# Clustering with the K-Means Method and Its Applications

Tuğçe ÇALIŞIR

# ABSTRACT

In this talk, we consider the task of clustering a collection of vectors into groups or clusters with the nearest mean (cluster centers or cluster centroid), as measured by the distance between pairs of vectors. We describe a clustering method called the k-means algorithm.

The talk includes the following contents in the given order.

1. Definition of Clustering
2. Some Examples
3. Clustering Objective
4. Optimiztion Methods
5. K-Means Algorithm
6. Convergence
7. Elbow Method
8. The Code of K-Means Algorithm

# DEFINITION OF CLUSTERING

Clustering is a problem of dividing data into a limited number of related classes so that items in the same group are as similar as possible and items in different groups are as different as possible.

For this purpose, suppose we have N n-vectors, $x_1, \ldots, x_N$ . The goal of clustering is to group or partition the vectors (if possible) into k groups or clusters, with the vectors in each group close to each other.

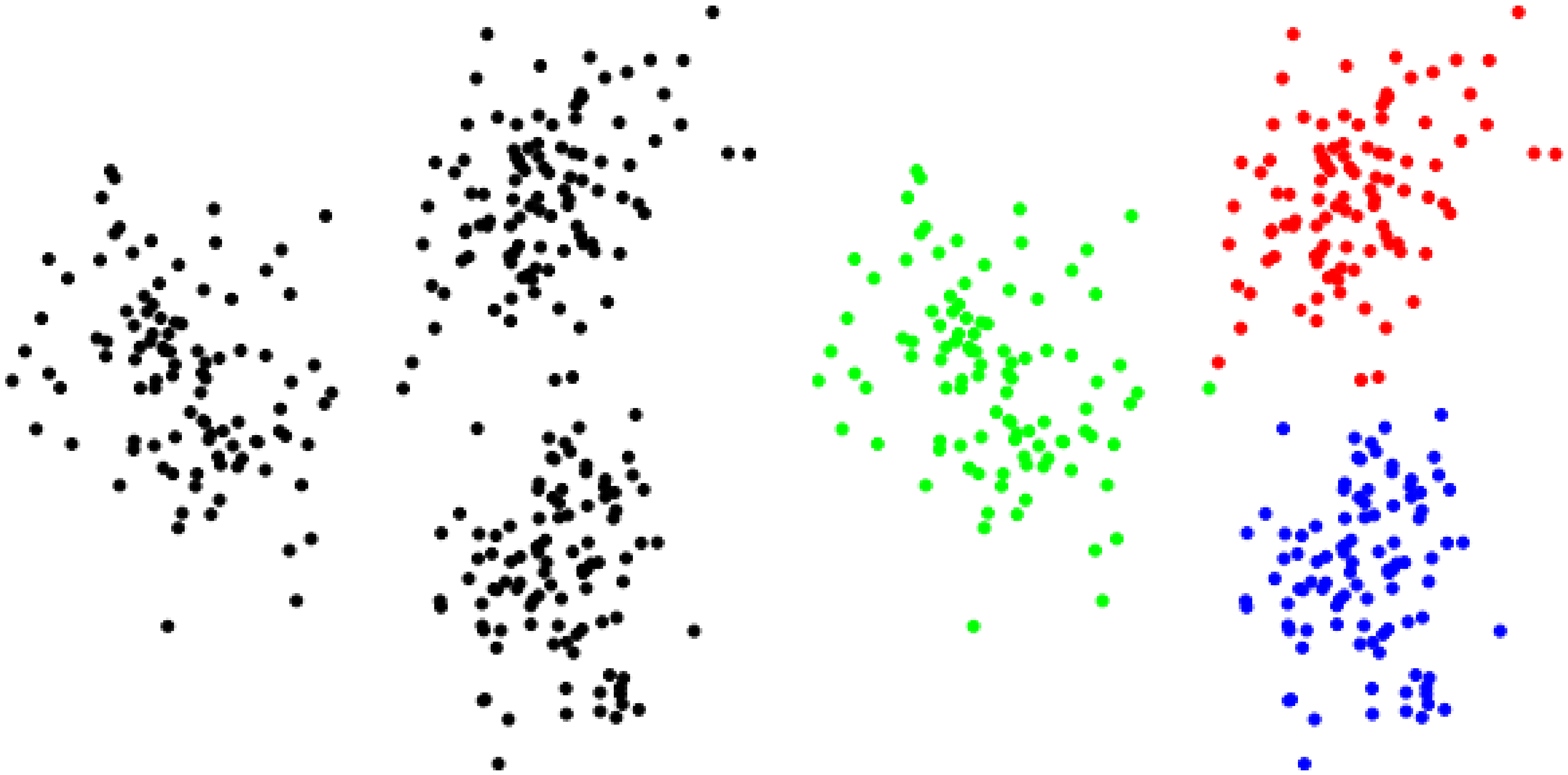

**Figure 1.** *300 points in a plane are clustered into 3 groups.*

# EXAMPLES

- topic discovery and document classification
  - $x_i$ is word count histogram for document i
- patient clustering
  - $x_i$ are patient attributes, test results, symptoms
- customer market segmentation
  - $x_i$ is purchase history and other attributes of customer i
- color compression of images
  - $x_i$ are RGB pixel values
- financial sectors
  - $x_i$ are n-vectors of financial attributes of company i

# CLUSTERING OBJECTIVE

**Specifying the cluster assignments**

The groups 1, . . . , k, and specify a clustering or assignment of the N given vectors to groups using an N-vector c, where $c_i \in [1,k]$ is the group (number) that the vector $x_i$ is assigned.

