

Clustering Analysis-II

Man-Kwan Shan

Dept. of Computer Science

National Cheng-Chi Univ.

Types of Input Dataset in Clustering Analysis

Types of Input Dataset in Cluster Analysis

- Relational data
- Transactional data
- Sequence
- Time series
- Spatial data
- Mobility Data (maps)
- Textual data
- Tree
- Graph, Social Network
- Image, Video
- Audio, Music

Clustering of Relation Data

- Two different types of data
 - data matrix (two-mode matrix)
 - object-by-attributes structure
 - n objects with p attributes
 - n by p matrix
 - dissimilarity matrix(one-mode matrix)
 - object-by-object structure
 - a collection of proximity for all pairs of n objects
 - n by n matrix

Data Matrix *(two-mode matrix)*

Object ID	ID	Weight	Height	Gender	Birthday	Age
1	111755001	45	165	F	1975/06/06	47
2	111755029	60	170	M	1982/08/30	40
3	111755009	48	160	F	1982/10/29	40
4	111755098	66	180	M	1995/03/16	27
5	111655034	63	175	M	1990/01/15	32

h

p

Dissimilarity Matrix (one-mode matrix)

n

	1	2	3	4	5
1	0	2.3	3.4	1.2	3.7
2	2.3	0	2.0	1.8	2.2
3	3.4	2.0	0	4.2	0.7
4	1.2	1.8	4.2	0	4.4
5	3.7	2.2	0.7	4.4	0

Input Data Type of Clustering Algorithm

- K-means 2 mode partition-based

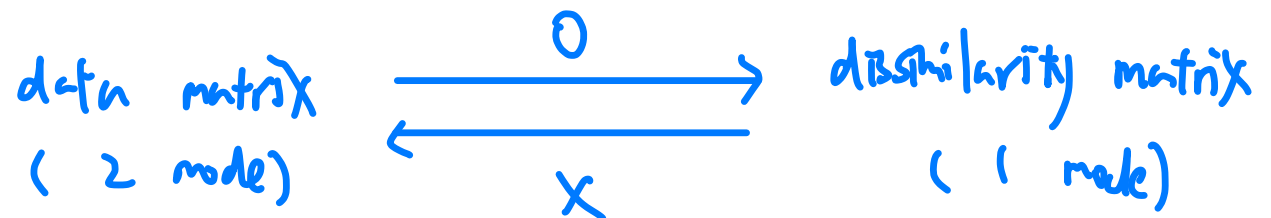
- Single-Link, Complete-Link, Average-Link 1 mode
- BIRCH 2 mode
- Chameleon 1 mode hierarchical

- DBScan 1 mode density-based

- GMM 2 mode model-based

情境

- Case 1
 - Clustering program: two mode
 - Data: one mode
 - Can not transform one mode into two mode
- Case 2
 - Clustering program: one mode
 - Data: two mode
 - Transform two mode into one mode based on similarity or distance



Attribute Type of Relational Data

Types of Attributes (variables, features)

- Interval-scaled variables: weight, height 有大小關係和倍數
- Ordinal: freshman, sophomore, junior, senior 沒倍數關係
- Ratio-scaled variables 有倍數關係但非線性
- Binary variables: male, female
- Nominal: R, G, B 無大小關係

Interval-scaled Variables

- Interval-scaled variable:
 - continuous measurement of a roughly linear scale
e.g.: weight, height, temperature
- Measurement unit may affect clustering analysis
 - smaller units → larger range
e.g. height=1.7 m, weight=60 Kg
height=170 cm, weight=60 Kg
 - Normalization to [0,1]
 - Min-max normalization: $x' = \frac{x - \min_A}{\max_A - \min_A}$
 - Z-score normalization: $x' = \frac{x - \text{mean}_A}{\text{std}_A}$

Interval-scaled Variables(cont.)

- Distance between object i, j of p interval attributes
 - Euclidean distance (L2)

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2}$$

- Manhattan (city block) distance (L1)

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

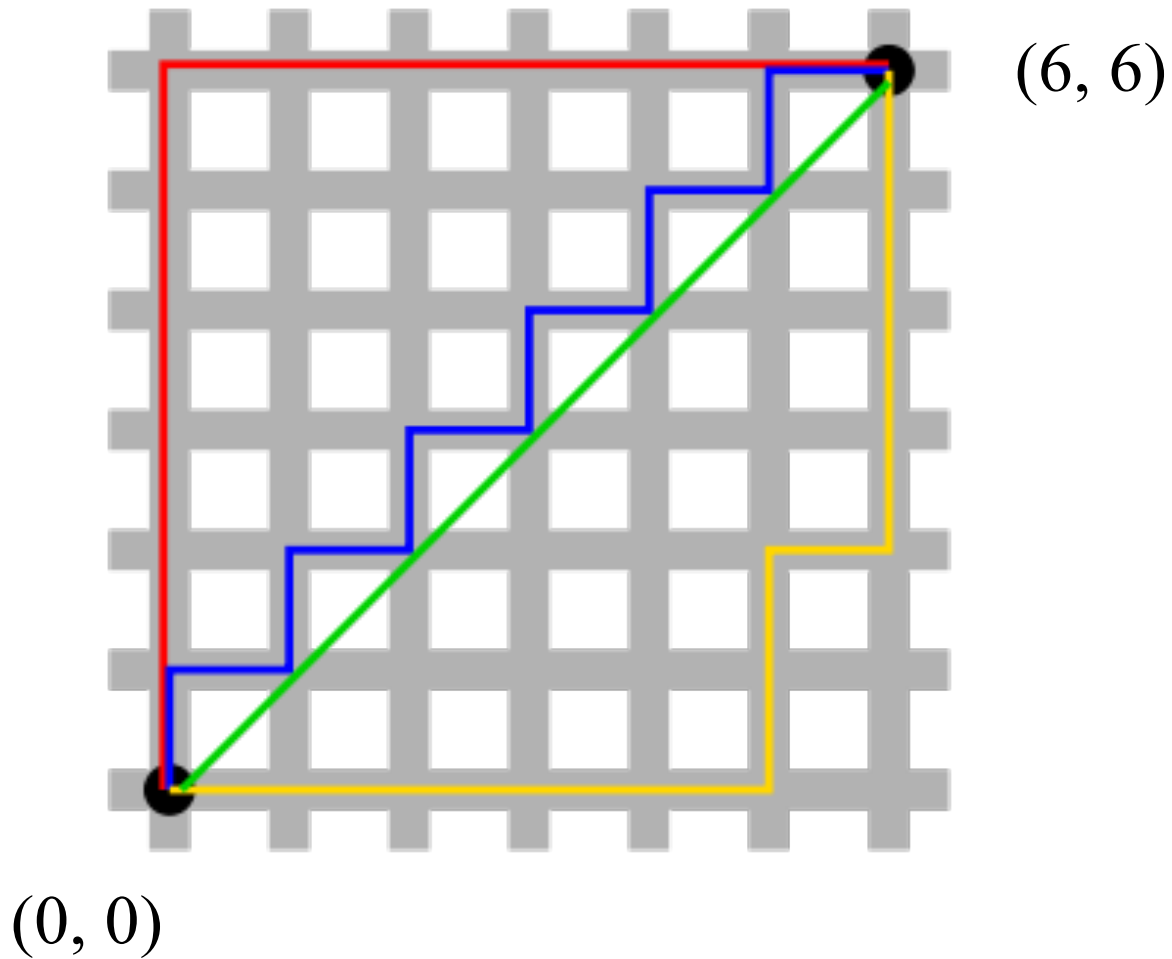
- Minkowski distance

$$d(i, j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q}$$

- Weighted Manhattan distance

$$d(i, j) = \sqrt[q]{w_1 |x_{i1} - x_{j1}|^q + w_2 |x_{i2} - x_{j2}|^q + \dots + w_p |x_{ip} - x_{jp}|^q}$$

Manhattan (city block) distance



Ordinal Variables

- Ordinal variables
 - relative ordering
 - e.g. assistant, associate, full professor
 - can be obtained by splitting value range of interval-scaled variable into a finite no. of classes
 - distance between objects i, j
 - step 1: replace by corresponding rank
 - step 2: normalize onto $[0, 1]$
 - computed using interval scaled distance

Ratio-scaled Variables

- Ratio-scaled variable
 - positive measurement on a nonlinear scale
 - e.g. Richter scale : as measured with the amplitude of the seismic waves to an arbitrary, minor amplitude.
 - Earthquake that registers 5.0 on the Richter scale has a shaking amplitude 10 times that of an earthquake that registered 4.0
 - e.g. decibel (dB) in acoustics
- Approaches for distance between objects i, j
 1. treat like interval-scaled variables
 2. apply transformation and treat as interval value
 3. treat as continuous ordinal variable

Binary Variables

- Binary variable
 - symmetric
 - e.g. male, female
 - asymmetric
 - e.g. HIV positive, HIV negative

Distance between Binary Vectors (cont.)

- Example: patient record

Name	Gender	fever	cough	Test1	Test2	Test3	Test4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N
:	:	:	:	:	:	:	:

- gender: symmetric, others: asymmetric

Similarity Between Binary Vectors

- To compute the similarity between two objects, p and q , having only binary attributes.

