Machine Learning

Dr. Mehran Safayani

safayani@iut.ac.ir

safayani.iut.ac.ir

https://www.aparat.com/mehran.safayani

https://github.com/safayani/machine_learning_course

Department of Electrical and computer engineering, Isfahan university of technology, Isfahan, Iran

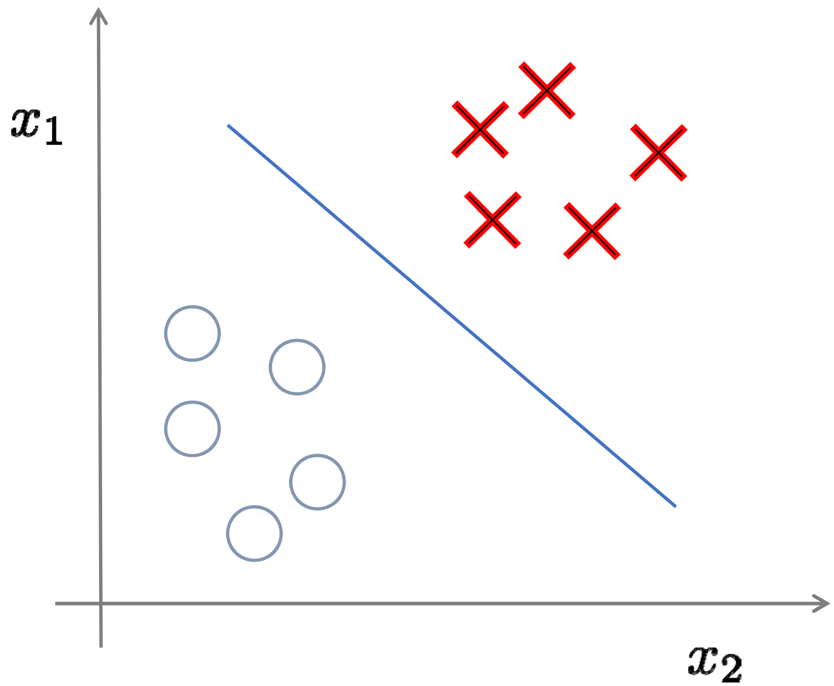
Machine Learning

Clustering (K-means)

Mehran Safayani

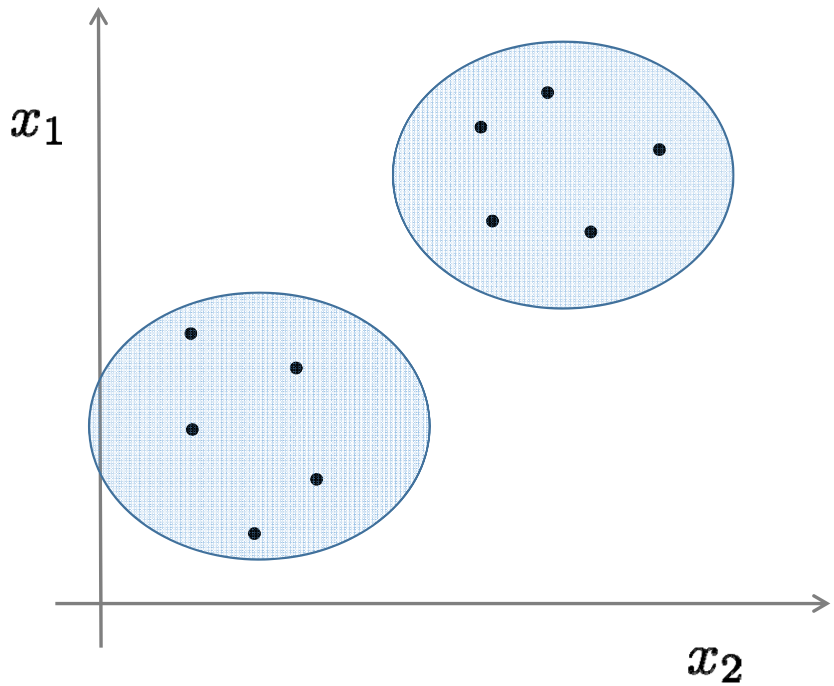
Slides adapted from Andrew NG, David Sontag

Supervised learning



Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

Unsupervised learning



Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$

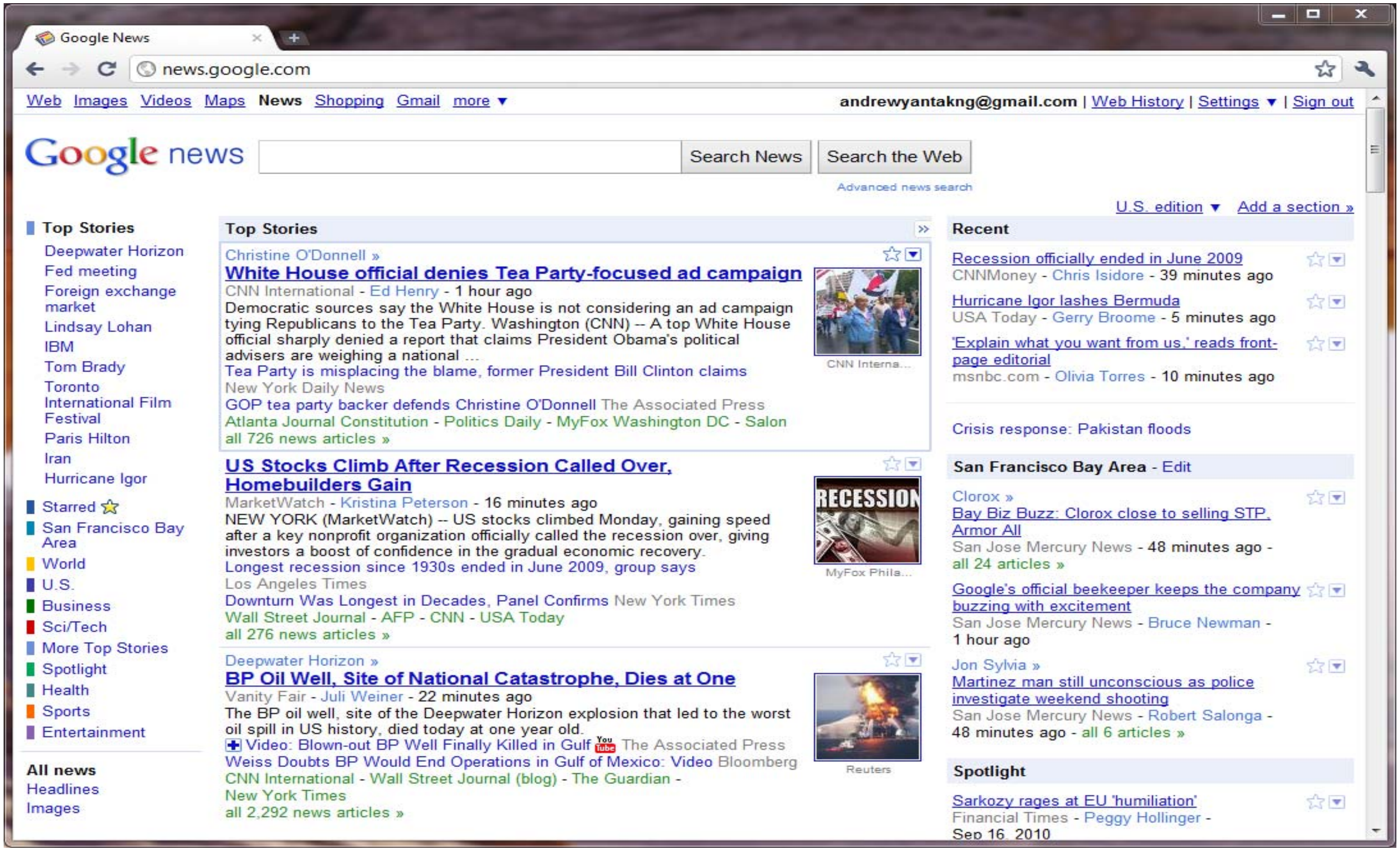
clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

Clustering Example (Image Segmentation)

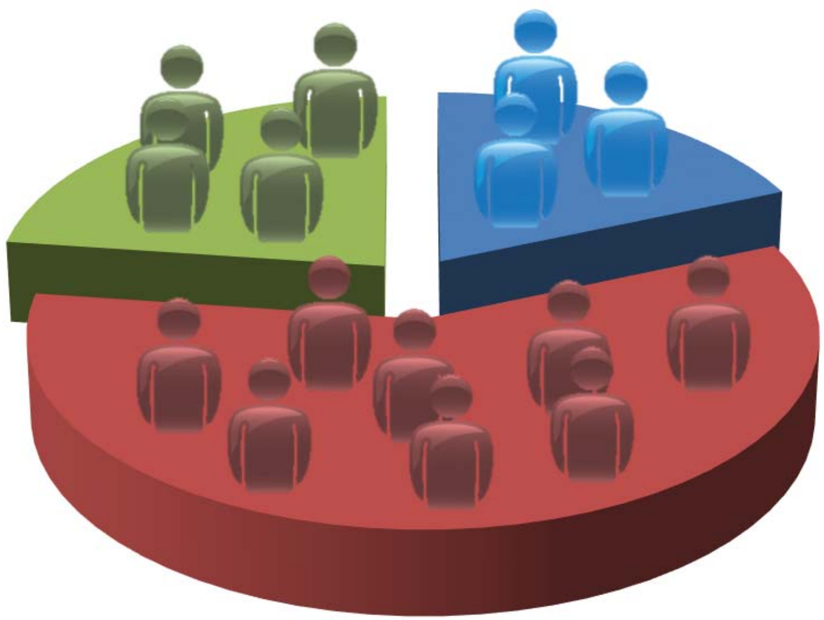
[Slide from James Hayes]

$x_{ij} = \begin{bmatrix} R \\ G \\ B \end{bmatrix}$

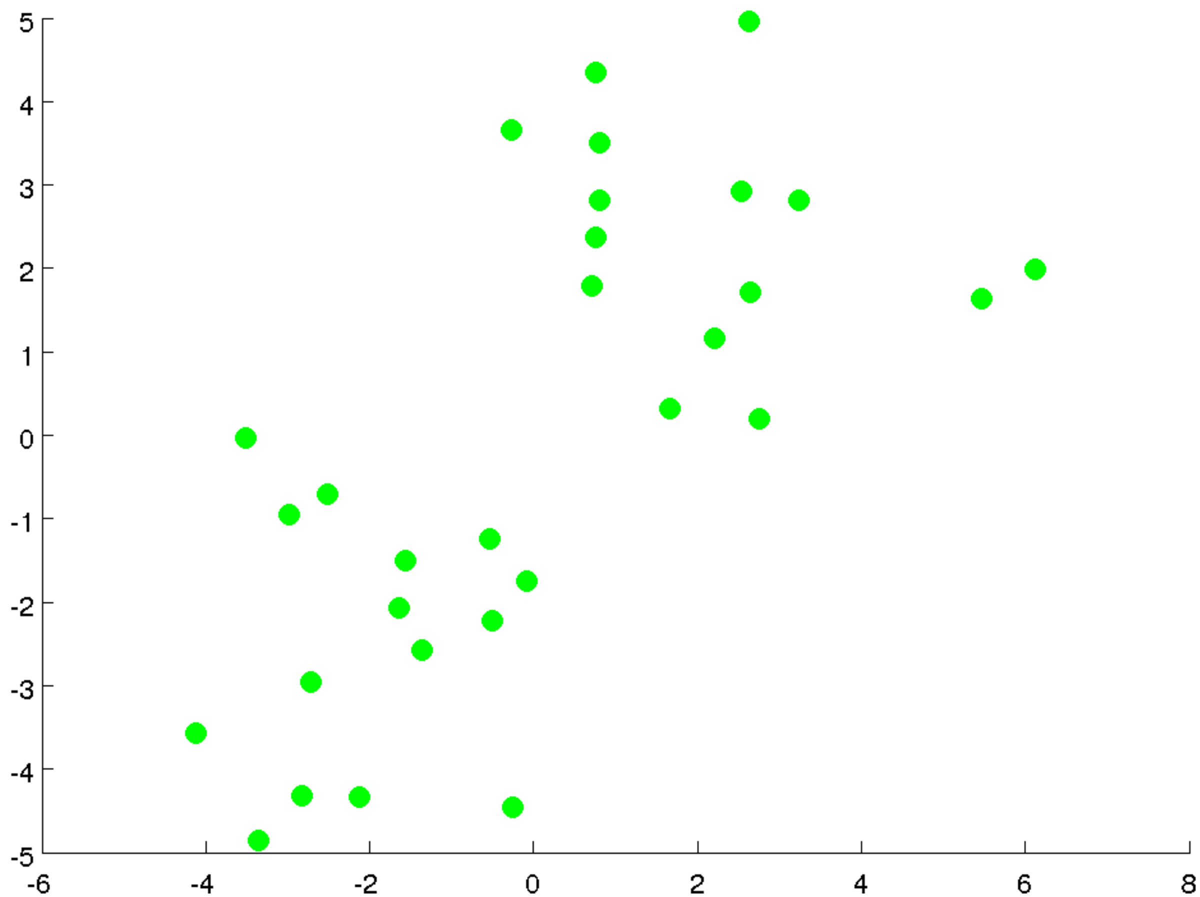
Clustering Example (News clustering)