The index sets in terms of the group assignment vector c as

$$G_j = \{i \mid c_i = j\}$$

which means that $G_j$ is the set of all indices i for which $c_i = j$.

**Group representatives**

Each of the groups is associated with a group representative n-vector, which are denoted by $z_1, \ldots, z_k$.

# Example 1

Suppose that $N = 5$ vectors and $k = 3$ groups, $c = (3, 1, 1, 1, 2)$.

That means that the first vector $x_1$ is assigned to group 3 which denotes as $z_3$.
The fifth vector $x_5$ is assigned to group 2 which denotes as $z_2$.
The other vectors ($x_2$, $x_3$, $x_4$) are associated to group one ($z_1$).
If the index is set in terms of the group assignment vector, we have

$$G_1 = \{2, 3, 4\},\ G_2 = \{5\},\ G_3 = \{1\}.$$

Where $G_j$ is the set of indices corresponding to group j.

Find cluster center z and assignments c to minimize the sum of squared distances of data points x to their assigned cluster centers

$$J^{clust} = \sum_{i=1}^{N} ||x_i - z_{c_i}||^2$$

for j = 1,. . . ,k and $c_i$ is the group that $x_i$ is in: $i \in G_j$ where $G_j \subset \{1,. . . ,N\}$ is the group j.

# OPTIMIZATION METHODS

1. Partitioning the vectors given the representatives (Fix centers, optimize assignments)
2. Choosing representatives given the partition (Fix assignments, optimize means)

# Partitioning the Vectors Given the Representatives

Suppose that the group representatives $z_1, \ldots, z_k$ are fixed, and the group assignments that minimize $J^{clust}$ are found by assigning each vector to its nearest representative. Each $z_j$ for z = 1, . . . ,k ,assigned to minimize mean square distance

$$|| x_i - z_{c_i} || = \min(j=1,\ldots,k) \; || x_i - z_j ||$$

so the value of $J^{clust}$ is given by

$$J_j = (1/N) \sum_{i=1}^{N} ||x_i - z_j||^2$$

# Choosing Representatives Given the Partition

Suppose that the group assignments $c_1, \ldots, c_N$ are fixed, the group representatives are found in order to minimize the value of $J^{clust}$.

Re-arranging of the sum of N terms into k sums, each associated with one group

$$J^{clust} = J_1 + \ldots + J_k,$$

where

$$J_j = (1/N) \sum_{i \in G_j} ||x_i - z_j||^2$$

is the contribution to the objective $J^{clust}$ from any $i \in G_j$, i.e., for any vector $x_i$ in group j.

The choice of group representative $z_j$ affects the term $J_j$ in $J^{clust}$. So each $z_j$ is chosen to minimize $J_j$. Thus the vector $z_j$ minimizes the mean square distance to the vectors in group j. $z_j$ is computed as the average (or mean or centroid) of the vectors $x_i$ in its group:

$$z_j = \left(\frac{1}{|G_j|}\right) \sum_{\mathrm{i} \in G_j} x_j ,$$

where $|G_j|$ is the number of elements in the set $G_j$, in the size of the group j.

# THE K-MEANS ALGORITHM

If both the group assignments and the group representatives are not fixed, then each depends on the other. The algorithm must be created to iterate between the two choices. This means that it repeatedly alternates between updating the group assignments, and then updating the representatives. In each step, the objective $J^{clust}$ gets better (i.e., goes down) unless the step does not change the choice. Iterating between choosing the group representatives and choosing the group assignments is the k-means algorithm for clustering a collection of vectors.

Given a list of N vectors $x_1, \ldots, x_N$ , and an initial list of k group representative vectors $z_1, \ldots, z_k$

Repeat until convergence

1. Partition the vectors into k groups. For each vector $i = 1, \ldots, N$, assign $x_i$ to the group associated with the nearest representative.

2. Update representatives. For each group $j = 1, \ldots, k$, set $z_j$ to be the mean of the vectors in group j.

# Example 2

We apply the algorithm on a set of N = 8 points and k = 3 groups.

The set is G = {(1,5),(2,7),(3,3),(4,8),(5,7),(6,1),(8,4),(7,3)}.

First, randomly picking group representative.

Initial group representatives are: $G_1$(1,5), $G_2$(5,7), $G_3$(3,3)

The group centers are fixed for the first iteration.

