Intelligent Systems

Escuela Técnica Superior de Informática

Universitat Politècnica de València

Block 2 Chapter 4:
Clustering
Unsupervised learning: K-means algorithm

Índice

Índice

Introduction

Now we shall investigate ***unsupervised learning*** or ***clustering***, which use *unlabeled samples*.

According to [Anderberg,1973], the aim of clustering is:

> To group objects in classes so the objects belonging to the same class have a high degree of ***natural association***, while the other classes are relatively different. The purpose is to create classes that are relatively different from each other.

In other words:

> To find ***natural groupings*** of a set of input patterns so that the descriptions of these objects can be done in terms of classes or groups with strong internal similarities.

We need a distance (similarity) function to assess the quality of a clustering result (high similarity within a cluster and low similarity between clusters).

Two types of clustering: ***Partitional*** and ***Hierarchical***.

Índice

Partitional Clustering

General problem:

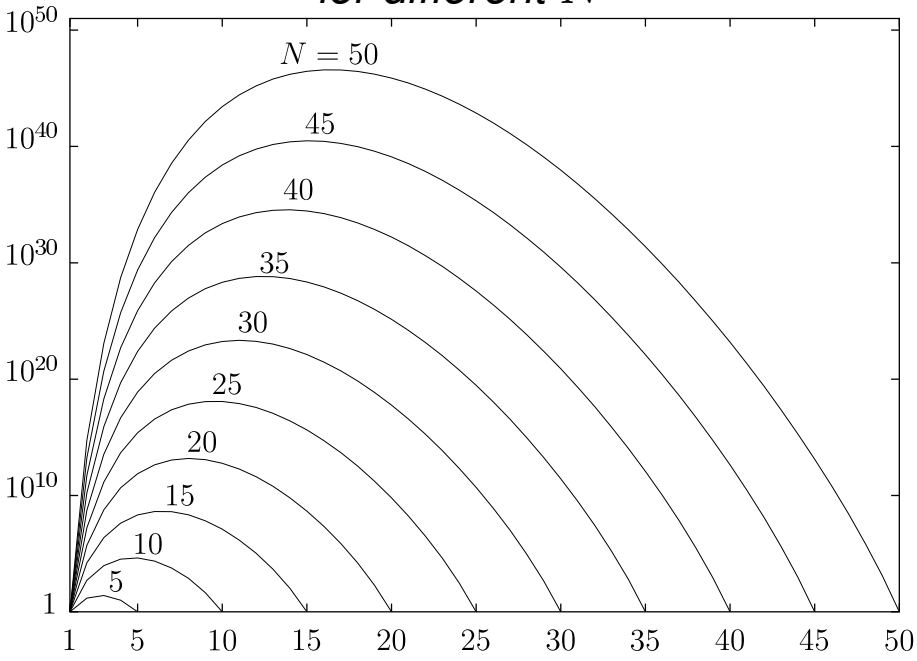
Given a set of N objects or observations $(x_1, x_2, \ldots, x_N)$, where each object is a d-dimensional real vector, the objective is to find a partition Π of the N objects into C **classes or clusters**; i.e.: $\Pi = \{X_1, \ldots, X_C\}$, where $X_1, \ldots, X_C$ denote the C clusters. In order to do so, we assume there is available a ***criterion function*** J to evaluate the quality of a partition Π. Thus, the clustering problem can be seen as a search problem:

$$\Pi^* = \underset{\Pi=\{X_1,\ldots,X_C\}}{\arg\min} J(\Pi) \quad \text{(best partition, the one that minimizes } J \text{ for each partition)} \qquad (1)$$

Difficulty:

The number of partitions to evaluate is too high even for small values of N and C (see right). Searching for global optimal solutions through enumeration techniques (explicitly or implicitly) is not feasible except for a few particular cases.

Number of partitions with respect to C for different N



Solution:

Search for suboptimal solutions obtained by approximate algorithms.