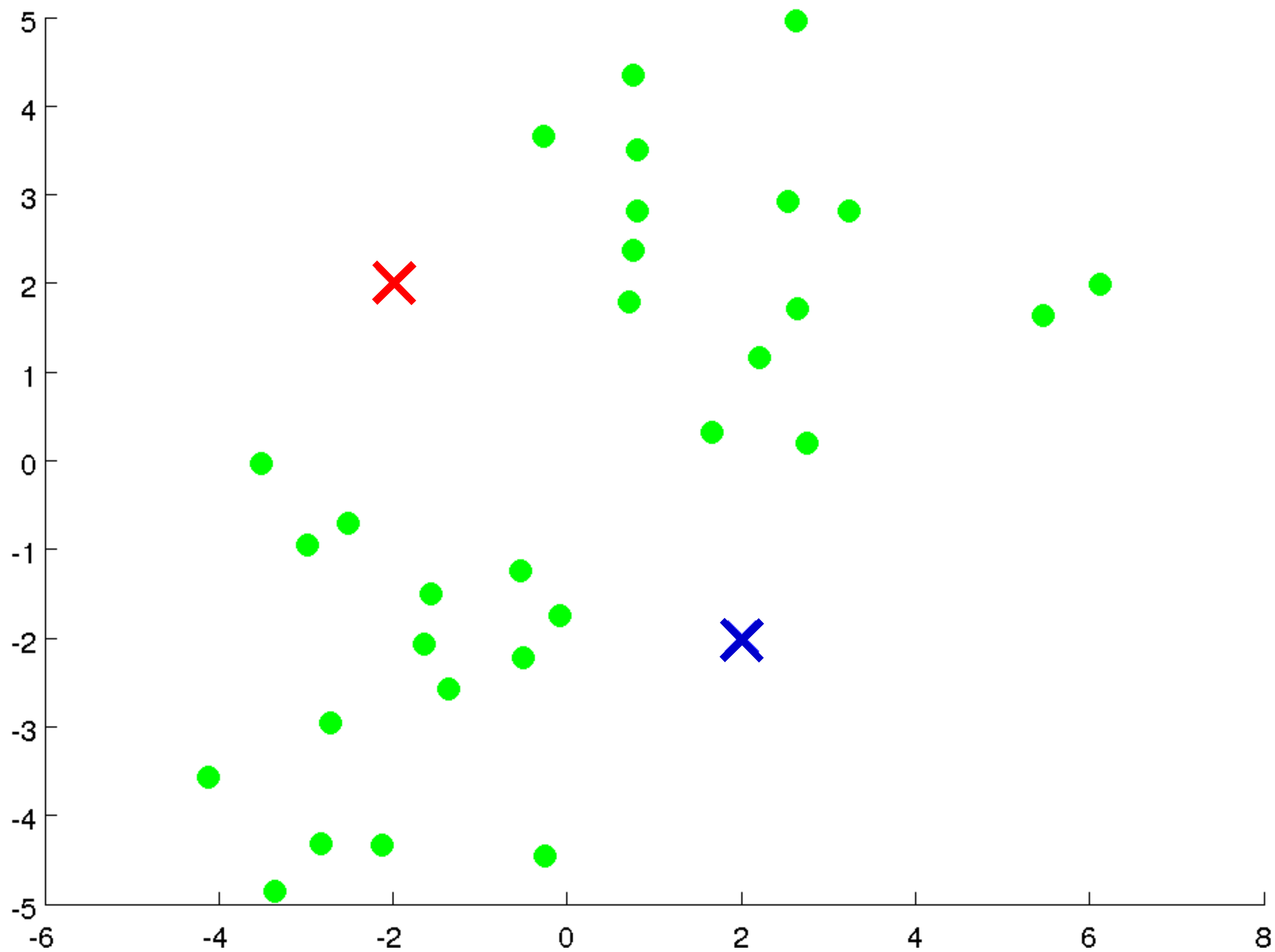


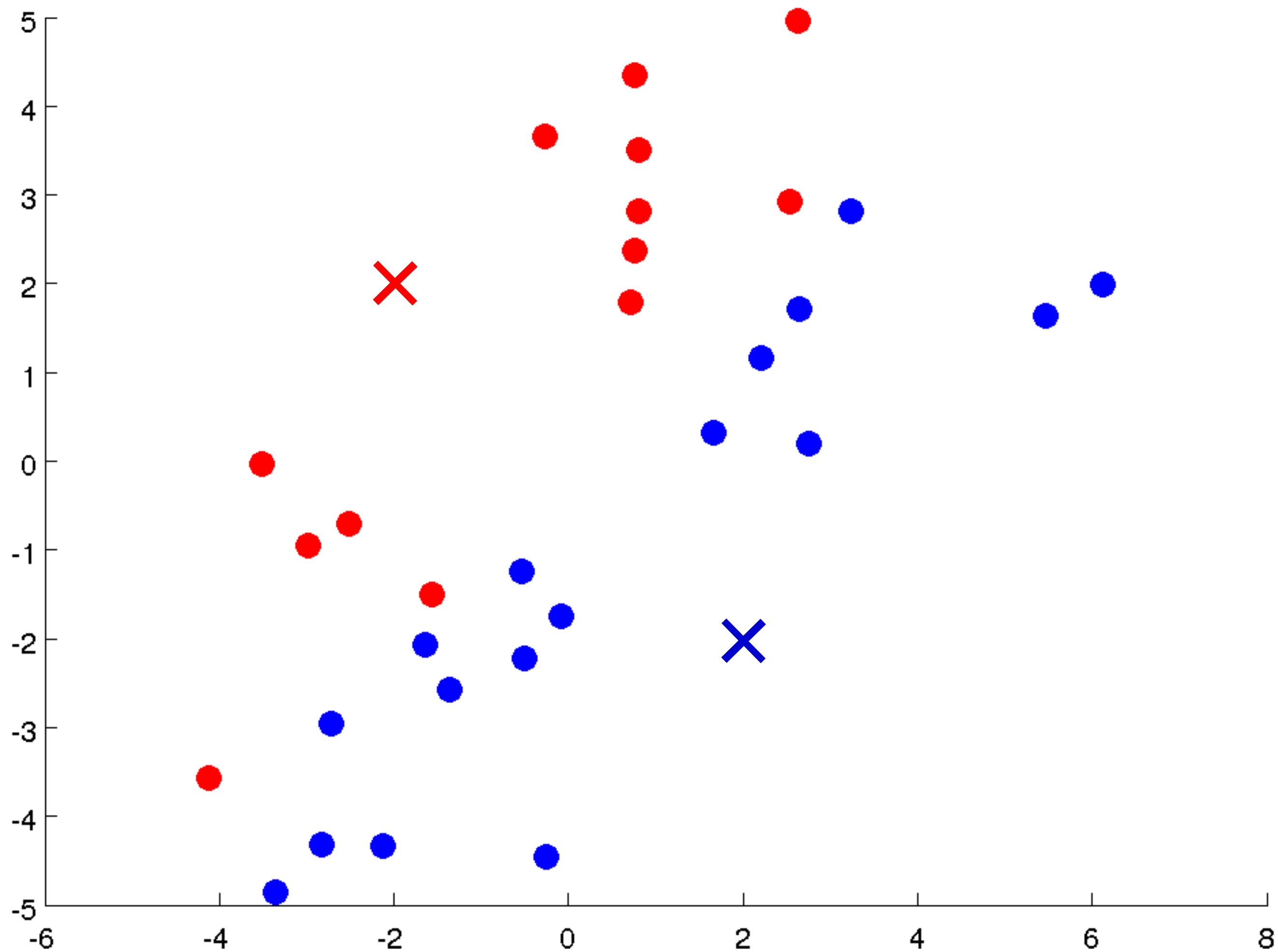
Clustering Example (Market Segmentation)