Each element of the set given to the representatives is partitioned.

The assignments of the set are found and optimized.

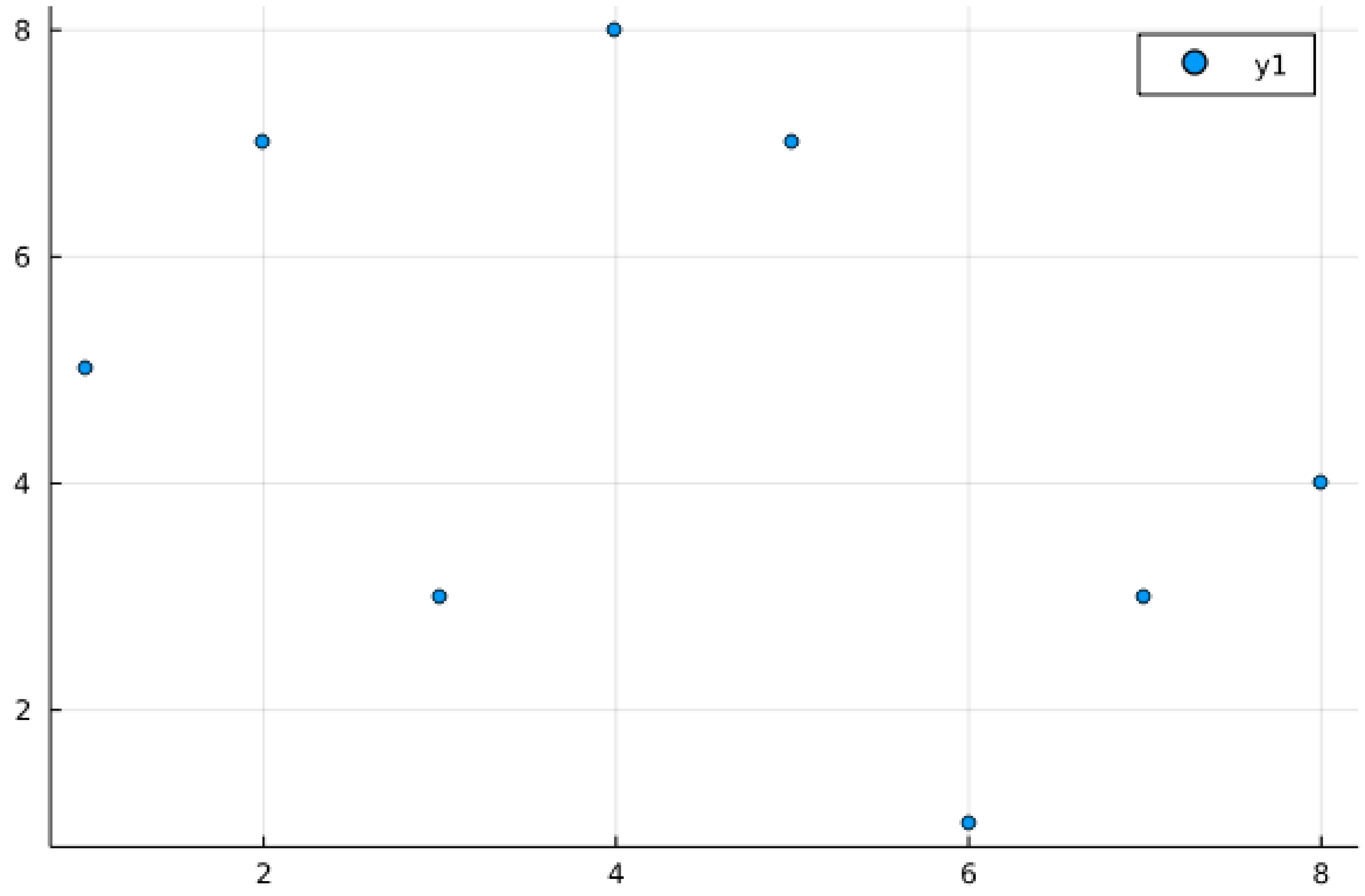


**Figure 2.** 8 points shown in a plane.

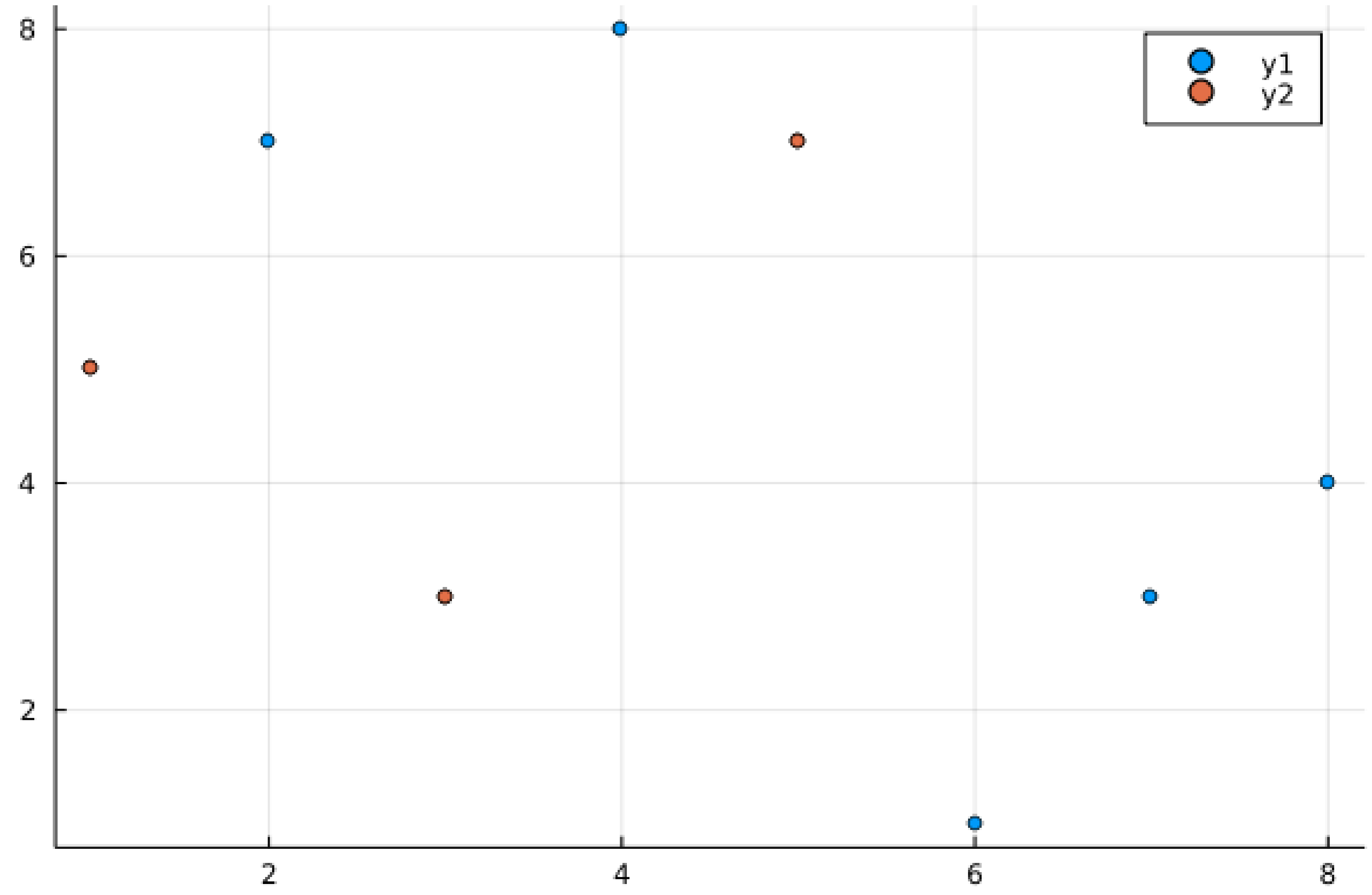


**Figure 3.** Set of N = 8 points and k = 3 groups are shown. Orange points represent the initial group representatives.

# Iteration 1

| | Dist. Mean 1 | Dist. Mean 2 | Dist. Mean 3 | Cluster |
|---|---|---|---|---|
| Point | (1,5) | (5,7) | (3,3) | |
| G1(1,5) | 0.00 | 4.472136 | 2.828427 | 1 |
| G2(2,7) | 2.236068 | 3.00 | 4.123106 | 1 |
| G3(3,3) | 2.828427 | 4.472136 | 0.00 | 3 |
| G4(4,8) | 4.242641 | 1.414214 | 5.099019 | 2 |
| G5(5,7) | 4.472136 | 0.00 | 4.472136 | 2 |
| G6(6,1) | 6.403124 | 6.082762 | 3.605551 | 3 |
| G7(8,4) | 7.071068 | 5.00 | 5.099019 | 2 |
| G8(7,3) | 6.324555 | 4.472136 | 4.00 | 3 |

$G_1 = \{(1,5),(2,7)\}$, $G_2 = \{(4,8),(5,7),(8,4)\}$, $G_3 = \{(3,3),(6,1),(7,3)\}$
$J^{clust}=2.090996$

Next we need to recompute the new cluster centers. We do so, by taking the mean of all points in each cluster.

For cluster 1

We have $G_1(1,5)$, $G_2(2,7)$.