- Compute similarities using the following quantities

M_{01} = number of attributes where p was 0 and q was 1

M_{10} = number of attributes where p was 1 and q was 0

M_{00} = number of attributes where p was 0 and q was 0

M_{11} = number of attributes where p was 1 and q was 1

- Simple Matching

SMC = number of matches / number of attributes

$$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

→ insignificant

↓ take M_{00} away from numerator & denominator

- Jaccard Coefficients

J = number of non-zero-matches / number of not-both-zero attributes

$$= M_{11} / (M_{01} + M_{10} + M_{11}) = |P \cap Q| / |P \cup Q|$$

SMC versus Jaccard: Example

$p = 0\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$ $\{b, g, j\}$

$q = 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0$ $\{a, g\}$

$a\ b\ c\ d\ e\ f\ g\ h\ i\ j$

$M_{01} = 1$ (number of attributes where i was 0 and j was 1)

$M_{10} = 2$ (number of attributes where i was 1 and j was 0)

$M_{00} = 6$ (number of attributes where i was 0 and j was 0)

$M_{11} = 1$ (number of attributes where i was 1 and j was 1)

$$\begin{aligned} \text{SMC} &= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{00} + M_{11}) \\ &= (1 + 6) / (1 + 2 + 6 + 1) = 0.7 \end{aligned}$$

 M_{00} dilutes the numerator & denominator!

$$\begin{aligned} J &= (M_{11}) / (M_{01} + M_{10} + M_{11}) = 1 / (1 + 2 + 1) = 0.25 \\ &= |P \cap Q| / |P \cup Q| \end{aligned}$$

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|},$$

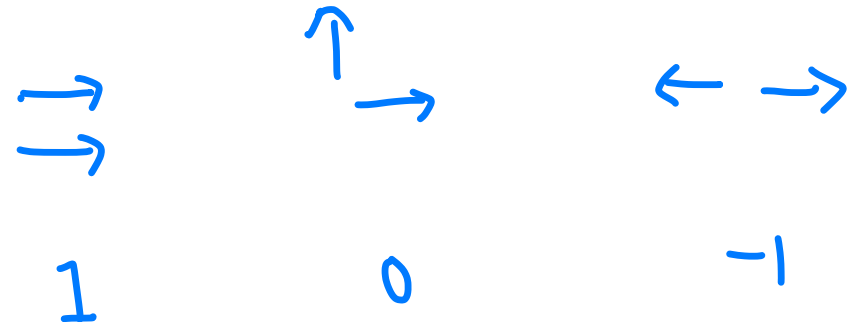
where

- indicates vector dot product and $\|d\|$ is the length of vector d .

- Example:

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$



$$d_1 \cdot d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|d_1\| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

Distance Between Binary Vectors

- distance between objects i, j
 - **symmetric**
 - e.g. male, female
 - $(M_{01} + M_{10}) / (M_{01} + M_{10} + M_{11} + M_{00})$
 - **asymmetric** 1 (presence) is more important than 0 (absence)
 - e.g. HIV positive, HIV negative
 - $(M_{01} + M_{10}) / (M_{01} + M_{10} + M_{11})$

Distance between Binary Vectors (cont.)

- Example: patient record

Name	Gender	fever	cough	Test1	Test2	Test3	Test4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N
:	:	:	:	:	:	:	:

- gender: symmetric, others: asymmetric
- let Y, P = 1, N=0

$$d(Jack, Mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(Jack, Jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(Jim, Mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

Jack 101000

Mary 101010

Jack 101000

Jim 110000

Jim 110000

Mary 101010

Nominal Variables

- Nominal variable
 - generalization of binary variable, take on more than two states
 - e.g. color
 - distance between objects i, j

$$d(i, j) = \frac{p - m}{p} = \frac{\text{\#(unmatched variables)}}{\text{\#(variables)}}$$

where $p : \text{\#(variables)}$, $m : \text{\#(matched variables)}$

- can be encoded by asymmetric binary variables
 - e.g. color variable $C(R, G, B) \Rightarrow$ color variables R, G, B

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(i, j) \geq 0$ for all i & j and

$d(i, j) = 0$ only if $i = j$. (Positive definiteness)

2. $d(i, j) = d(j, i)$ for all i and j . (Symmetry)

3. $d(i, k) \leq d(i, j) + d(j, k)$ for all i, j , and k .

(Triangle Inequality)

where $d(i, j)$ is the distance (dissimilarity) between points (data objects), i and j .

- metric:** distance that satisfies these properties

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

2. $s(p, q) = s(q, p)$ for all p and q . (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), p and q .

Distance of K-means

K-means Objective Function

- Sum of Squared Errors(SSE)

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} \text{dist}^2(c_i, x)$$

C_i : i th cluster
 c_i : centroid of the i th cluster

- The centroid that minimize the SSE of the cluster is the mean

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

**What if the distance is cosine
or L1 distance,
rather than Euclidean distance ?**



Gradient Descent for Euclidean Distance

$$\frac{\partial}{\partial c_k} \text{SSE} = \frac{\partial}{\partial c_k} \sum_{i=1}^K \sum_{x \in C_i} (c_i - x)^2$$

1. c_k only affects
the i th cluster

2. $\frac{\partial}{\partial c_k} (c_k - x_k)^2$

$$= \frac{\partial (c_k - x_k)^2}{\partial (c_k - x_k)} \frac{\partial (c_k - x_k)}{\partial c_k}$$



$$= \sum_{i=1}^K \sum_{x \in C_i} \frac{\partial}{\partial c_k} (c_i - x)^2$$

$$= \sum_{x \in C_k} 2 \times (c_k - x_k) = 0$$

$$\sum_{x \in C_k} 2 \times (c_k - x_k) = 0 \Rightarrow m_k c_k = \sum_{x \in C_k} x_k \Rightarrow c_k = \frac{1}{m_k} \sum_{x \in C_k} x_k$$

m_k : number of objects in the k -th cluster

mean of
the cluster

Gradient Descent for L1 Distance

sum of
absolute
errors



$$\begin{aligned}\frac{\partial}{\partial c_k} \text{SAE} &= \frac{\partial}{\partial c_k} \sum_{i=1}^K \sum_{x \in C_i} |c_i - x| \\ &= \sum_{i=1}^K \sum_{x \in C_i} \frac{\partial}{\partial c_k} |c_i - x| \\ &= \sum_{x \in C_k} \frac{\partial}{\partial c_k} |c_k - x| = 0\end{aligned}$$

$$\sum_{x \in C_k} \frac{\partial}{\partial c_k} |c_k - x| = 0 \Rightarrow \sum_{x \in C_k} \text{sign}(x - c_k) = 0$$

c_k : **median** of objects in the k -th cluster

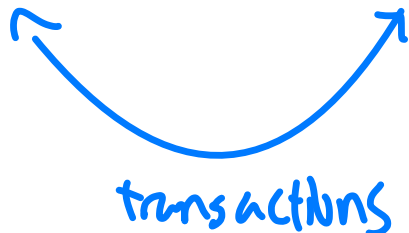
Distance between Transactions

ROCK

- ROCK(RObust Clustering using linKs, '99)
 - Clustering category data
- Major ideas
 - Use links to measure similarity/proximity
 - Not distance-based
 - Computational complexity:
- Algorithm: sampling-based clustering
 - Draw random sample
 - Cluster with links
 - Label data in disk

Similarity Measure in ROCK

- How to cluster set objects
 - Example: how to cluster the following 14 transactions into two clusters
 - Approach:
 - Step 1: Generate proximity matrix
 - Step 2: Hierarchical clustering
 - How to measure the similarity (distance) between two item-sets ?
 - Jaccard coefficient
 - $T_1 = \{a, b, c\}$, $T_2 = \{c, d, e\}$



Similarity Measure in ROCK

- Jaccard coefficient measures for categorical data may not work well
- Example: Two groups (clusters) of transactions
 - C_1 (concept 1) **<a, b, c, d, e>**: {a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}, {b, c, d}, {b, c, e}, {b, d, e}, {c, d, e}
 - C_2 . (concept 2) **<a, b, f, g>**: {a, b, f}, {a, b, g}, {a, f, g}, {b, f, g}
- Jaccard co-efficient may lead to wrong clustering result
 - C_1 : 0.2 ({a, b, c}, {b, d, e}) to 0.5 ({a, b, c}, {a, b, d})
 - C_1 & C_2 : could be as high as 0.5 ({a, b, c}, {a, b, f})

Link Measure in ROCK

- Links: # of common neighbors (link threshold Jaccard ≥ 0.5)
 - $C_1 \langle a, b, c, d, e \rangle$: $\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \{a, c, e\}, \{a, d, e\}, \{b, c, d\}, \{b, c, e\}, \{b, d, e\}, \{c, d, e\}$
 - $C_2 \langle a, b, f, g \rangle$: $\{a, b, f\}, \{a, b, g\}, \{a, f, g\}, \{b, f, g\}$
 - $\text{link}(\{a, b, c\}_1, \{c, d, e\}_1) = 4$, 4 common neighbors
 - $\{a, c, d\}, \{a, c, e\}, \{b, c, d\}, \{b, c, e\}$
 - $\text{link}(\{a, b, c\}_1, \{a, b, f\}_2) = 3$, 3 common neighbors
 - $\{a, b, d\}, \{a, b, e\}, \{a, b, g\}$
 - $\text{link}(\{a, f, g\}_2, \{a, b, g\}_2) = 2$, 2 common neighbors
 - $\{a, b, f\}, \{b, f, g\}$
 - $\text{link}(\{a, f, g\}_2, \{a, b, c\}_1) = 0$, 0 common neighbors
- Link is a better measure than Jaccard coefficient

Distance between Sequences ?

$$d(\text{abbc}, \text{babb}) = ?$$

Distance between Time Series ?

$$d(<300, 250, 280, 350>, \\ <290, 280, 350, 300>) = ?$$

Cluster Evaluation

Cluster Evaluation

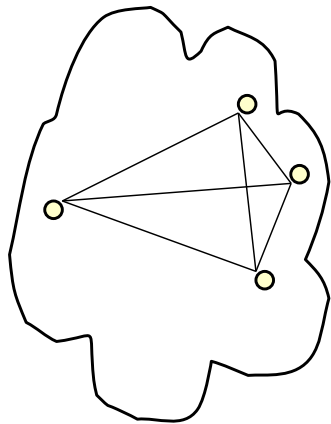
- Cluster evaluation is not well-developed.
- Sometimes, cluster analysis is conducted as a part of an exploratory data analysis. Hence cluster evaluation seems to be unnecessary.
- There are different types of clusters. Each type requires different evaluation measure.
- Nonetheless, cluster evaluation should be part of any cluster analysis.
 - To avoid finding patterns in noise
 - To determine the clustering tendency of dataset
 - To determine the number of clusters
 - To compare clustering algorithms
 - To compare clustering results (internally, externally)

Cluster Evaluation (cont.)