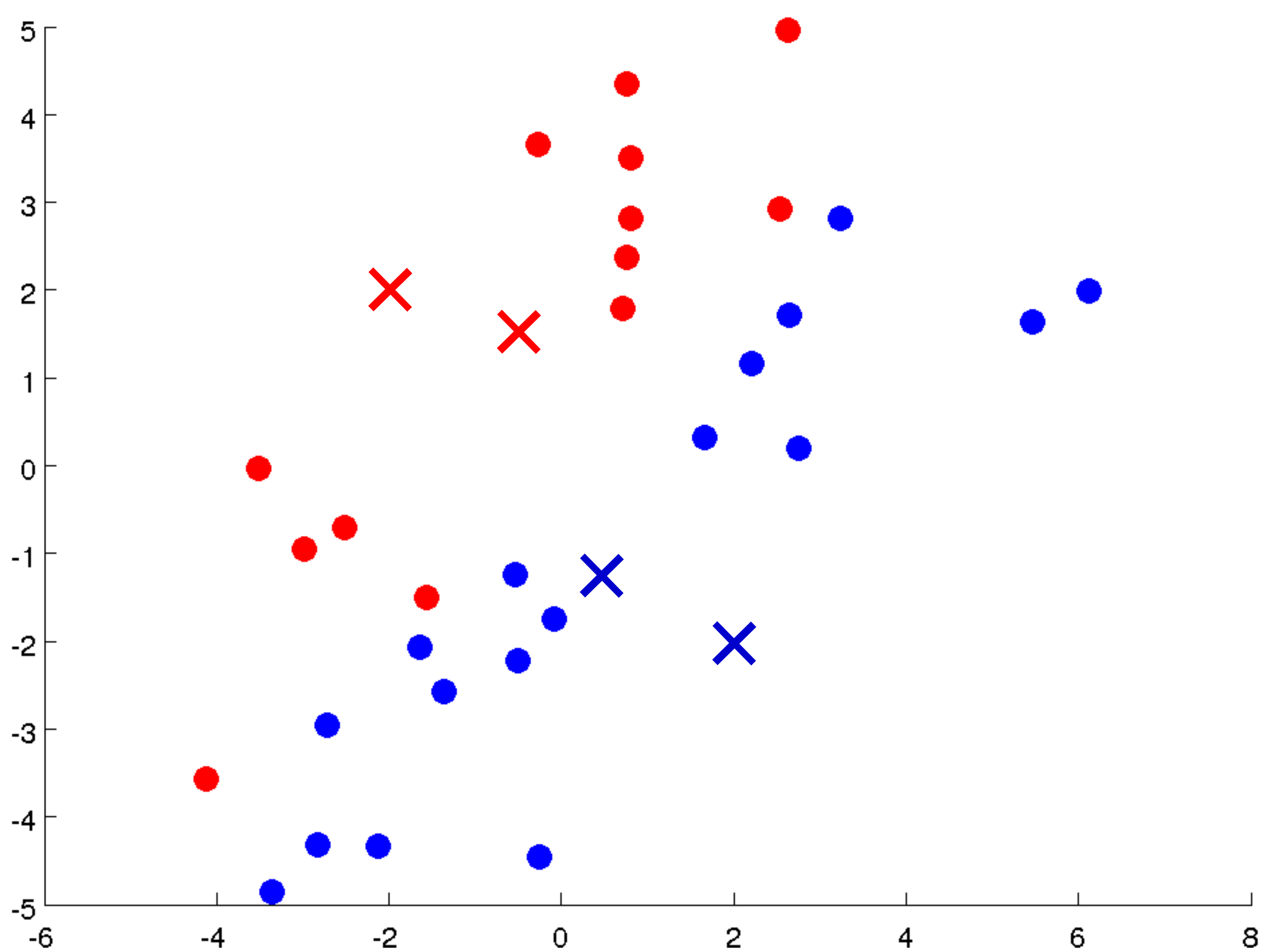


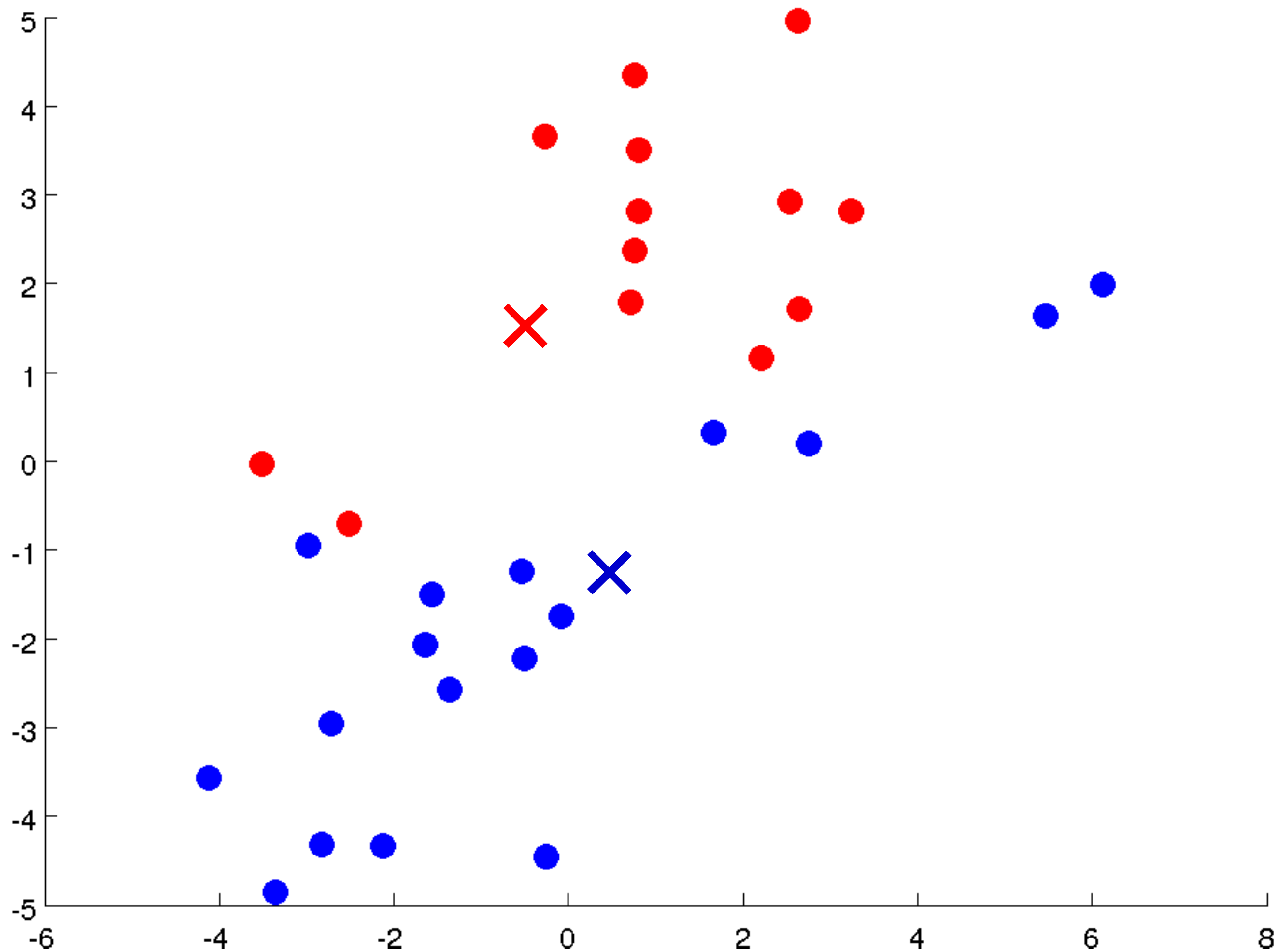
K-means Clustering Algorithm



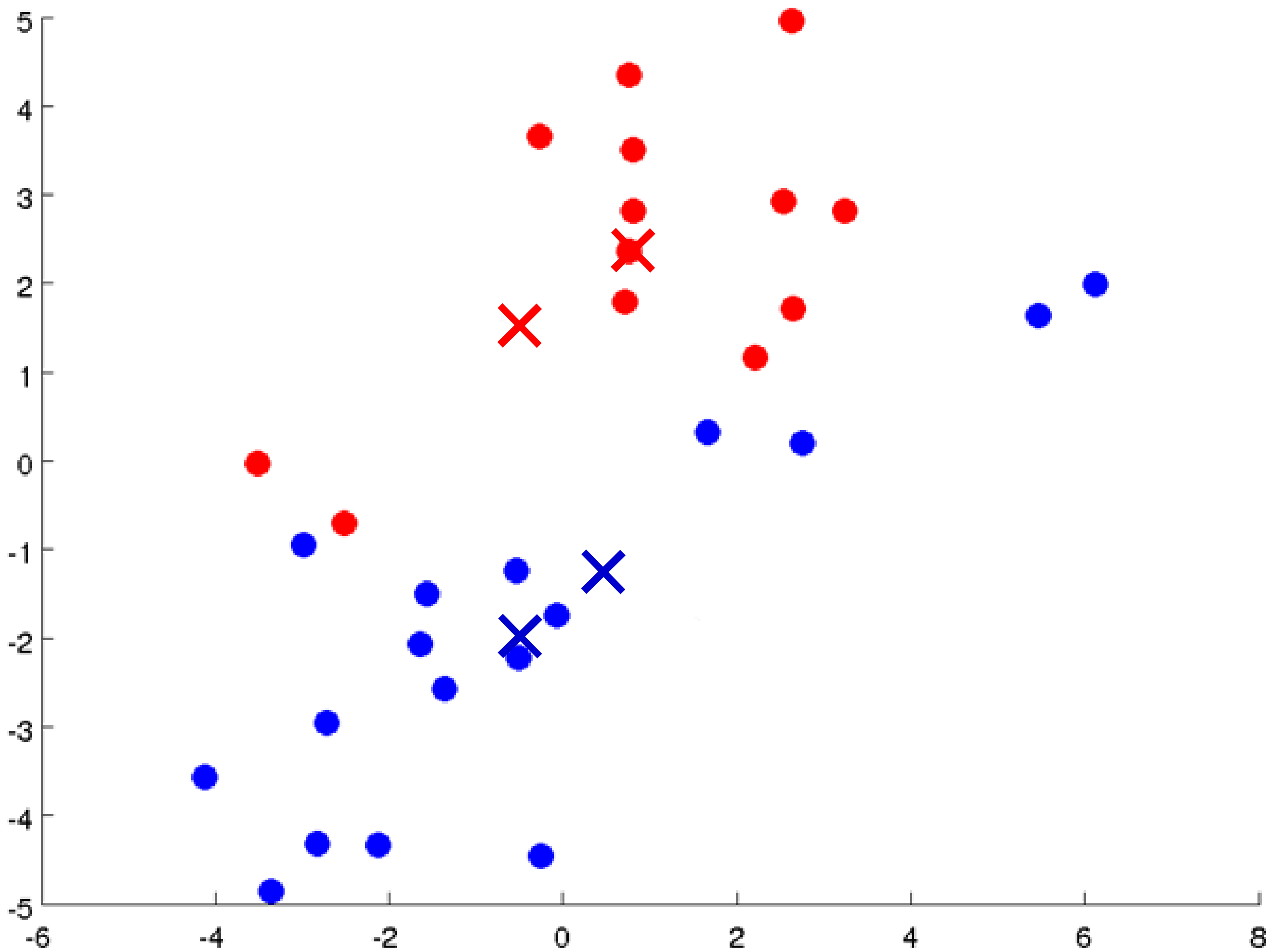

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

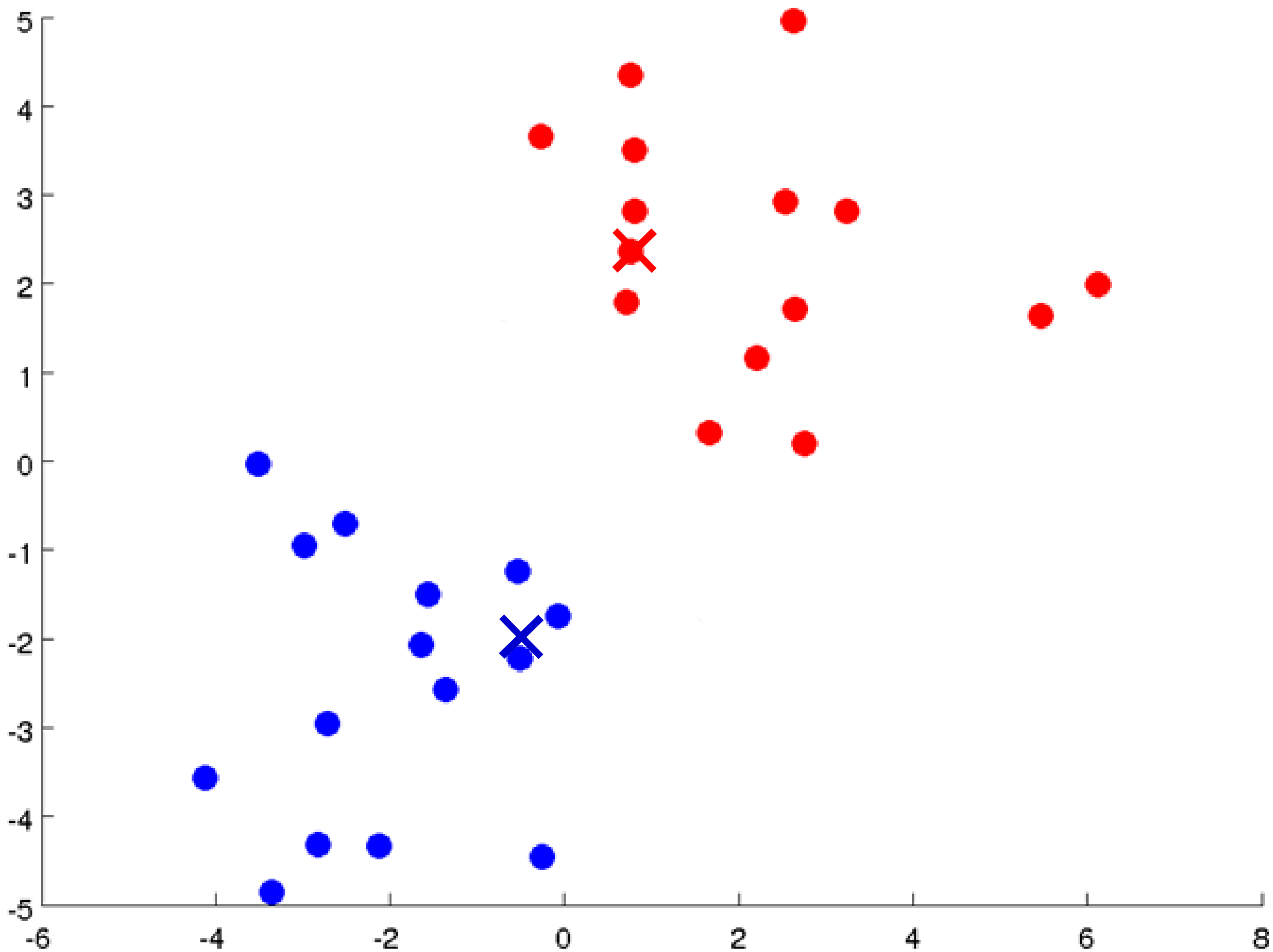

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

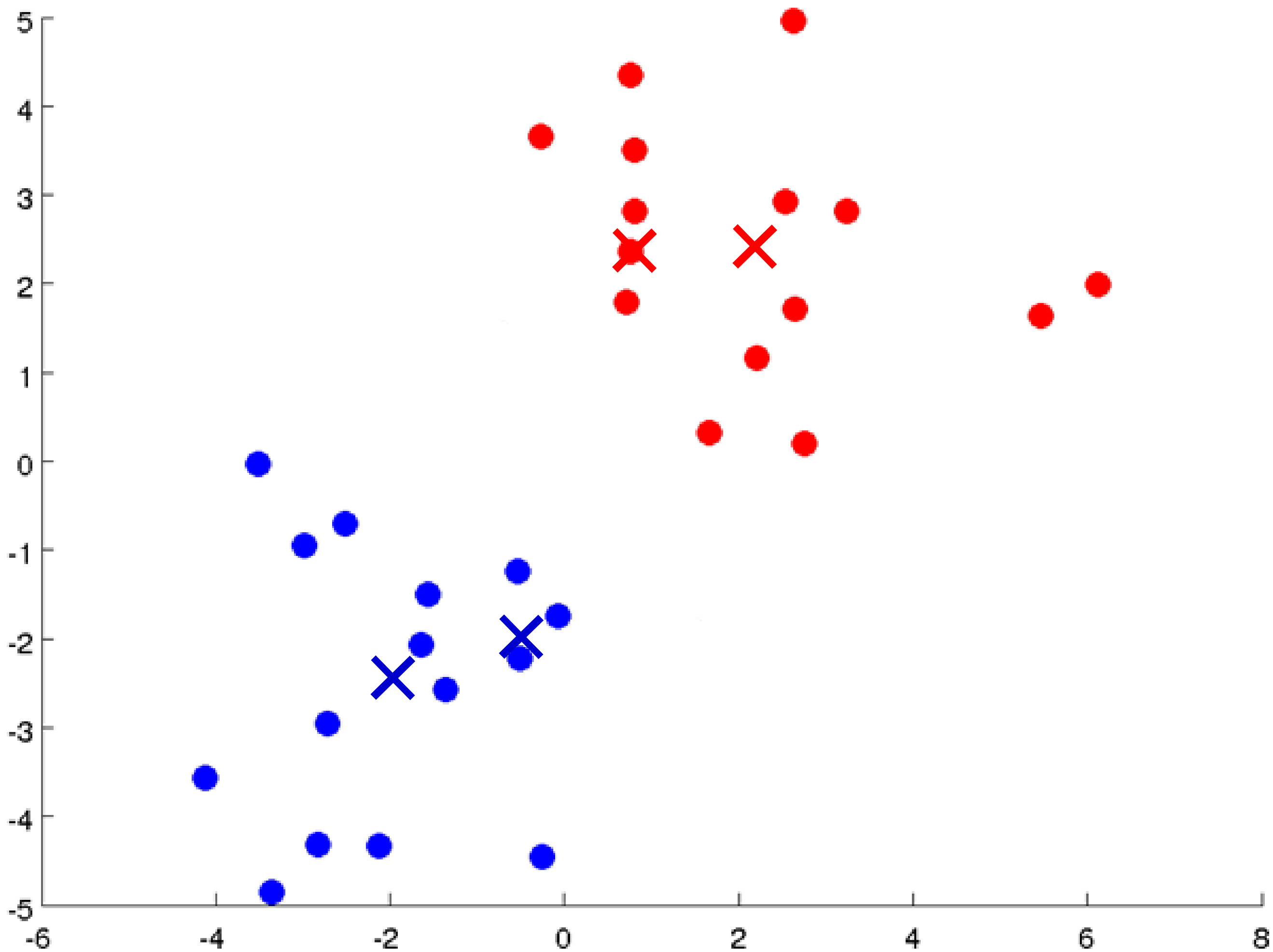


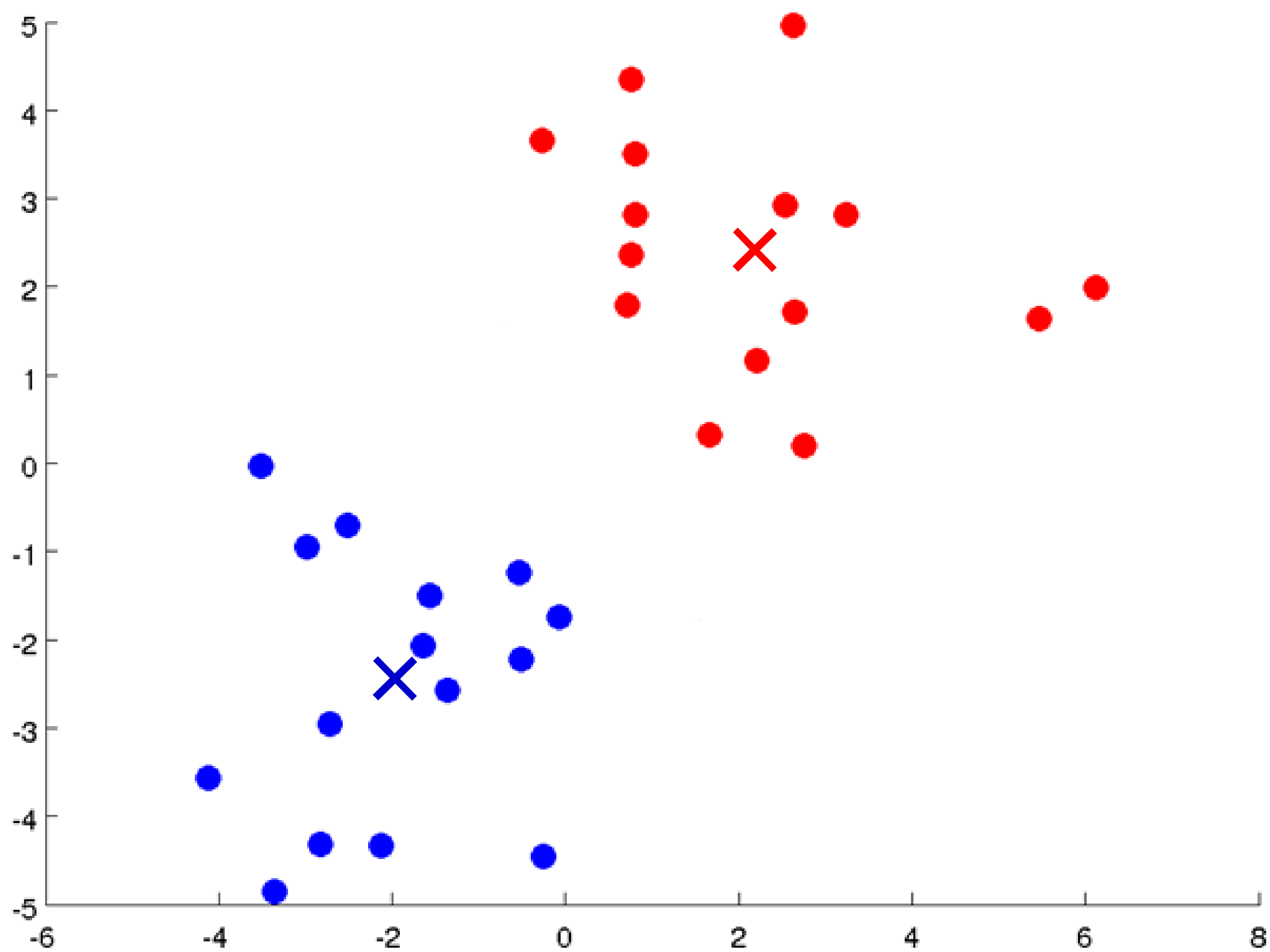
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8



5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

K-means algorithm

Input:

- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$x^{(i)} \in \mathbb{R}^n$ (drop $x_0 = 1$ convention)

K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat {

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

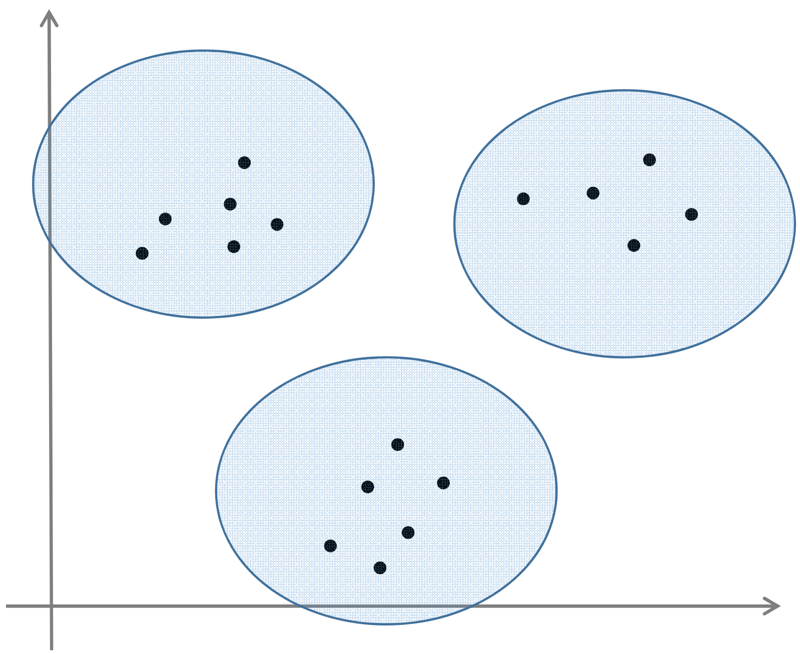
$c^{(i)} = \arg\min_j \|x^i - \mu_j\|^2$

for k = 1 to K

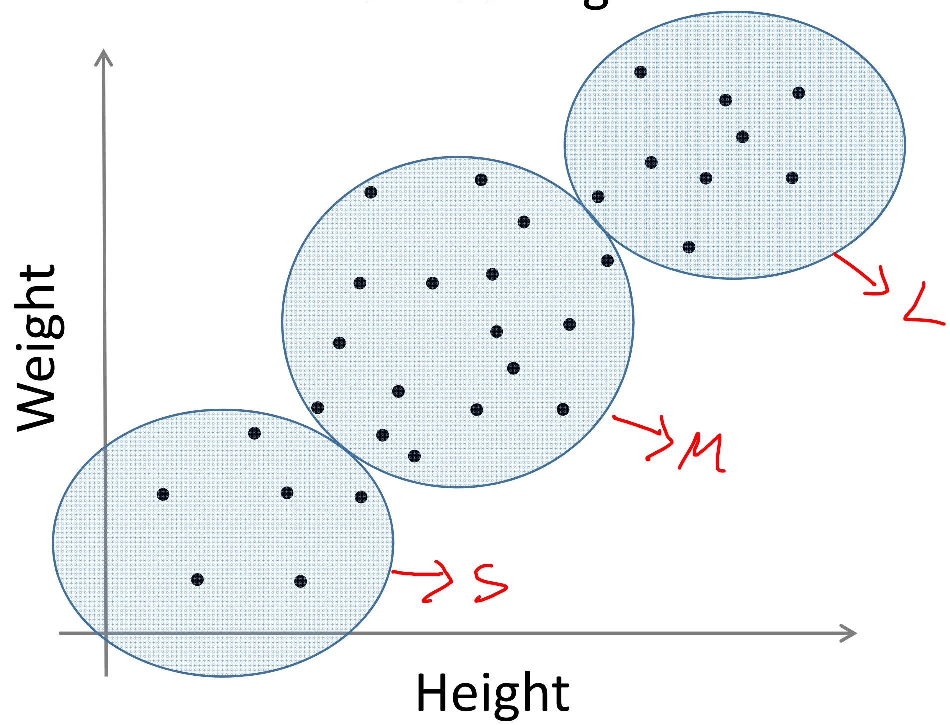
μ_k := average (mean) of points assigned to cluster k

}