Partitional clustering: "Sum of Squared Errors" (SSE) criterion

The SSE is a criterion to select a partition Π of N objects into C clusters. The objective is to select the partition that minimizes SSE.

The SSE value of a partition Π of N objects $(x_1, x_2, \ldots, x_N)$ into C clusters, $\Pi = \{X_1, \ldots, X_C\}$, is computed as follows:

- Each cluster X_c contains a subset of the N objects (disjoint sets)
- For each cluster X_c, compute the squared errors for all $\boldsymbol{x} \in X_c$: $SE_x = \|\boldsymbol{x} - \boldsymbol{m}_c\|^2$.
  $\boldsymbol{m}_c$ is the mean of points in X_c:
  $$\boldsymbol{m}_c = \frac{1}{|X_c|} \sum_{\boldsymbol{x} \in X_c} \boldsymbol{x} \tag{2}$$
  $\boldsymbol{m}_c$ is interpreted as the natural prototype of X_c. Each object $\boldsymbol{x} \in X_c$, is interpreted as a "distorted version" of $\boldsymbol{m}_c$ and the distortion of $\boldsymbol{x}$ is modeled by the *error vector* $\boldsymbol{x} - \boldsymbol{m}_c$.
- For each cluster X_c, J_c is the function for cluster X_c. $J_c = WCSE_c$ (within-cluster squared errors)
  $$J_c = WCSE_c = \sum_{\boldsymbol{x} \in X_c} \|\boldsymbol{x} - \boldsymbol{m}_c\|^2 \tag{3}$$
- The SSE value of a partition Π is computed as:
  $$SSE_\Pi = J(X_1, \ldots, X_C) = \sum_c J_c = \sum_c WCSE_c \tag{4}$$

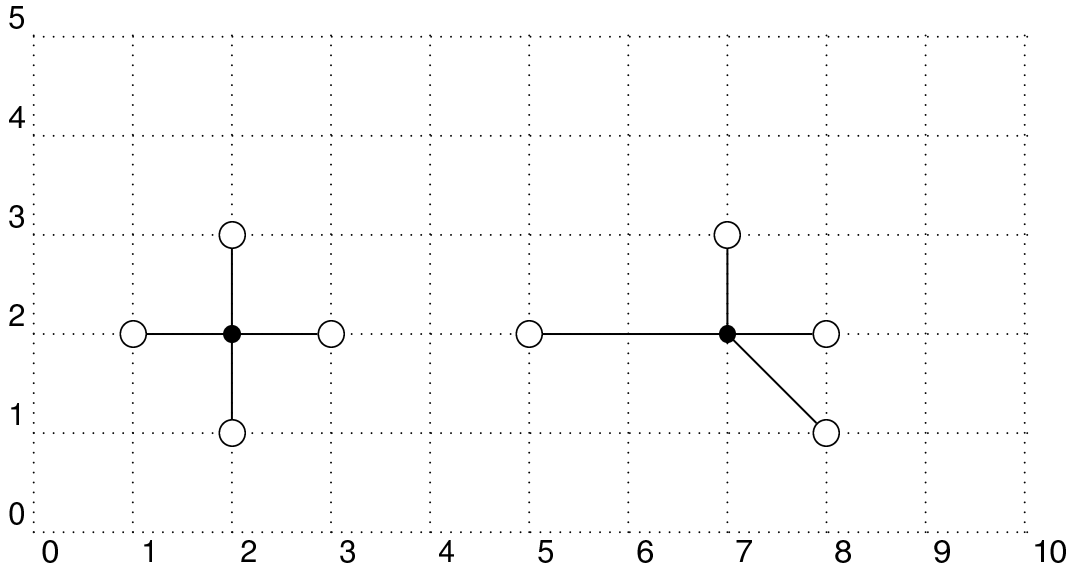
SSE criterion (continuation)

Now, we would have to calculate the SSE value for every partition (if they are explicitly given) and take the one that minimizes SSE (Π^*, best partition)

SSE measures the sum (or average) of the squares of the magnitudes of the error vectors, and, obviously, it is a criterion to minimize.

The *mean* of each cluster, $\boldsymbol{m}_c$, is the point that represents the object of this cluster with the lowest SSE. $\boldsymbol{m}_c$ is the center of the cluster X_c (called *centroid*).

Partitional clustering example (partition I)



SSE$= 12$

Partition I (Π_1): 8 samples $(x_1, \ldots, x_N)$, 2 clusters in partition $\Pi_1 = \{X_1, X_2\}$

$$X_1 = \{x_1, \ldots, x_4\} : x_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, x_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, x_3 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, x_4 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

$$X_2 = \{x_5, \ldots, x_8\} : x_5 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}, x_6 = \begin{pmatrix} 7 \\ 3 \end{pmatrix}, x_7 = \begin{pmatrix} 8 \\ 1 \end{pmatrix}, x_8 = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$$

$$\boldsymbol{m}_1 = \frac{1}{4} \sum_{x \in X_1} x = \frac{1}{4} \begin{pmatrix} 8 \\ 8 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \qquad \boldsymbol{m}_2 = \frac{1}{4} \sum_{x \in X_2} x = \frac{1}{4} \begin{pmatrix} 28 \\ 8 \end{pmatrix} = \begin{pmatrix} 7 \\ 2 \end{pmatrix}$$

Partitional clustering example (partition I, continuation)

Calculations for cluster X_1:

$$SE_{x1} = \|\boldsymbol{x1} - \boldsymbol{m}_1\|^2 = \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)^2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}^2 = 1$$

$$SE_{x2} = \|\boldsymbol{x2} - \boldsymbol{m}_1\|^2 = \left(\begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)^2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}^2 = 1$$

$$SE_{x3} = \|\boldsymbol{x3} - \boldsymbol{m}_1\|^2 = \ldots = 1$$

$$SE_{x4} = \|\boldsymbol{x4} - \boldsymbol{m}_1\|^2 = \ldots = 1$$

$$J_{X1} = WCSE_{X1} = \sum_{\boldsymbol{x} \in X_1} \|\boldsymbol{x} - \boldsymbol{m}_1\|^2 = 4 \quad \text{(squared errors within cluster } X_1\text{)}$$

Partitional clustering example (partition I, continuation)

Calculations for cluster X_2:

$$SE_{x5} = \|\boldsymbol{x5} - \boldsymbol{m}_2\|^2 = \left(\begin{pmatrix} 5 \\ 2 \end{pmatrix} - \begin{pmatrix} 7 \\ 2 \end{pmatrix}\right)^2 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}^2 = 4$$

$$SE_{x6} = \|\boldsymbol{x6} - \boldsymbol{m}_2\|^2 = \left(\begin{pmatrix} 7 \\ 3 \end{pmatrix} - \begin{pmatrix} 7 \\ 2 \end{pmatrix}\right)^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^2 = 1$$