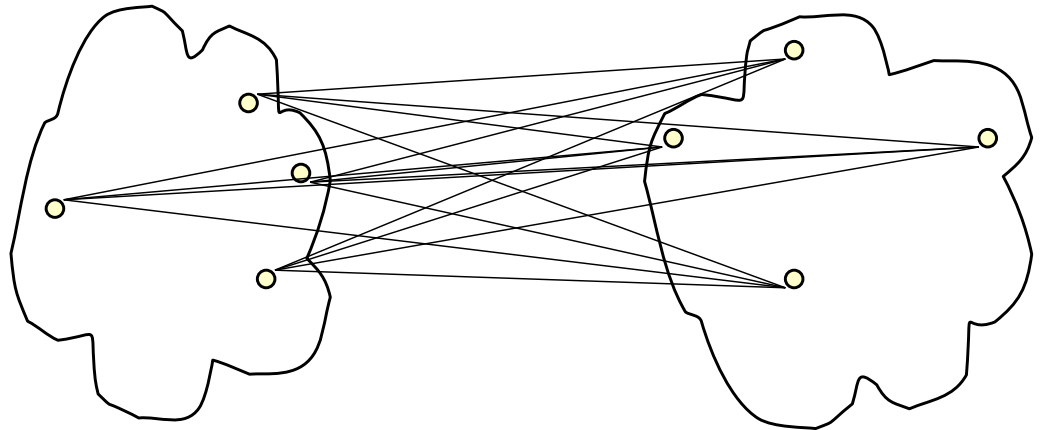
- Evaluation measures
 - Unsupervised (**internal** indices): to measure the goodness of a clustering structure without respect to external information.
 - Cluster **cohesion**
 - Cluster **separation**
 - Supervised (**external** indices): to measure the extent to which cluster labels match externally supplied class labels.

Unsupervised Cluster Evaluation

- Graph-based view : focusing on pairwise relationships
 - overall validity = $\sum_{i=1}^K w_i \times \text{validity}(C_i)$ (w_i : weights)
 - cohesion(C_i) = $\sum_{x,y \in C_i} \text{proximity}(x,y)$
 - separation(C_i, C_j) = $\sum_{x \in C_i, y \in C_j} \text{proximity}(x,y)$



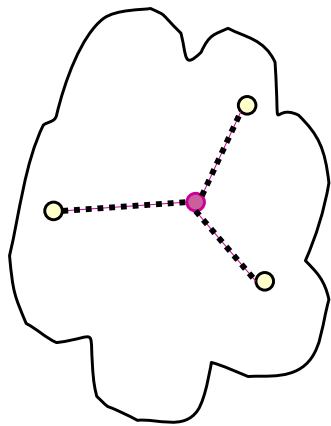
cohesion



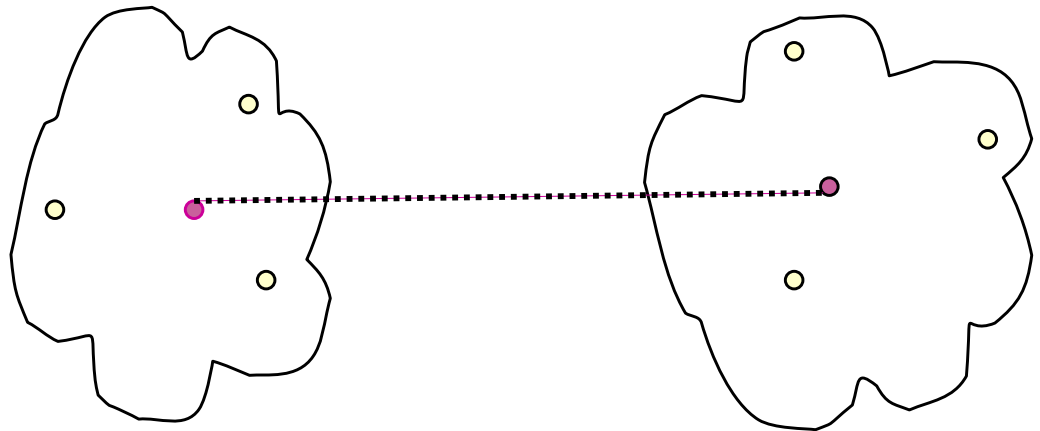
separation

Unsupervised Cluster Evaluation (cont.)

- Prototype-based view : focusing on representative points
 - $overall\ validity = \sum_{i=1}^K w_i \times validity(C_i)$
 - $cohesion(C_i) = \sum_{x \in C_i} proximity(x, C_i)$
 - $separation(C_i, C_j) = proximity(C_i, C_j)$
 - $separation(C_i) = proximity(C_i, c)$, c is overall centroid



cohesion



separation

Evaluating Individual Object

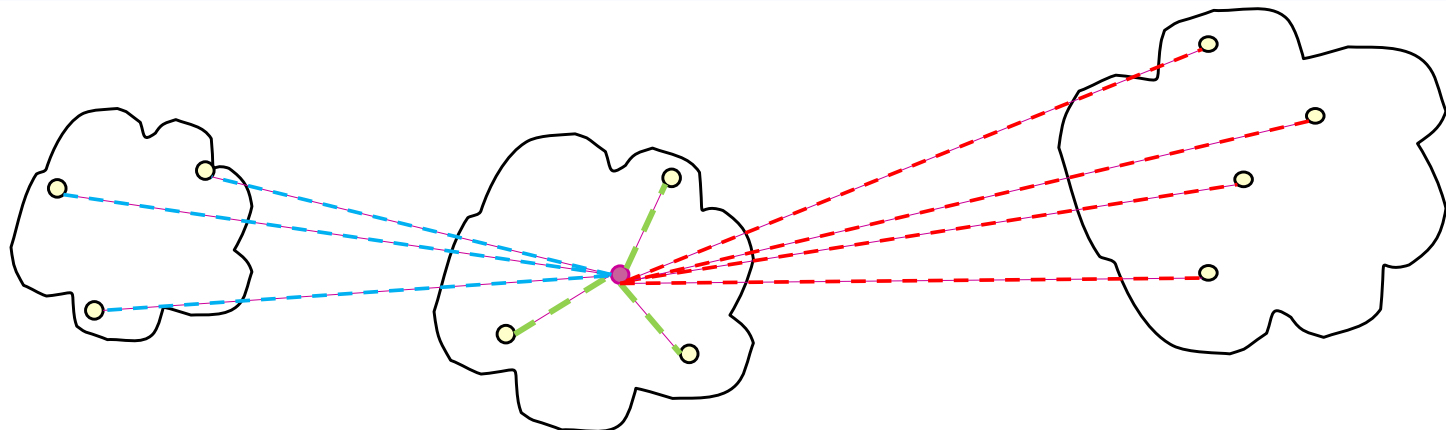
- Evaluation individual object within a cluster
 - **Silhouette coefficient** : a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)

– Silhouette coefficient for an object i , $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$, $[-1, 1]$

- a_i : average distance to all objects within cluster

b_i : minimum distance to other clusters

where distance to other cluster is the average distance



Silhouette Coefficient:

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

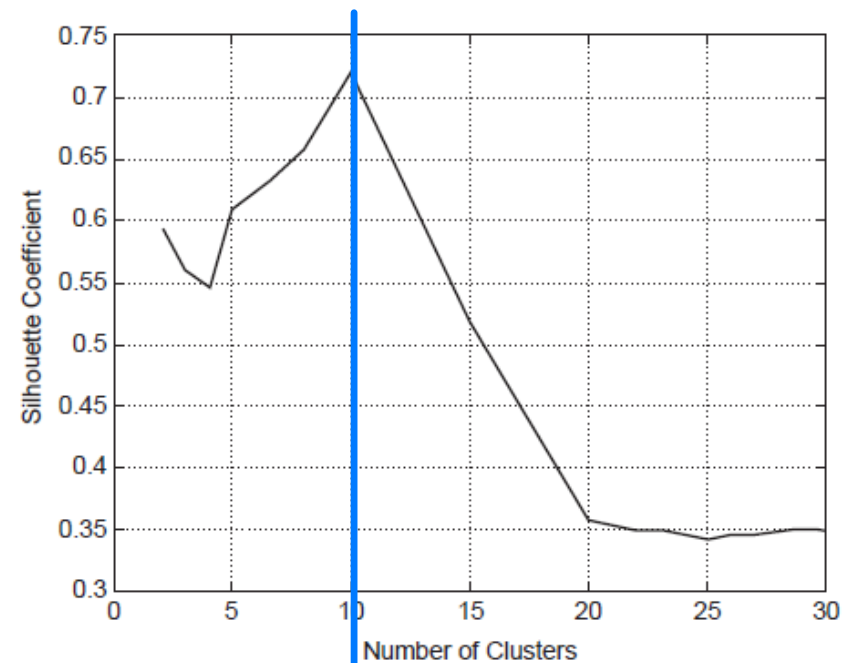
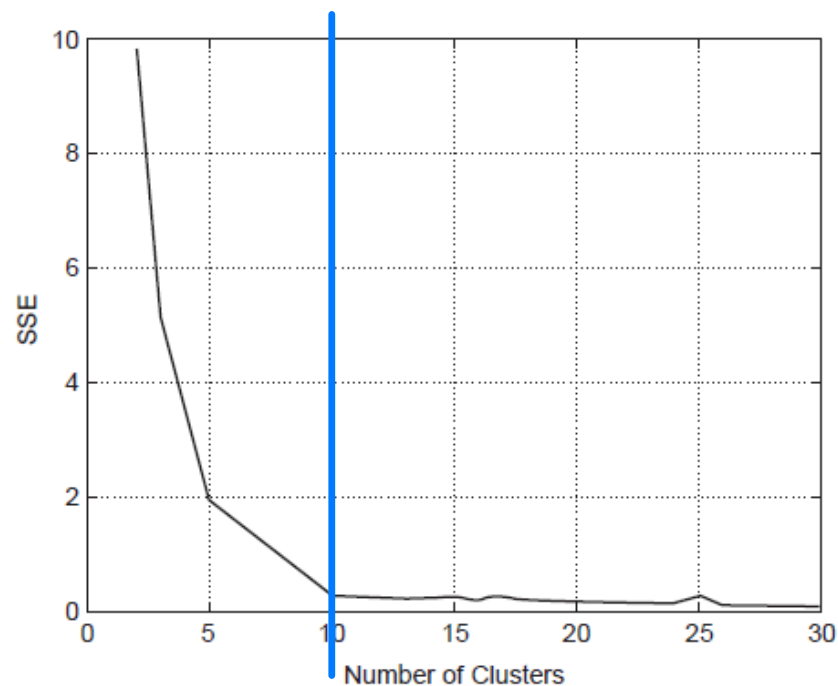
① $b_i > a_i :$