K-means for non-separated clusters



T-shirt sizing



K-means optimization objective

$c^{(i)}$ = index of cluster (1,2,…,K) to which example $x^{(i)}$ is currently assigned

μ_k = cluster centroid k ($\mu_k \in \mathbb{R}^n$)

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

Optimization objective:

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

$$\min_{\substack{c^{(1)}, \ldots, c^{(m)}, \\ \mu_1, \ldots, \mu_K}} J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat {

for i = 1 to m

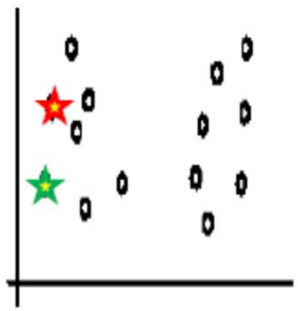
$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

for k = 1 to K

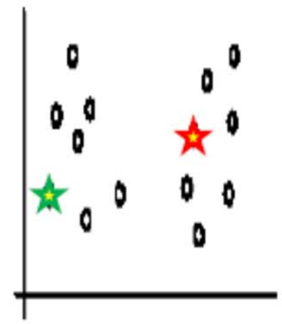
μ_k := average (mean) of points assigned to cluster k

}

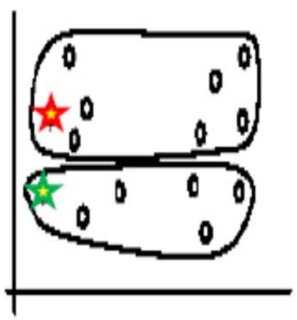
Sensitivity to initial seeds



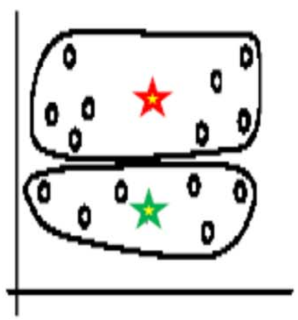
Random selection of seeds (centroids)



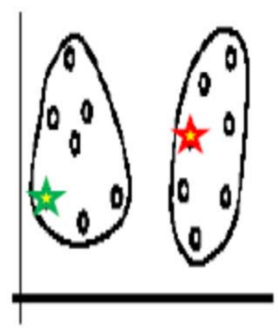
Random selection of seeds (centroids)



