Representative-based Clustering

Clustering

- Clustering is the process of grouping similar items together in representative-based clusters
- The items are partitioned in k groups
- The dataset is described as n points in d-dimensional space, $\mathbf{D} = \{\mathbf{x}_i\}_{i=1}^{n}$
- The centroid is a point that represents the summary of the cluster

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Clustering continued

- The centroid is defined as $\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} \mathbf{x}_j$
- n_i represents the number of points in C_i
- Brute-force algorithms generate all n points into all possible partitions into k clusters
- These clusters are evaluated to determine some optimization score, the best score is retained

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Clustering continued

- The Stirling numbers of the second kind produces the exact number of partitions of n points
- The points are organized into k nonempty and disjointed groups

$$S(n,k) = \frac{1}{k!} \sum_{t=0}^{k} (-1)^t \binom{k}{t} (k-t)^n$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Clustering continued

- Since any point can be assigned to any of the k clusters, it is possible to have $O(k^n / k!)$ clusters
- This proves that brute-force is not a possible solution for clustering
- We overcome these issues with two approaches, K-means and expectation-maximization algorithms

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Scoring Function

$$SSE(\mathcal{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2$$

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \{SSE(\mathcal{C})\}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Two Steps of K-means

- Cluster assignment
- Centroid update
- Each point is assigned to the closest mean
- Each point is assigned to cluster C_{j*}

$$j^* = \arg\min_{i=1}^{k} \left\{ \left\| \mathbf{x}_j - \mu_i \right\|^2 \right\}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

K-means Algorithm

K-MEANS $(\mathbf{D}, k, \epsilon)$:

$t = 0$

Randomly initialize k centroids: $\mu_1^t, \mu_2^t, \ldots, \mu_k^t \in \mathbb{R}^d$

repeat

  $t \leftarrow t + 1$

  $C_j \leftarrow \emptyset$ for all $j = 1, \cdots, k$

  `// Cluster Assignment Step`

  foreach $\mathbf{x}_j \in \mathbf{D}$ **do**

    $j^* \leftarrow \arg\min_i \left\{ \left\| \mathbf{x}_j - \mu_i^{t-1} \right\|^2 \right\}$ `// Assign` $\mathbf{x}_j$ `to closest centroid`

    $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$

  `// Centroid Update Step`

  foreach $i = 1$ *to* k **do**

    $\mu_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$

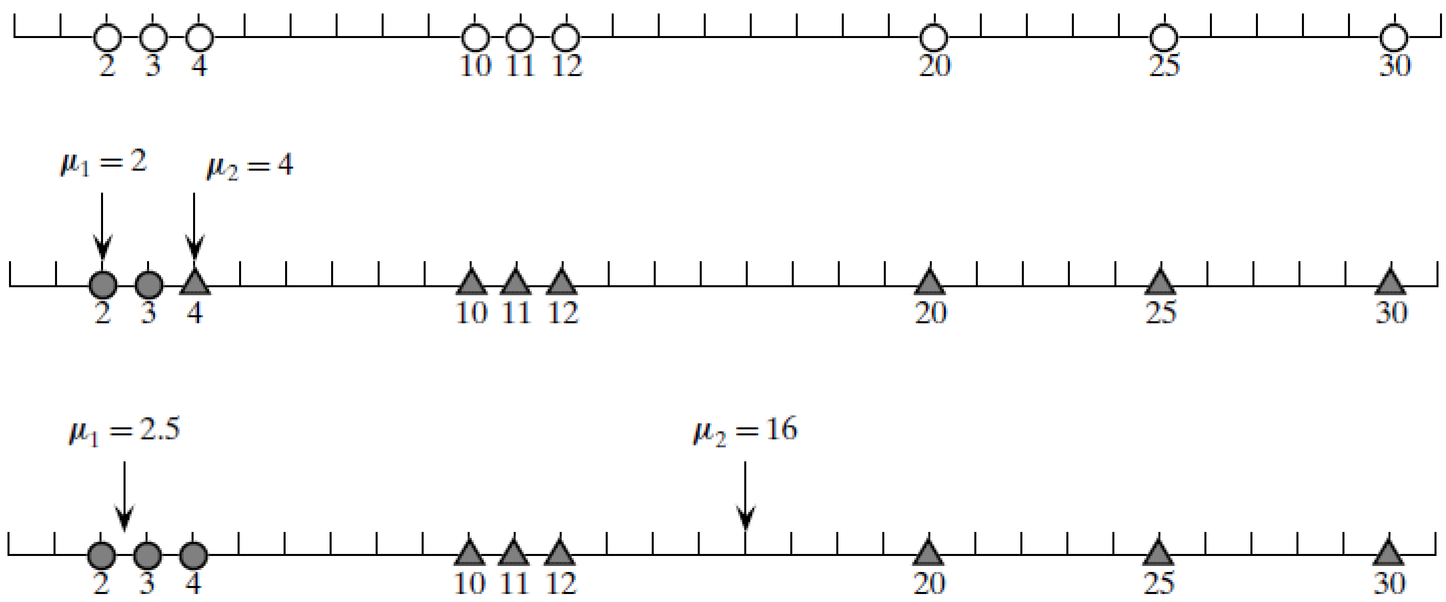
until $\sum_{i=1}^k \left\| \mu_i^t - \mu_i^{t-1} \right\|^2 \le \epsilon$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.1

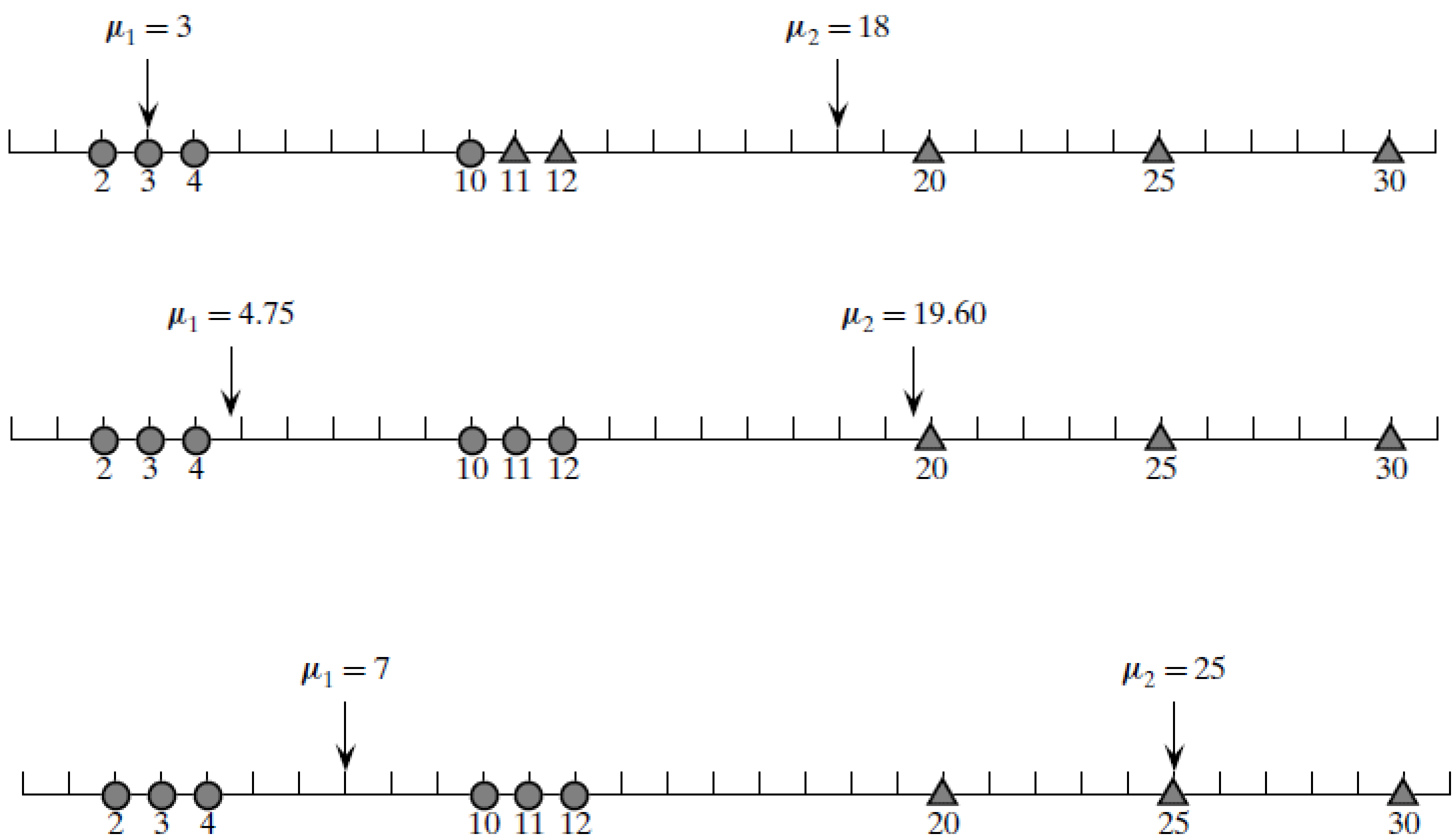
Given: $k = 2$ $\mu_1 = 2$ $\mu_2 = 4$

$C_1 = \{2, 3\}$ $C_2 = \{4, 10, 11, 12, 20, 25, 30\}$



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.1 continues

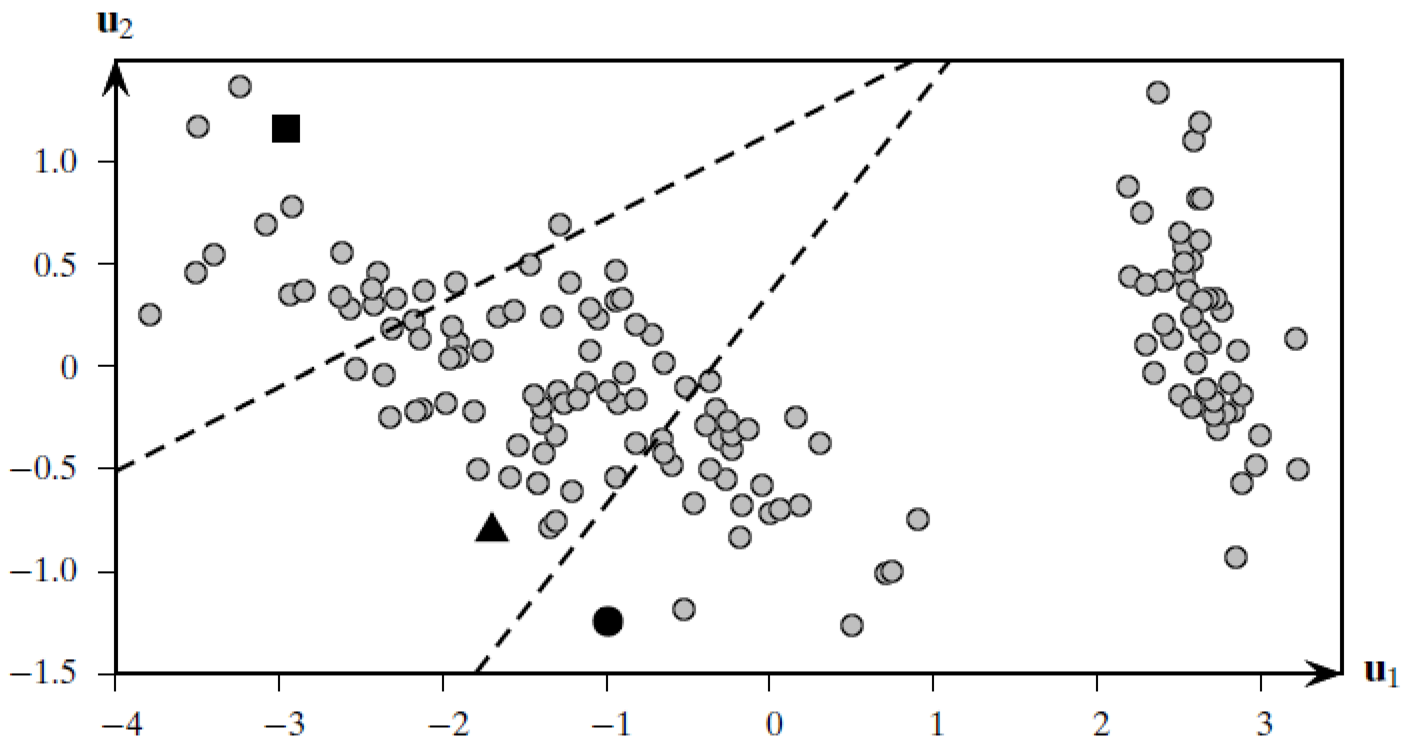


Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.2

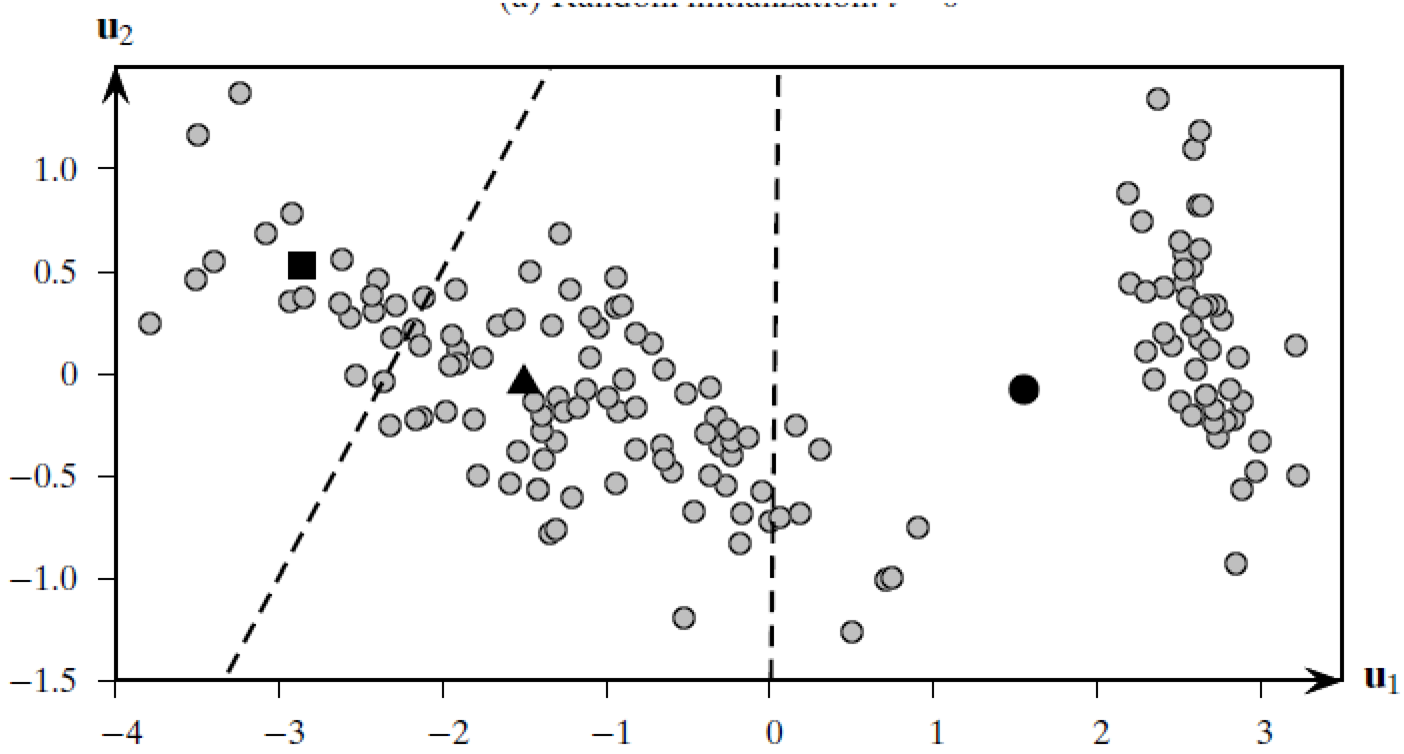
Given: $n = 150$ $k = 3$ $\mu_1 = (-0.98, -1.24)^T$

$\mu_2 = (-2.96, 1.16)^T$ $\mu_3 = (-1.69, -0.80)^T$



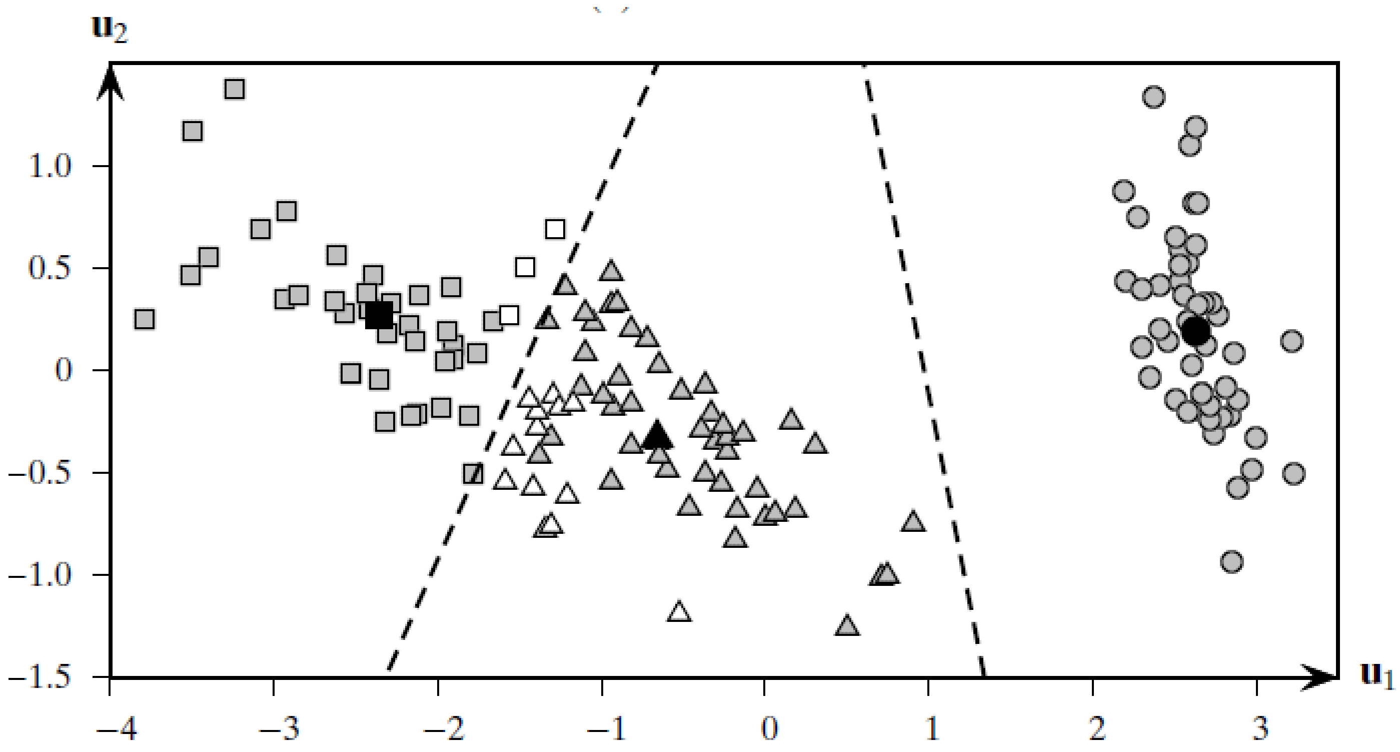
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.2 continue



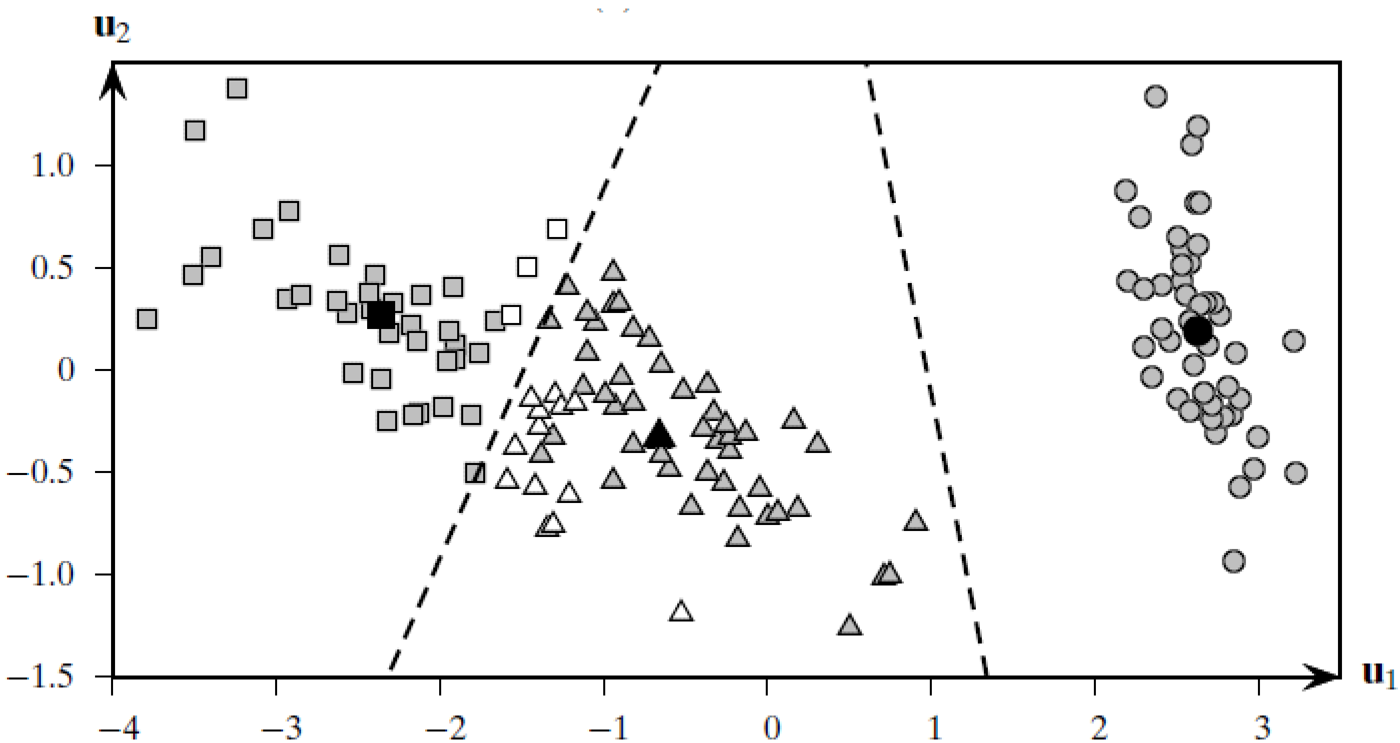
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.2 continue



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.2 continue



$\mu_1 = (2.64, 0.19)^T$ $\mu_2 = (-2.35, 0.27)^T$ $\mu_3 = (-0.66, -0.33)^T$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Representative-based Clustering

The End

Kernel K-means

Kernel K-means

- Kernel K-means allow for non-linear boundaries
- This technique detects nonconvex clusters
- The technique maps data points high-dimensional space using non-linear mapping
- The kernel trick allow for feature space to be explored by the function using dot product

Kernel K-means continued

$$\mu_i^{\phi} = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} \phi(\mathbf{x}_j)$$

$$\min_{\mathcal{C}} SSE(\mathcal{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \left\| \phi(\mathbf{x}_j) - \mu_i^{\phi} \right\|^2$$

$$SSE(\mathcal{C}) = \sum_{j=1}^{n} K(\mathbf{x}_j, \mathbf{x}_j) - \sum_{i=1}^{k} \frac{1}{n_i} \sum_{\mathbf{x}_a \in C_i} \sum_{\mathbf{x}_b \in C_i} K(\mathbf{x}_a, \mathbf{x}_b)$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Computing the Mean

$$
\begin{aligned}
\left\|\phi(\mathbf{x}_j)-\boldsymbol{\mu}_i^{\phi}\right\|^2 &= \left\|\phi(\mathbf{x}_j)\right\|^2 - 2\phi(\mathbf{x}_j)^T\boldsymbol{\mu}_i^{\phi} + \left\|\boldsymbol{\mu}_i^{\phi}\right\|^2 \\
&= \phi(\mathbf{x}_j)^T\phi(\mathbf{x}_j) - \frac{2}{n_i}\sum_{\mathbf{x}_a\in C_i}\phi(\mathbf{x}_j)^T\phi(\mathbf{x}_a) + \frac{1}{n_i^2}\sum_{\mathbf{x}_a\in C_i}\sum_{\mathbf{x}_b\in C_i}\phi(\mathbf{x}_a)^T\phi(\mathbf{x}_b) \\
&= K(\mathbf{x}_j,\mathbf{x}_j) - \frac{2}{n_i}\sum_{\mathbf{x}_a\in C_i}K(\mathbf{x}_a,\mathbf{x}_j) + \frac{1}{n_i^2}\sum_{\mathbf{x}_a\in C_i}\sum_{\mathbf{x}_b\in C_i}K(\mathbf{x}_a,\mathbf{x}_b) \qquad (1
\end{aligned}
$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Closest Cluster Mean

$$C^*(\mathbf{x}_j) = \arg\min_i \left\{ \left\| \phi(\mathbf{x}_j) - \mu_i^\phi \right\|^2 \right\}$$

$$= \arg\min_i \left\{ K(\mathbf{x}_j, \mathbf{x}_j) - \frac{2}{n_i} \sum_{\mathbf{x}_a \in C_i} K(\mathbf{x}_a, \mathbf{x}_j) + \frac{1}{n_i^2} \sum_{\mathbf{x}_a \in C_i} \sum_{\mathbf{x}_b \in C_i} K(\mathbf{x}_a, \mathbf{x}_b) \right\}$$

$$= \arg\min_i \left\{ \frac{1}{n_i^2} \sum_{\mathbf{x}_a \in C_i} \sum_{\mathbf{x}_b \in C_i} K(\mathbf{x}_a, \mathbf{x}_b) - \frac{2}{n_i} \sum_{\mathbf{x}_a \in C_i} K(\mathbf{x}_a, \mathbf{x}_j) \right\}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Kernel K-means Algorithm

KERNEL-KMEANS($\mathbf{K}, k, \epsilon$):

$t \leftarrow 0$

$\mathcal{C}^t \leftarrow \{C_1^t, \ldots, C_k^t\}$ `// Randomly partition points into` k `clusters`

repeat

    $t \leftarrow t + 1$

    foreach $C_i \in \mathcal{C}^{t-1}$ **do** `// Compute squared norm of cluster means`

        $\text{sqnorm}_i \leftarrow \frac{1}{n_i^2} \sum_{\mathbf{x}_a \in C_i} \sum_{\mathbf{x}_b \in C_i} K(\mathbf{x}_a, \mathbf{x}_b)$

    foreach $\mathbf{x}_j \in \mathbf{D}$ **do** `// Average kernel value for` $\mathbf{x}_j$ `and` C_i

        foreach $C_i \in \mathcal{C}^{t-1}$ **do**

            $\text{avg}_{ji} \leftarrow \frac{1}{n_i} \sum_{\mathbf{x}_a \in C_i} K(\mathbf{x}_a, \mathbf{x}_j)$

    `// Find closest cluster for each point`

    foreach $\mathbf{x}_j \in \mathbf{D}$ **do**

        foreach $C_i \in \mathcal{C}^{t-1}$ **do**

            $d(\mathbf{x}_j, C_i) \leftarrow \text{sqnorm}_i - 2 \cdot \text{avg}_{ji}$

        $j^* \leftarrow \arg\min_i \left\{ d(\mathbf{x}_j, C_i) \right\}$

        $C_{j^*}^t \leftarrow C_{j^*}^t \cup \{\mathbf{x}_j\}$ `// Cluster reassignment`

    $\mathcal{C}^t \leftarrow \left\{ C_1^t, \ldots, C_k^t \right\}$

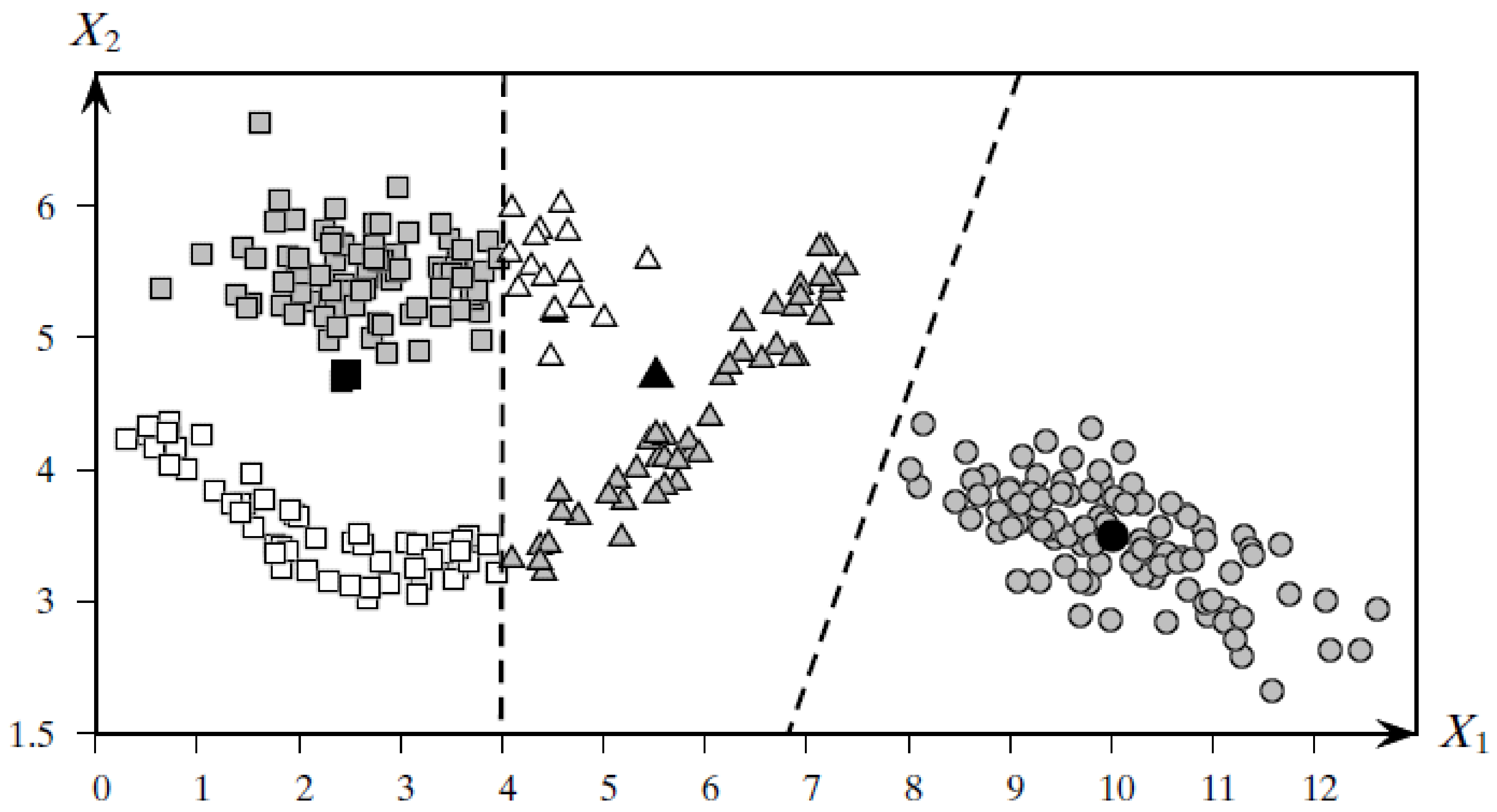
until $1 - \frac{1}{n} \sum_{i=1}^{k} \left| C_i^t \cap C_i^{t-1} \right| \le \epsilon$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Computational Complexity

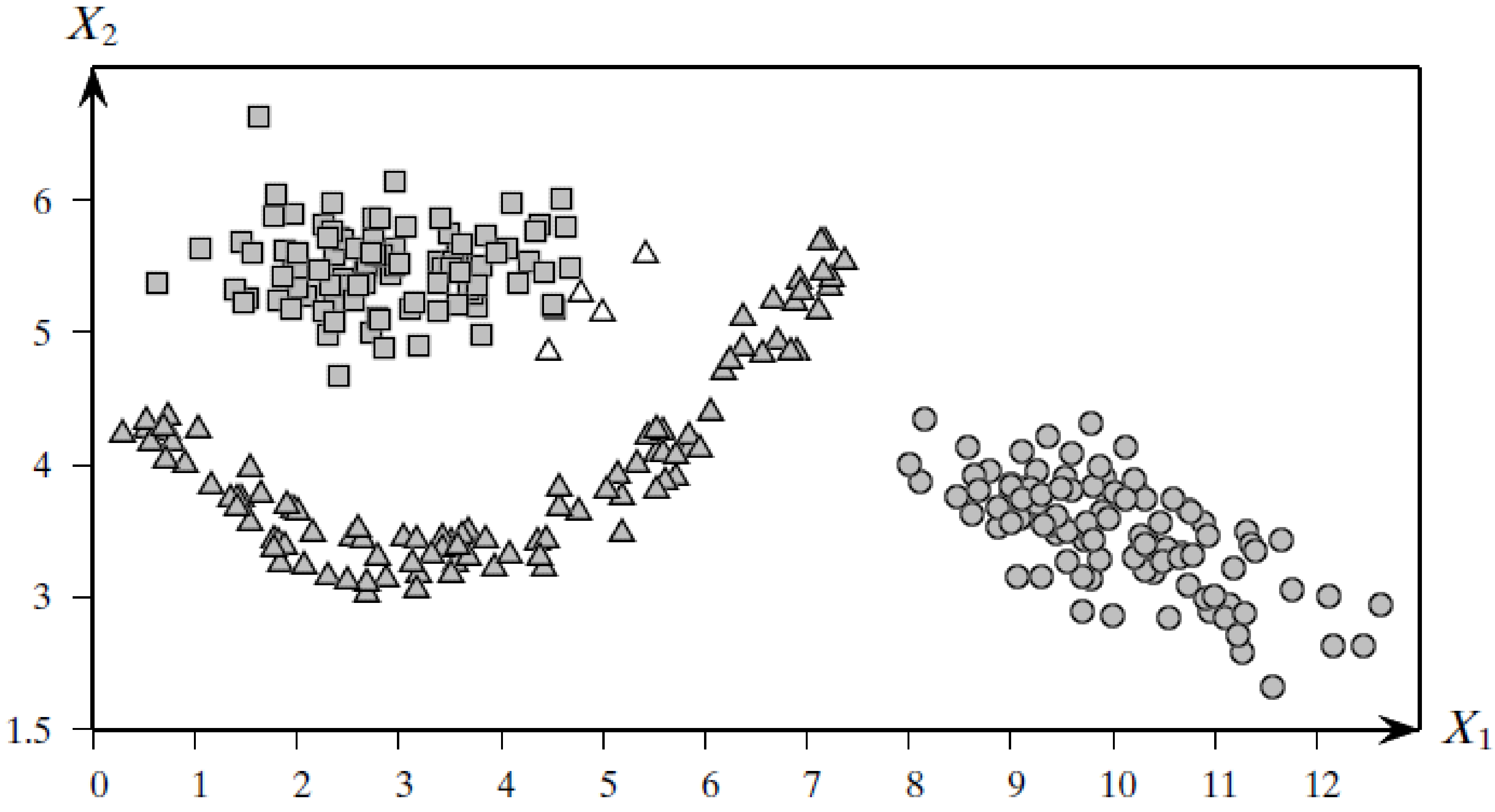
- K-means can be operationally expensive
- The It take the $O(n^2)$ to compute the average kernel value for all clusters.
- It take the $O(n^2)$ to compute the average kernel cluster for each k
- Total complexity is therefore $O(tn^2)$,where t is the number of iterations
- The I/O complexity is $O(t)$ scans of the kernel matrix

Example 13.3



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.3 continued



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Kernel K-means

The End