$$S_i = \frac{b_i - a_i}{b_i} = 1 - \frac{a_i}{b_i} > 0$$

② $b_i = a_i :$ $S_i = 0$

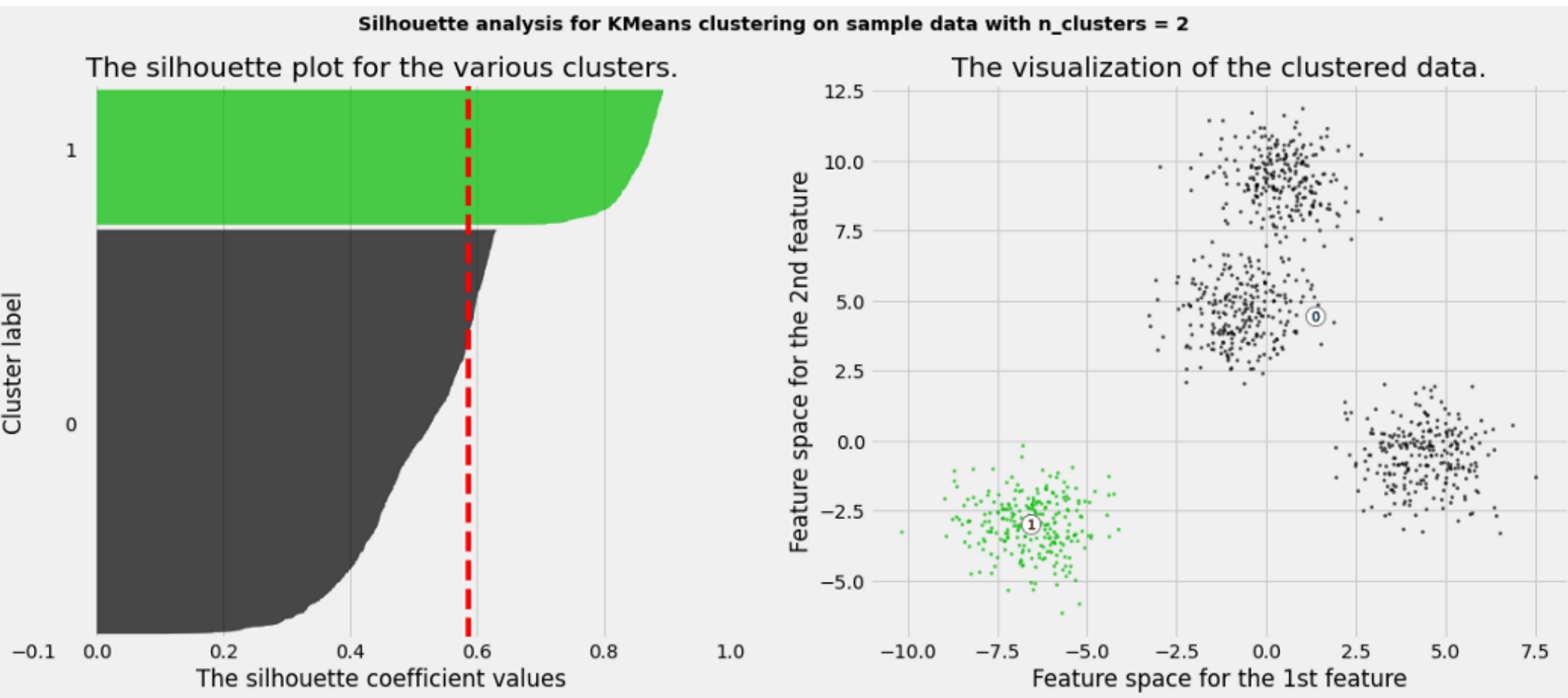
③ $b_i < a_i :$ $\frac{b_i - a_i}{a_i} = \frac{b_i}{a_i} - 1 < 0$

Determining the Number of Clusters



Elbow method

Determining the Number of Clusters (cont.)

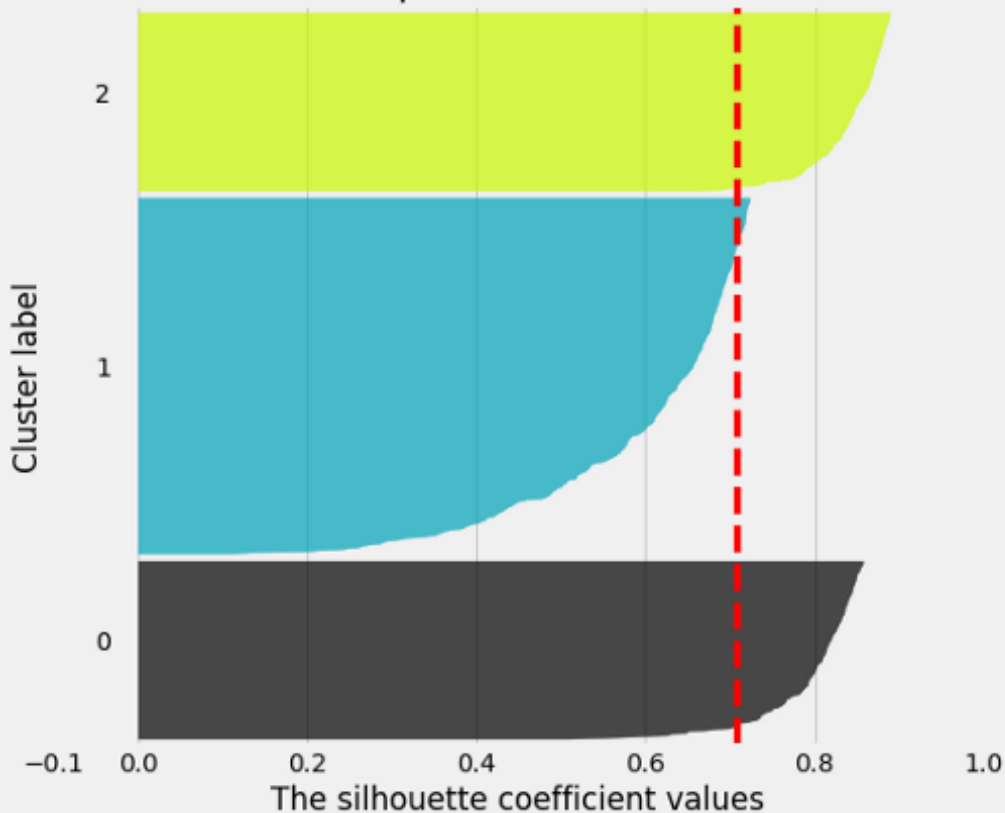


ref: <https://medium.com/mlearning-ai/stop-using-the-elbow-method-to-compute-optimal-clusters-in-k-means-clustering-1f572c587d2e>

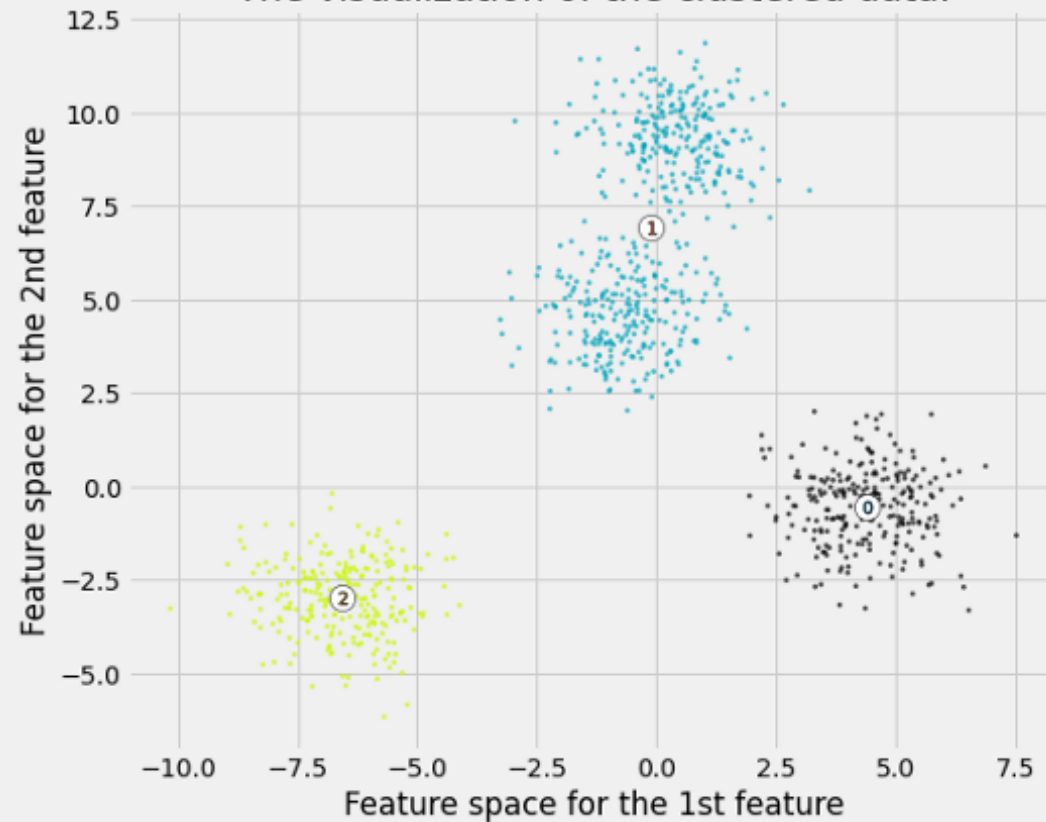
Determining the Number of Clusters (cont.)

Silhouette analysis for KMeans clustering on sample data with $n_clusters = 3$

The silhouette plot for the various clusters.