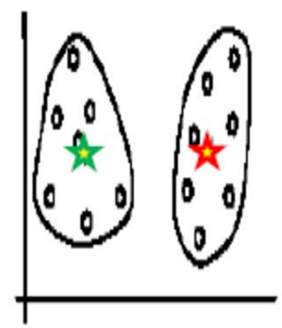
Iteration 1



Iteration 2



Iteration 1



Iteration 2

Random initialization

For i = 1 to 100 {

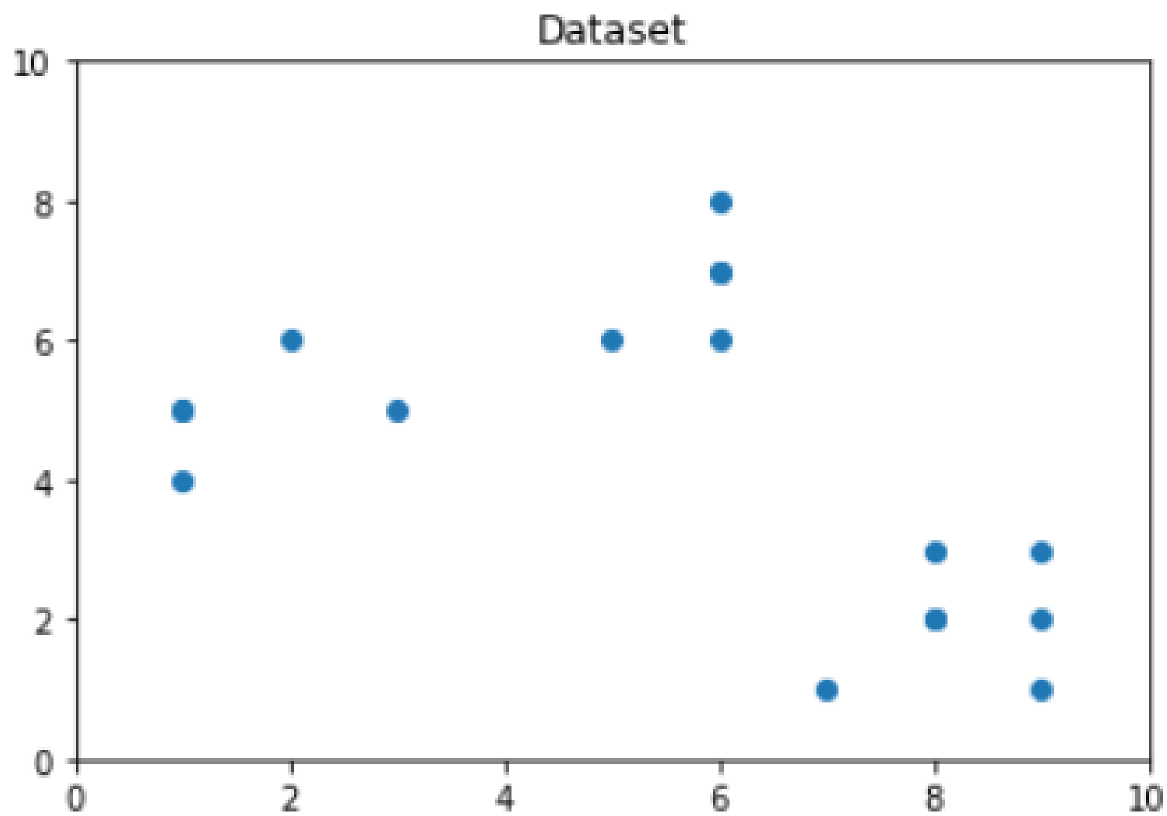
Randomly initialize K-means.
Run K-means. Get $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$.
Compute cost function (distortion)

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

}

Pick clustering that gave lowest cost $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$

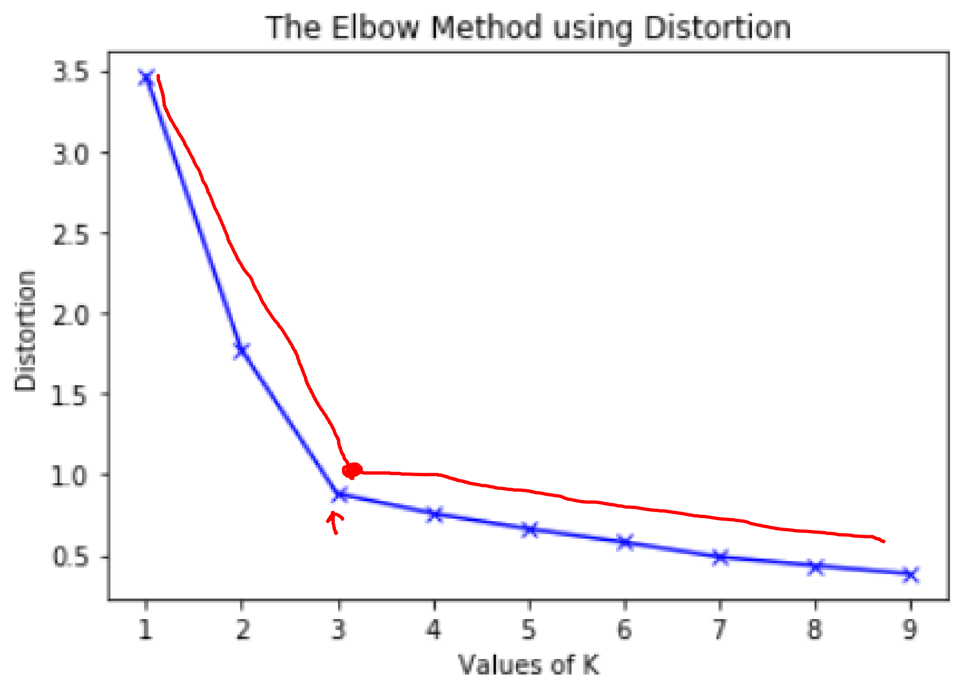
What is the right value of K?

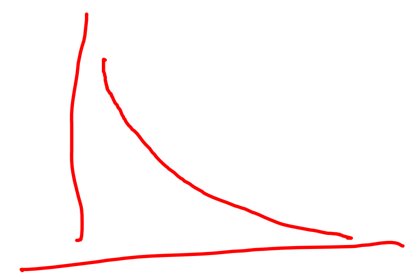


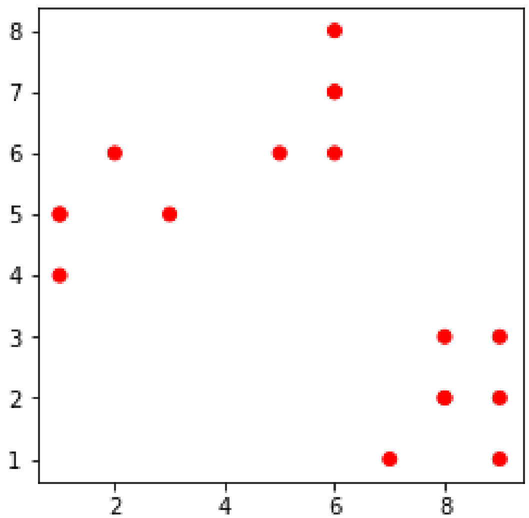
Choosing the value of K

Elbow method:

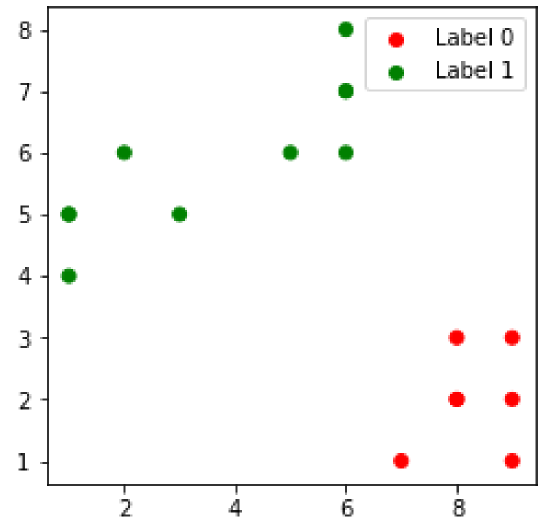
Distortion: It is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.



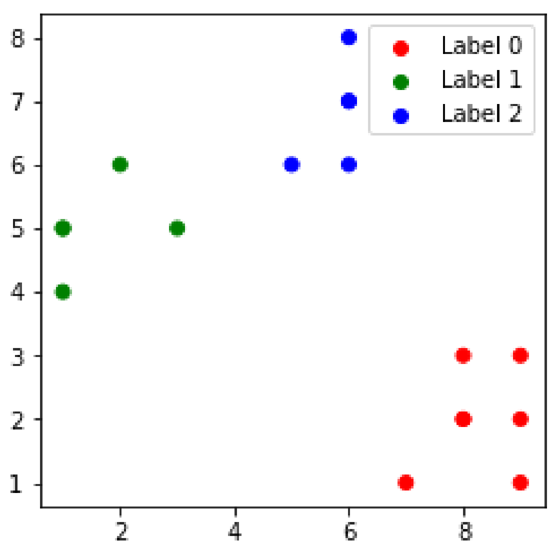




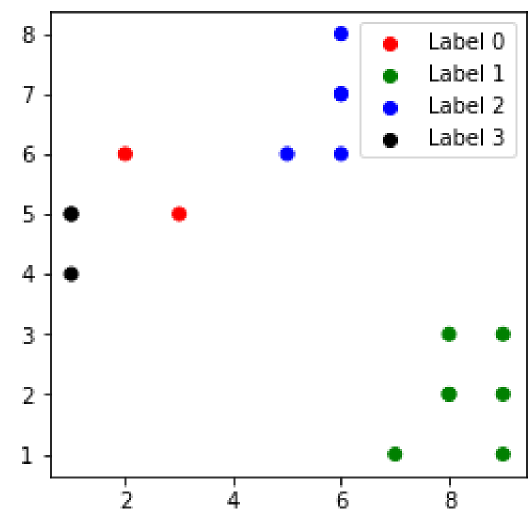
K=1


Label 0
Label 1

K = 2


Label 0
Label 1
Label 2

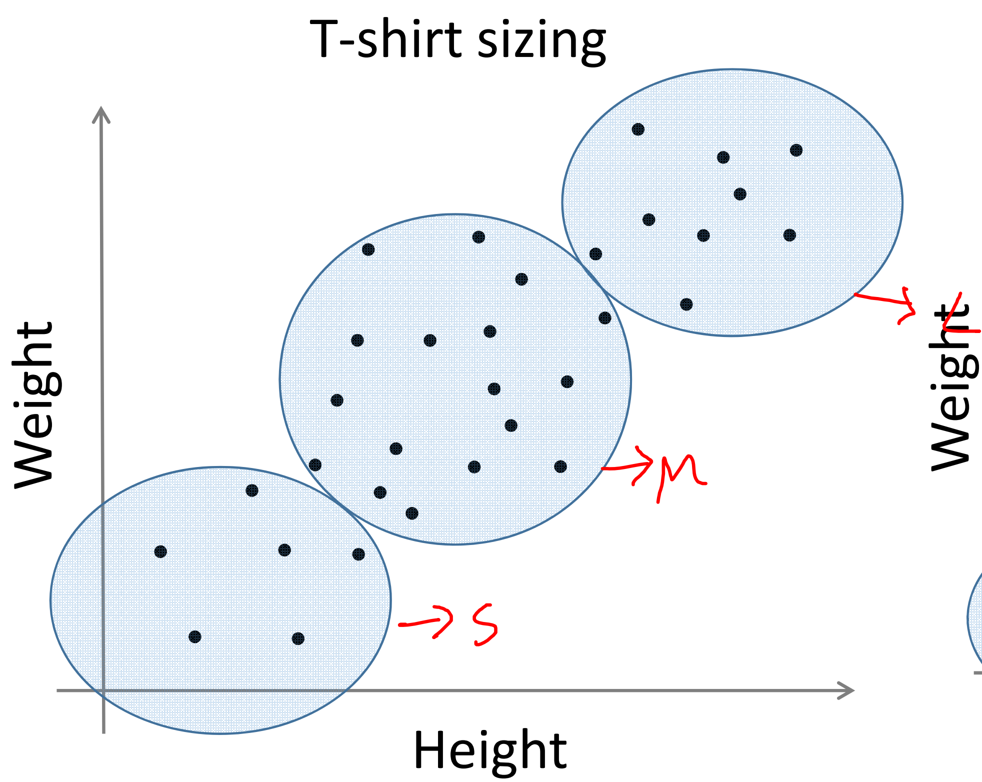
K = 3

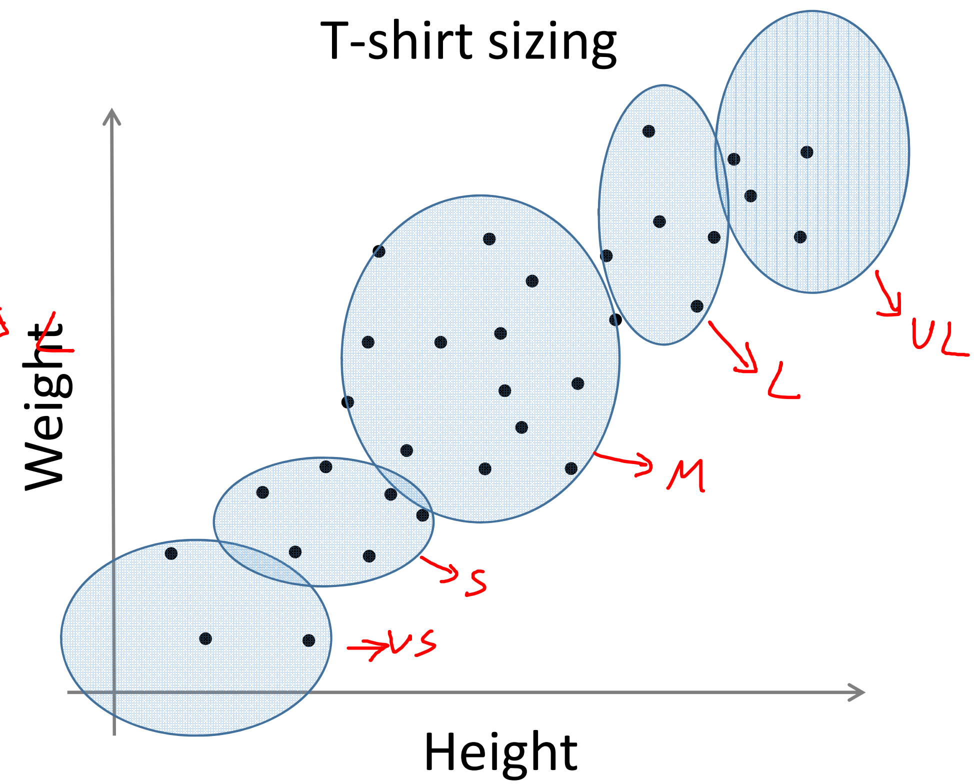

Label 0
Label 1
Label 2
Label 3

K = 4

K-means for non-separated clusters

Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.