$$SE_{x7} = \|\boldsymbol{x7} - \boldsymbol{m}_2\|^2 = \ldots = 2$$

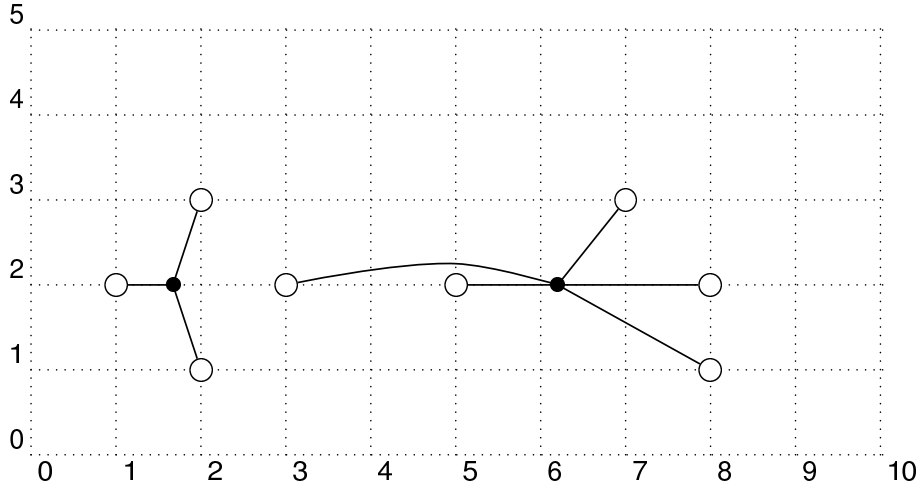
$$SE_{x8} = \|\boldsymbol{x8} - \boldsymbol{m}_2\|^2 = \ldots = 1$$

$$J_{X2} = WCSE_{X2} = \sum_{\boldsymbol{x} \in X_2} \|\boldsymbol{x} - \boldsymbol{m}_2\|^2 = 8 \quad \text{(squared errors within cluster } X_2\text{)}$$

Sum of squared errors for all clusters in partition Π_1:

$$J(\Pi_1) = J(X_1, X_2) = SSE_{\Pi_1} = \sum_c WCSE_c = 12$$

Partitional clustering example (partition II)



SSE> 12

Partition II (Π_2): 8 samples $(x_1, \ldots, x_N)$, 2 clusters in partition $\Pi_2 = \{X_1, X_2\}$

$$X_1 = \{x_1, \ldots, x_3\} : x_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, x_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, x_3 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

$$X_2 = \{x_4, \ldots, x_8\} : x_4 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, x_5 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}, x_6 = \begin{pmatrix} 7 \\ 3 \end{pmatrix}, x_7 = \begin{pmatrix} 8 \\ 1 \end{pmatrix}, x_8 = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$$

$$\boldsymbol{m}_1 = \frac{1}{3} \sum_{x \in X_1} x = \frac{1}{3} \begin{pmatrix} 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 1.66 \\ 2 \end{pmatrix} \qquad \boldsymbol{m}_2 = \frac{1}{5} \sum_{x \in X_2} x = \frac{1}{5} \begin{pmatrix} 31 \\ 10 \end{pmatrix} = \begin{pmatrix} 6.2 \\ 2 \end{pmatrix}$$

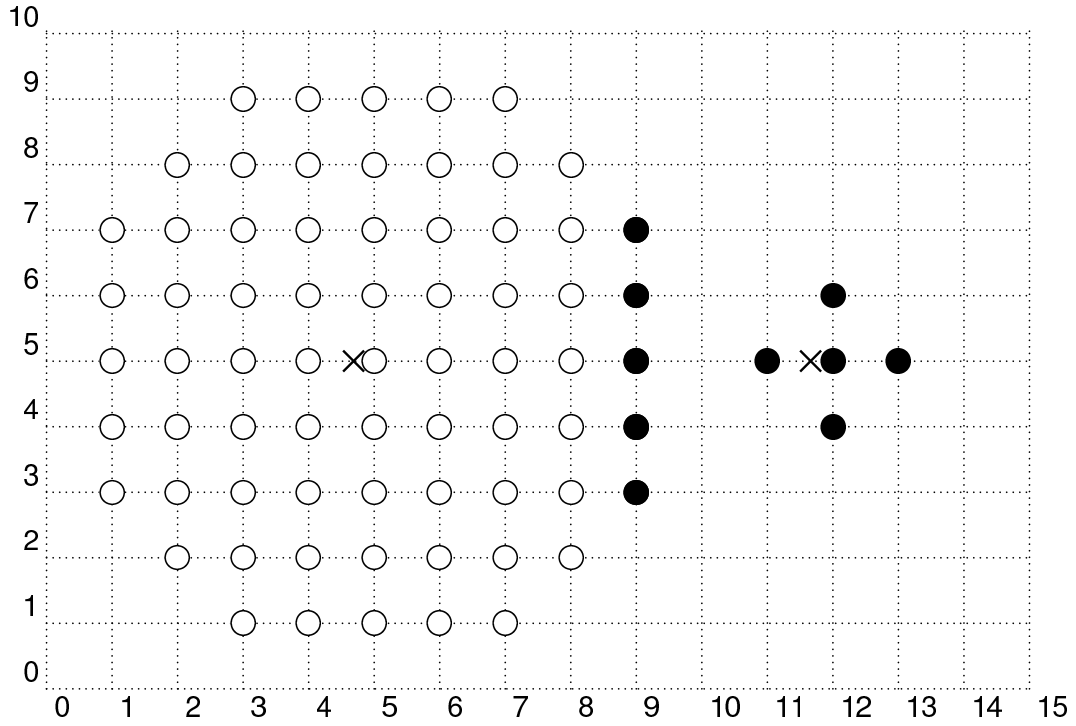
Exercise: calculate the values for $WCSE_{X1}$ and $WCSE_{X2}$ in partition Π_2 and check that the value of $SSE_{\Pi 2}$ is higher than 12.

Goodness of the SSE criterion

The SSE criterion is appropriate only when the objects form ***hyper-spherics clusters of similar size.***

If the sizes of the clusters are too different, it may happen that the natural grouping does not give the minimum SSE and so we would not find the natural clusters.

Therefore, the centroid representation alone works well if the clusters are of the hyper-spherical shape. If clusters are elongated or are of other shapes, centroids are not sufficient.



Índice

Incremental computation of SSE after moving x from cluster X_i to X_j

$$X_i' = X_i - \{\boldsymbol{x}\} \qquad X_j' = X_j + \{\boldsymbol{x}\}$$

$$\boldsymbol{m}_i' = \boldsymbol{m}_i - \frac{\boldsymbol{x} - \boldsymbol{m}_i}{n_i - 1} \qquad \boldsymbol{m}_j' = \boldsymbol{m}_j + \frac{\boldsymbol{x} - \boldsymbol{m}_j}{n_j + 1}$$

$$J_i' = J_i - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 \qquad J_j' = J_j + \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2$$

$$\triangle J = \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$

The movement will be successful if the increment of SSE is negative; that is:

$$\frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 < \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 \tag{5}$$

These equations allow to minimize the SSE by using successive refinements from an initial partition.

SSE optimization: algorithm C-means
(K-means in English bibliography)

Input:

- $X = (x_1, x_2, \ldots, x_N)$, the N objects/observations
- C, the number of clusters
- $\Pi_0 = \{X_1, \ldots, X_C\}$, a randomly initial partition that classifies the N objects into C clusters; these will be the initial centroids (cluster centers)

Output:

- Π^*, the best partition
- $\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_C$, the cluster centers for Π^*
- J (optionally, it depends on the C-means version)

Goal:

The data given by X is clustered by C-means algorithm, which aims to partition the N objects into C clusters such that the sum of the squared errors from points to the assigned cluster is minimized.

Algorithm *K-means* ("correct" version [Duda & Hart])

Input: $X = (x_1, x_2, \ldots, x_N)$; C; $\Pi_0 = \{X_1, \ldots, X_C\}$; (random initial partition)

Output: $\Pi^* = \{X_1, \ldots, X_C\}$; $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_C$; $J(\Pi^*)$

for $c = 1$ **to** C **do** $\boldsymbol{m}_c = \frac{1}{n_c} \sum_{\boldsymbol{x} \in X_c} \boldsymbol{x}$ **endfor** (compute initial centroids for the clusters in input partition)

compute $J = J(\Pi_0)$

repeat

$transfers =$ false

forall $\boldsymbol{x} \in X$ (let $i : \boldsymbol{x} \in X_i$) **do** (for all the objects ... assume object x belongs to cluster Xi)

if $n_i > 1$ **then** (if cluster Xi has more than one object, not only x)

$j^* = \arg\min_{j \neq i} \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2$ (study x in any other cluster different from Xi; take the min. cluster X_{j*})

$\triangle J = \frac{n_{j^*}}{n_{j^*} + 1} \|\boldsymbol{x} - \boldsymbol{m}_{j^*}\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$ (increment of SSE, difference bt. X_{j*} and current cluster)

if $\triangle J < 0$ **then**

$transfers =$ true (move x to X_{j*})

$\boldsymbol{m}_i = \boldsymbol{m}_i - \frac{\boldsymbol{x} - \boldsymbol{m}_i}{n_i - 1}$ $\quad \boldsymbol{m}_{j^*} = \boldsymbol{m}_{j^*} + \frac{\boldsymbol{x} - \boldsymbol{m}_{j^*}}{n_{j^*} + 1}$

$X_i = X_i - \{\boldsymbol{x}\}$ $\quad X_{j^*} = X_{j^*} + \{\boldsymbol{x}\}$

$J^* = J + \triangle J$ $\quad J = J^*$ (J value of the best partition so far ...)

endif

endif

endforall

until $\neg transfers$ Cost per iteration: $O(N \cdot C \cdot D)$, $N = |X|$, D = dimension

Example of C-means application (Duda & Hart) (1)

$X = \{1, 3, 4.5\} \subset \mathbb{R}$; $x_1 = (1), x_2 = (3), x_3 = (4.5)$
$C = 2$ clusters; $X_1 = \{x_1, x_2\}$, $X_2 = \{x_3\}$ $\Pi^0 = \{\{1, 3\}, \{4.5\}\}$
$\boldsymbol{m}_1 = 2$ $J_1 = 2$ $\boldsymbol{m}_2 = 4.5$ $J_2 = 0$ $J(\Pi^0) = 2$

$x = x_1 = 1$ (Cluster X_1, $i = 1$, $n_1 = 2$: two objects in X_1)

A) we have only one more cluster (X_2) so $\arg\min_{j \neq i}$ only analyzes cluster $j = 2$ ($n_j = n_2 = 1$).

B)

$$j^* = \arg\min_{j \neq i} \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 = \arg\min_{j \neq i} \frac{n_2}{n_2 + 1} \|\boldsymbol{x_1} - \boldsymbol{m}_2\|^2 =$$

$$\frac{1}{2} \|1 - 4.5\|^2 = 6.125$$

C)

$$\triangle J = \frac{n_{j^*}}{n_{j^*} + 1} \|\boldsymbol{x} - \boldsymbol{m}_{j^*}\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 =$$

$$\frac{n_2}{n_2 + 1} \|\boldsymbol{x_1} - \boldsymbol{m}_2\|^2 - \frac{n_1}{n_1 - 1} \|\boldsymbol{x_1} - \boldsymbol{m}_1\|^2 =$$

$$6.125 - \frac{2}{1} \|1 - 2\|^2 = 6.125 - 2 = 4.125$$

D) x_1 is not transferred to X_2

Example of C-means application (Duda & Hart) (2)

$x = x_2 = 3$ (Cluster X_1, $i = 1$, $n_1 = 2$: two objects in X_1)

A) we have only one more cluster (X_2) so $\arg\min_{j \neq i}$ only analyzes cluster $j = 2$ ($n_j = n_2 = 1$).

B)

$$j^* = \arg\min_{j \neq i} \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 = \arg\min_{j \neq i} \frac{n_2}{n_2 + 1} \|\boldsymbol{x_2} - \boldsymbol{m}_2\|^2 = \frac{1}{2} \|3 - 4.5\|^2 = 1.125$$

C)

$$\triangle J = \frac{n_{j^*}}{n_{j^*} + 1} \|\boldsymbol{x} - \boldsymbol{m}_{j^*}\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 = \frac{n_2}{n_2 + 1} \|\boldsymbol{x_2} - \boldsymbol{m}_2\|^2 - \frac{n_1}{n_1 - 1} \|\boldsymbol{x_2} - \boldsymbol{m}_1\|^2 = 1.125 - \frac{2}{1}\|3 - 2\|^2 = 1.125 - 2 = -0.875$$

D) $transfers = true$, we move x_2 to cluster X_2

E)

$$\boldsymbol{m}_i = \boldsymbol{m}_i - \frac{\boldsymbol{x} - \boldsymbol{m}_i}{n_i - 1} = \boldsymbol{m}_1 - \frac{\boldsymbol{x_2} - \boldsymbol{m}_1}{n_1 - 1} = 2 - \frac{3 - 2}{2 - 1} = 1 \quad (m_1 = 1)$$

$$\boldsymbol{m}_{j^*} = \boldsymbol{m}_{j^*} + \frac{\boldsymbol{x} - \boldsymbol{m}_{j^*}}{n_{j^*} + 1} = \boldsymbol{m}_2 + \frac{\boldsymbol{x_2} - \boldsymbol{m}_2}{n_2 + 1} = 4.5 + \frac{3 - 4.5}{1 + 1} = 3.75 \quad (m_2 = 3.75)$$

Example of C-means application (Duda & Hart) (3)

F) We update the clusters: $X_1 = \{x_1\}, X_2 = \{x_2, x_3\}$

G) $J = J + \triangle J = 2 - 0.875 = 1.125$

$x = x_3 = 4.5$ (Cluster X_2, $i = 2$, $n_2 = 2$: two objects in X_2)

A) we have only one more cluster (X_1) so $\arg\min_{j \neq i}$ only analyzes cluster $j = 1$ ($n_j = n_1 = 1$).

B)

$$j^* = \arg\min_{j \neq i} \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 = \arg\min_{j \neq i} \frac{n_1}{n_1 + 1} \|\boldsymbol{x_3} - \boldsymbol{m}_1\|^2 = \frac{1}{2} \|4.5 - 1\|^2 = 6.125$$

C)

$$\triangle J = \frac{n_{j^*}}{n_{j^*} + 1} \|\boldsymbol{x} - \boldsymbol{m}_{j^*}\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 = \frac{n_1}{n_1 + 1} \|\boldsymbol{x_3} - \boldsymbol{m}_1\|^2 - \frac{n_2}{n_2 - 1} \|\boldsymbol{x_3} - \boldsymbol{m}_2\|^2 = 6.125 - \frac{2}{1} \|4.5 - 3.75\|^2 = 6.125 - 1.125 = 5$$

D) x_3 is not transferred to X_1

Example of C-means application (Duda & Hart) (4)

The algorithm makes one more iteration of the loop 'repeat ... until' because $transfers$ is true; in the next iteration of the algorithm no modifications are produced.

Final result: $\Pi^* = \{X_1, X_2\}, X_1 = \{x_1\}, X_2 = \{x_2, x_3\}, J(\Pi^*) = SSE_{\Pi^*} = 1.125$

SSE optimization: another version of C-means

Algorithm *C-means* ("popular" version)
Input: X; C; $\Pi = \{X_1, \ldots, X_C\}$;
Output: $\Pi^* = \{X_1, \ldots, X_C\}$; $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_C$
repeat
  $transfers =$ false
  for $c = 1$ **to** C **do** $\boldsymbol{m}_c = \frac{1}{n_c} \sum_{\boldsymbol{x} \in X_c} \boldsymbol{x}$ **endfor**
  forall $\boldsymbol{x} \in X$ (let $i : \boldsymbol{x} \in X_i$) **do**
    if $n_i > 1$ **then**
      $j^* = \arg\min_{1 \le j \le C} d(\boldsymbol{x}, \boldsymbol{m}_j)$ (Euclidean distance from x to each centroid; $d(\boldsymbol{x}, \boldsymbol{m}_j) = \|\boldsymbol{x} - \boldsymbol{m}_j)\|$)
      (study x in all clusters, including current cluster, and take the min. cluster)
      if $j^* \neq i$ **then** (if the min. cluster is different from the current cluster, move x to the min. cluster)
        $transfers =$ true
        $X_i = X_i - \{\boldsymbol{x}\}$; $X_{j^*} = X_{j^*} + \{\boldsymbol{x}\}$
      endif
    endif
  endforall
until $\neg transfers$ (Cost per iteration: $O(N \cdot C \cdot D)$, $N = |X|$, D = cost of $d(\cdot, \cdot)$)

The "popular" version does not apply increments/decrements to the mean (centroid) nor the SSE. It simply takes the Euclidean distance from x to each centroid (not the squared Euclidean distance), moves x to the minimum cluster and updates the objects in the current and minimum cluster. When all x are studied, it re-computes the mean points of all the clusters again.

Optimality of algorithms C-means

Exercise: Consider the same example in page 17. Apply the popular version of the C-means algorithm. The results are:

$\Pi^* = \{X_1, X_2\}, X_1 = \{x_1, x_2\}, X_2 = \{x_3\}, J(\Pi^*) = SSE_{\Pi^*} = 2.0$; that is:

$\Pi^* = \Pi^0, J(\Pi^*) = J(\Pi^0)$

- None of the two C-*means* versions guarantee a global minimum of the SSE
- The version presented in the book Duda & Hart obtains a *local* minimum
- The "popular" version does not guarantee the local minimization is some cases

The SSE criterion and Vectorial Quantification

A common way to represent clusters is the centroid representation: use the centroid of each cluster as the representative of the cluster $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_C$.

The next criteria to minimize are equivalent:

$$J(X_1, \ldots, X_C) = \sum_c \sum_{\boldsymbol{x} \in X_c} \|\boldsymbol{x} - \boldsymbol{m}_c\|^2 \tag{6}$$

$$J(X_1, \ldots, X_C; \boldsymbol{r}_1, \ldots, \boldsymbol{r}_C) = \sum_c \sum_{\boldsymbol{x} \in X_c} \|\boldsymbol{x} - \boldsymbol{r}_c\|^2 \tag{7}$$

$$J(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_C) = \sum_{\boldsymbol{x}} \min_c \|\boldsymbol{x} - \boldsymbol{r}_c\|^2 \tag{8}$$

Justification:

- (7) is equivalent to (6) because for all partitions $X_1, \ldots, X_C$, the cluster representatives $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_C$ that minimize (7) are the mean of the clusters.
- (7) is equivalent to (8) because for all representative set $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_C$, the partition that minimizes (7) is the one in which each object is assigned the nearest cluster representative.
- (8) is known as the problem of *vectorial quantification design* in *information theory*

Another interpretation of SSE

SSE can also be defined without using the *mean* of the clusters:

$$J(X_1, \ldots, X_C) = \frac{1}{2} \sum_c n_c \, \bar{s}_c \tag{9}$$

where n_c is the number of objects in X_c and $\bar{s}_c$ is the mean of the squared Euclidean distances for each pair of objects in the cluster.

$$\bar{s}_c = \frac{1}{n_c^2} \sum_{\boldsymbol{x}, \boldsymbol{x}' \in X_c} \|\boldsymbol{x} - \boldsymbol{x}'\|^2 \tag{10}$$

Thus, SSE can also be interpreted as a weighted average of intra-cluster squared distances.

Under this interpretation, it is possible to redefine $\bar{s}_c$ and obtain criteria similar to SSE (valid even for *non-vectorial* data):

$$\bar{s}_c = \frac{1}{n_c^2} \sum_{x, x' \in X_c} d(x, x') \qquad \bar{s}_c = \frac{1}{n_c^2} \operatorname{máx}_{x, x' \in X_c} d(x, x') \tag{11}$$