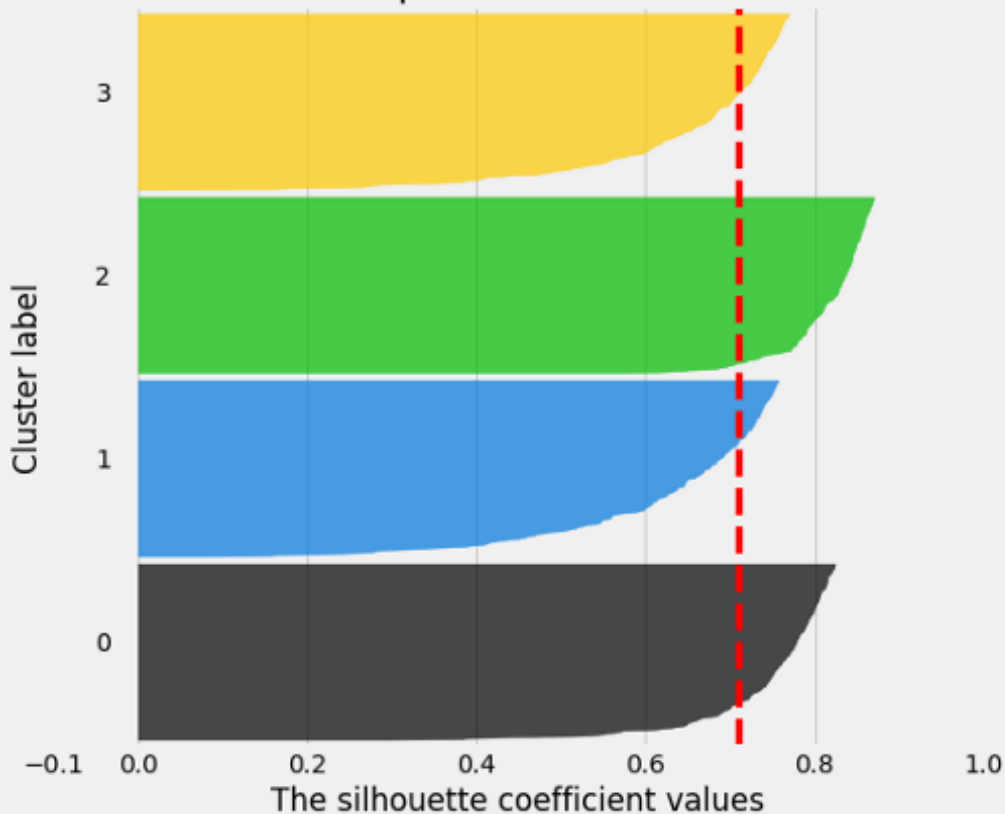
The visualization of the clustered data.



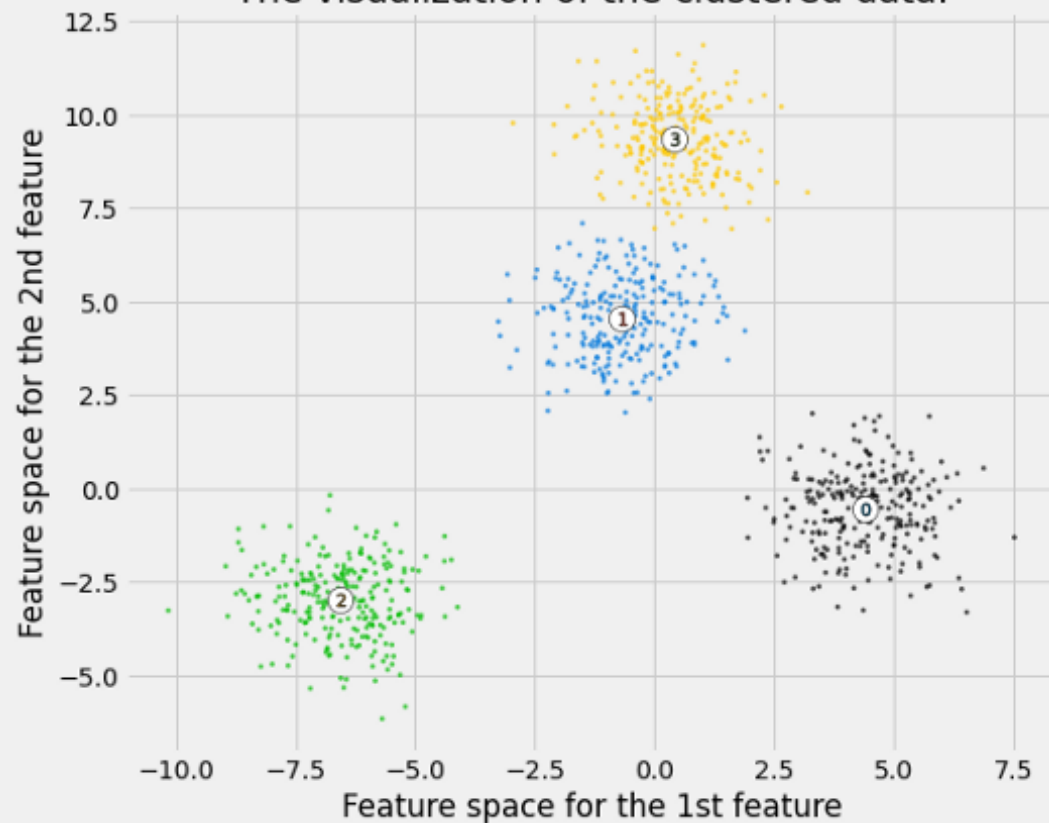
Determining the Number of Clusters (cont.)

Silhouette analysis for KMeans clustering on sample data with $n_clusters = 4$

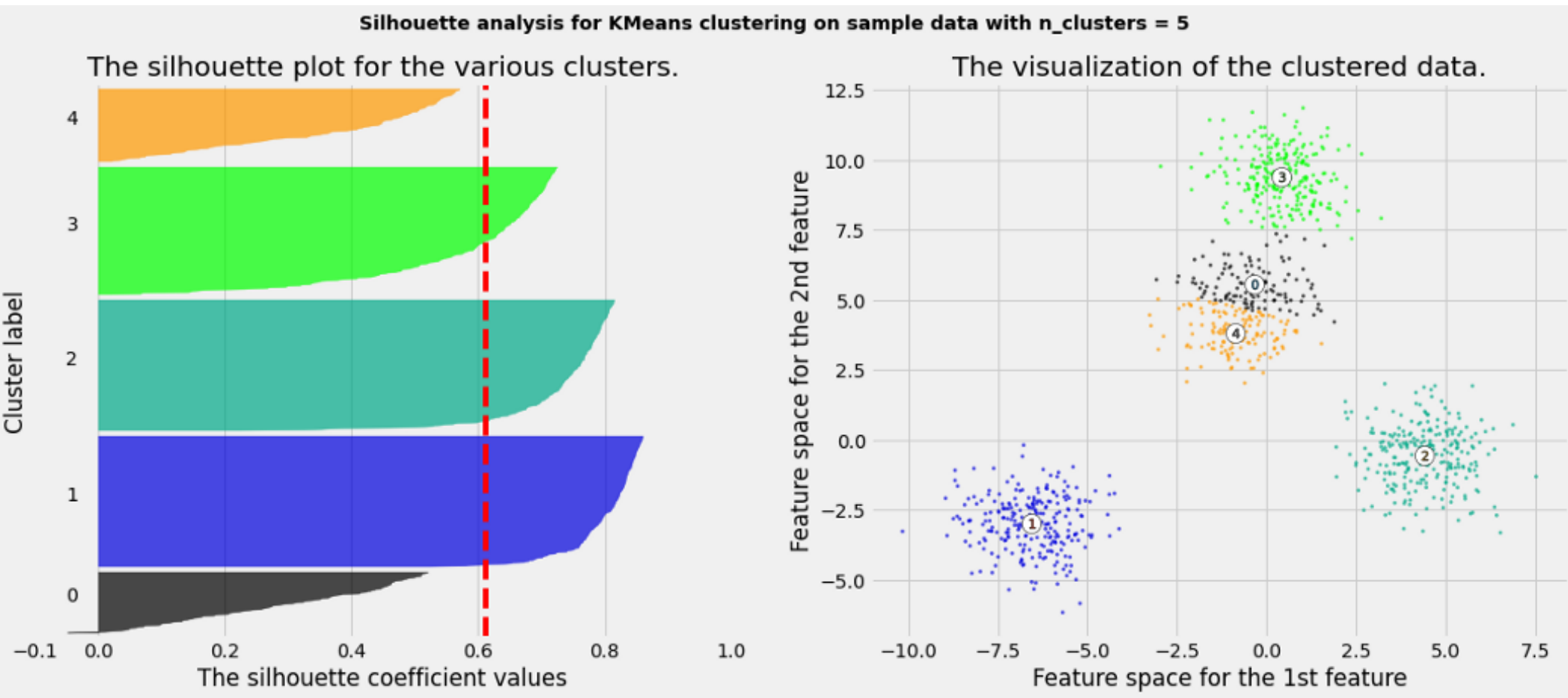
The silhouette plot for the various clusters.



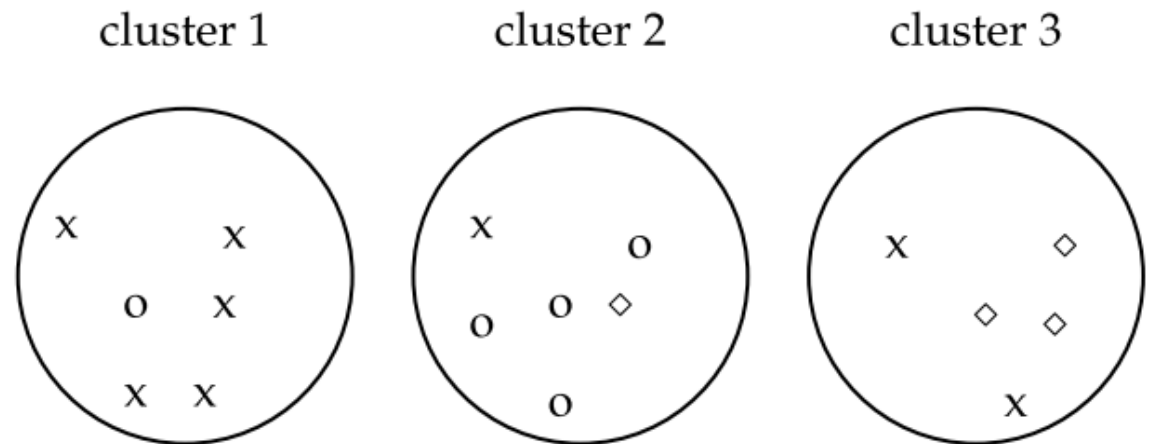
The visualization of the clustered data.



Determining the Number of Clusters (cont.)



Supervised Cluster Evaluation



	purity	NMI	RI	F_5
lower bound	0.0	0.0	0.0	0.0
maximum	1	1	1	1
value	0.71	0.36	0.68	0.46

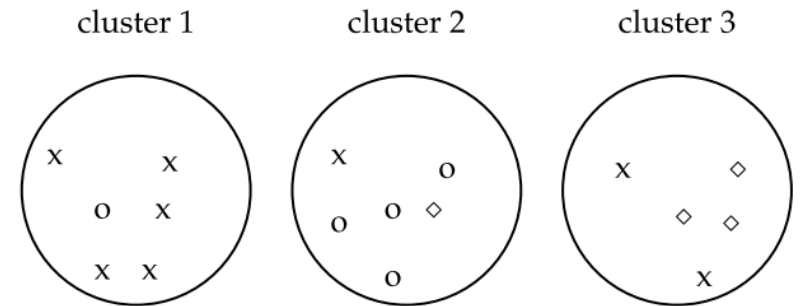
(ref: Introduction to Information Retrieval by Manning et al.)

