Hierarchical Clustering

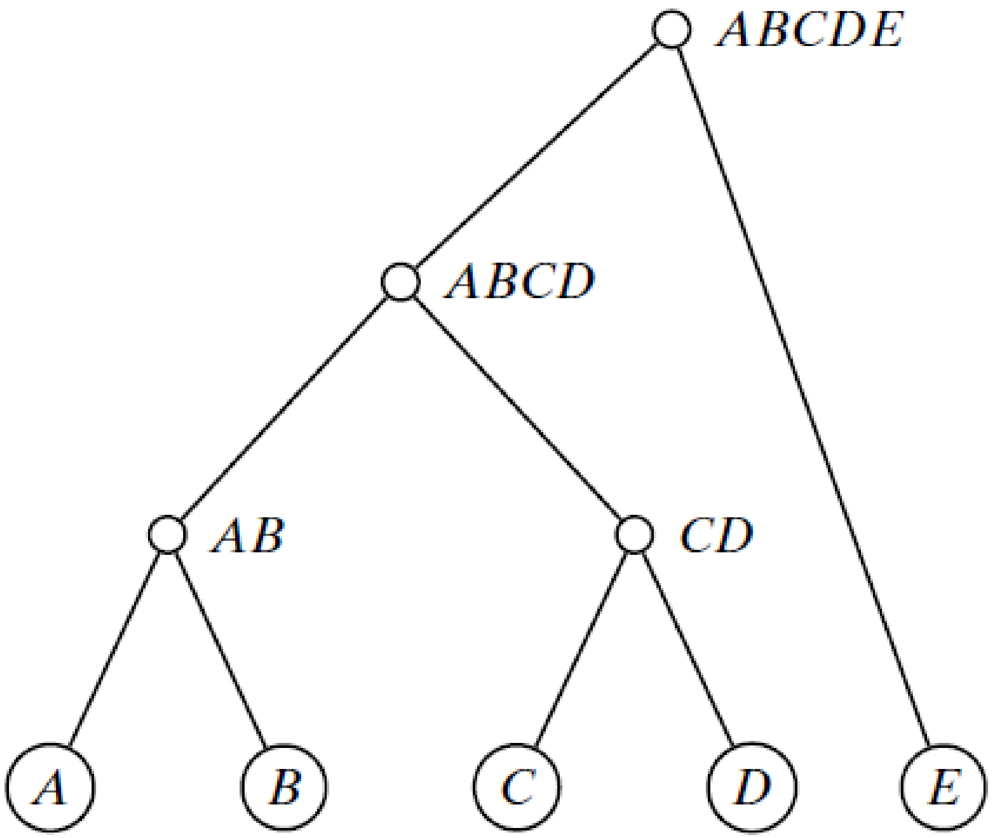
Hierarchical Clustering

- The goal of hierarchical clustering is to take n points in d-dimensional space to create nested partitions that can be visualized in a tree or hierarchy.
- Such clusters are referred to as dendrograms
- This bottom up approach builds from a single point in a leaf to all points in the root

Preliminaries

- Clusters are a partition of the entire dataset
- The hierarchy is built from the clusters that contain single points nested in the cluster that contains all points
- The dendrogram is a rooted binary tree that encompasses the nested structure

Example 14.1



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

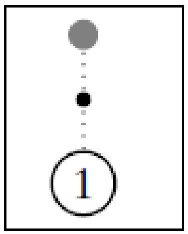
Number of Hierarchical Clusterings

- The hierarchical tree will have some number of nodes and one less number of edges
- A rooted binary tree will have some number of leaves and one less number of internal nodes
- The total number of dendrograms that can be derived from n leaves is the product

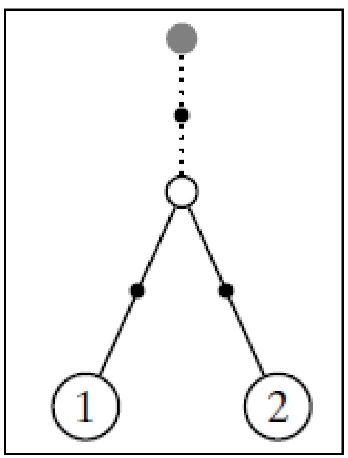
$$\prod_{m=1}^{n-1}(2m-1) = 1 \times 3 \times 5 \times 7 \times \cdots \times (2n-3) = (2n-3)!!$$

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

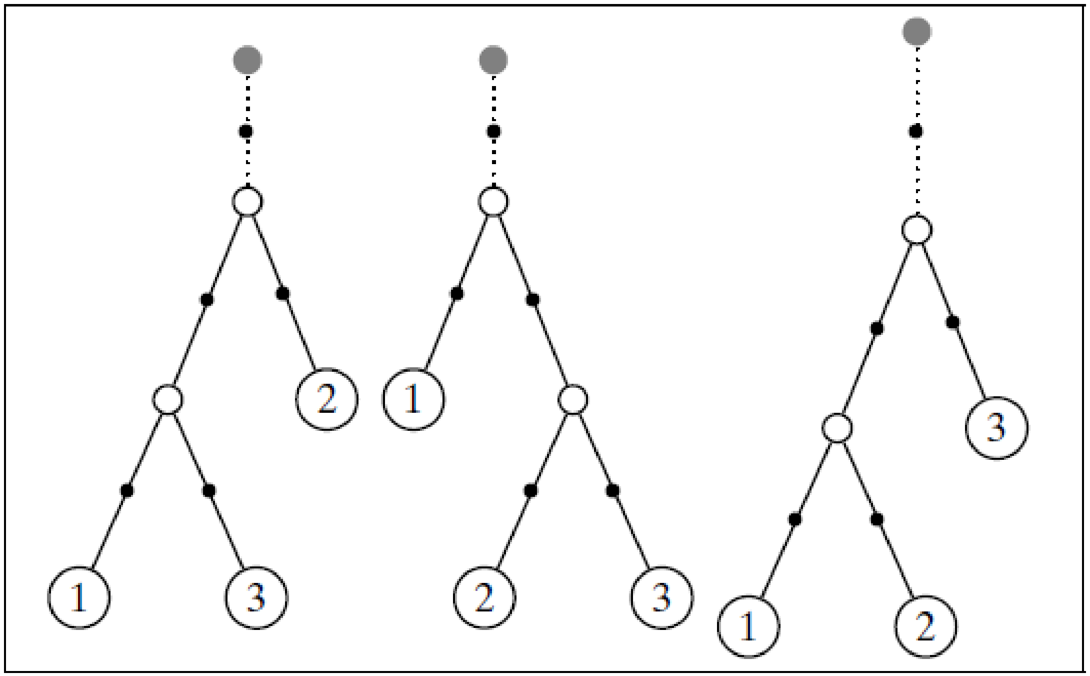
Example 14.2



(a) $n = 1$



(b) $n = 2$



(c) $n = 3$

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

Agglomerative Hierarchical Clustering

AGGLOMERATIVECLUSTERING$(\mathbf{D}, k)$:

$\mathcal{C} \leftarrow \{C_i = \{\mathbf{x}_i\} \mid \mathbf{x}_i \in \mathbf{D}\}$ `// Each point in separate cluster`

$\boldsymbol{\Delta} \leftarrow \{\delta(\mathbf{x}_i, \mathbf{x}_j) \colon \mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}\}$ `// Compute distance matrix`

repeat

- Find the closest pair of clusters $C_i, C_j \in \mathcal{C}$
- $C_{ij} \leftarrow C_i \cup C_j$ `// Merge the clusters`
- $\mathcal{C} \leftarrow \big(\mathcal{C} \setminus \{C_i, C_j\}\big) \cup \{C_{ij}\}$ `// Update the clustering`
- Update distance matrix $\boldsymbol{\Delta}$ to reflect new clustering

until $|\mathcal{C}| = k$

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

Hierarchical Clustering

The End

Distance Between Clusters

Distances

Euclidean Distance

$$\delta(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \left(\sum_{i=1}^{d} (x_i - y_i)^2\right)^{1/2}$$

Single Link

$$\delta(C_i, C_j) = \min\{\delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

Complete Link

$$\delta(C_i, C_j) = \max\{\delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

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

Average Distance

Group Average

$$\delta(C_i, C_j) = \frac{\sum_{\mathbf{x}\in C_i} \sum_{\mathbf{y}\in C_j} \delta(\mathbf{x}, \mathbf{y})}{n_i \cdot n_j}$$

Mean Distance

$$\delta(C_i, C_j) = \delta(\mu_i, \mu_j)$$

Where

$$\mu_i = \frac{1}{n_i} \sum_{\mathbf{x}\in C_i} \mathbf{x}.$$

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

Minimum Variance: Ward's Method

$$SSE_i = \sum_{\mathbf{x} \in C_i} \left\| \mathbf{x} - \mu_i \right\|^2$$

$$\begin{aligned} SSE_i &= \sum_{\mathbf{x} \in C_i} \left\| \mathbf{x} - \mu_i \right\|^2 \\ &= \sum_{\mathbf{x} \in C_i} \mathbf{x}^T \mathbf{x} - 2 \sum_{\mathbf{x} \in C_i} \mathbf{x}^T \mu_i + \sum_{\mathbf{x} \in C_i} \mu_i^T \mu_i \\ &= \left(\sum_{\mathbf{x} \in C_i} \mathbf{x}^T \mathbf{x} \right) - n_i \mu_i^T \mu_i \end{aligned}$$

$$SSE = \sum_{i=1}^{m} SSE_i = \sum_{i=1}^{m} \sum_{\mathbf{x} \in C_i} \left\| \mathbf{x} - \mu_i \right\|^2$$ For Clustering

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

Ward's Measure

$$\delta(C_i, C_j) = \Delta SSE_{ij} = SSE_{ij} - SSE_i - SSE_j$$

$$\begin{aligned}
\delta(C_i, C_j) &= \Delta SSE_{ij} \\
&= \sum_{\mathbf{z} \in C_{ij}} \|\mathbf{z} - \boldsymbol{\mu}_{ij}\|^2 - \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 - \sum_{\mathbf{y} \in C_j} \|\mathbf{y} - \boldsymbol{\mu}_j\|^2 \\
&= \sum_{\mathbf{z} \in C_{ij}} \mathbf{z}^T\mathbf{z} - n_{ij}\boldsymbol{\mu}_{ij}^T\boldsymbol{\mu}_{ij} - \sum_{\mathbf{x} \in C_i} \mathbf{x}^T\mathbf{x} + n_i\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i - \sum_{\mathbf{y} \in C_j} \mathbf{y}^T\mathbf{y} + n_j\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j \\
&= n_i\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + n_j\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j - (n_i + n_j)\boldsymbol{\mu}_{ij}^T\boldsymbol{\mu}_{ij}
\end{aligned}$$

$$\textstyle\sum_{\mathbf{z} \in C_{ij}} \mathbf{z}^T\mathbf{z} = \sum_{\mathbf{x} \in C_i} \mathbf{x}^T\mathbf{x} + \sum_{\mathbf{y} \in C_j} \mathbf{y}^T\mathbf{y}. \qquad \boldsymbol{\mu}_{ij} = \frac{n_i\boldsymbol{\mu}_i + n_j\boldsymbol{\mu}_j}{n_i + n_j}$$

$$\boldsymbol{\mu}_{ij}^T\boldsymbol{\mu}_{ij} = \frac{1}{(n_i + n_j)^2}\left(n_i^2\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + 2n_in_j\boldsymbol{\mu}_i^T\boldsymbol{\mu}_j + n_j^2\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j\right)$$

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

Ward's Measure

$$
\begin{aligned}
\delta(C_i, C_j) &= \Delta SSE_{ij} \\
&= n_i \mu_i^T \mu_i + n_j \mu_j^T \mu_j - \frac{1}{(n_i + n_j)} \left(n_i^2 \mu_i^T \mu_i + 2 n_i n_j \mu_i^T \mu_j + n_j^2 \mu_j^T \mu_j \right) \\
&= \frac{n_i (n_i + n_j) \mu_i^T \mu_i + n_j (n_i + n_j) \mu_j^T \mu_j - n_i^2 \mu_i^T \mu_i - 2 n_i n_j \mu_i^T \mu_j - n_j^2 \mu_j^T \mu_j}{n_i + n_j} \\
&= \frac{n_i n_j \left(\mu_i^T \mu_i - 2 \mu_i^T \mu_j + \mu_j^T \mu_j \right)}{n_i + n_j} \\
&= \left(\frac{n_i n_j}{n_i + n_j} \right) \left\| \mu_i - \mu_j \right\|^2
\end{aligned}
$$