Clustering Evaluation metrics

- **Silhouette Coefficient**

$$s = \frac{b - a}{max(a, b)}$$

The Silhouette Coefficient is defined for each sample and is composed of two scores:

a: The mean distance between a sample and all other points in the same cluster.

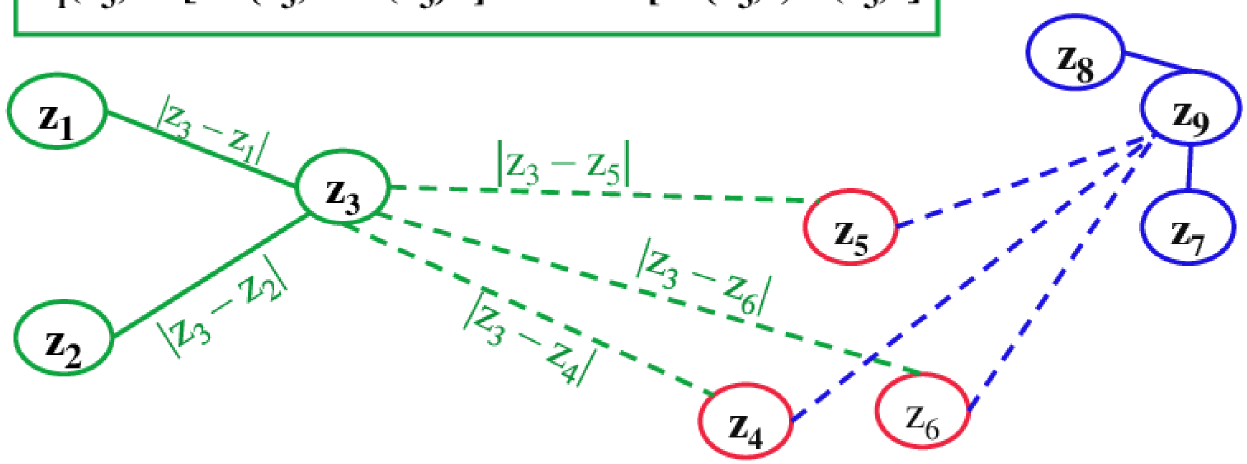
b: The mean distance between a sample and all other points in the next nearest cluster.

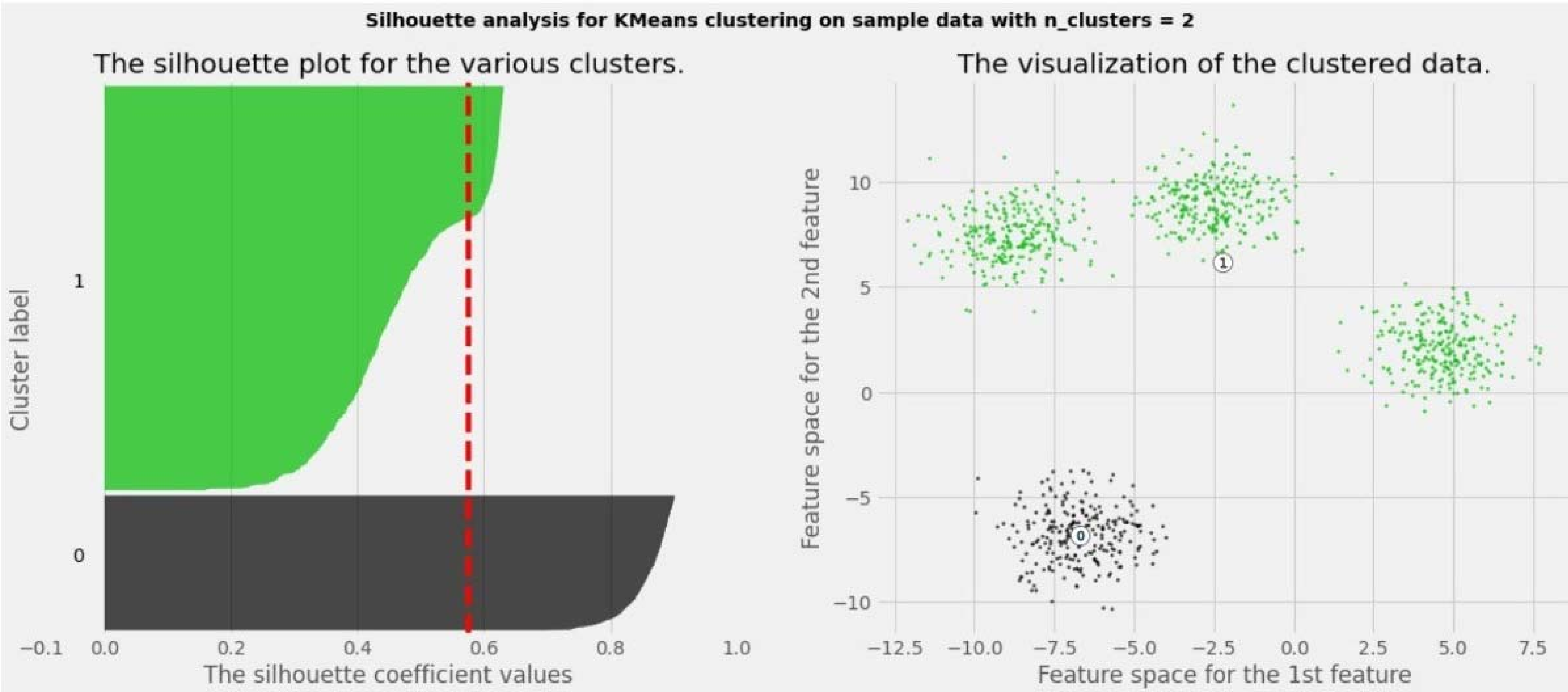
The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample. The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters. The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

$$d(z_3) = [\,|z_3 - z_1| + |z_3 - z_2|\,] / 2$$

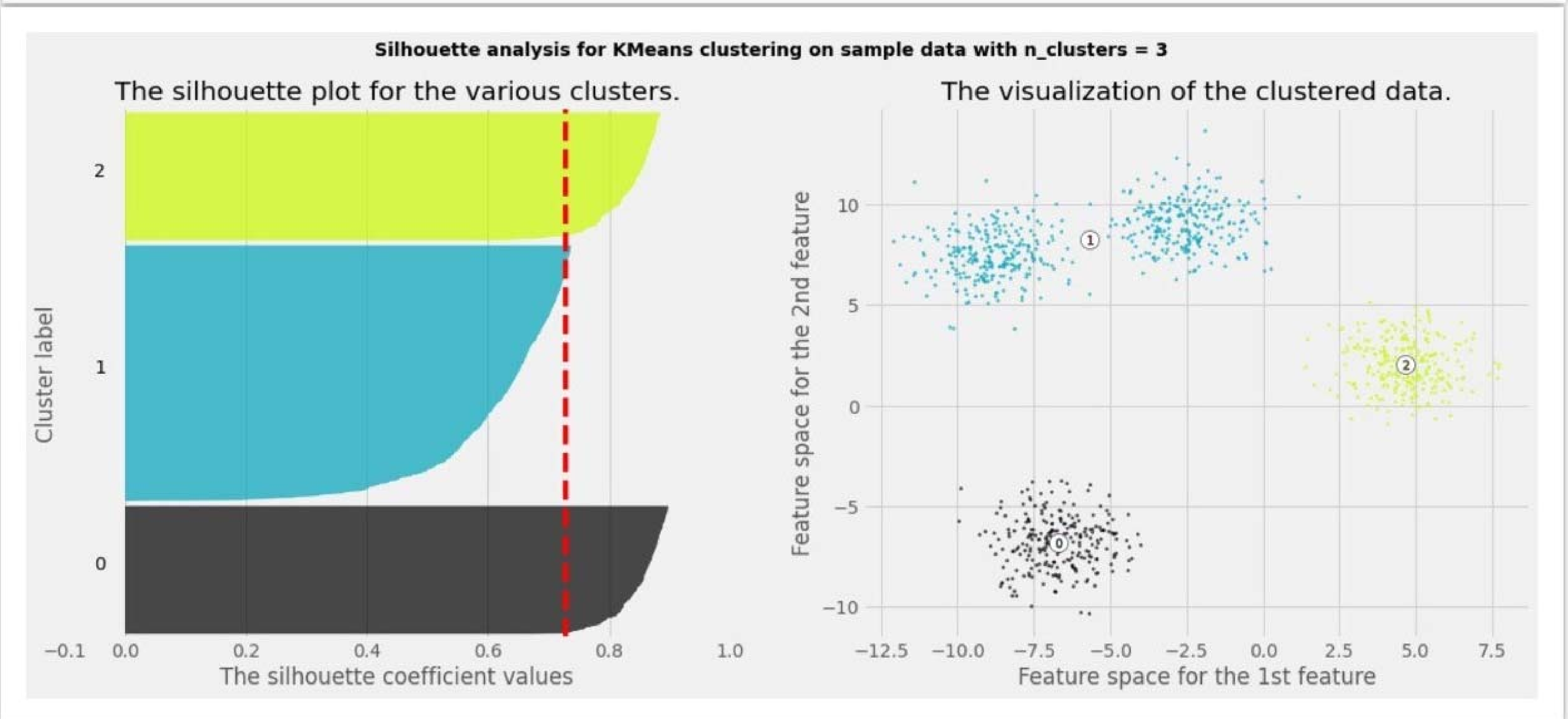
$$d'(z_3) = [\,|z_3 - z_4| + |z_3 - z_5| + |z_3 - z_6|\,] / 3$$