Supervised Cluster Evaluation: Purity

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_K\}$$

$$\mathbb{C} = \{c_1, c_2, \dots, c_J\}$$



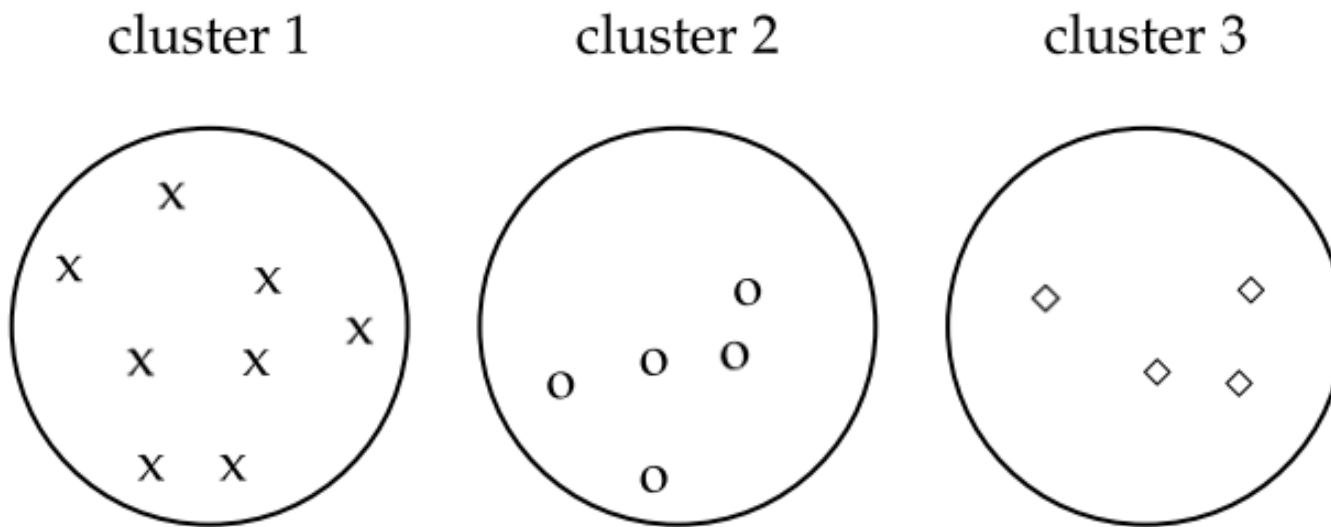
confusion
matrix.

$$\begin{aligned} \text{purity}(\Omega, \mathbb{C}) &= \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j| \\ &= \frac{1}{(6 + 6 + 5)} * (5 + 4 + 3) \\ &= 0.71 \end{aligned}$$

	1	2	3	
x	5	1	2	
o	1	4	0	
◇	0	1	3	
	6	6	5	

Supervised Cluster Evaluation: Purity (cont.)

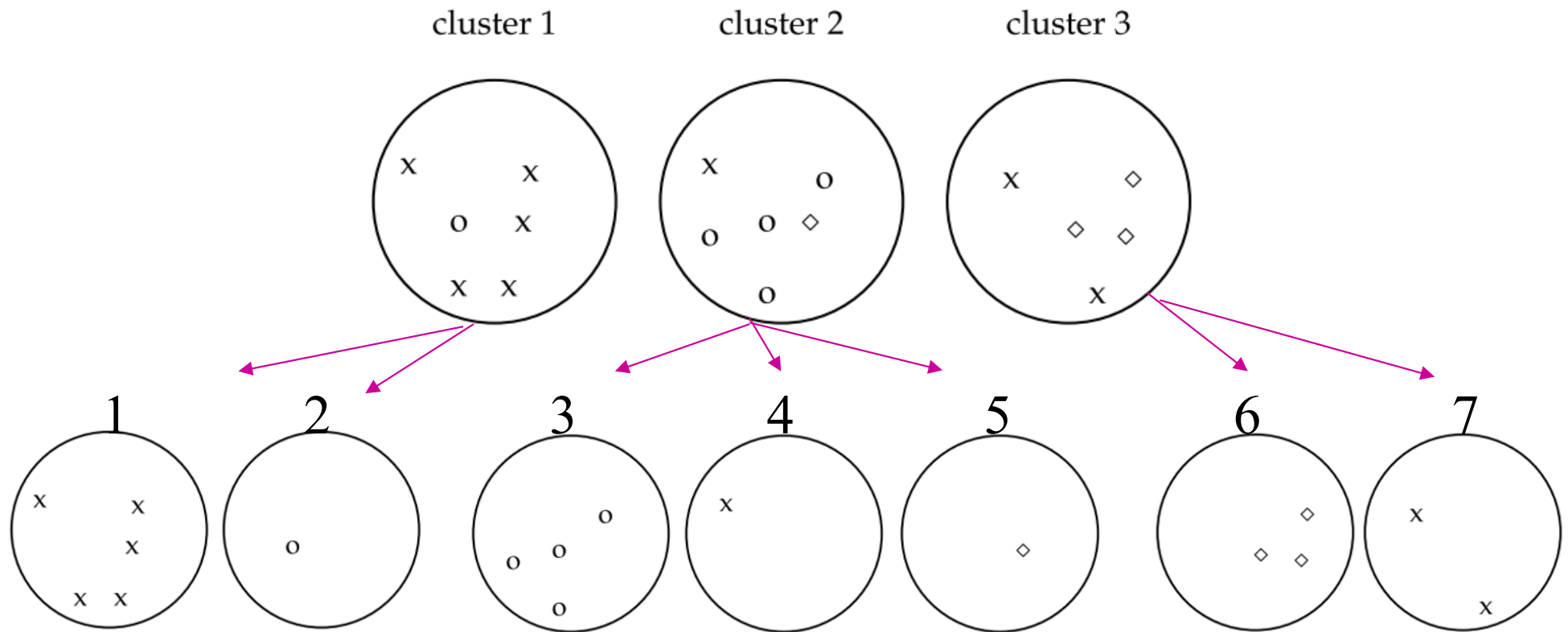
- Condition for Purity = 1



$$\begin{aligned}\text{purity}(\Omega, \mathbb{C}) &= \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j| \\ &= \frac{1}{(8 + 5 + 4)} * (8 + 5 + 4) \\ &= 1\end{aligned}$$

Supervised Cluster Evaluation: Purity (cont.)

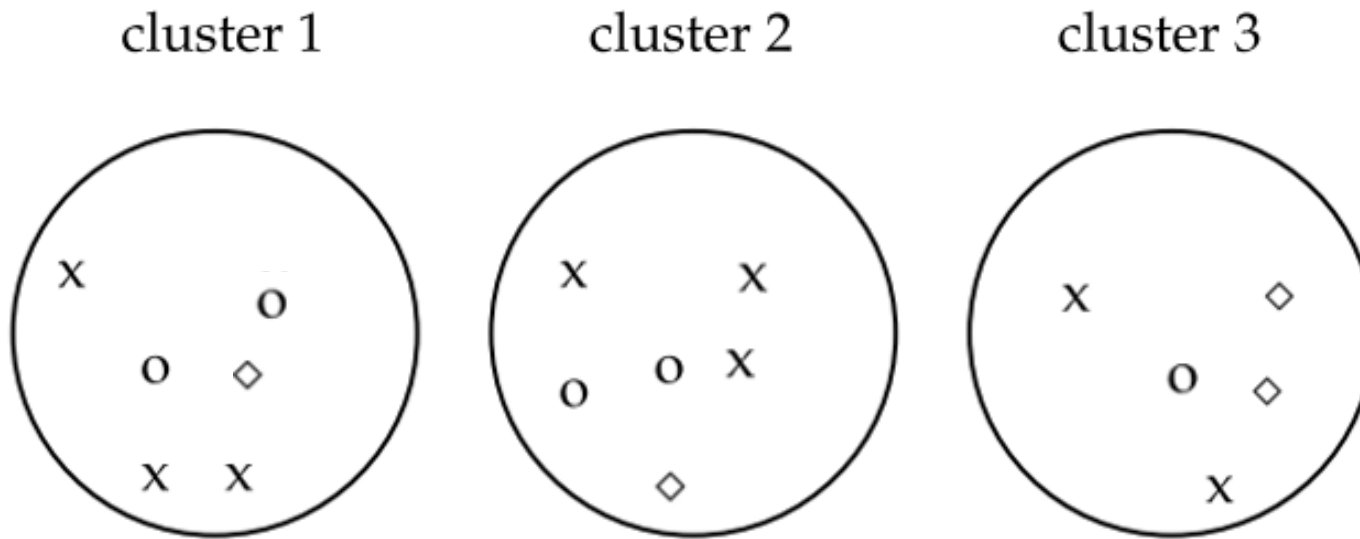
- Other condition for Purity = 1



$$\begin{aligned}
 \text{purity}(\Omega, \mathbf{C}) &= \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j| \\
 &= \frac{1}{17} * (5 + 1 + 4 + 1 + 1 + 3 + 2) = 1
 \end{aligned}$$

Supervised Cluster Evaluation: Purity (cont.)

- Purity for worse clustering



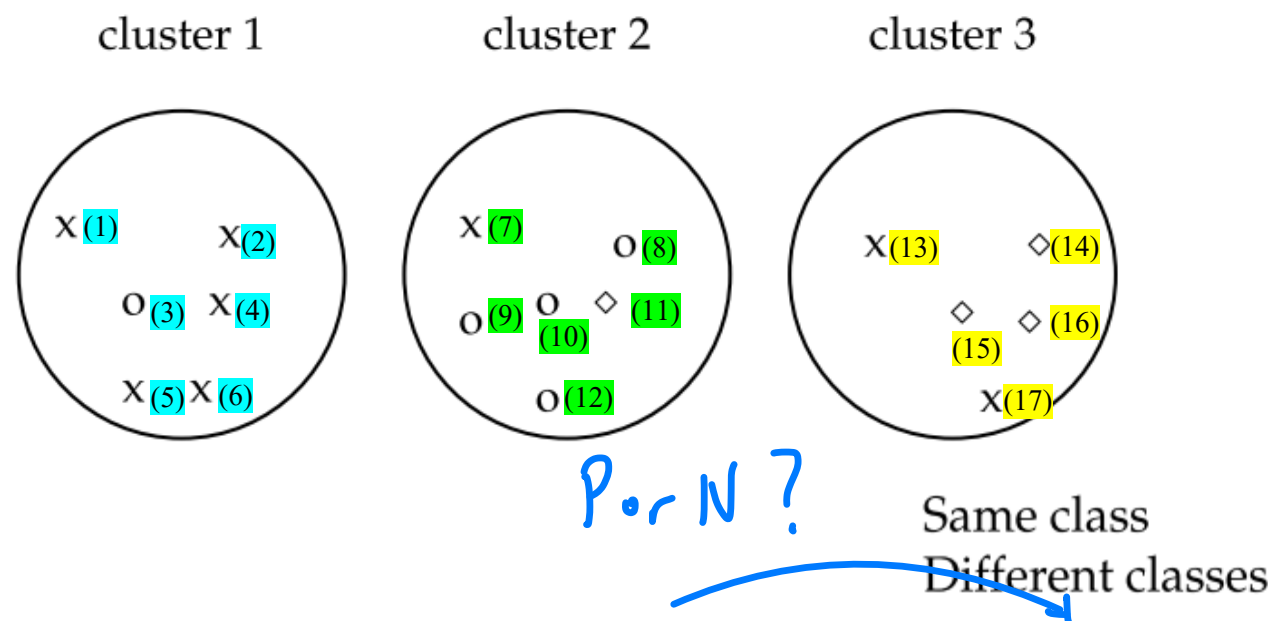
$$\begin{aligned}\text{purity}(\Omega, \mathbb{C}) &= \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j| \\ &= \frac{1}{17} * (3 + 3 + 2) = 0.47\end{aligned}$$

- **Purity:** Purity is a measure of the extent to which clusters contain a single class.^[38] Its calculation can be thought of as follows: For each cluster, count the number of data points from the most common class in said cluster. Now take the sum over all clusters and divide by the total number of data points. Formally, given some set of clusters M and some set of classes D , both partitioning N data points, purity can be defined as:

$$\frac{1}{N} \sum_{m \in M} \max_{d \in D} |m \cap d|$$

This measure doesn't penalize having many clusters, and more clusters will make it easier to produce a high purity. A purity score of 1 is always possible by putting each data point in its own cluster. Also, purity doesn't work well for imbalanced data, where even poorly performing clustering algorithms will give a high purity value. For example, if a size 1000 dataset consists of two classes, one containing 999 points and the other containing 1 point, then every possible partition will have a purity of at least 99.9%.

Supervised Cluster Evaluation: Rand Index



Same cluster	Different clusters
TP = 20	FN = 24
FP = 20	TN = 72

C(17,2)	Pair (x,y)		Same Cluster	Same Class	
1	1	2	T	T	TP
2	1	3	T	F	FP
3	1	7	F	T	FN
4	1	8	F	F	TN
5	1	11	F	F	TN
...					
136	16	17	T	F	FP

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

$$RI = \frac{20+72}{(20+20+ 24+72)} = 0.68$$

Predicted Class

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Conclusions

- Quality factors
 - Similarity measure and its implementation
 - Euclidean distance
 - City block distance
 - Cosine
 - ...
 - definition and representation of cluster chosen
 - Means
 - Medoids
 - Nearest, Fareast
 - clustering algorithm

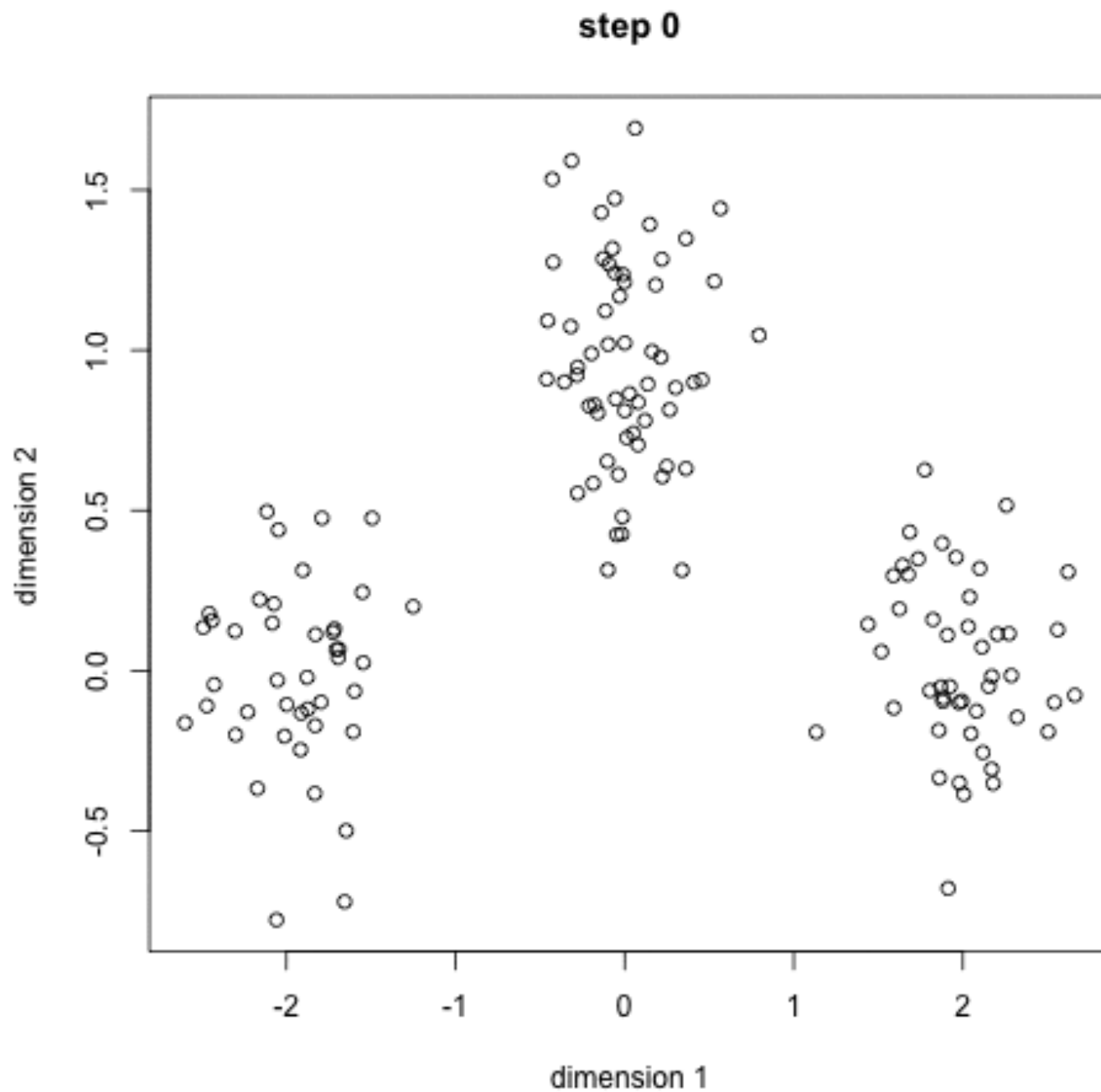
Conclusions (cont.)

- Quality factors
 - Similarity measure and its implementation
 - definition and representation of cluster chosen
 - clustering algorithm
 - K-means, K-medoids
 - Single-Link, Complete-Link, Average-Link, Chameleon
 - DBSCAN
 - EM