$$
\delta(\mu_i, \mu_j) = \left\| \mu_i - \mu_j \right\|^2
$$

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

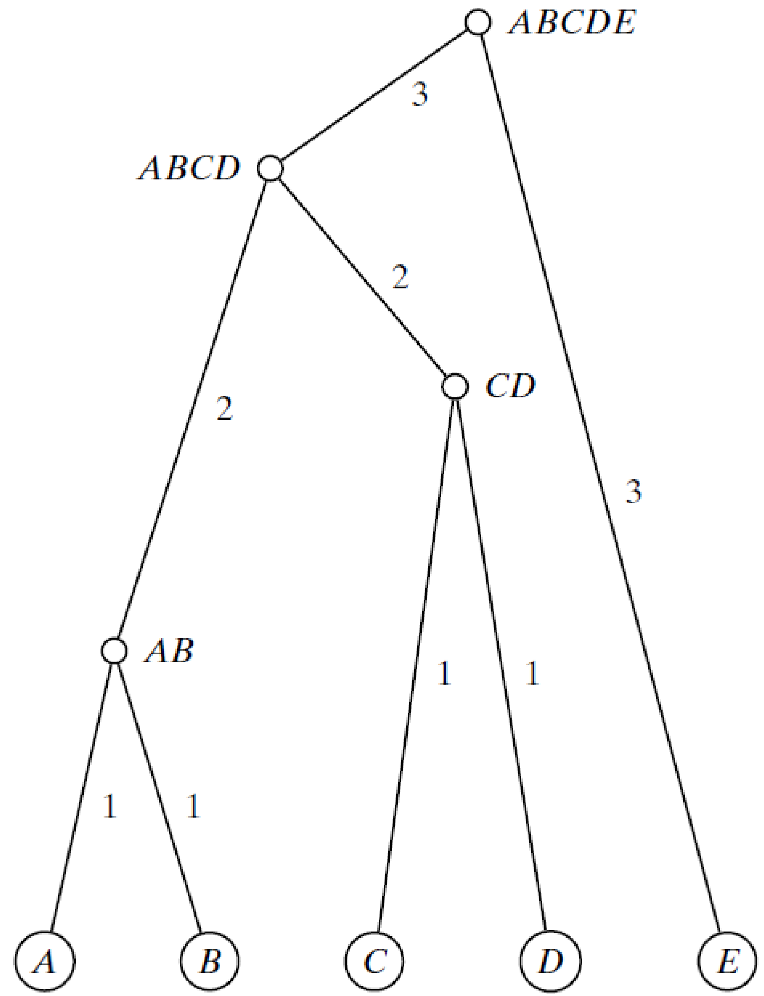
Example 14.3

δ	B	C	D	E
A	1	3	2	4
B		3	2	3
C			1	3
D				5

δ	C	D	E
AB	3	2	3
C		1	3
D			5

δ	CD	E
AB	2	3
CD		3

δ	E
ABCD	3



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

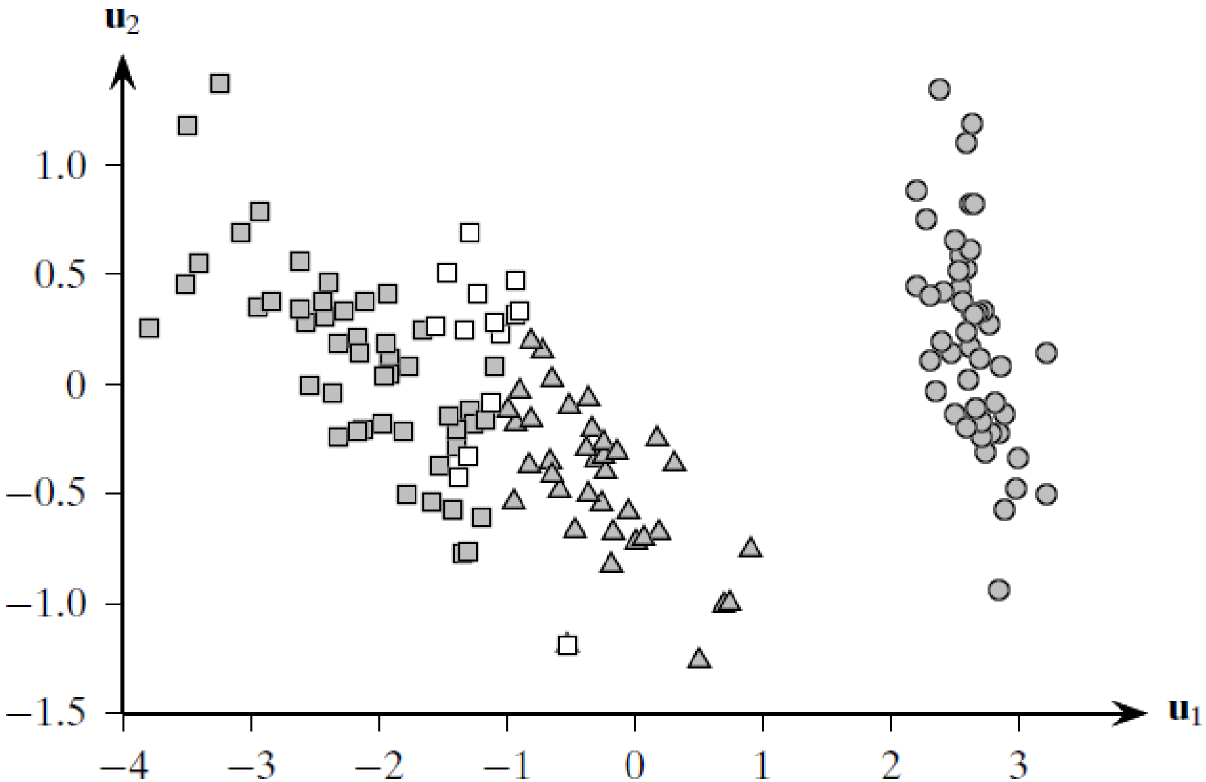
Updating Distance Matrix

$$\delta(C_{ij}, C_r) = \alpha_i \cdot \delta(C_i, C_r) + \alpha_j \cdot \delta(C_j, C_r) + \beta \cdot \delta(C_i, C_j) + \gamma \cdot \left|\delta(C_i, C_r) - \delta(C_j, C_r)\right|$$

Measure	α_i	α_j	β	γ
Single link	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete link	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Group average	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Mean distance	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i \cdot n_j}{(n_i+n_j)^2}$	0
Ward's measure	$\frac{n_i+n_r}{n_i+n_j+n_r}$	$\frac{n_j+n_r}{n_i+n_j+n_r}$	$\frac{-n_r}{n_i+n_j+n_r}$	0

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

Example 14.4



	iris-setosa	iris-virginica	iris-versicolor
C_1 (circle)	50	0	0
C_2 (triangle)	0	1	36
C_3 (square)	0	49	14

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

Computational Complexity

- The initial time to create the pairwise distance matrix is $O(n^2)$
- After each merge step the distances must be recalculated
- A heap data structure is created to store the minimum distances in $O(n^2)$
- Updating the heap takes $O(\log n)$ per operation
- Thus the total computations takes $O(n^2 \log n)$

Distance Between Clusters

The End

Density-based Clustering

Density-Based Clustering

- K-means and expectation-maximization method are great for mining ellipsoid and convex cluster, however, fail in non-convex space
- In non-convex space it is possible to have two points in different clusters closer than two points in the same cluster
- Density-based clustering overcomes this problem

Density-Based Scan

- DB clustering relies on the density of points in a cluster rather than the distance between the points
- DB clusters can be thought of as balls, the radius of the ball is referred to as ϵ
- An ϵ -neighborhood is defined as

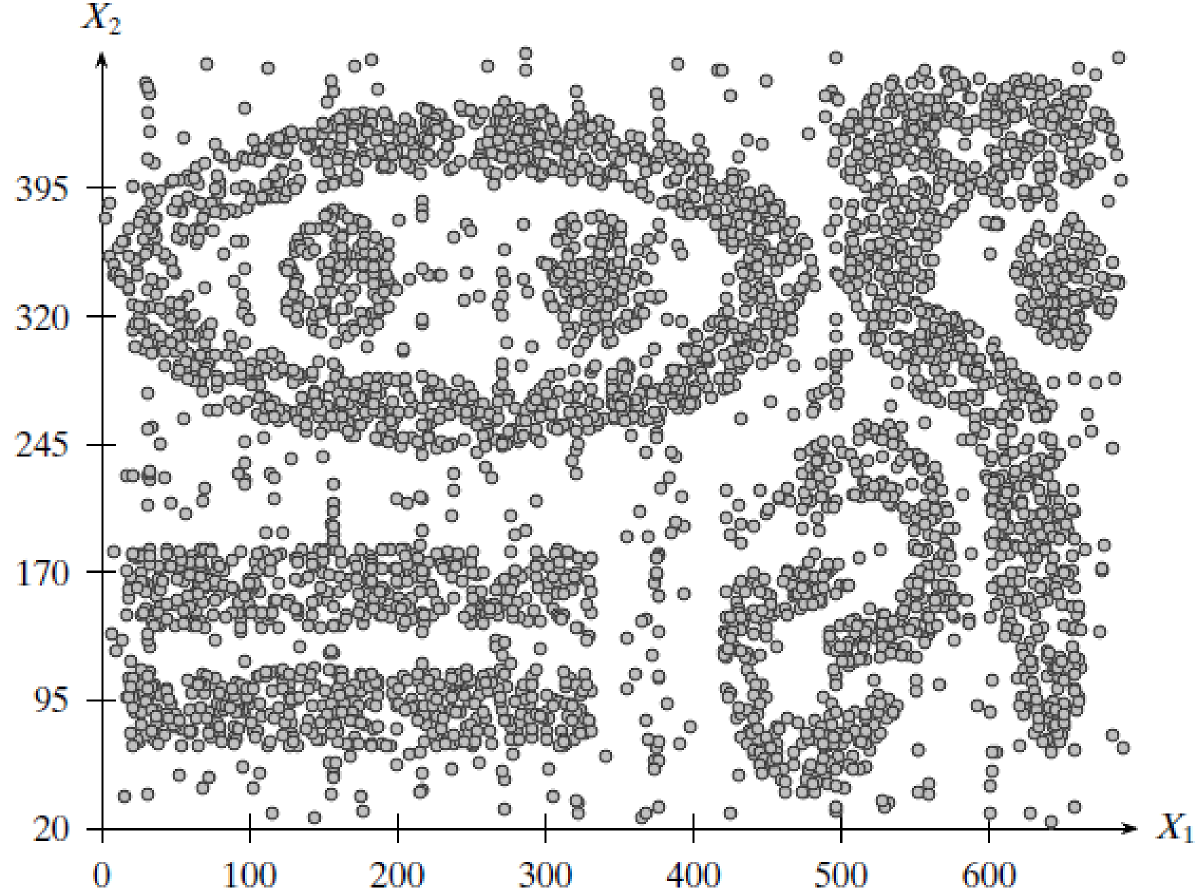
$$N_\epsilon(\mathbf{x}) = B_d(\mathbf{x}, \epsilon) = \{\mathbf{y} \mid \delta(\mathbf{x}, \mathbf{y}) \leq \epsilon\}$$

- The delta element is a distance measure, it can be any distance metric

Density-Based Scan

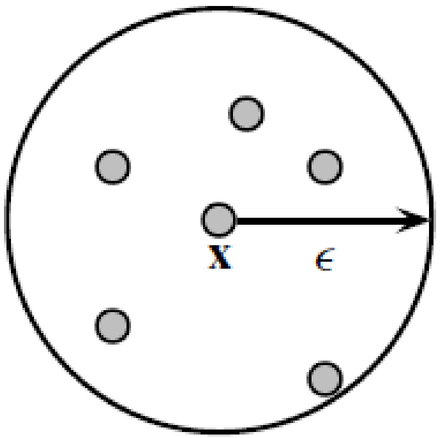
- A core point, is any point in the ϵ -neighborhood when minpts exist
- A minpts is user-defined density threshold
- If a point does not meet the minimum threshold of minpts, but it still belongs to the neighborhood, it is referred to as border point
- A point is a noise point, if it is not either a core or border point

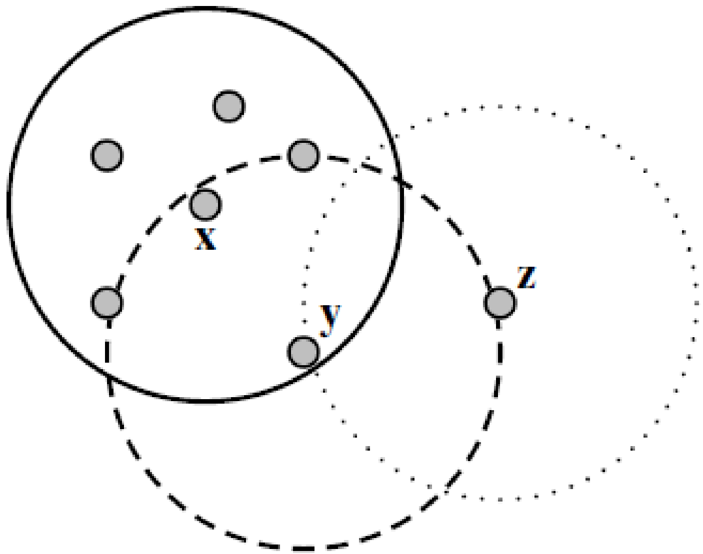
Density-based Dataset



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

Point Types





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

Point Types

- A point is directly density reachable to another point, if the other point is a core point
- A point is density reachable to another point, if there are a set of core points between the points
- Two points are density connected if they are density reachable to a core point
- A set of density connected points form a density-based cluster

Density-based Clustering Algorithm

DBSCAN $(\mathbf{D}, \epsilon, minpts)$:

$Core \leftarrow \emptyset$

foreach $\mathbf{x}_i \in \mathbf{D}$ **do** // Find the core points

  Compute $N_\epsilon(\mathbf{x}_i)$

  $id(\mathbf{x}_i) \leftarrow \emptyset$ // cluster id for $\mathbf{x}_i$

  if $N_\epsilon(\mathbf{x}_i) \geq minpts$ **then** $Core \leftarrow Core \cup \{\mathbf{x}_i\}$

$k \leftarrow 0$ // cluster id

foreach $\mathbf{x}_i \in Core$, *such that* $id(\mathbf{x}_i) = \emptyset$ **do**

  $k \leftarrow k + 1$

  $id(\mathbf{x}_i) \leftarrow k$ // assign $\mathbf{x}_i$ to cluster id k

  DensityConnected $(\mathbf{x}_i, k)$

$\mathcal{C} \leftarrow \{C_i\}_{i=1}^{k}$, where $C_i \leftarrow \{\mathbf{x} \in \mathbf{D} \mid id(\mathbf{x}) = i\}$

$Noise \leftarrow \{\mathbf{x} \in \mathbf{D} \mid id(\mathbf{x}) = \emptyset\}$

$Border \leftarrow \mathbf{D} \setminus \{Core \cup Noise\}$

return $\mathcal{C}, Core, Border, Noise$

DensityConnected $(\mathbf{x}, k)$:

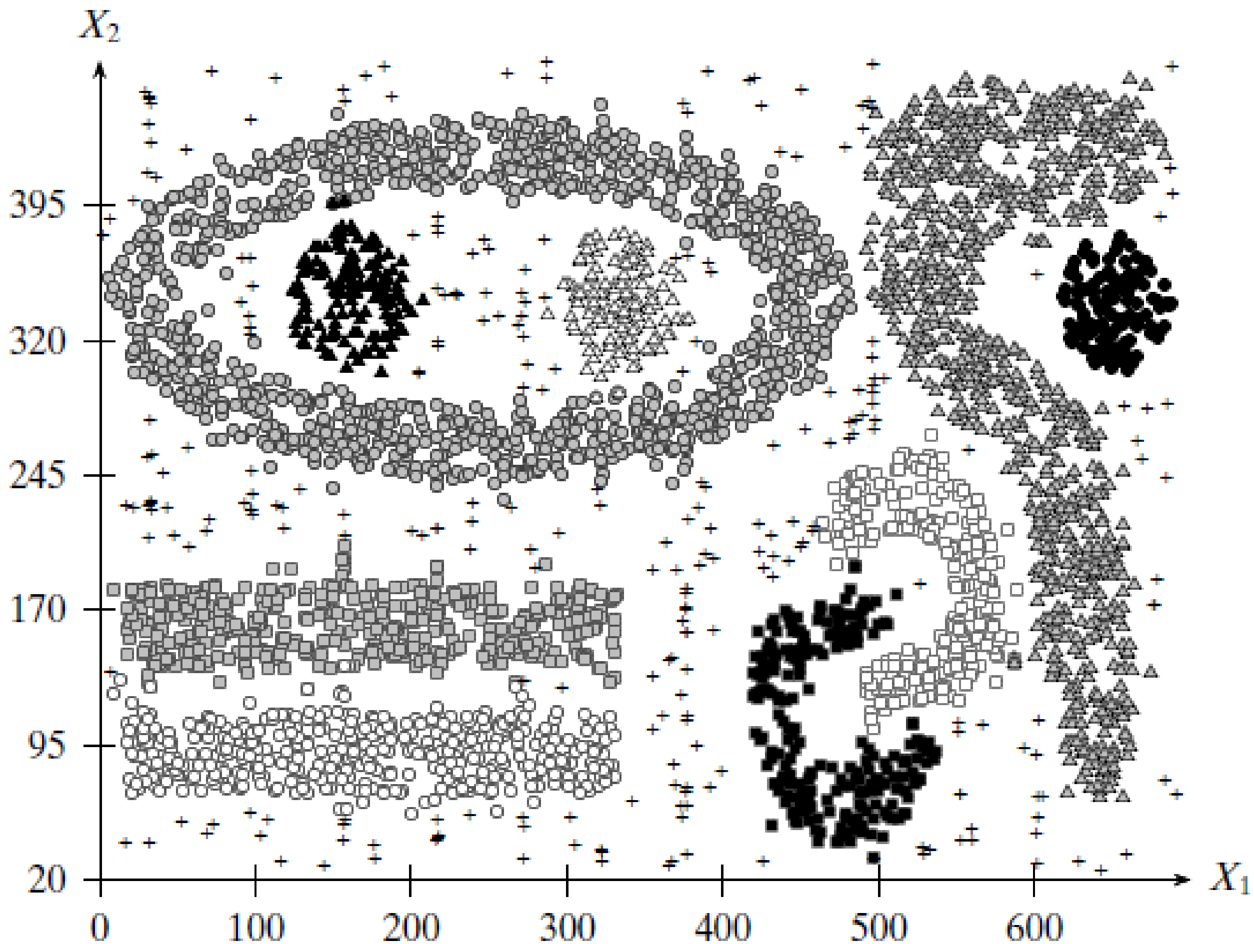
foreach $\mathbf{y} \in N_\epsilon(\mathbf{x})$ **do**