$$\mathbf{s_i(z_3) = [\,d'(z_3) - d(z_3)\,] / MAX\,[d'(z_3)\,,\,d(z_3)\,]}$$

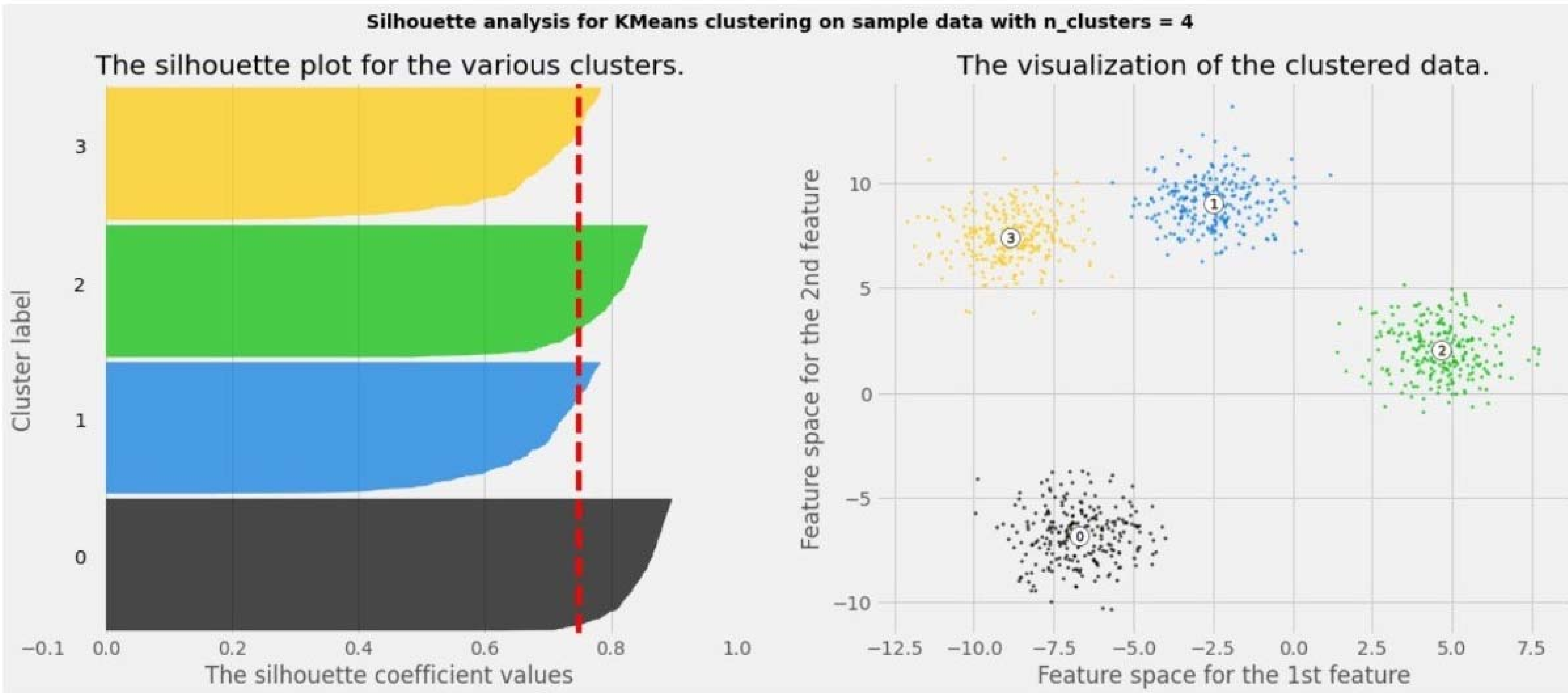




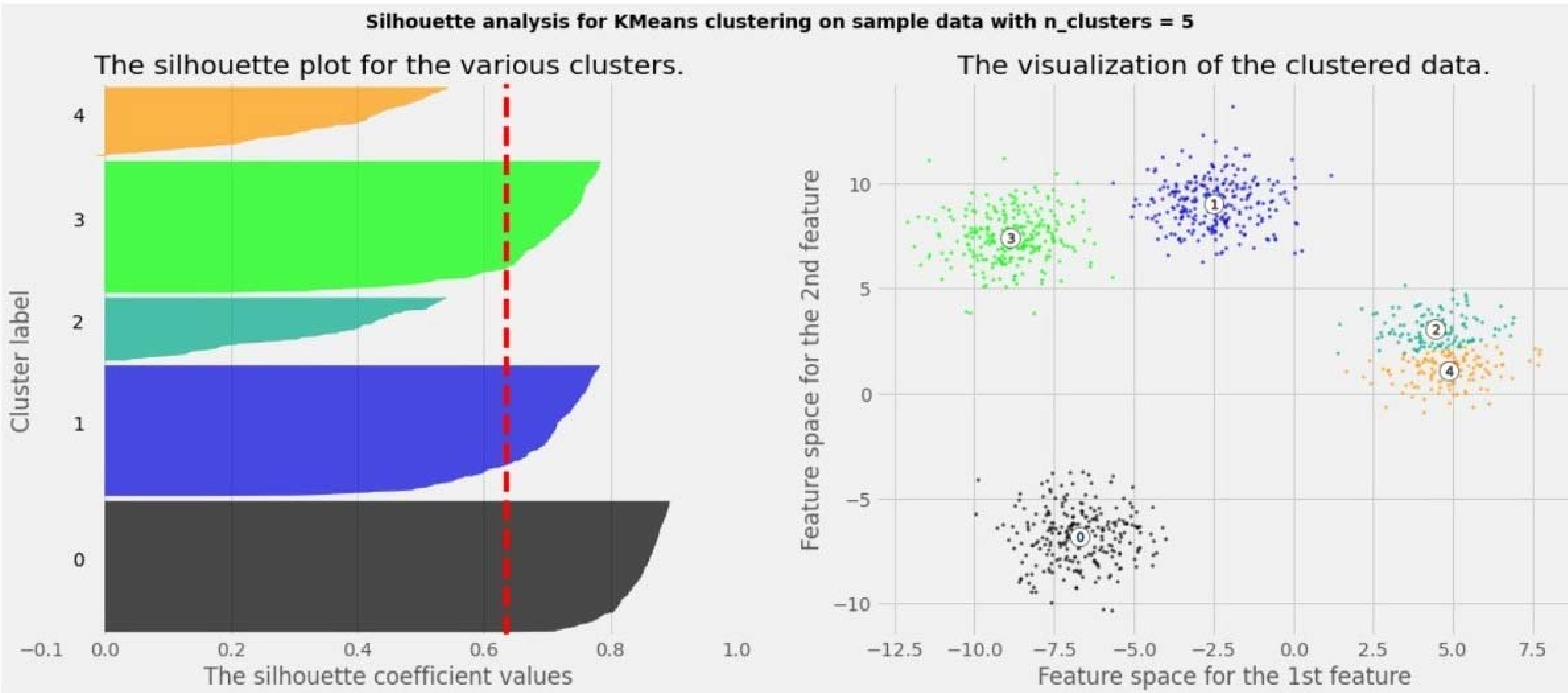
•The thickness of the silhouette plot for the cluster with cluster_label=1 when n_clusters=2, is bigger in size owing to the grouping of the 3 sub-clusters into one big cluster.



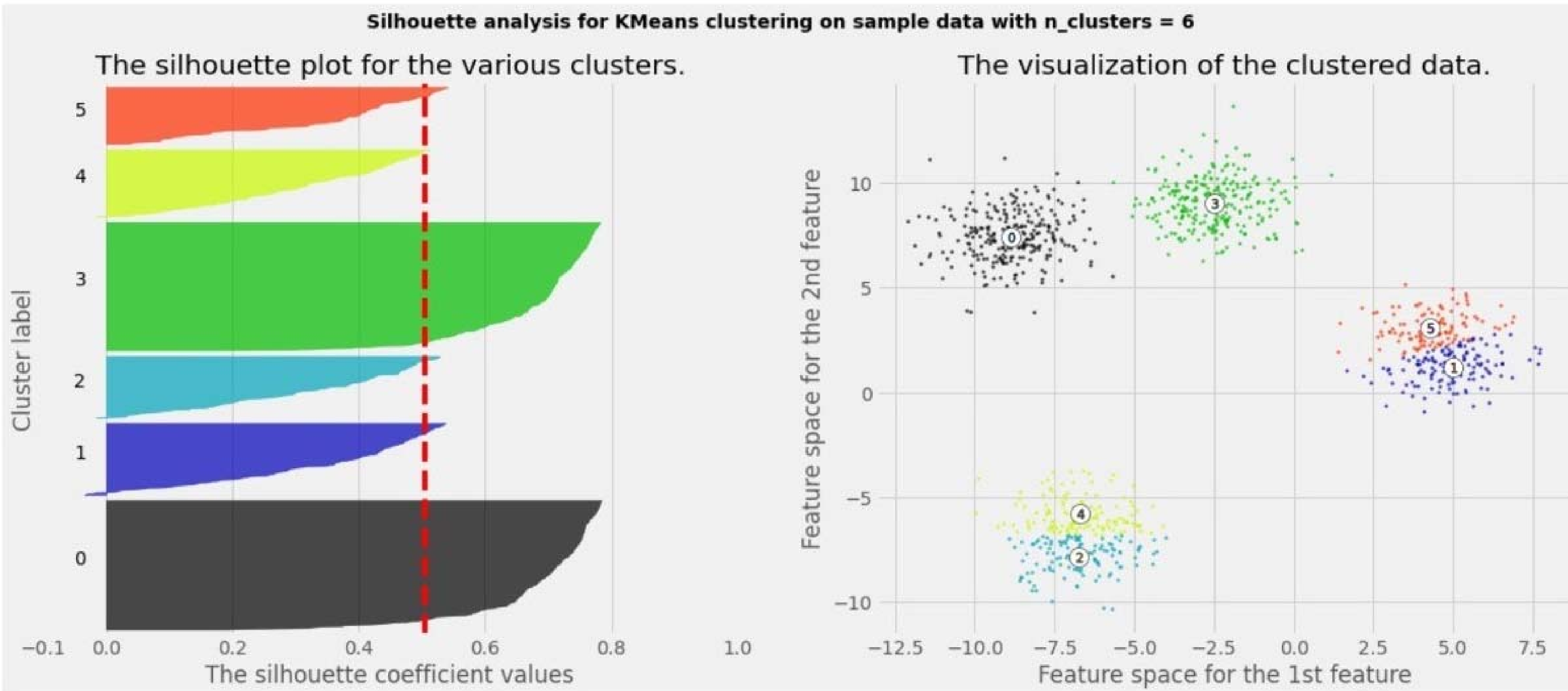
•The silhouette plot shows that the n_cluster value of **3** is a bad pick, as all the points in the cluster with cluster_label=1 are below-average silhouette scores.



•For n_clusters=4, all the plots are more or less of similar thickness and hence are of similar sizes, as can be considered as **best 'k'.**

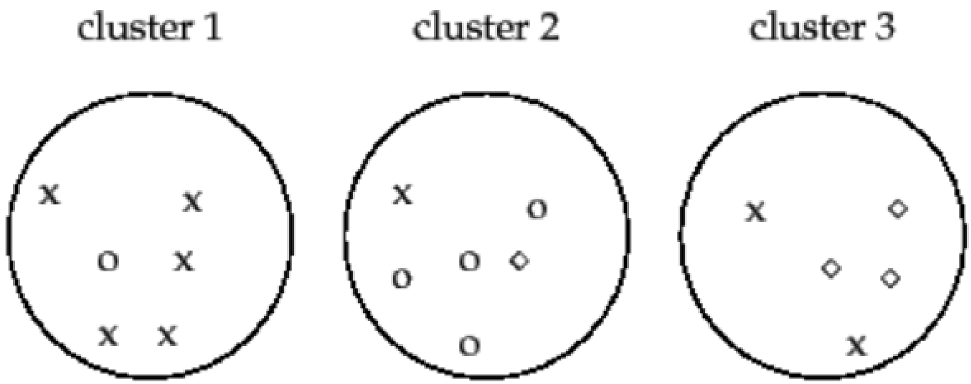


•The silhouette plot shows that the n_cluster value of **5** is a bad pick, as all the points in the cluster with cluster_label=2 and 4 are below-average silhouette scores.



•The silhouette plot shows that the n_cluster value of **6** is a bad pick, as all the points in the cluster with cluster_label=1,2,4 and 5 are below-average silhouette scores, and also due to the presence of outliers.

0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
silhouette score
2
3
4
5
6
7
Number of Clusters (k)



► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

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

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ is the set of classes.

Clustering Evaluation metrics

Given a set of n elements $S = \{o_1, \ldots, o_n\}$ and two partitions of S to compare, $X = \{X_1, \ldots, X_r\}$, a partition of *S* into *r* subsets, and $Y = \{Y_1, \ldots, Y_s\}$, a partition of *S* into *s* subsets, define the following:

- a, the number of pairs of elements in S that are in the **same** subset in X and in the **same** subset in Y
- b, the number of pairs of elements in S that are in **different** subsets in X and in **different** subsets in Y
- c, the number of pairs of elements in S that are in the **same** subset in X and in **different** subsets in Y
- d, the number of pairs of elements in S that are in **different** subsets in X and in the **same** subset in Y

The Rand index, R, is:[1][2]

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Intuitively, $a+b$ can be considered as the number of agreements between X and Y and $c+d$ as the number of disagreements between X and Y.

Since the denominator is the total number of pairs, the Rand index represents the *frequency of occurrence* of agreements over the total pairs, or the probability that X and Y will agree on a randomly chosen pair.

$\binom{n}{2}$ is calculated as $n(n-1)/2$.

Clustering Evaluation metrics

The contingency table [edit]

Given a set S of n elements, and two groupings or partitions (*e.g.* clusterings) of these elements, namely $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each entry n_{ij} denotes the number of objects in common between X_i and Y_j : $n_{ij} = |X_i \cap Y_j|$.

$X \diagdown Y$	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	

Definition [edit]

The original Adjusted Rand Index using the Permutation Model is

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where n_{ij}, a_i, b_j are values from the contingency table.