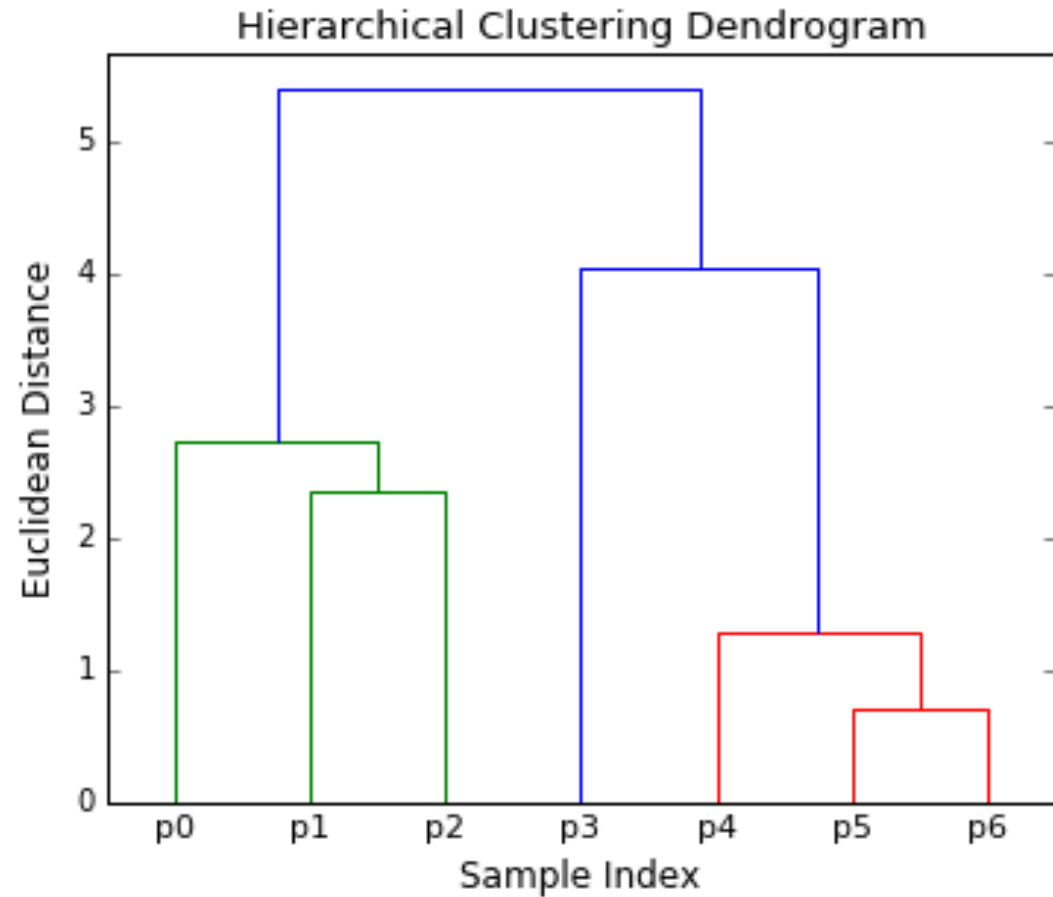
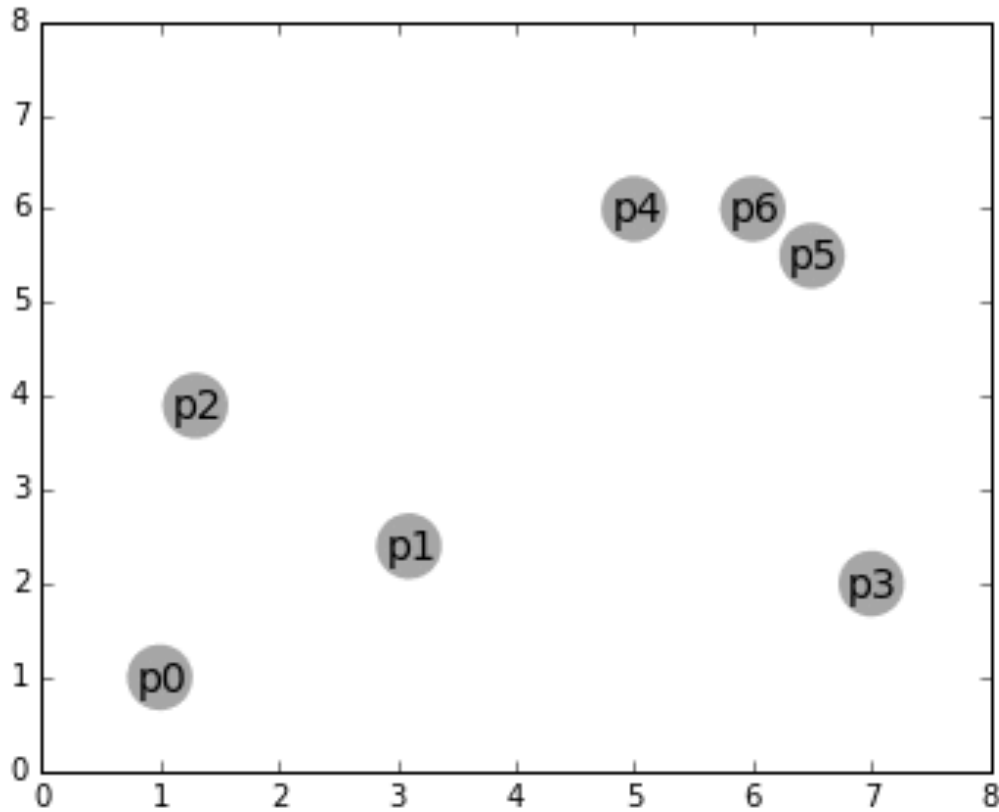
Conclusions (cont.)

- Which clustering algorithm ?
 - Type of clustering produced.
 - Biological Taxonomy: Hierarchical clustering
 - Characteristics of clusters: shape, sizes, densities
 - Characteristics of data set and attributes
 - K-means: data matrix
 - Noise & outlier
 - Number of objects (scalability)
 - Number of attributes
 - Cluster interpretation
 - Algorithmic considerations: ordered, parameters

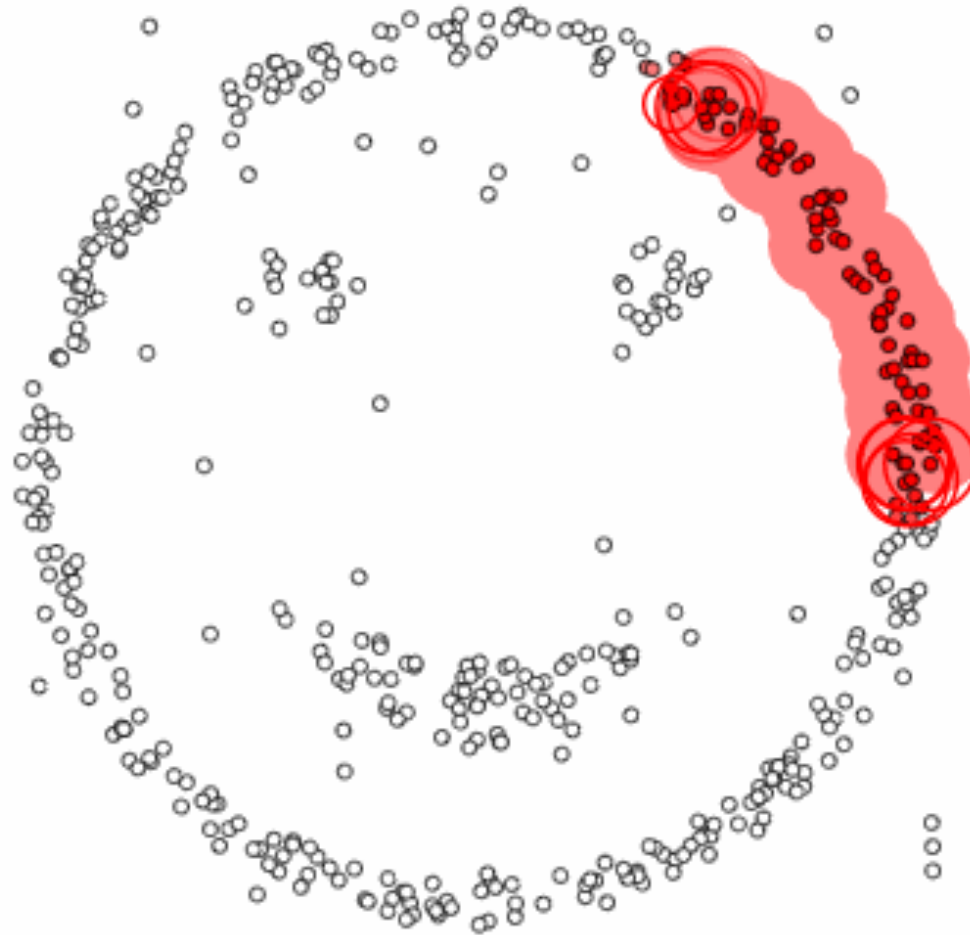
K-means



Hierarchical Clustering



DBSCAN



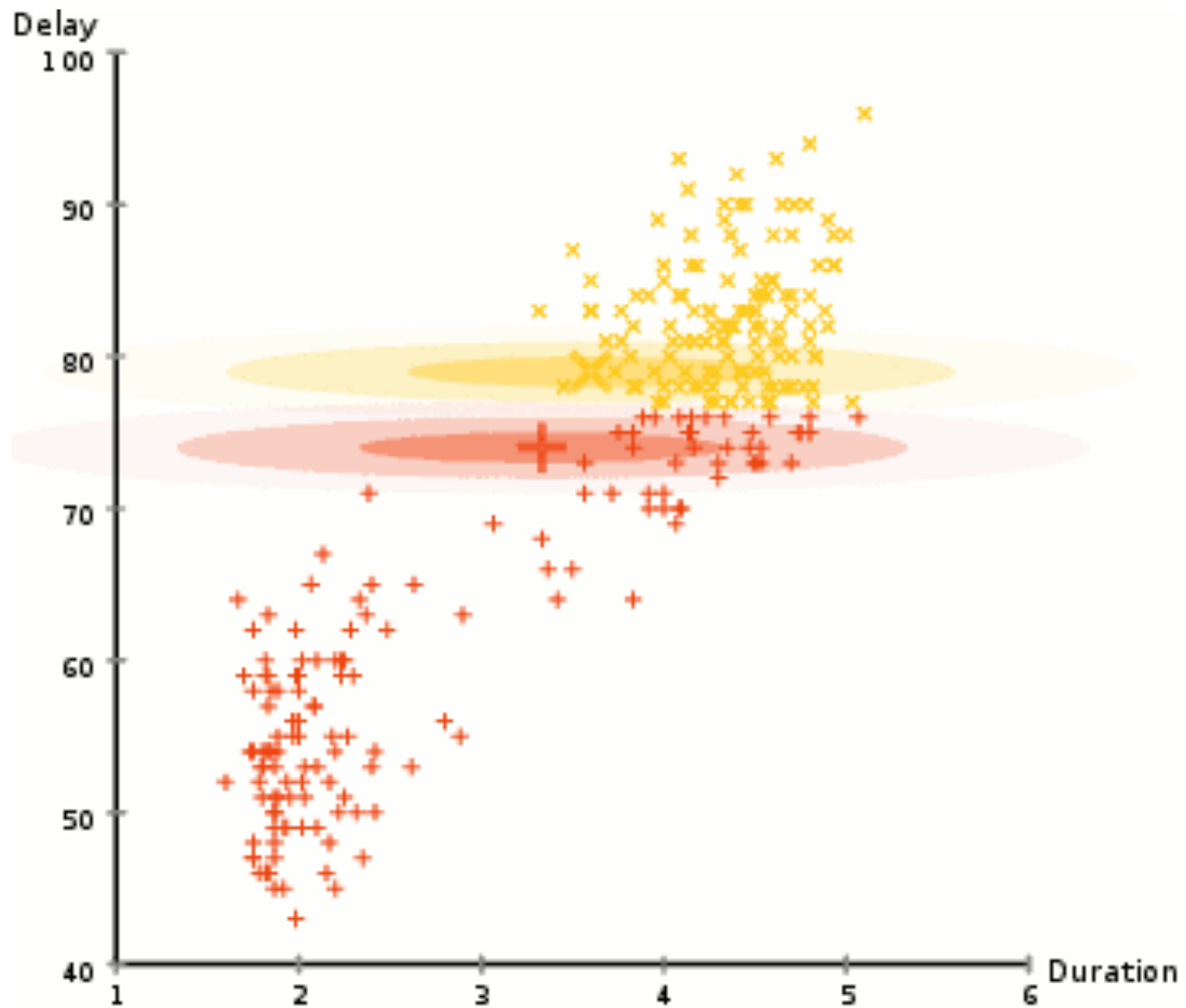
epsilon = 1.00
minPoints = 4

Restart



Pause

GMM



MiniBatchKMeans AffinityPropagation MeanShift SpectralClustering Ward AgglomerativeClustering DBSCAN OPTICS Birch GaussianMixture