New cluster center is $(1+2,5+7)/2 = (1.5,6)$

For cluster 2

We have $G_4(4,8)$, $G_5(5,7)$, $G_7(8,4)$

New cluster center is $(4+5+8,8+7+4)/3 = (5.666667,6.333333)$

For cluster 3

We have $G_3(3,3)$, $G_6(6,1)$, $G_8(7,3)$

New cluster center is $(3+6+7,3+1+3)/3 = (5.333333,2.333333)$

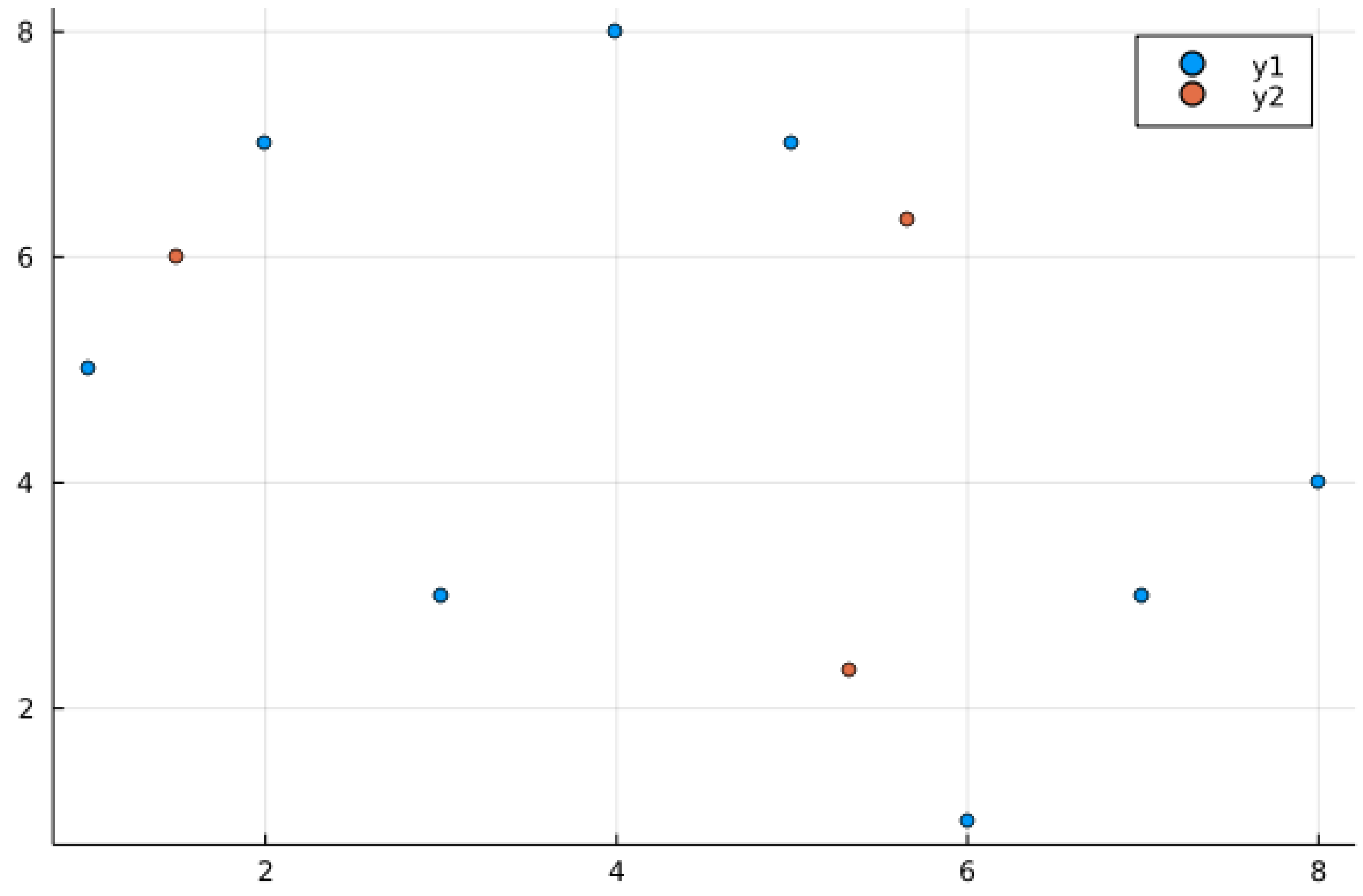


**Figure 4.** First iteration of k-means algorithm.

# Iteration 2

| | Dist. Mean 1 | Dist. Mean 2 | Dist. Mean 3 | Cluster |
|---|---|---|---|---|
| Point | (1.5,6) | (5.666667,6.333333) | (5.333333,2.333333) | |
| G1(1,5) | 1.118034 | 4.853407 | 5.088112 | 1 |
| G2(2,7) | 1.118034 | 3.726780 | 5.734883 | 1 |
| G3(3,3) | 3.354102 | 4.268749 | 2.426703 | 3 |
| G4(4,8) | 3.201562 | 2.357023 | 5.821417 | 2 |
| G5(5,7) | 3.640055 | 0.942809 | 4.678557 | 2 |
| G6(6,1) | 6.726812 | 5.343739 | 1.490712 | 3 |
| G7(8,4) | 6.800735 | 3.299831 | 3.144661 | 3 |
| G8(7,3) | 6.264982 | 3.590109 | 1.795055 | 3 |

$G_1$ ={(1,5),(2,7)}, $G_2$ ={(4,8),(5,7)},$G_3$ ={(3,3),(6,1),(7,3),(8,4)}

$J^{clust}$=1.799129

Next we need to recompute the new cluster centers. We do so, by taking the mean of all points in each cluster.