	Y_1	Y_2	Y_3	$Rowsums$
X_1	3	0	1	4
X_2	1	2	1	4
X_3	0	2	2	4
$Columnsums$	4	4	4	

1. $\sum_{ij}\binom{n_{ij}}{2} = \binom{3}{2}+\binom{0}{2}+\binom{1}{2}+\binom{1}{2}+\binom{2}{2}+\binom{1}{2}+\binom{0}{2}+\binom{2}{2}+\binom{2}{2} = 3+0+0+0+1+0+0+1+1=6$
2. $\sum_{i}\binom{a_i}{2} = \binom{4}{2}+\binom{4}{2}+\binom{4}{2} = 6+6+6=18$
3. $\sum_{j}\binom{b_j}{2} = \binom{4}{2}+\binom{4}{2}+\binom{4}{2} = 6+6+6=18$

$$ARI = \frac{6-[18\times 18]/\binom{12}{2}}{\frac{1}{2}[18+18]-[18\times 18]/\binom{12}{2}} = \frac{6-4.909091}{18-4.909091} = 0.08333333$$

Normalized Mutual Information

- Normalized Mutual Information:

$$NMI(Y,C) = \frac{2 \times I(Y;C)}{[H(Y) + H(C)]}$$

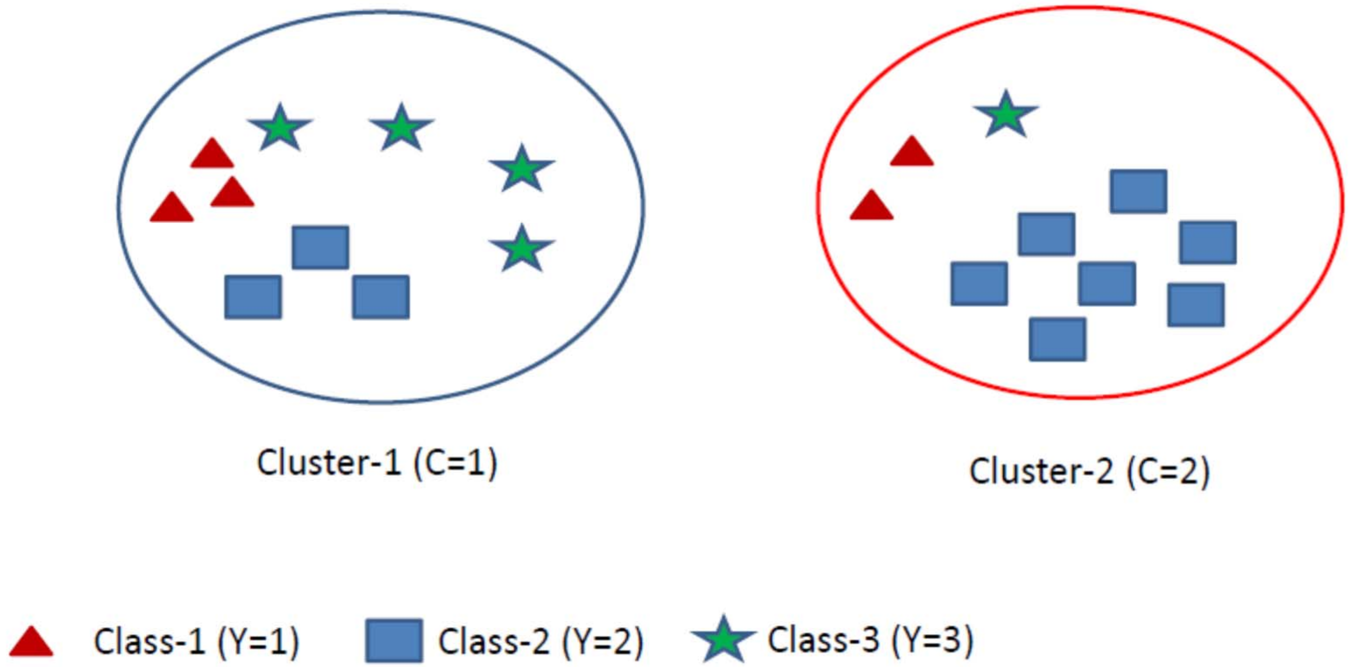
where,

1) Y = class labels
2) C = cluster labels
3) H(.) = Entropy
4) I(Y;C) = Mutual Information b/w Y and C

Note: All logs are base-2.

Calculating NMI for Clustering

- Assume m=3 classes and k=2 clusters



H(Y) = Entropy of Class Labels

- P(Y=1) = 5/20 = ¼
- P(Y=2) = 5/20 = ¼
- P(Y=3) = 10/20 = ½
- H(Y) = $-\frac{1}{4}\log\left(\frac{1}{4}\right)-\frac{1}{4}\log\left(\frac{1}{4}\right)-\frac{1}{2}\log\left(\frac{1}{2}\right)=1.5$

This is calculated for the entire dataset and can be calculated prior to clustering, as it will not change depending on the clustering output.

H(C) = Entropy of Cluster Labels

- P(C=1) = 10/20 = 1/2
- P(C=2) = 10/20 = ½
- H(Y) $=-\frac{1}{2}\log\left(\frac{1}{2}\right)-\frac{1}{2}\log\left(\frac{1}{2}\right)=1$

This will be calculated every time the clustering changes. You can see from the figure that the clusters are balanced (have equal number of instances).

I(Y;C)= Mutual Information

- Mutual information is given as:
  - $I(Y; C) = H(Y) - H(Y|C)$
  - We already know H(Y)
  - H(Y|C) is the entropy of class labels within each cluster, how do we calculate this??

Mutual Information tells us the reduction in the entropy of class labels that we get if we know the cluster labels. (Similar to Information gain in deicison trees)

H(Y|C): conditional entropy of class labels for clustering C

- Consider Cluster-1:
  - P(Y=1|C=1)=3/10 (three triangles in cluster-1)
  - P(Y=2|C=1)=3/10 (three rectangles in cluster-1)
  - P(Y=3|C=1)=4/10 (four stars in cluster-1)
  - Calculate conditional entropy as:

$$H(Y|C=1) = -P(C=1) \sum_{y \in \{1,2,3\}} P(Y=y|C=1) \log(P(Y=y|C=1))$$

$$= -\frac{1}{2} \times [\frac{3}{10} \log\left(\frac{3}{10}\right) + \frac{3}{10} \log(\frac{3}{10}) + \frac{4}{10} \log(\frac{4}{10})] = 0.7855$$

H(Y|C): conditional entropy of class labels for clustering C

- Now, consider Cluster-2:
  - P(Y=1|C=2)=2/10 (two triangles in cluster-1)
  - P(Y=2|C=2)=7/10 (seven rectangles in cluster-1)
  - P(Y=3|C=2)=1/10 (one star in cluster-1)
  - Calculate conditional entropy as:

$$H(Y|C=2) = -P(C=2)\sum_{y\in\{1,2,3\}} P(Y=y|C=2)\log(P(Y=y|C=2))$$

$$=-\frac{1}{2}\times[\frac{2}{10}\log\left(\frac{2}{10}\right)+\frac{7}{10}\log(\frac{7}{10})+\frac{1}{10}\log(\frac{1}{10})] = 0.5784$$

I(Y;C)

- Finally the mutual information is:

$$
\begin{aligned}
I(Y;C) &= H(Y) - H(Y|C) \\
&= 1.5 - [0.7855 + 0.5784] \\
&= 0.1361
\end{aligned}
$$

The NMI is therefore,

$$NMI(Y,C) = \frac{2 \times I(Y;C)}{[H(Y) + H(C)]}$$

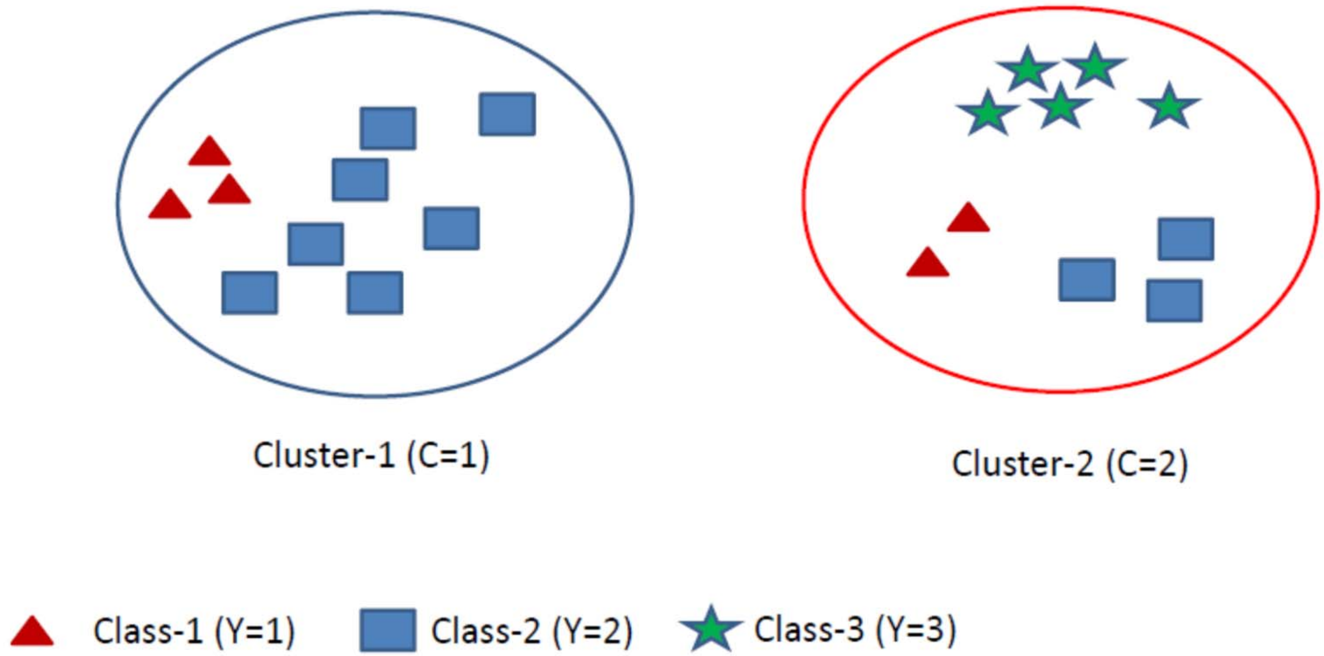
$$NMI(Y,C) = \frac{2 \times 0.1361}{[1.5 + 1]} = 0.1089$$

NMI

- NMI is a good measure for determining the quality of clustering.
- It is an external measure because we need the class labels of the instances to determine the NMI.
- Since it’s normalized we can measure and compare the NMI between different clusterings having different number of clusters.

NMI for Clustering

- Calculate the NMI:



H(Y|C): conditional entropy of class labels for clustering C

- Consider Cluster-1:
  - P(Y=1|C=1)=3/10 (three triangles in cluster-1)
  - P(Y=2|C=1)=7/10 (seven rectangles in cluster-1)
  - P(Y=3|C=1)=0/10 (no stars in cluster-1)
  - Calculate conditional entropy as:

$$H(Y|C=1) = -P(C=1) \sum_{y \in \{1,2,3\}} P(Y=y|C=1)\log(P(Y=y|C=1))$$

$$= -\frac{1}{2} \times [\frac{3}{10}\log\left(\frac{3}{10}\right) + \frac{0}{10}\log(\frac{0}{10}) + \frac{7}{10}\log(\frac{7}{10})] = 0.4406$$

We used 0log(0)=0

H(Y|C): conditional entropy of class labels for clustering C

- Now, consider Cluster-2:
  - P(Y=1|C=2)=2/10 (two triangles in cluster-1)
  - P(Y=2|C=2)=3/10 (three rectangles in cluster-1)
  - P(Y=3|C=2)=5/10 (five stars in cluster-1)
  - Calculate conditional entropy as:

$$H(Y|C=2) = -P(C=2) \sum_{y\in\{1,2,3\}} P(Y=y|C=2)\log(P(Y=y|C=2))$$

$$= -\frac{1}{2} \times [\frac{2}{10}\log\left(\frac{2}{10}\right) + \frac{3}{10}\log(\frac{3}{10}) + \frac{5}{10}\log(\frac{5}{10})] = 0.7427$$

I(Y;C)

- Finally the mutual information is:

$$\begin{aligned} I(Y;C) &= H(Y) - H(Y|C) \\ &= 1.5 - [0.4406 + 0.7427] \\ &= 0.3167 \end{aligned}$$

The NMI is therefore,

$$NMI(Y,C) = \frac{2 \times I(Y;C)}{[H(Y) + H(C)]}$$

$$NMI(Y,C) = \frac{2 \times 0.3167}{[1.5 + 1]} = 0.2533$$

Comments

- NMI for the second clustering is higher than the first clustering. It means we would prefer the second clustering over the first.
  - You can see that one of the clusters in the second case contains all instances of class-3 (stars).
- If we have to compare two clustering that have different number of clusters we can still use NMI.

References and further readings

- Andrew NG., Machine Learning Course, Coursera, slide: Clustering

- [David Sontag, Clustering, lecture 14, New York university,](http://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf) May 2020, http://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf