total cost of swapping

Total Cost = 26



Swapping O and O_{ramdom}

If quality is improved.

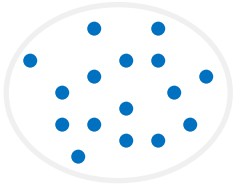
Do loop

Until no change

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Centroid, Radius and Diameter of a cluster (for numerical data sets)

- Centroid:
  - the "middle" of a cluster $c_m = \frac{\sum_{i=1}^{n} t_{ip}}{n}$



- Radius:
  - square root of average distance from any point of the cluster to its centroid

    $R_m = \sqrt{\frac{\sum_{i=1}^{n}(t_{ip}-c_m)^2}{n}}$
- Diameter:
  - square root of average mean squared distance between all pairs of points in the cluster $\mathrm{D}_m = \sqrt{\frac{\sum_{i=1}^{n}\sum_{i=1}^{n}(t_{ip}-t_{iq})^2}{n(n-1)}}$

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

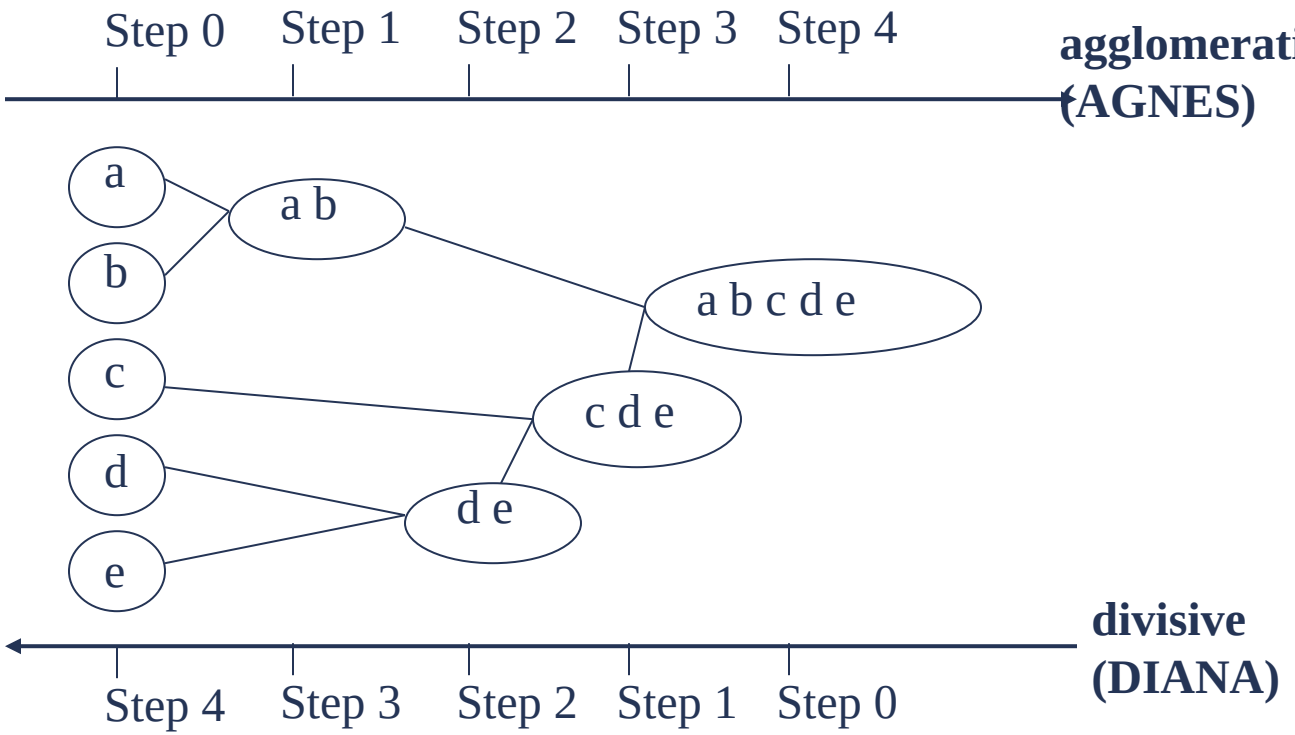
Distance between Clusters

- Single link:
  - smallest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = \min(t_{ip}, t_{jq})$
- Complete link:
  - largest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = max(t_{ip}, t_{jq})$
- Average:
  - avg distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = mean(t_{ip}, t_{jq})$
- Centroid/Medoid:
  - distance between the centroids/medoids of two clusters, i.e., $dist(K_i, K_j) = mean(c_i, c_j)$

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Hierarchical Clustering

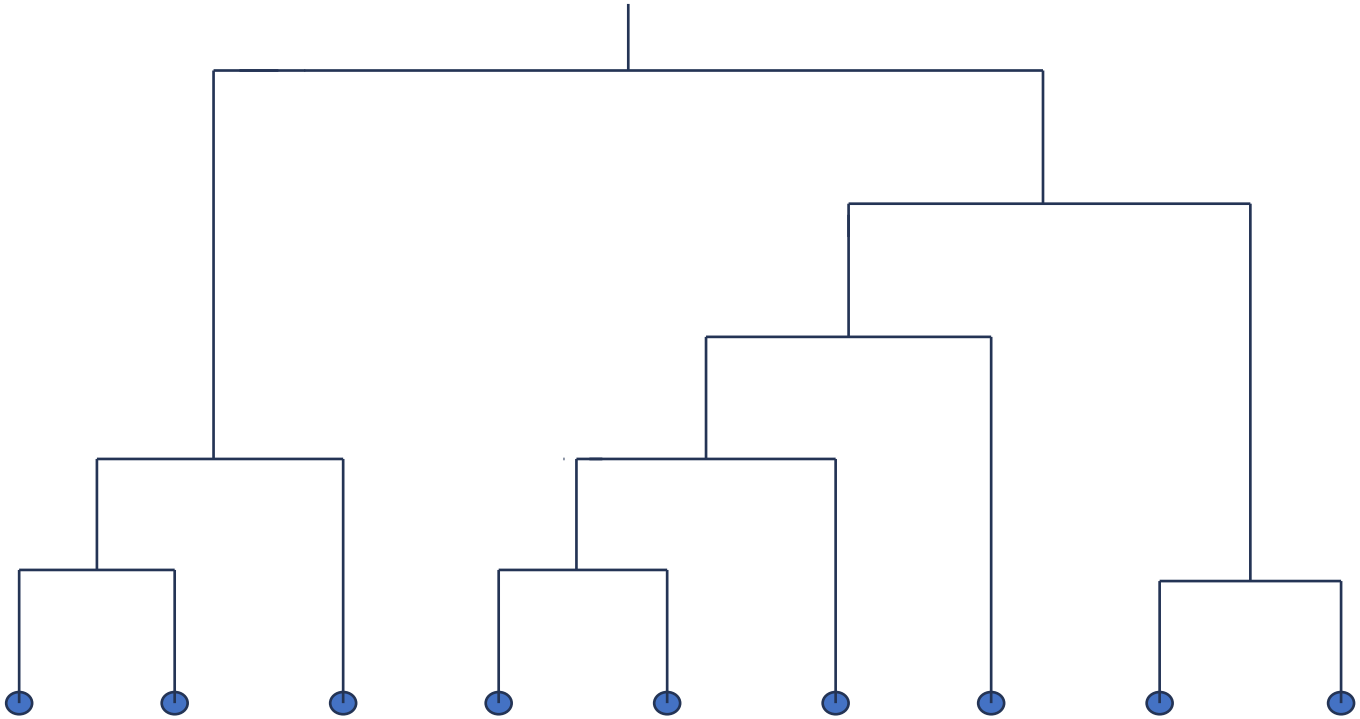
- Use distance matrix as clustering criteria.
  - This method does not require the number of clusters k as an input, but needs a termination condition



[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Dendrogram: Shows How Clusters are Merged

- Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram
  - A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster



[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

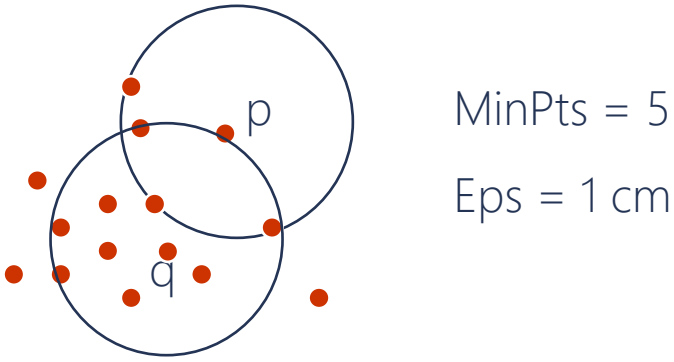
Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN
  - OPTICS
  - DENCLUE
  - CLIQUE

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Density-Based Clustering: Basic Concepts

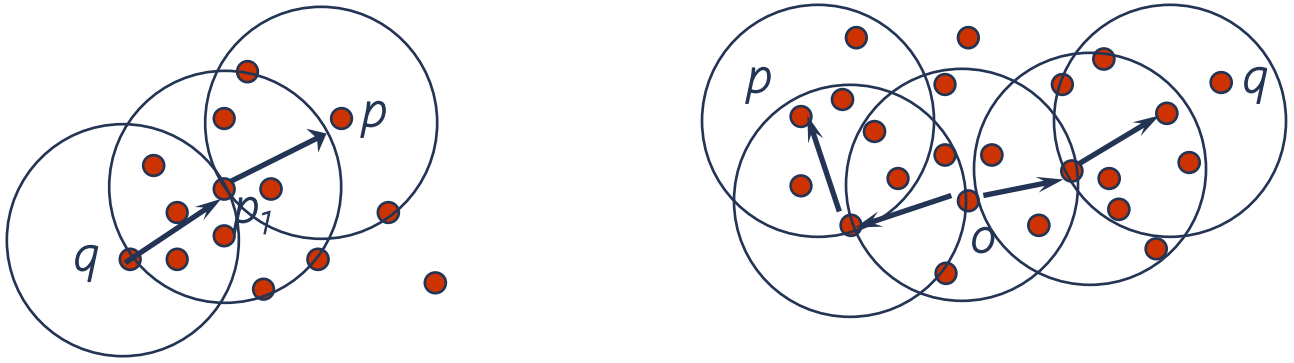
- Two parameters:
  - Eps: Maximum radius of the neighborhood
  - $MinPts$: Minimum number of points in an Eps-neighborhood of that point
- $N_{Eps}(q)$: $\{p\ belongs\ to\ D \mid dist(p,q) \leq Eps\}$
- Directly density-reachable: A point p is directly density-reachable from a point q w.r.t. Eps, $MinPts$ if
  - p belongs to $N_{Eps}(q)$
  - core point condition:$|N_{Eps}(q)| \geq$ MinPts



[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Density-Reachable and Density-Connected

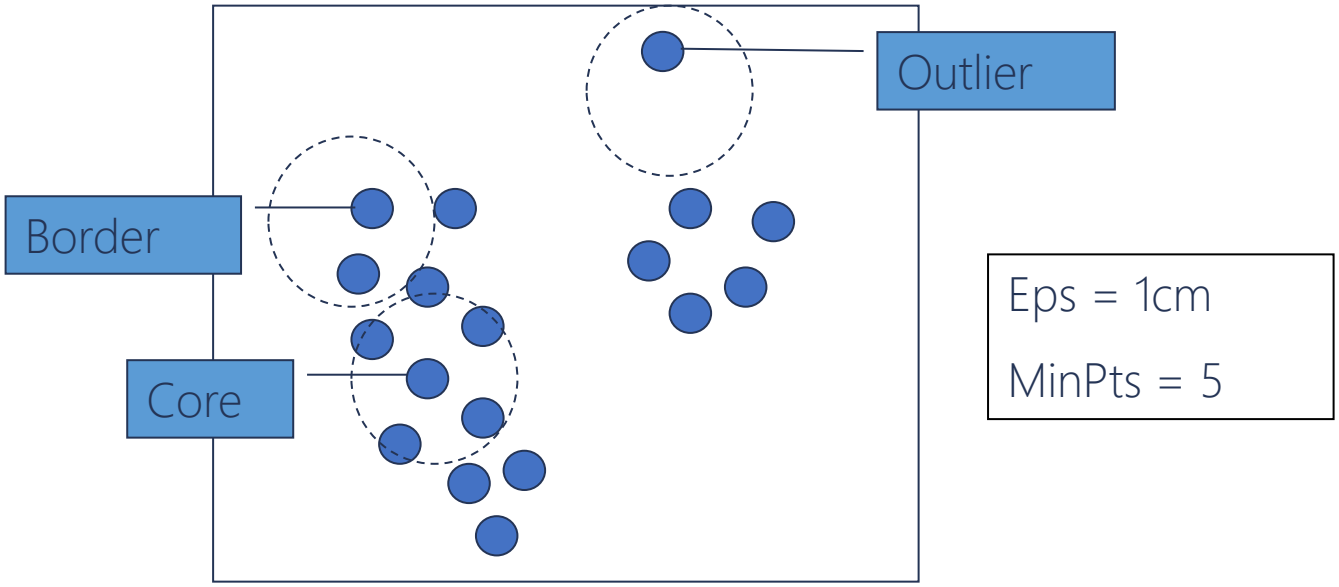
- Density-reachable:
  - A point p is density-reachable from a point q w.r.t. Eps, $MinPts$ if there is a chain of points p_1, …, p_n, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i
- Density-connected
  - A point p is density-connected to a point q w.r.t. Eps, $MinPts$ if there is a point o such that both, p and q are density-reachable from o w.r.t. Eps and $MinPts$