For cluster 1

We have $G_1$ ={(1,5),(2,7)},

New cluster center is (1+2,5+7)/2 = (1.5,6)

For cluster 2

We have $G_2$ ={(4,8),(5,7)}.

New cluster center is (4+5,8+7)/2 = (4.5,7.5)

For cluster 3

We have $G_3$ ={(3,3),(6,1),(7,3),(8,4)}.

New cluster center is (3+6+7+8,3+1+3+4)/4 = (6.0,2.75)

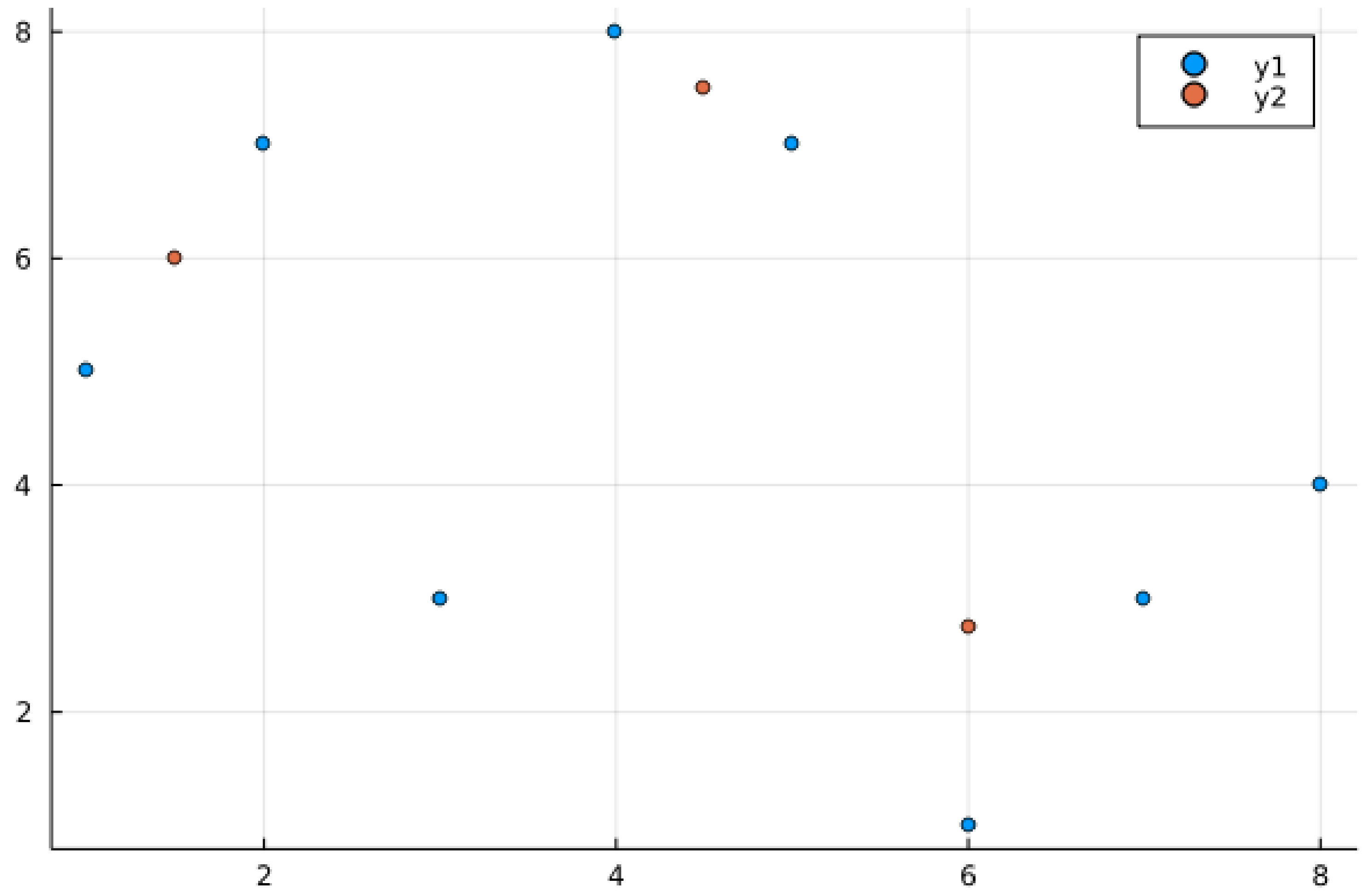


**Figure 5.** Second iteration of k-means algorithm.

# CONVERGENCE

The fact that $J^{clust}$ decreases in each step.

This means the k-means algorithm converges (after enough iteration) to the group representative $z_j$.

Iterations stop when $z_j$ stops changing.

Depending on the initial choice of representatives, the algorithm can converge to different final partitions, with different objective values.

The k-means algorithm is a heuristic one, which means it cannot guarantee that the partition it finds minimizes our objective $J^{clust}$. For this reason, it is common to run the k-means algorithm several times, with different initial representatives, and choose the one among them with the smallest final value of $J^{clust}$.

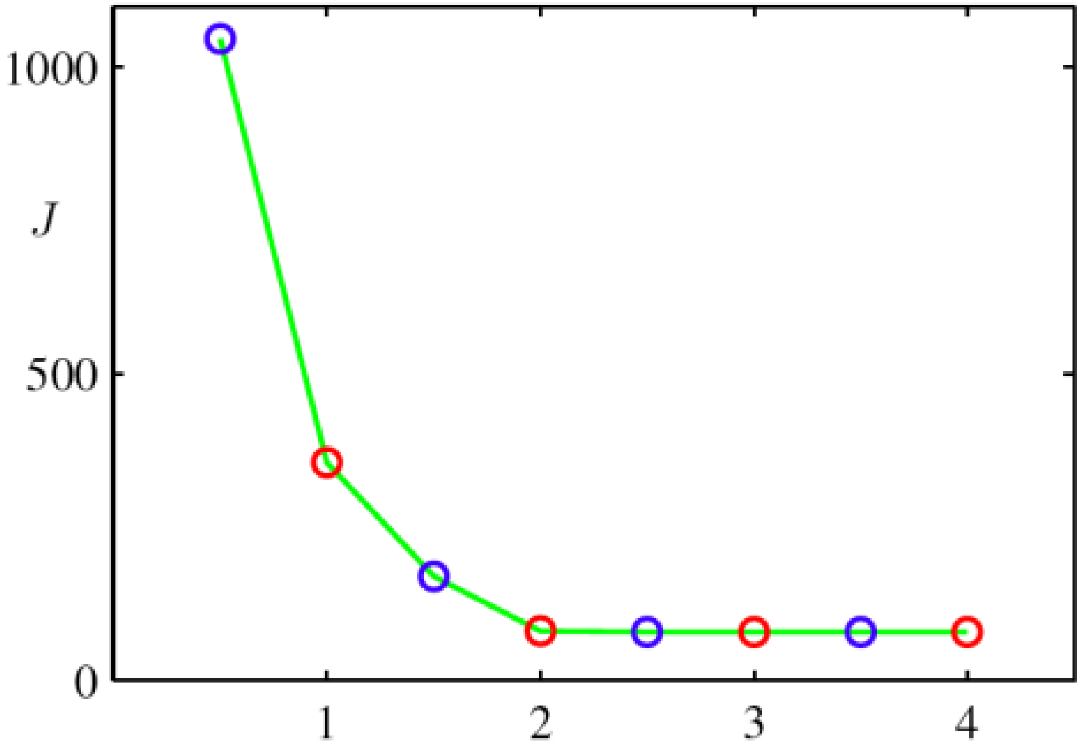


**Figure 6.** Convergence of k-means clustering.

# ELBOW METHOD

In cluster analysis, the elbow method is a heuristic one used in determining the number of clusters.

The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

In this way, the optimal number of clusters is found.

There are two methods to find the optimal number of clusters. These are:

1. Within Cluster Sum of Squares

2. Between Cluster Sum of Squares

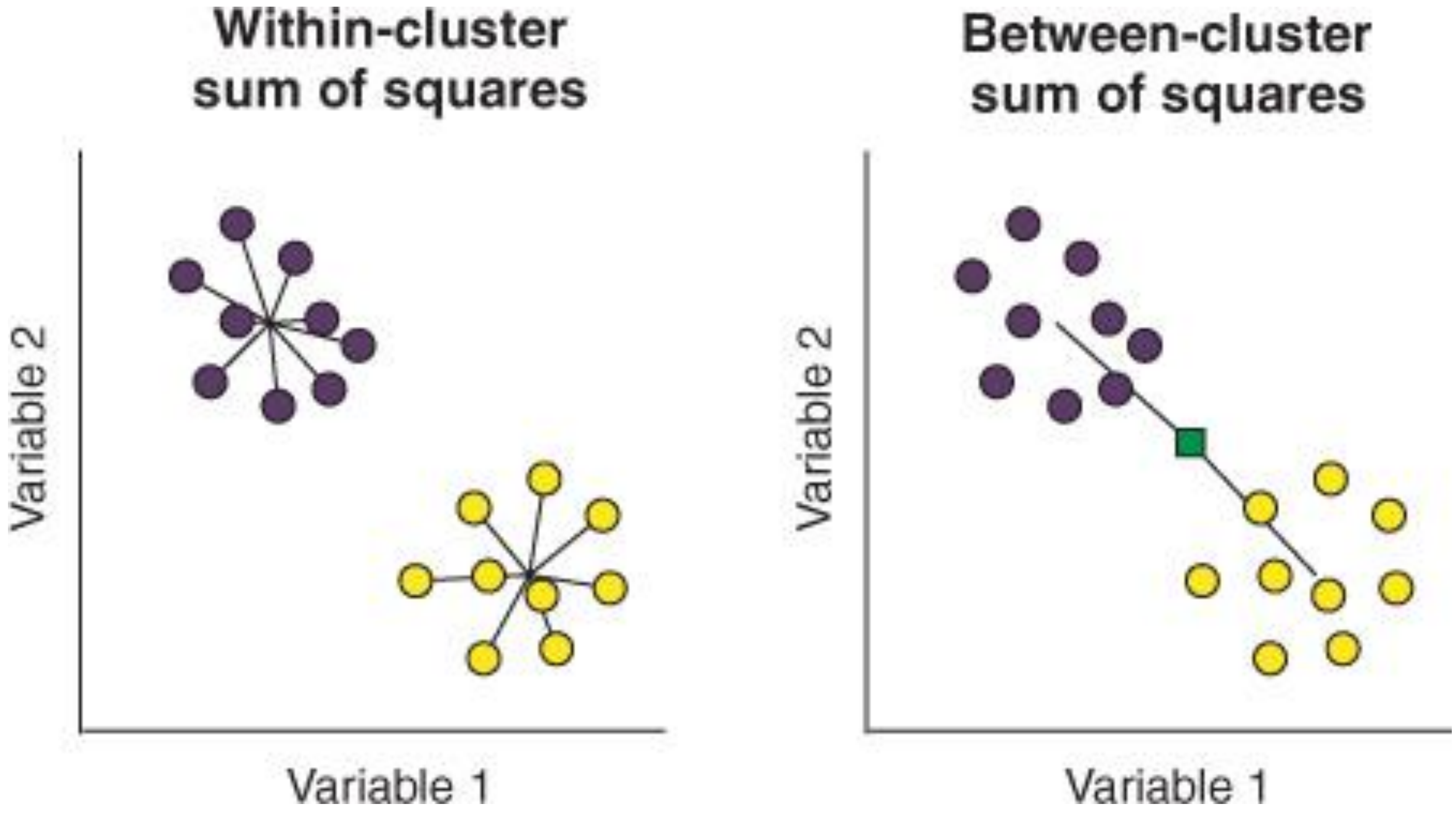


**Figure 7.** Between and within cluster sum of squares methos.

# Within Cluster Sum of Squares

Cluster Cohesion: Measures how closely related are objects in a cluster.

Cluster cohesion is measured by the within cluster sum of squares method which is

$$WCSS = \sum_{j=1}^{k} \sum_{i \in G_j} ||x_i - z_{c_i}||^2$$

$z_{c_i}$ is the representative vector associated with data vector $x_i$ is in: i ∈ $G_j$ where $G_j$ ⊂ {1,. . . ,N} is the set of indices corresponding to group j.

# Between Cluster Sum of Squares

Cluster Separation: Measures how distinct or well-separated a cluster is from other clusters.

Separation is measured by the between cluster sum of squares which is

$$BCSS = \sum_{i \in G_j} |G_j|.||z - z_{c_i}||^2$$

Where $G_j \subset \{1,. . . ,N\}$ is the set of indices corresponding to group j. $|G_j|$ is standard mathematical notation for the number of elements in the set $G_j$, i.e., the size of group j.

$z_{c_i}$ is the representative vector associated with i elements in group $j=c_i$.

z is sample’s mean.

# THE CODE OF K-MEANS ALGORITHM

1: kmeans(x,k)

2: length of x

3: size of each x elements

4: vector for store the distance of each point to the nearest representative

5: vector for store representatives

6: the array that stores the assignments of N integers between 1 and k

7: stopping condition

8: for j=1:k

9: Cluster j representative (average of points in cluster j)

10: end

11: for i = 1:N

12: For each x, the distance to the nearest representative and its group index

13: end

14: clustering objective and finding $J^{clust}$ value

15: if J stopped decreasing, terminate

16: return assignment, reps

17: end

18: end

19: end

# References


- https://online.stat.psu.edu/stat508/book/export/html/648
- On the number of groups in clustering Aurélie Fischer ∗
- Introduction to Applied Linear Algebra Vectors, Matrices, and Least Squares Stephen Boyd Lieven Vandenberghe
- FINDING THE REPRESENTATIVE IN A CLUSTER USING CORRELATION CLUSTERING 1 Dávid NAGY* , 2 Laszló ASZALÓS, 3 Tamás MIHÁLYDEÁK
- CSC 411 Lecture 14:Clustering Ethan Fetaya, James Lucas and Emad Andrews
- K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median Edy Umargono1 , Jadmiko Endro Suseno2 and Vincensius Gunawan S. K.2
- https://online.stat.psu.edu/stat508/book/export/html/648
- https://www.kdnuggets.com/2019/05/golden-goose-cohort-analysis.html/2