[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

- Relies on a density-based notion of cluster: A cluster is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

DBSCAN: The Algorithm

- Arbitrary select a point p
- Retrieve all points density-reachable from p w.r.t. Eps and $MinPts$
- If p is a core point, a cluster is formed
- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database
- Continue the process until all of the points have been processed
- If a spatial index is used, the computational complexity of DBSCAN is $O(n \cdot log(n))$, where n is the number of database objects. Otherwise, the complexity is $O(n^2)$

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Neural Network Approaches

- Neural network approaches
  - Represent each cluster as an exemplar, acting as a "prototype" of the cluster
  - New objects are distributed to the cluster whose exemplar is the most similar according to some distance measure
- Typical methods
  - SOM (Soft-Organizing feature Map)
  - Competitive learning
    - Involves a hierarchical architecture of several units (neurons)
    - Neurons compete in a "winner-takes-all" fashion for the object currently being presented

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Self-Organizing Feature Map (SOM)

- SOMs, also called topological ordered maps, or Kohonen Self-Organizing Feature Map (KSOMs)
- It maps all the points in a high-dimensional source space into a 2 to 3-d target space, s.t., the distance and proximity relationship (i.e., topology) are preserved as much as possible
- Similar to k-means: cluster centers tend to lie in a low-dimensional manifold in the feature space
- Clustering is performed by having several units competing for the current object
  - The unit whose weight vector is closest to the current object wins
  - The winner and its neighbors learn by having their weights adjusted
- SOMs are believed to resemble processing that can occur in the brain
- Useful for visualizing high-dimensional data in 2- or 3-D space

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Determine the Number of Clusters

- Empirical method
  - # of clusters: $k \approx \sqrt{\frac{n}{2}}$ for a dataset of n points, e.g., n = 200, k = 10
- Elbow method
  - Use the turning point in the curve of sum of within cluster variance w.r.t the number of clusters
- Cross validation method
  - Divide a given data set into m parts
  - Use m – 1 parts to obtain a clustering model
  - Use the remaining part to test the quality of the clustering
    - E.g., For each point in the test set, find the closest centroid, and use the sum of squared distance between all points in the test set and the closest centroids to measure how well the model fits the test set
  - For any k > 0, repeat it m times, compare the overall quality measure w.r.t. different k's, and find number of clusters that fits the data the best

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Measuring Clustering Quality

- Three kinds of measures:
  - External
  - Internal
  - Relative
- External: supervised, employ criteria not inherent to the dataset
  - Compare a clustering against prior or expert-specified knowledge (i.e., the ground truth) using certain clustering quality measure
  - Entropy, Normalized Mutual Information (NMI)
- Internal: unsupervised, criteria derived from data itself
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are, e.g., Silhouette coefficient
- Relative: directly compare different clustering, usually those obtained via different parameter settings for the same algorithm

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

Entropy-Based Measure: Conditional Entropy

- Entropy of clustering C: $H(C) = -\sum_{i=1}^{r} p_{C_i} \cdot \log(p_{C_i})$
  - $p_{C_i} = \frac{n_i}{n}$, the probability of cluster C_i
- Entropy of partitioning T: $H(T) = -\sum_{i=1}^{r} p_{T_i} \cdot \log(p_{T_i})$
- Entropy of T w.r.t. cluster C_i : $H(T|C_i) = -\sum_{j=1}^{k} \left(\frac{n_{ij}}{n_i}\right) \log(\frac{n_{ij}}{n_i})$
- Conditional entropy of T w.r.t. clustering C:
  - $H(T|C) = -\sum_{j=1}^{k} \left(\frac{n_i}{n}\right) H(T|C_i) = -\sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij} \cdot \log(\frac{p_{ij}}{p_{C_i}})$
  - $H(T|C) = -\sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij} \cdot \log(p_{ij}) + \sum_{i=1}^{r} (\log(p_{C_i}) \cdot \sum_{j=1}^{k} p_{ij})$
  - $= -\sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij} \cdot \log(p_{ij}) + \sum_{i=1}^{r} (p_{C_i} \cdot \log(p_{C_i})) = H(C,T) - H(C)$
  - The more a cluster's members are split into different partitions, the higher the conditional entropy
  - For a perfect clustering, the conditional entropy value is 0, where the worst possible conditional entropy value is log k

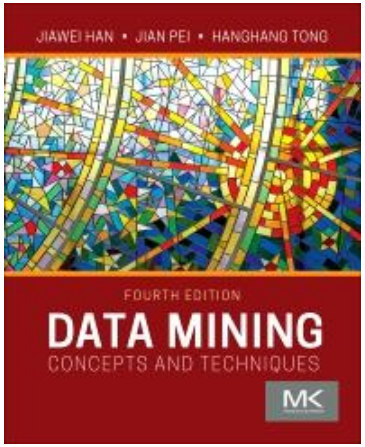
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

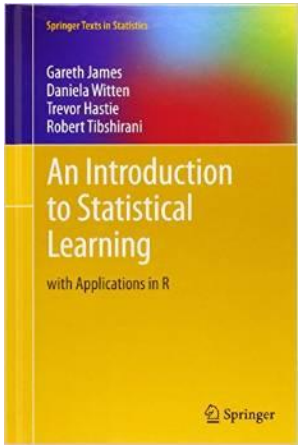
Other types of clustering

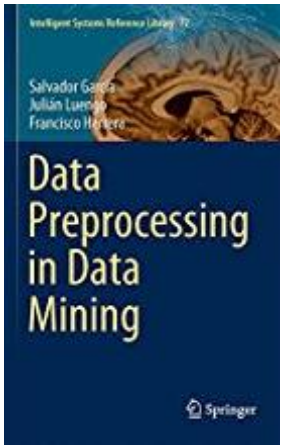
- Spectral Clustering
- Clustering Graphs and Network Data
- Constraint-Based Cluster
  - Must-link vs. cannot link constraints
    - $must - link(x, y)$: x and y should be grouped into one cluster
  - Constraints can be defined using variables, e.g.,
    - $cannot - link(x, y)\ if\ dist(x, y) > d$

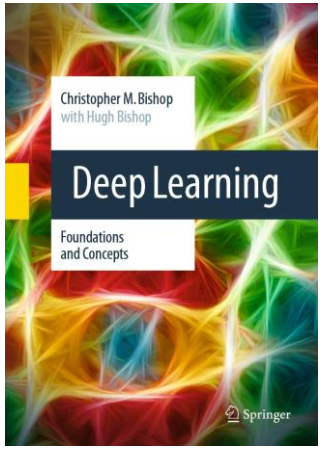
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022..

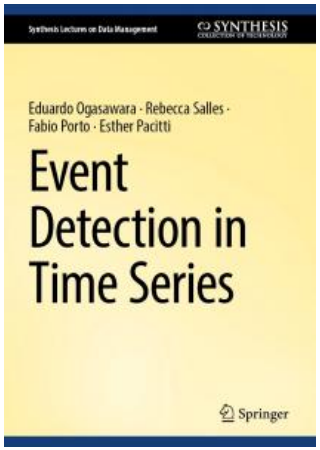
Main References











[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

[2] G. M. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R. Springer Nature, 2021.

[3] S. Garcia, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining. Springer, 2014.

[4] C. M. Bishop and H. Bishop, Deep Learning: Foundations and Concepts. Springer Nature, 2023.

[5] E. Ogasawara, R. Salles, F. Porto, and E. Pacitti, Event Detection in Time Series, 1st ed. in Synthesis Lectures on Data Management. Cham: Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-75941-3.

Slides and videos at: https://eic.cefet-rj.br/~eogasawara/data-mining/