  $id(\mathbf{y}) \leftarrow k$ // assign $\mathbf{y}$ to cluster id k

  if $\mathbf{y} \in Core$ **then** DensityConnected $(\mathbf{y}, k)$

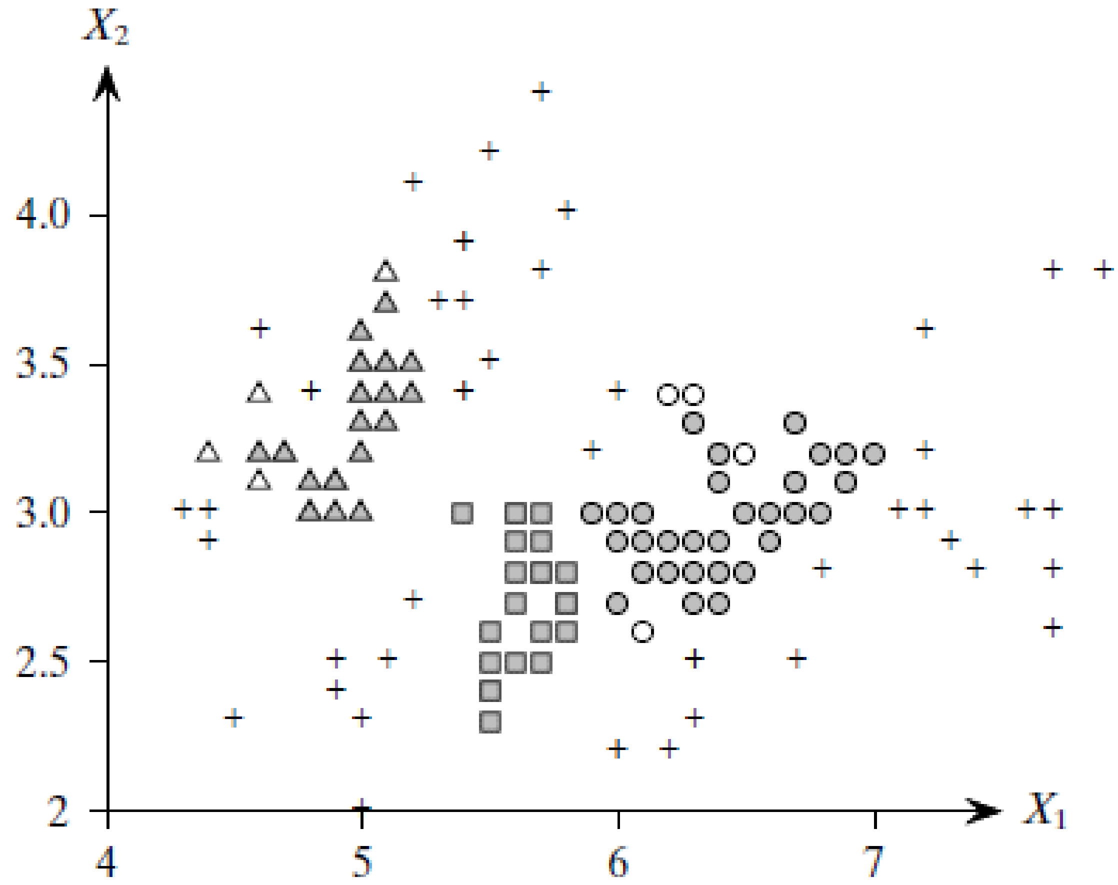
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example15.2



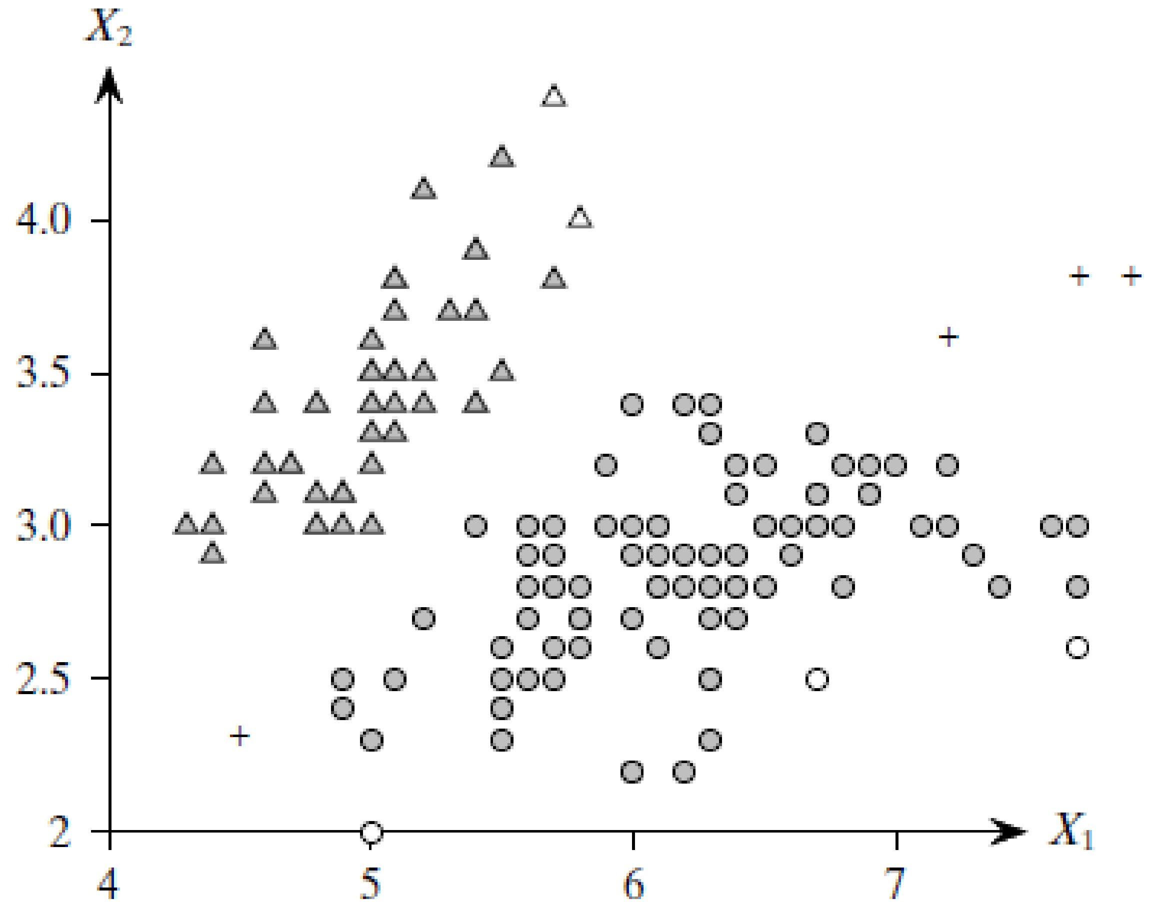
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.3



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

Example 15.3



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

Computational Complexity

- The greatest computational complexity is generating the ϵ -neighborhood for each point
- If the dimensionality is not too excessive, the special structure can be generated in $O(n \log n)$ time
- If the dimensionality is excessive, the special structure generation can be $O(n^2)$

Density-based Clustering

The End

Kernel Density Estimation

Univariate Density Estimation

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$$

$$\hat{f}(x) = \frac{\hat{F}\left(x + \frac{h}{2}\right) - \hat{F}\left(x - \frac{h}{2}\right)}{h} = \frac{k/n}{h} = \frac{k}{nh}$$

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

Kernel Estimator

$$K(x) \geq 0 \quad K(-x) = K(x) \quad \int K(x)dx = 1.$$

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad K(z) = \begin{cases} 1 & \text{If } |z| \leq \frac{1}{2} \\ 0 & \text{Otherwise} \end{cases}$$

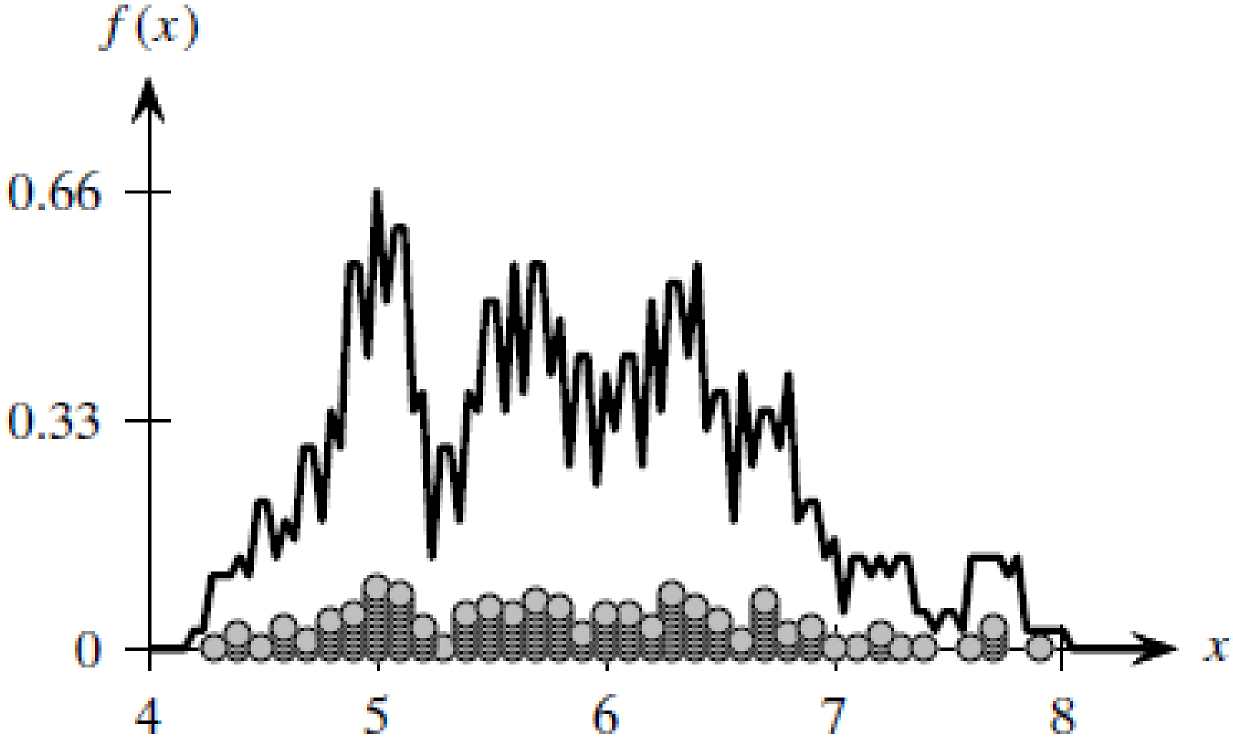
$$\left|\frac{x - x_i}{h}\right| \leq \frac{1}{2} \text{ implies that } -\frac{1}{2} \leq \frac{x_i - x}{h} \leq \frac{1}{2}, \text{ or}$$

$$-\frac{h}{2} \leq x_i - x \leq \frac{h}{2}, \text{ and finally}$$

$$x - \frac{h}{2} \leq x_i \leq x + \frac{h}{2}$$

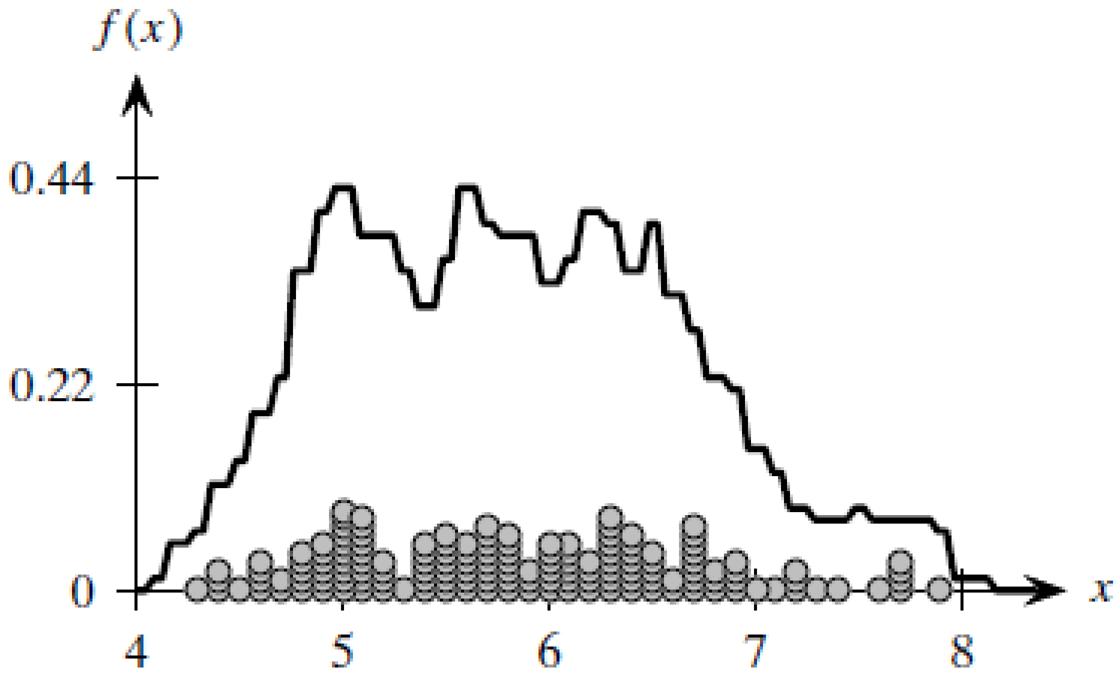
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.4



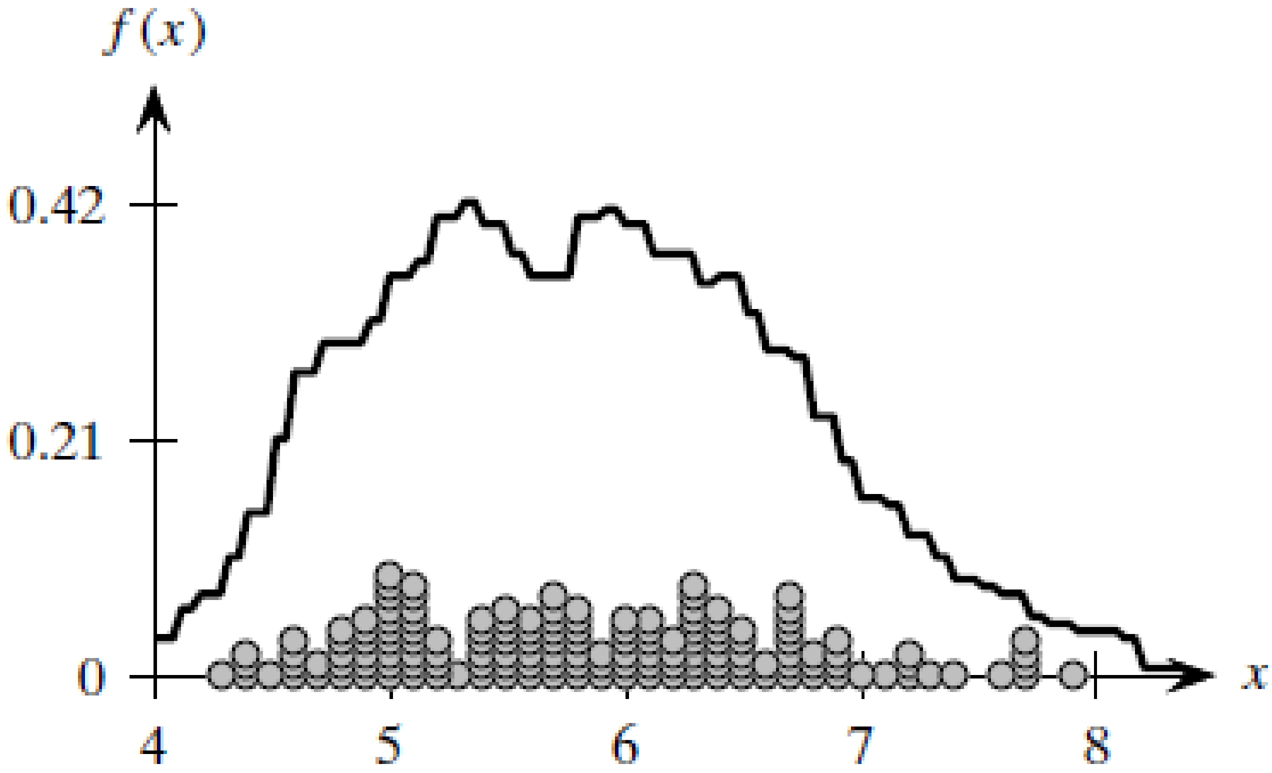
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.4



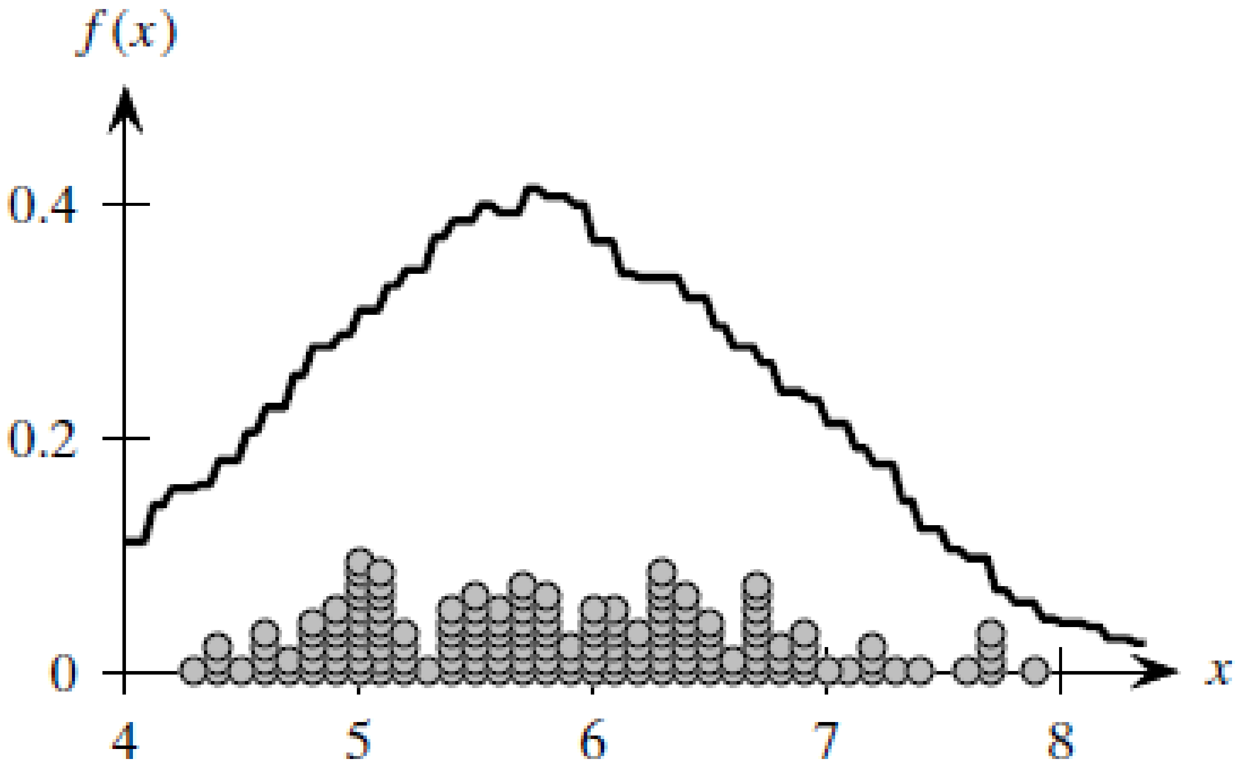
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.4



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

Example 15.4



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

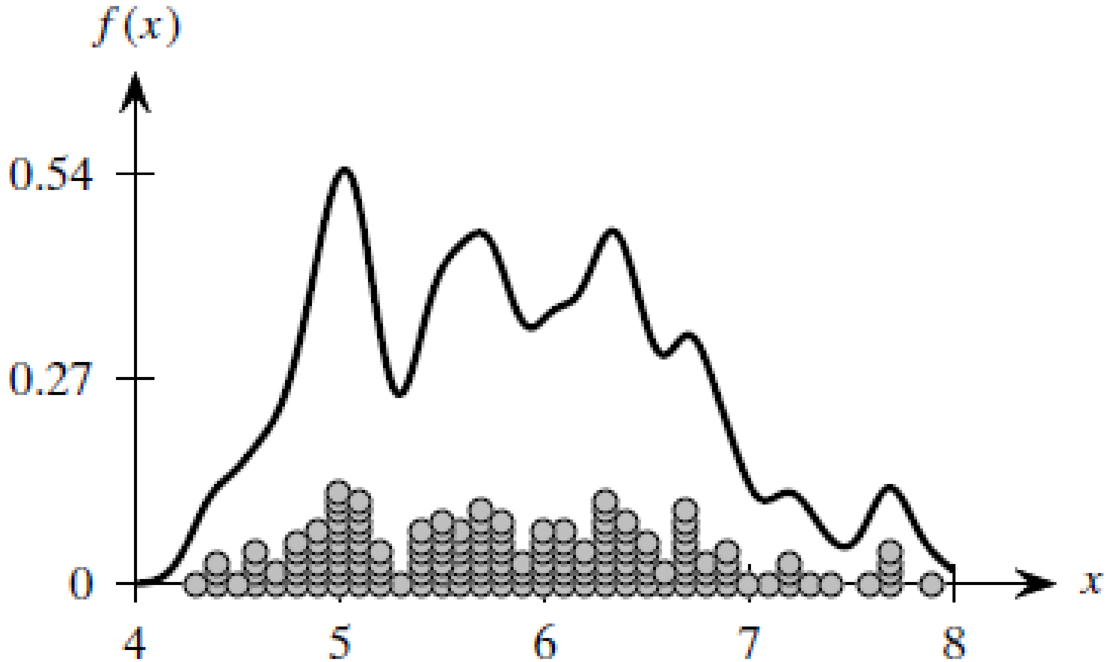
Gaussian Kernel

$$K(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$$

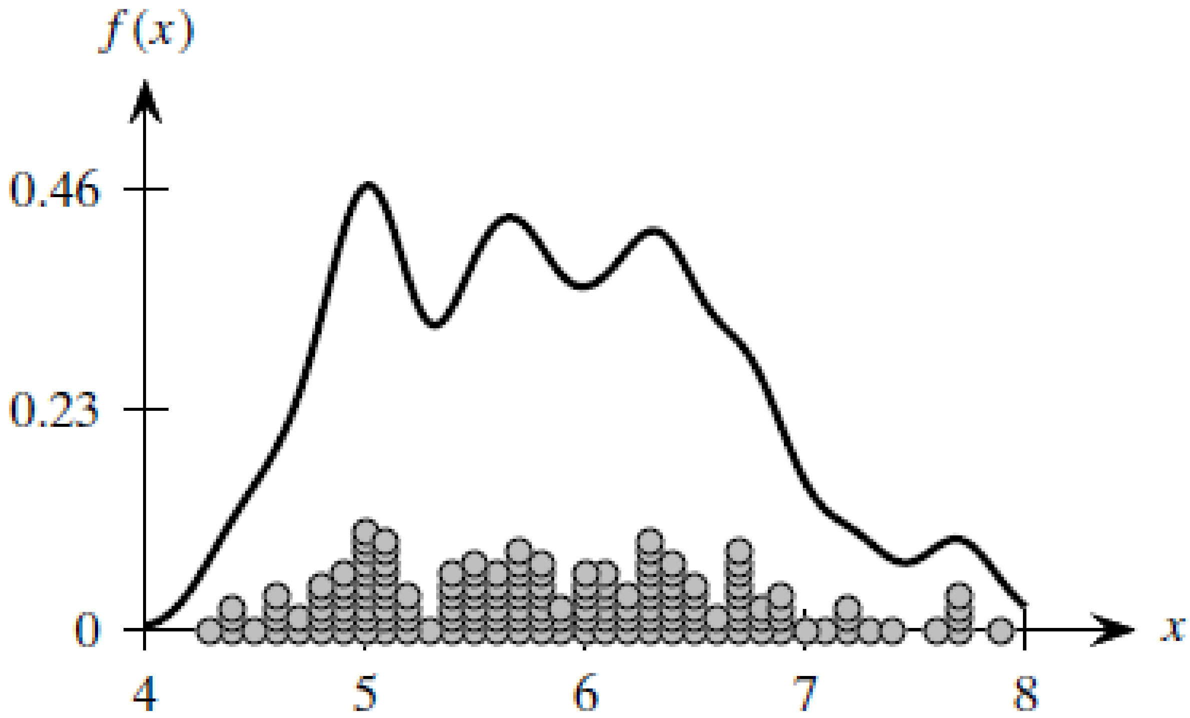
$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x - x_i)^2}{2h^2}\right\}$$

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

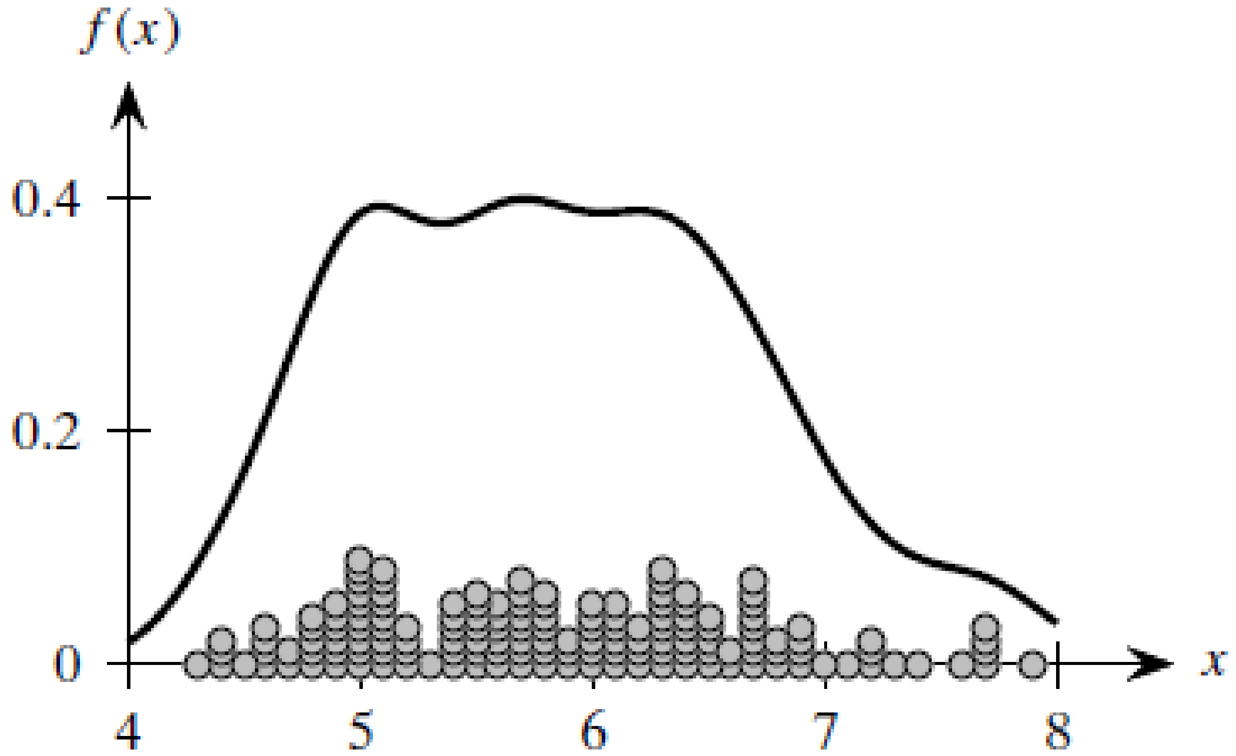
Example 15.5



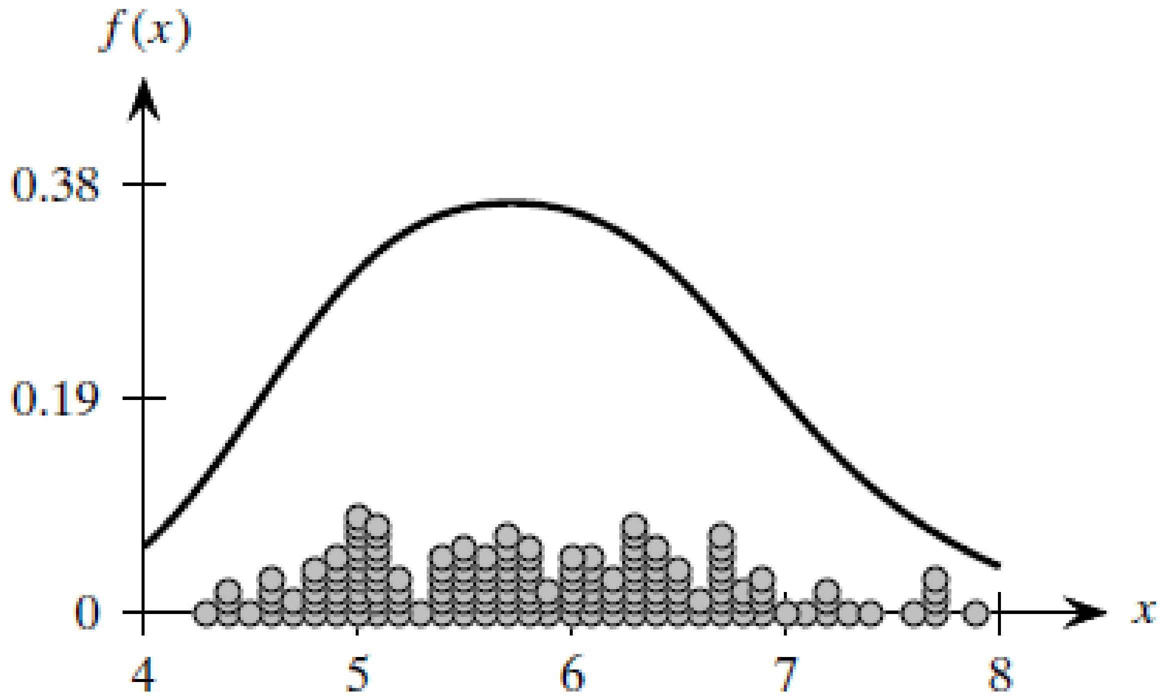
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*



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



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



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

Kernel Density Estimation

The End

Multivariate Density Estimation

Multivariate Density Estimation

$$\mathrm{vol}(H_d(h)) = h^d$$

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

$$\int K(\mathbf{z})d\mathbf{z} = 1.$$

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

Discrete Kernel

$$K(\mathbf{z}) = \begin{cases} 1 & \text{If } |z_j| \le \frac{1}{2}, \text{ for all dimensions } j = 1, \ldots, d \\ 0 & \text{Otherwise} \end{cases}$$

$$\mathbf{z} = \frac{\mathbf{x} - \mathbf{x}_i}{h} \qquad K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) = 1$$

$$\left|\frac{x_j - x_{ij}}{h}\right| \le \frac{1}{2} \qquad \frac{1}{n}$$

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

Gaussian Kernel

$$K(\mathbf{z}) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{\mathbf{z}^T\mathbf{z}}{2}\right\}$$

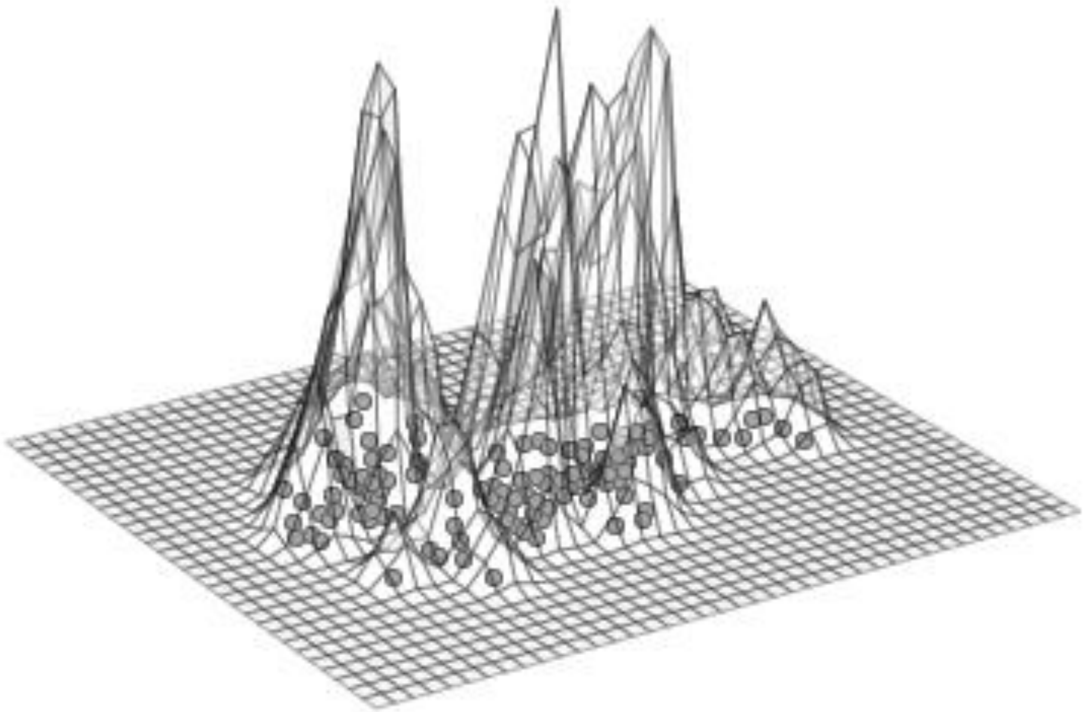
$$\mathbf{\Sigma} = \mathbf{I}_d$$

$$\mathbf{z} = \frac{\mathbf{x}-\mathbf{x}_i}{h}$$

$$K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{(\mathbf{x}-\mathbf{x}_i)^T(\mathbf{x}-\mathbf{x}_i)}{2h^2}\right\}$$

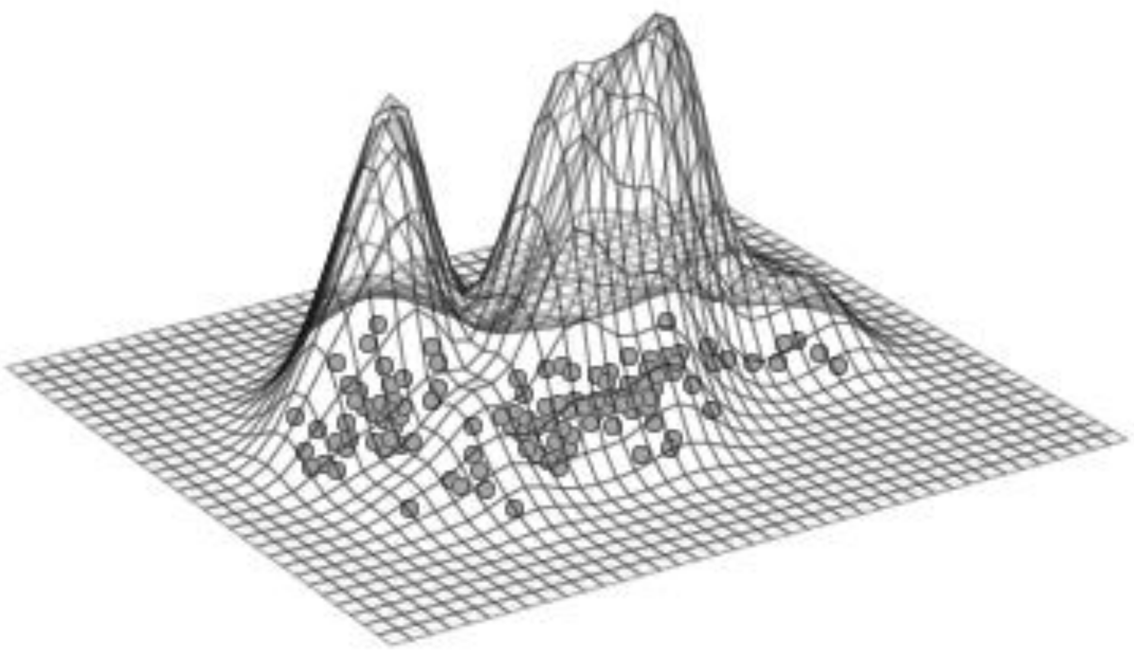
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.6



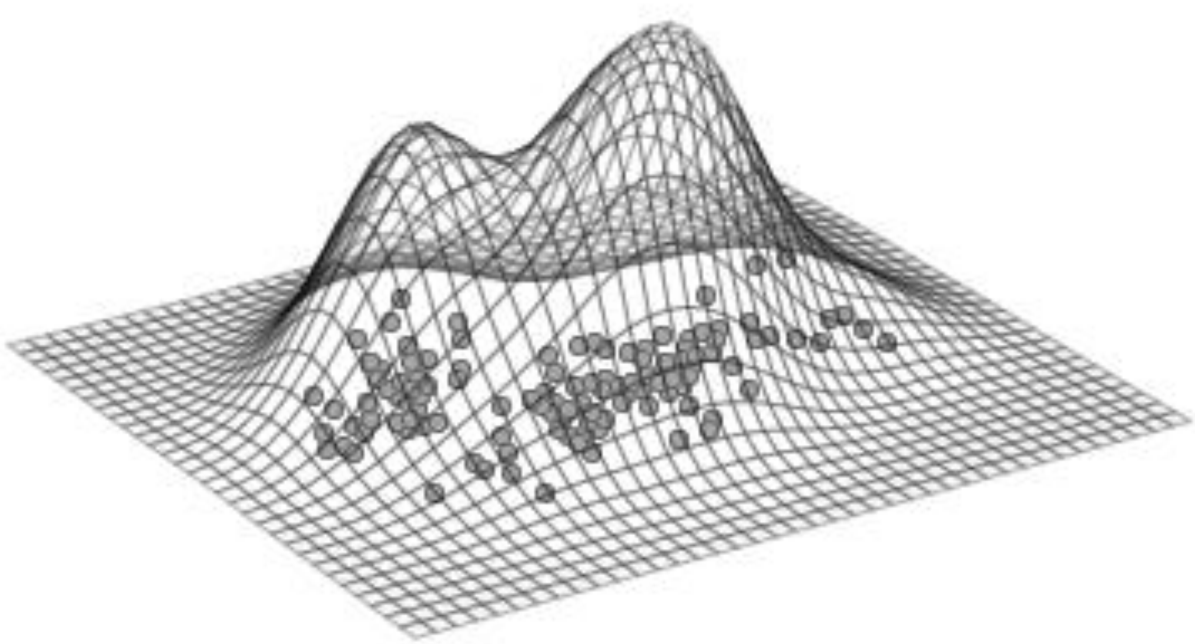
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.6



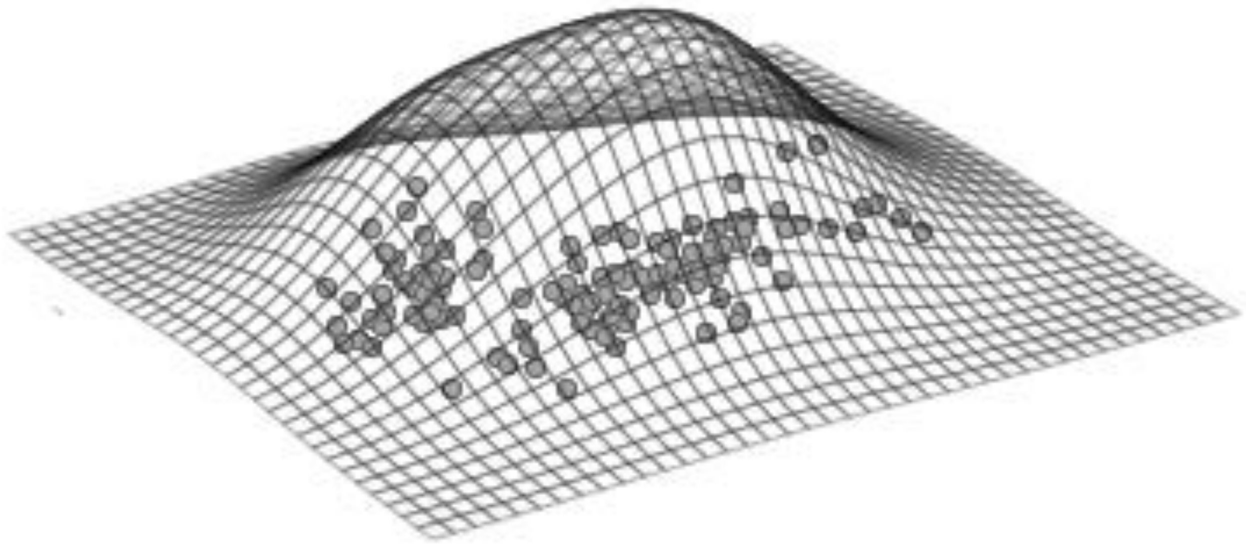
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.6



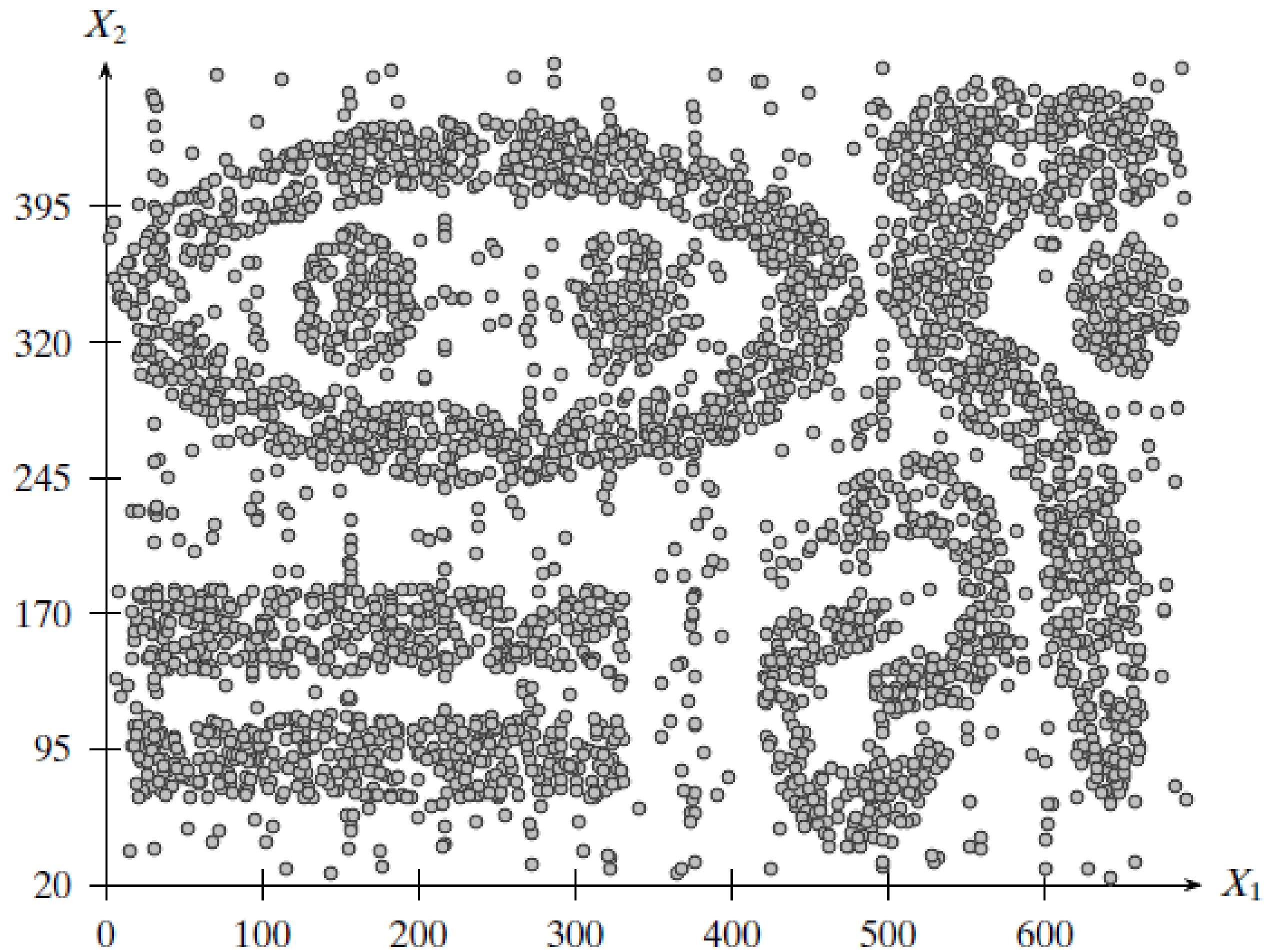
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.6



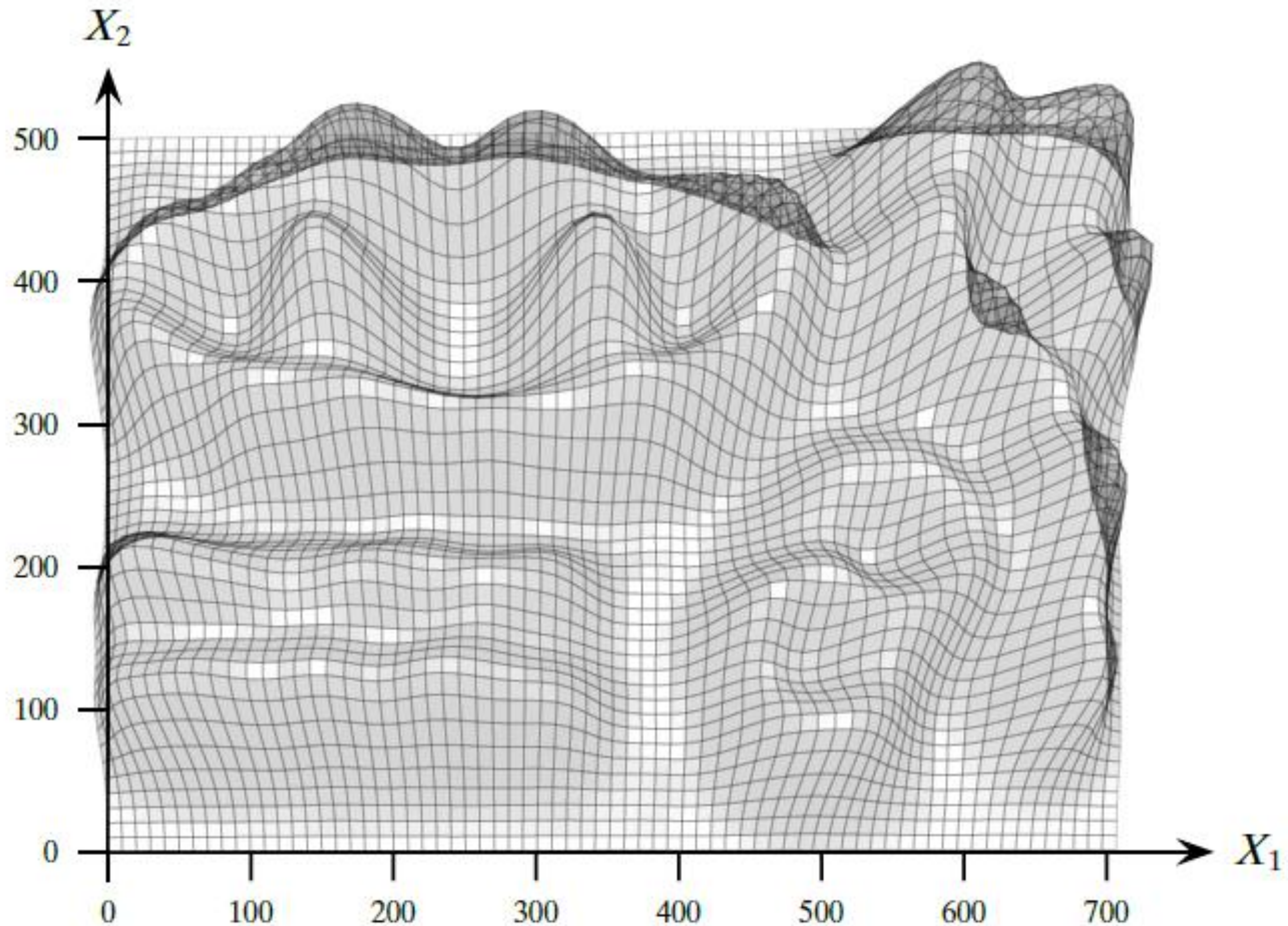
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.7



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

Insert Title Here



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

Nearest Neighbor Density Estimation

$$\hat{f}(\mathbf{x}) = \frac{k}{n\,\mathrm{vol}(S_d(h_{\mathbf{x}}))}$$

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

Multivariate Density Estimation

The End

Density Based Cluster

Density Attractors and Gradient

$$\nabla \hat{f}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{x}} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

$$\frac{\partial}{\partial \mathbf{x}} K(\mathbf{z}) = \left(\frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{\mathbf{z}^T\mathbf{z}}{2}\right\}\right) \cdot -\mathbf{z} \cdot \frac{\partial \mathbf{z}}{\partial \mathbf{x}}$$
$$= K(\mathbf{z}) \cdot -\mathbf{z} \cdot \frac{\partial \mathbf{z}}{\partial \mathbf{x}}$$

$$\mathbf{z} = \frac{\mathbf{x}-\mathbf{x}_i}{h}$$

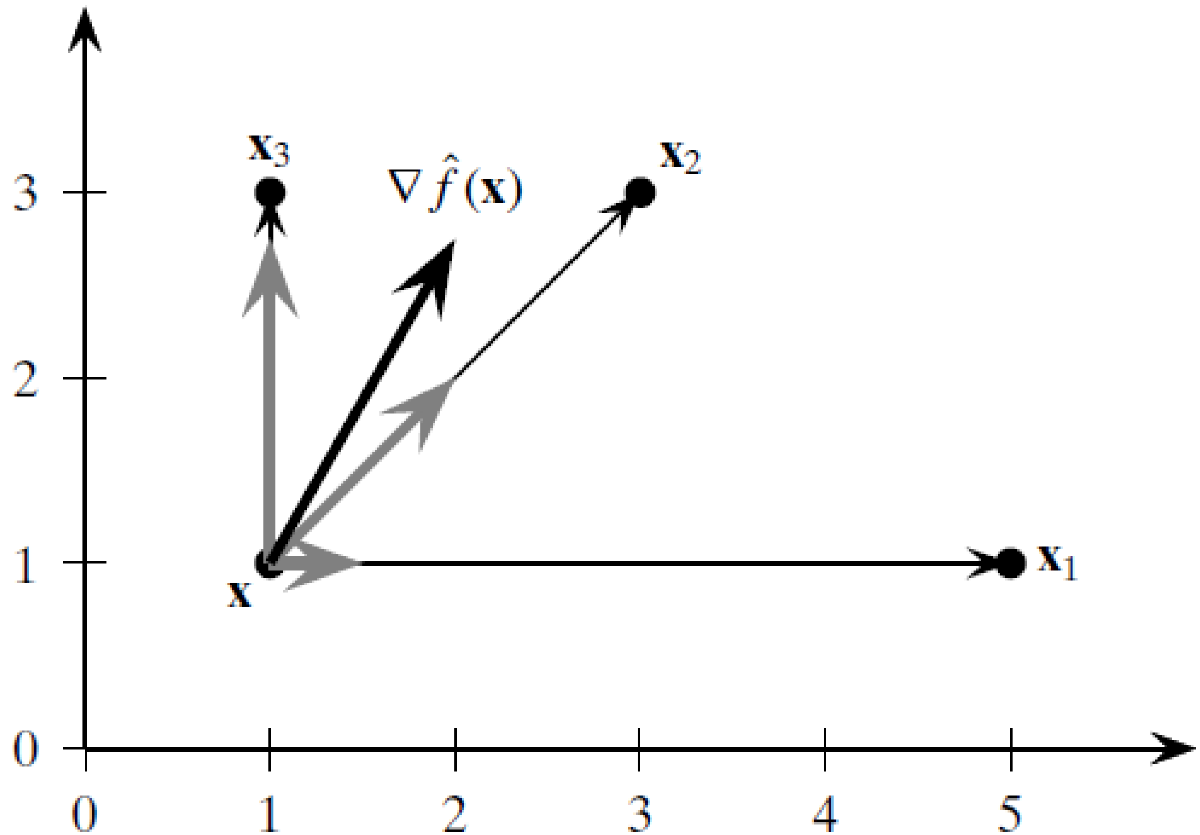
$$\frac{\partial}{\partial \mathbf{x}} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) = K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \cdot \left(\frac{\mathbf{x}_i-\mathbf{x}}{h}\right) \cdot \left(\frac{1}{h}\right)$$

$$\frac{\partial}{\partial \mathbf{x}}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) = \frac{1}{h}$$

$$\nabla \hat{f}(\mathbf{x}) = \frac{1}{nh^{d+2}} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \cdot (\mathbf{x}_i - \mathbf{x})$$

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

Gradient Vector



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

Gradient Ascent Method

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \delta \cdot \nabla \hat{f}(\mathbf{x}_t)$$

$$\nabla \hat{f}(\mathbf{x}) = \mathbf{0}$$

$$\frac{1}{nh^{d+2}} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \cdot (\mathbf{x}_i - \mathbf{x}) = \mathbf{0}$$

$$\mathbf{x} \cdot \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) = \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \mathbf{x}_i$$

$$\mathbf{x} = \frac{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \mathbf{x}_i}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}$$

$$\mathbf{x}_{t+1} = \frac{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_t-\mathbf{x}_i}{h}\right) \mathbf{x}_i}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_t-\mathbf{x}_i}{h}\right)}$$

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

Center-Defined Cluster

$$\hat{f}(\mathbf{x}^*) \geq \xi$$

$$\hat{f}(\mathbf{x}^*) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}^* - \mathbf{x}_i}{h}\right) \geq \xi$$

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

Density-based Cluster

A density-based cluster is an arbitrary-shaped cluster that is an element of the entire dataset if density attractors exist such that

- The point is an element of the cluster
- The density of the density attractor is greater than xi
- The points on the path are density reachable to any of the density attractors

DENCLUE Algorithm

DENCLUE $(\mathbf{D}, h, \xi, \epsilon)$:

$\mathcal{A} \leftarrow \emptyset$

foreach $\mathbf{x} \in \mathbf{D}$ **do** `// find density attractors`

  $\mathbf{x}^* \leftarrow \textsc{FindAttractor}(\mathbf{x}, \mathbf{D}, h, \epsilon)$

  if $\hat{f}(\mathbf{x}^*) \geq \xi$ **then**

    $\mathcal{A} \leftarrow \mathcal{A} \cup \{\mathbf{x}^*\}$

    $R(\mathbf{x}^*) \leftarrow R(\mathbf{x}^*) \cup \{\mathbf{x}\}$

$\mathcal{C} \leftarrow \{\text{maximal } C \subseteq \mathcal{A} \mid \forall \mathbf{x}_i^*, \mathbf{x}_j^* \in C, \mathbf{x}_i^* \text{ and } \mathbf{x}_j^* \text{ are density reachable}\}$

foreach $C \in \mathcal{C}$ **do** `// density-based clusters`

  foreach $\mathbf{x}^* \in C$ **do** $C \leftarrow C \cup R(\mathbf{x}^*)$

return $\mathcal{C}$

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

DENCLUE Algorithm

FINDATTRACTOR $(\mathbf{x}, \mathbf{D}, h, \epsilon)$:

$t \leftarrow 0$

$\mathbf{x}_t \leftarrow \mathbf{x}$

repeat

$$\mathbf{x}_{t+1} \leftarrow \frac{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_t - \mathbf{x}_i}{h}\right) \cdot \mathbf{x}_t}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_t - \mathbf{x}_i}{h}\right)}$$

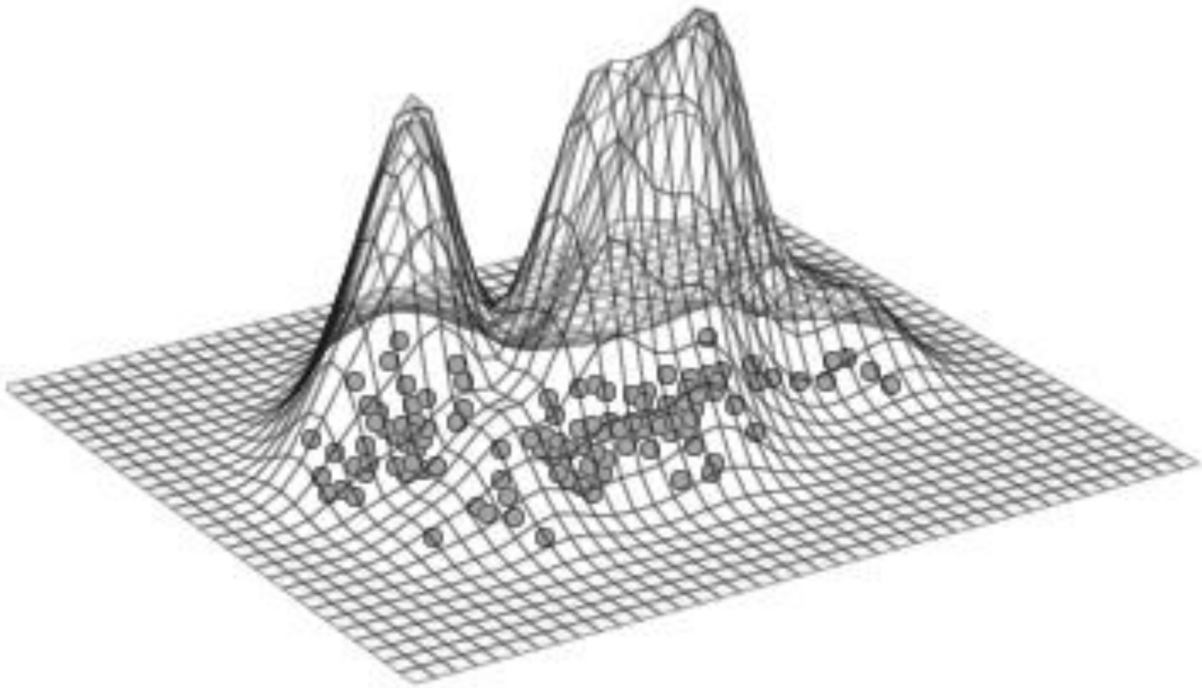
$t \leftarrow t + 1$

until $\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \leq \epsilon$

return $\mathbf{x}_t$

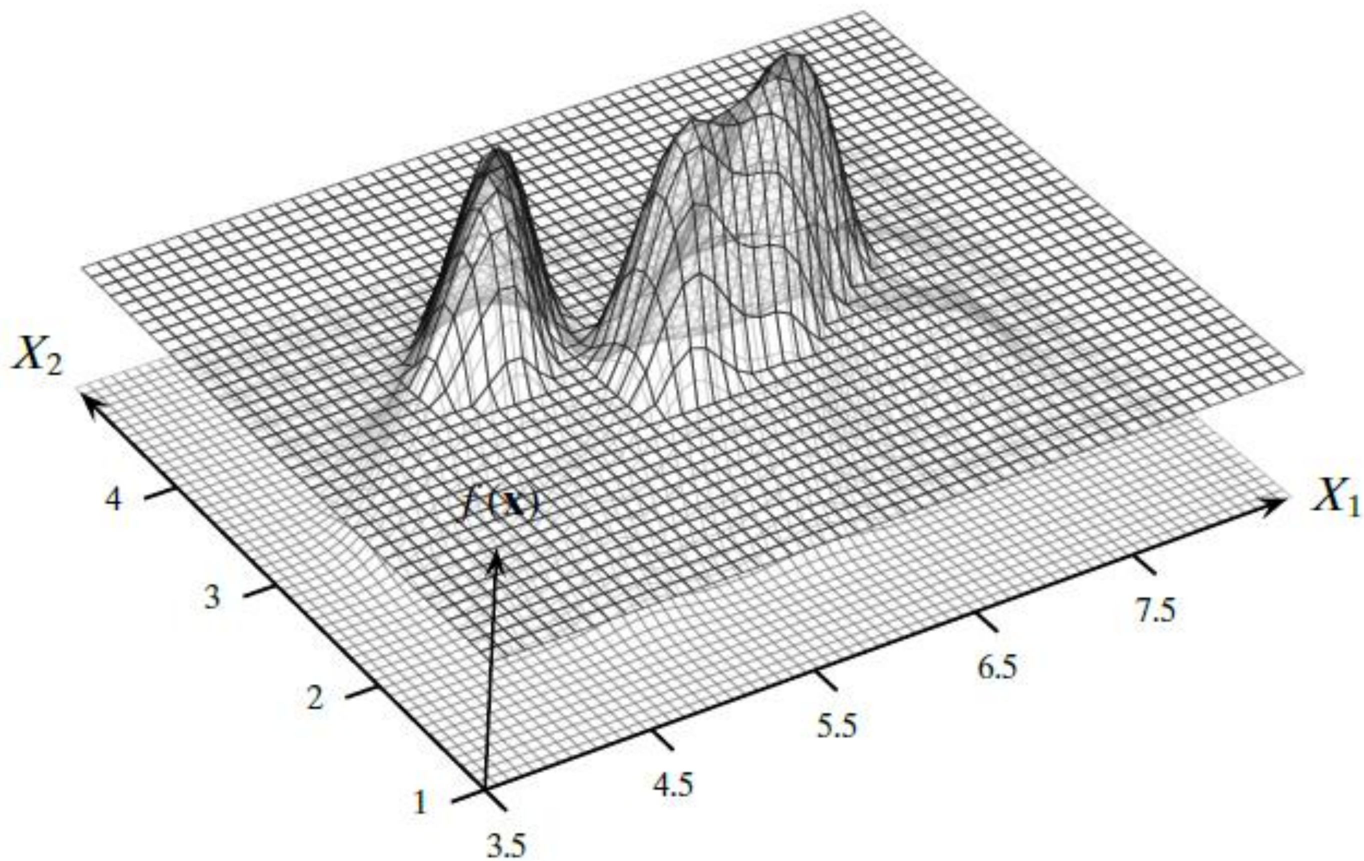
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.8



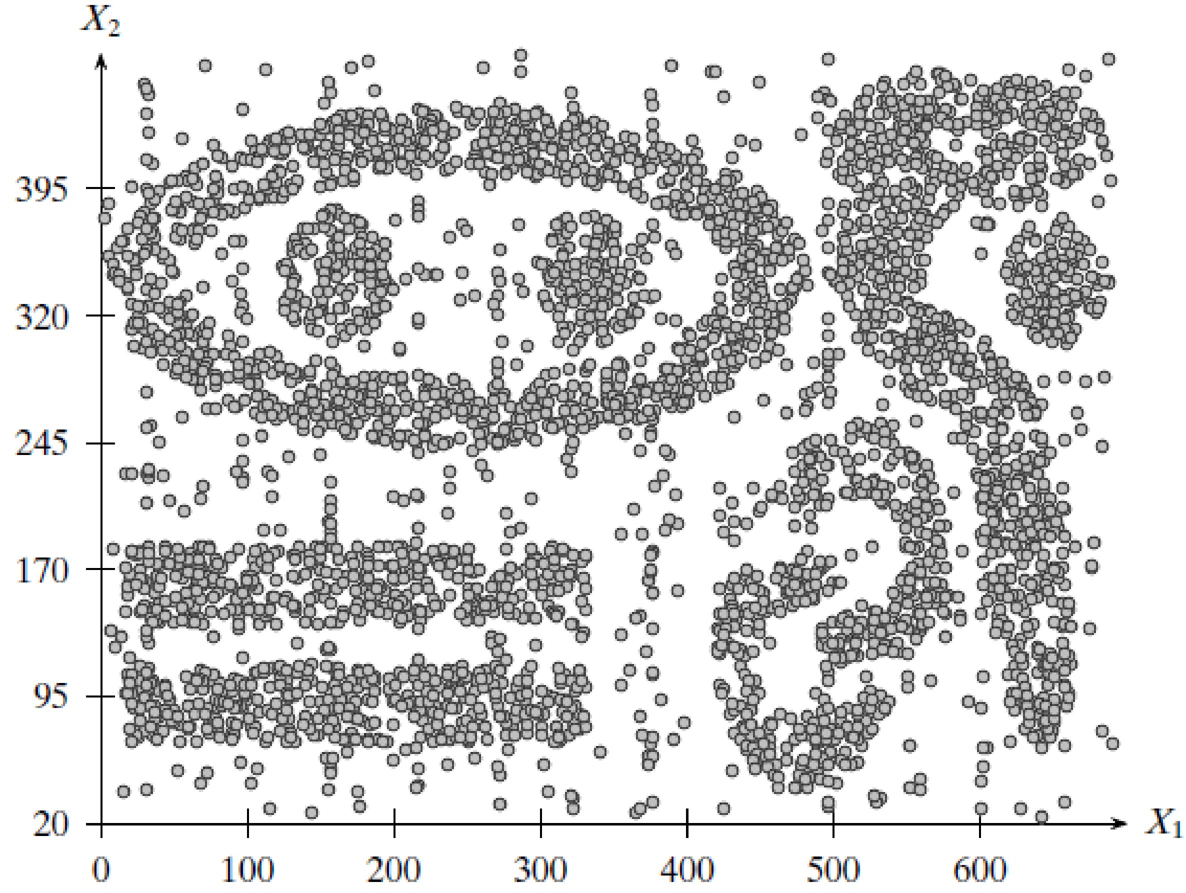
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.8



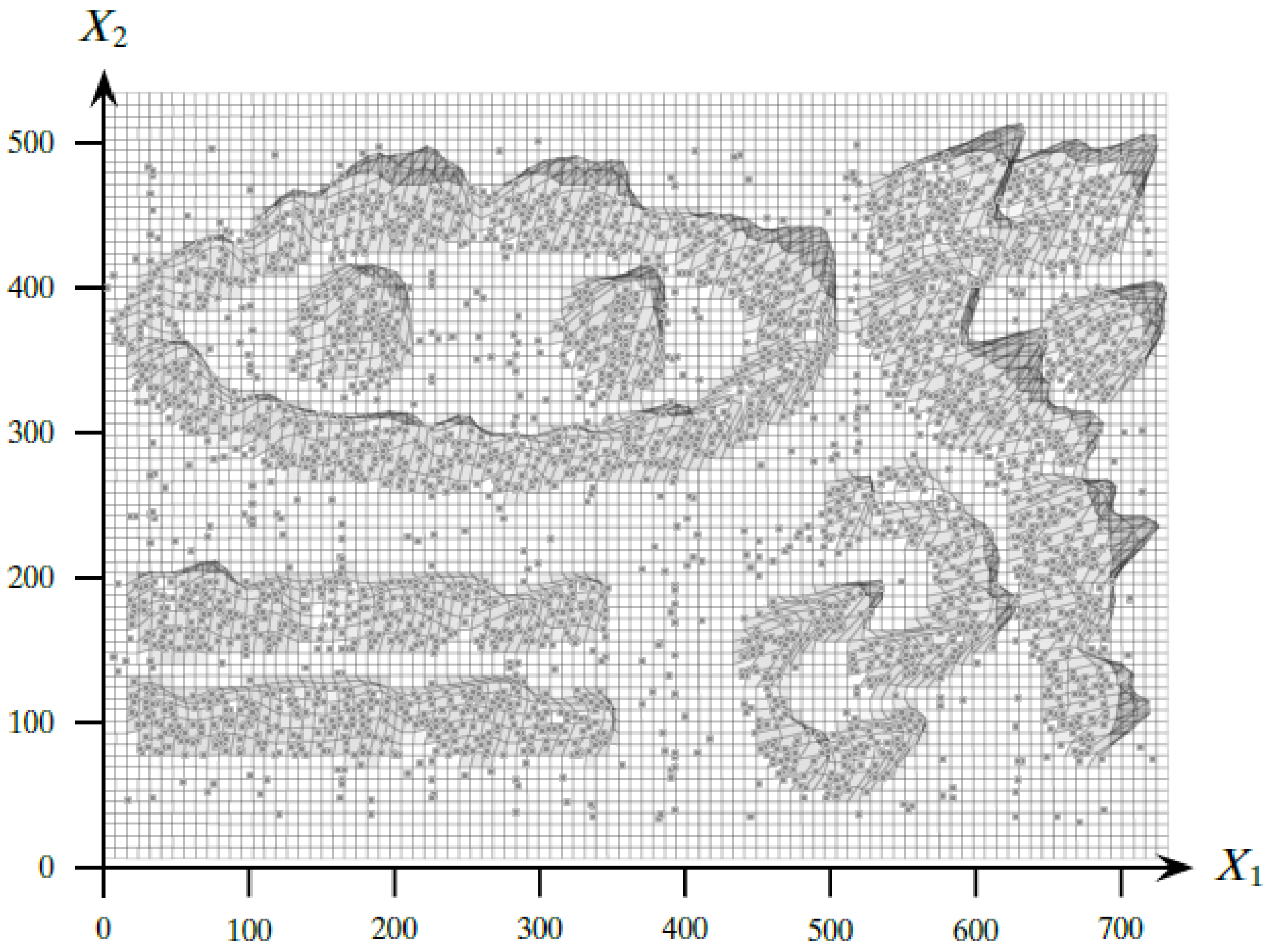
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 15.9



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

Example 15.9



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

Special Case of DENCLUE

- DENCLUE is the general case of kernel density estimate based clustering
- DBSCAN is a special case of kernel density estimate based clustering when where $h = \epsilon$ and $\xi = minpts$.
- The density attractors and the core points correspond
- The attractors are defined by a set of connected core points

Computational Complexity

- The significant cost of DENCLUE is the hill-climbing process
- Locating each attractor take $O(nt)$, where t is the number of iterations to climb the hill
- The total cost is $O(n^2t)$. to compute the attractors
- It take $O(m^2)$. to locate all the reachable attractors
- It takes $O(n)$ to obtain the final clusters

Density Based Cluster

The End

Cluster Validation

Cluster Validation

- The quality or goodness of clustering is assessed by cluster evaluation
- The sensitivity of the clusters are assessed by cluster stability
- The suitability of applying clustering is assessed by clustering tendency

Cluster Validation

- Validations measure not inherent within the dataset of External measures
- Validations measure derived from the dataset are Internal measures
- Different clustering measures comparing cluster parameters are Relative measures

External Measures

- The ground-truth clustering is where all points in the cluster have the same label
- Referring to the entire dataset is the ground-truth partition
- Each subset of the dataset are referred to as partitions
- External evaluation measure attempts to quantify which points from the same partition are included in the same cluster

External Measures

- External evaluation measures also quantifies how points from different partitions are grouped in different clusters
- These evaluations can be quantified explicitly by measurements
- These evaluations can be quantified implicitly by computations

External Measures

- External measure are derived from a contingency table that is defined by

$$\mathbf{N}(i, j) = n_{ij} = \left| C_i \cap T_j \right|$$

- The contingency table computed by examining the partition and cluster labels

Matching Based Measure

$$purity_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\}$$

$$purity = \sum_{i=1}^{r} \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^{r} \max_{j=1}^{k} \{n_{ij}\}$$

$$match = \arg\max_{M} \left\{ \frac{w(M)}{n} \right\}$$

$$prec_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\} = \frac{n_{ij_i}}{n_i}$$

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$$

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

Matching Based Measure

$$prec_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\} = \frac{n_{ij_i}}{n_i}$$

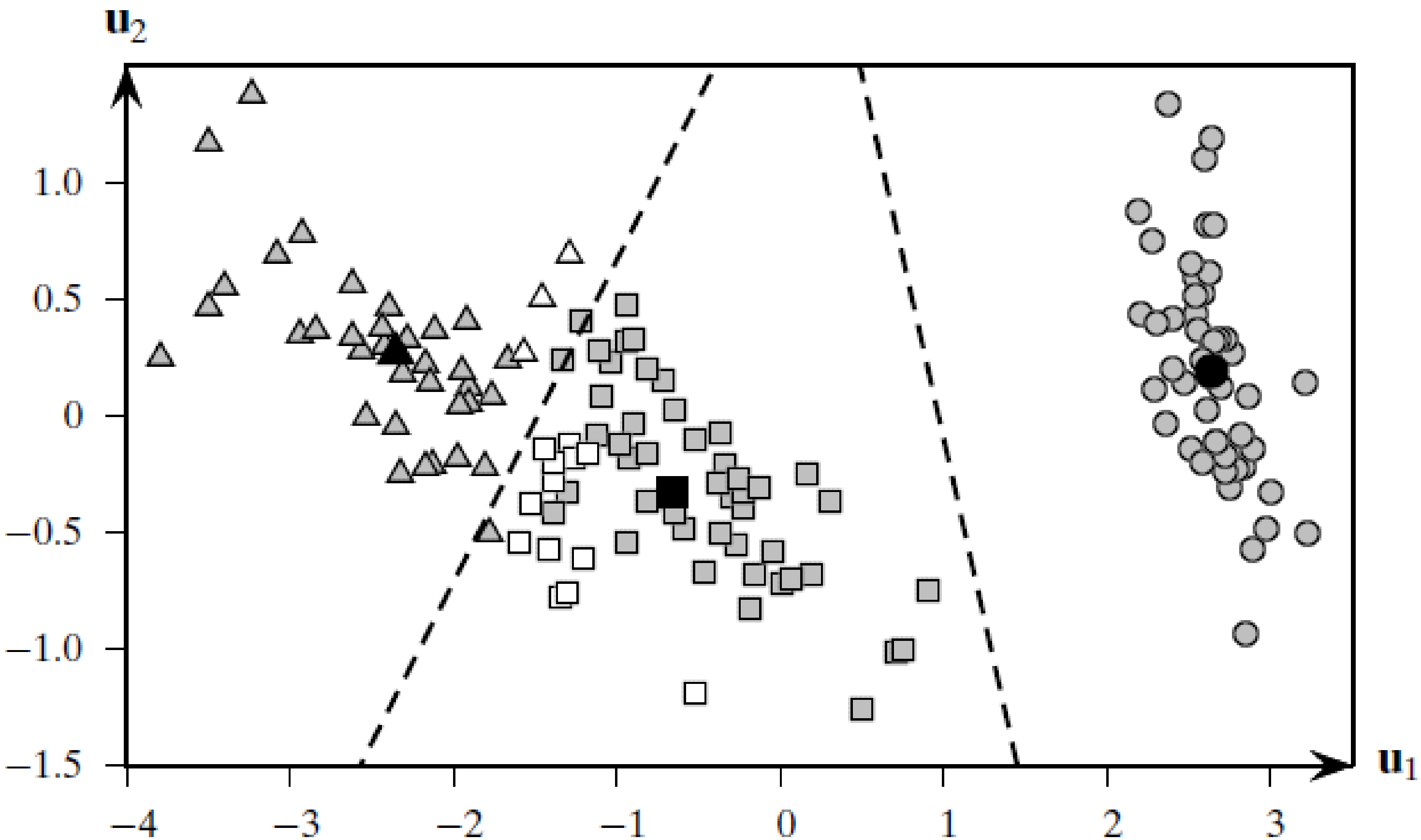
$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$$

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2\, n_{ij_i}}{n_i + m_{j_i}}$$

$$F = \frac{1}{r} \sum_{i=1}^{r} F_i$$

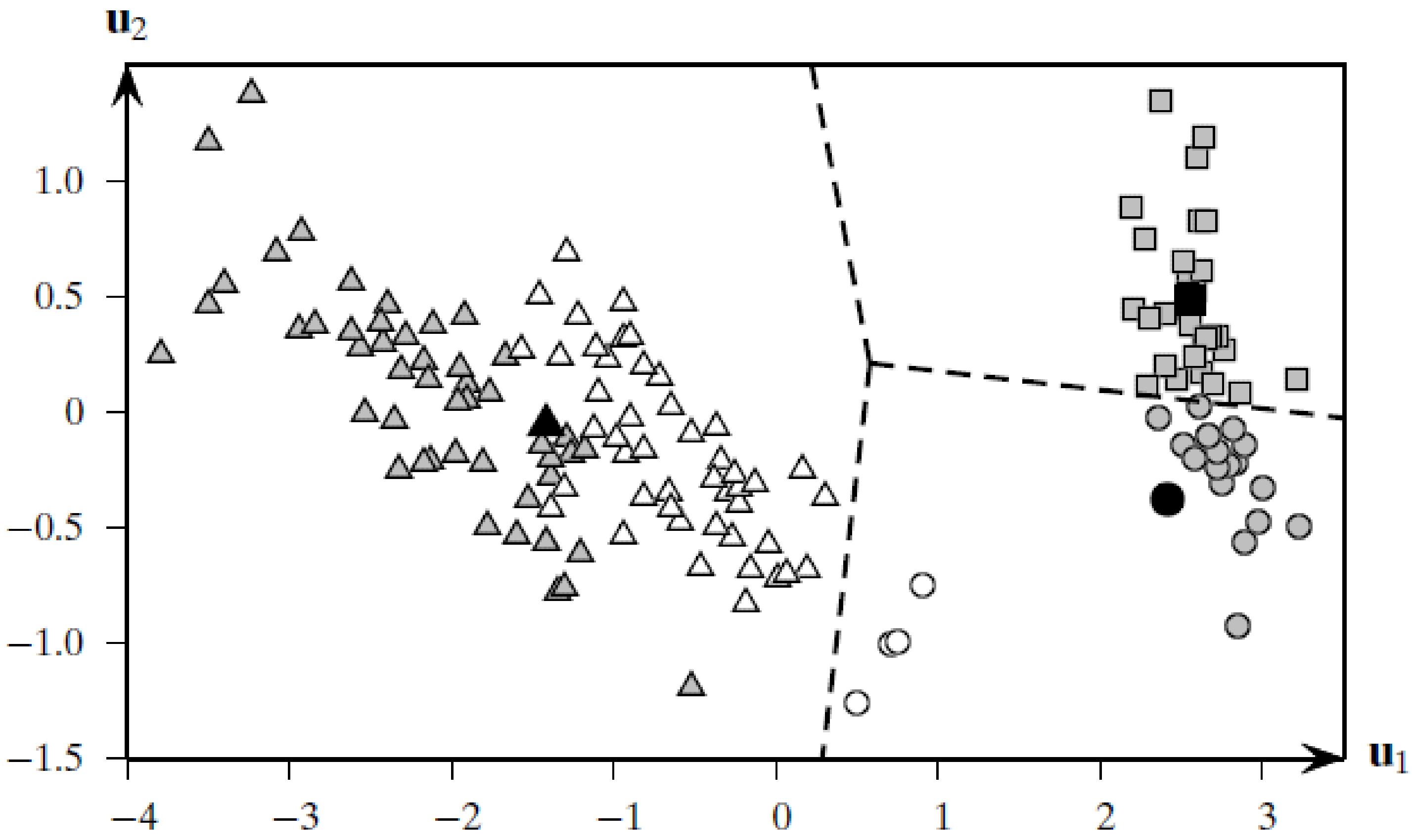
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 17.1



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

Example 17.1



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

Example 17.1

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 150$

$$purity = \frac{1}{150}(47 + 50 + 36) = \frac{133}{150} = 0.887$$

$$prec_1 = \frac{47}{61} = 0.77$$

$$recall_1 = \frac{47}{50} = 0.94$$

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

Example 17.1

$$F_1 = \frac{2 \cdot 0.77 \cdot 0.94}{0.77 + 0.94} = \frac{1.45}{1.71} = 0.85$$

$$F_1 = \frac{2 \cdot n_{12}}{n_1 + m_2} = \frac{2 \cdot 47}{61 + 50} = \frac{94}{111} = 0.85$$

$$F = \frac{1}{3}(F_1 + F_2 + F_3) = \frac{2.66}{3} = 0.88$$

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1	30	0	0	30
C_2	20	4	0	24
C_3	0	46	50	96
m_j	50	50	50	$n = 150$

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

Example 17.1

$$purity = \frac{1}{150}(30 + 20 + 50) = \frac{100}{150} = 0.67$$

$$match = \frac{1}{150}(30 + 4 + 50) = \frac{84}{150} = 0.56$$

	purity	*match*	*F*
(a) Good	0.887	0.887	0.885
(b) Bad	0.667	0.560	0.658

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

Cluster Validation

The End

External Measures, Part 1

Conditional Entropy

$$H(\mathcal{C}) = -\sum_{i=1}^{r} p_{C_i} \log p_{C_i}$$

$$H(\mathcal{T}) = -\sum_{j=1}^{k} p_{T_j} \log p_{T_j}$$

$$H(\mathcal{T}|C_i) = -\sum_{j=1}^{k} \left(\frac{n_{ij}}{n_i}\right) \log \left(\frac{n_{ij}}{n_i}\right)$$

$$H(\mathcal{T}|\mathcal{C}) = \sum_{i=1}^{r} \frac{n_i}{n} H(\mathcal{T}|C_i) = -\sum_{i=1}^{r}\sum_{j=1}^{k} \frac{n_{ij}}{n} \log \left(\frac{n_{ij}}{n_i}\right)$$

$$= -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log \left(\frac{p_{ij}}{p_{C_i}}\right)$$

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

Conditional Entropy

$$
\begin{aligned}
H(\mathcal{T}|\mathcal{C}) &= -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\left(\log p_{ij} - \log p_{C_i}\right) \\
&= -\left(\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log p_{ij}\right) + \sum_{i=1}^{r}\left(\log p_{C_i}\sum_{j=1}^{k} p_{ij}\right) \\
&= -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log p_{ij} + \sum_{i=1}^{r} p_{C_i}\log p_{C_i} \\
&= H(\mathcal{C},\mathcal{T}) - H(\mathcal{C})
\end{aligned}
$$

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

Normalized Mutual Information

$$I(\mathcal{C},\mathcal{T}) = \sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log\left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}}\right)$$

$$I(\mathcal{C},\mathcal{T}) = H(\mathcal{T}) - H(\mathcal{T}|\mathcal{C})$$

$$I(\mathcal{C},\mathcal{T}) = H(\mathcal{C}) - H(\mathcal{C}|\mathcal{T})$$

$$NMI(\mathcal{C},\mathcal{T}) = \sqrt{\frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{C})} \cdot \frac{I(\mathcal{C},\mathcal{T})}{H(\mathcal{T})}} = \frac{I(\mathcal{C},\mathcal{T})}{\sqrt{H(\mathcal{C}) \cdot H(\mathcal{T})}}$$

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

Variation of Information

$$
\begin{aligned}
VI(\mathcal{C},\mathcal{T}) &= (H(\mathcal{T}) - I(\mathcal{C},\mathcal{T})) + (H(\mathcal{C}) - I(\mathcal{C},\mathcal{T})) \\
&= H(\mathcal{T}) + H(\mathcal{C}) - 2I(\mathcal{C},\mathcal{T})
\end{aligned}
$$

$$VI(\mathcal{C},\mathcal{T}) = H(\mathcal{T}|\mathcal{C}) + H(\mathcal{C}|\mathcal{T})$$

$$VI(\mathcal{C},\mathcal{T}) = 2H(\mathcal{T},\mathcal{C}) - H(\mathcal{T}) - H(\mathcal{C})$$

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

Example 17.2

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1	0	47	14	61
C_2	50	0	0	50
C_3	0	3	36	39
m_j	50	50	50	$n = 100$

$$
\begin{aligned}
H(\mathcal{T}|C_1) &= -\frac{0}{61}\log_2\left(\frac{0}{61}\right) - \frac{47}{61}\log_2\left(\frac{47}{61}\right) - \frac{14}{61}\log_2\left(\frac{14}{61}\right) \\
&= -0 - 0.77\log_2(0.77) - 0.23\log_2(0.23) = 0.29 + 0.49 = 0.78
\end{aligned}
$$

$$
H(\mathcal{T}|\mathcal{C}) = \frac{61}{150}\cdot 0.78 + \frac{50}{150}\cdot 0 + \frac{39}{150}\cdot 0.39 = 0.32 + 0 + 0.10 = 0.42
$$

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

Example 17.2

$$H(\mathcal{T}) = -3\left(\frac{50}{150}\log_2\left(\frac{50}{150}\right)\right) = 1.585$$

$$\begin{aligned} H(\mathcal{C}) &= -\left(\frac{61}{150}\log_2\left(\frac{61}{150}\right) + \frac{50}{150}\log_2\left(\frac{50}{150}\right) + \frac{39}{150}\log_2\left(\frac{39}{150}\right)\right) \\ &= 0.528 + 0.528 + 0.505 = 1.561 \end{aligned}$$

$$\begin{aligned} I(\mathcal{C},\mathcal{T}) = &\frac{47}{150}\log_2\left(\frac{47\cdot 150}{61\cdot 50}\right) + \frac{14}{150}\log_2\left(\frac{14\cdot 150}{61\cdot 50}\right) + \frac{50}{150}\log_2\left(\frac{50\cdot 150}{50\cdot 50}\right) \\ &+ \frac{3}{150}\left(\log_2\frac{3\cdot 150}{39\cdot 50}\right) + \frac{36}{150}\log_2\left(\frac{36\cdot 150}{39\cdot 50}\right) \\ = &\, 0.379 - 0.05 + 0.528 - 0.042 + 0.353 = 1.167 \end{aligned}$$

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

Example 17.2

$$NMI(\mathcal{C},\mathcal{T}) = \frac{I(\mathcal{C},\mathcal{T})}{\sqrt{H(\mathcal{T})\cdot H(\mathcal{C})}} = \frac{1.167}{\sqrt{1.585\times 1.561}} = 0.742$$

$$VI(\mathcal{C},\mathcal{T}) = H(\mathcal{T}) + H(\mathcal{C}) - 2I(\mathcal{C},\mathcal{T}) = 1.585 + 1.561 - 2\cdot 1.167 = 0.812$$

	$H(\mathcal{T}\|\mathcal{C})$	NMI	VI
(a) Good	0.418	0.742	0.812
(b) Bad	0.743	0.587	1.200

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

Pairwise Measures

$$TP = \left|\{(\mathbf{x}_i, \mathbf{x}_j) : \ y_i = y_j \text{ and } \hat{y}_i = \hat{y}_j\}\right|$$

$$FN = \left|\{(\mathbf{x}_i, \mathbf{x}_j) : \ y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}\right|$$

$$FP = \left|\{(\mathbf{x}_i, \mathbf{x}_j) : \ y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}\right|$$

$$TN = \left|\{(\mathbf{x}_i, \mathbf{x}_j) : \ y_i \neq y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}\right|$$

$$N = TP + FN + FP + TN$$

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

Pairwise Measures

$$TP = \sum_{i=1}^{r}\sum_{j=1}^{k}\binom{n_{ij}}{2} = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}(n_{ij}-1)}{2} = \frac{1}{2}\left(\sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2 - \sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}\right)$$

$$= \frac{1}{2}\left(\left(\sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2\right) - n\right)$$

$$FN = \sum_{j=1}^{k}\binom{m_j}{2} - TP = \frac{1}{2}\left(\sum_{j=1}^{k}m_j^2 - \sum_{j=1}^{k}m_j - \sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2 + n\right)$$

$$= \frac{1}{2}\left(\sum_{j=1}^{k}m_j^2 - \sum_{i=1}^{r}\sum_{j=1}^{k}n_{ij}^2\right)$$

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

Pairwise Measures

$$FP = \sum_{i=1}^{r} \binom{n_i}{2} - TP = \frac{1}{2}\left(\sum_{i=1}^{r} n_i^2 - \sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}^2\right)$$

$$TN = N - (TP + FN + FP) = \frac{1}{2}\left(n^2 - \sum_{i=1}^{r} n_i^2 - \sum_{j=1}^{k} m_j^2 + \sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}^2\right)$$

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

External Measures, Part 1

The End

External Measures, Part 2

Pairwise Measures

$$Jaccard = \frac{TP}{TP + FN + FP}$$

$$Rand = \frac{TP + TN}{N}$$

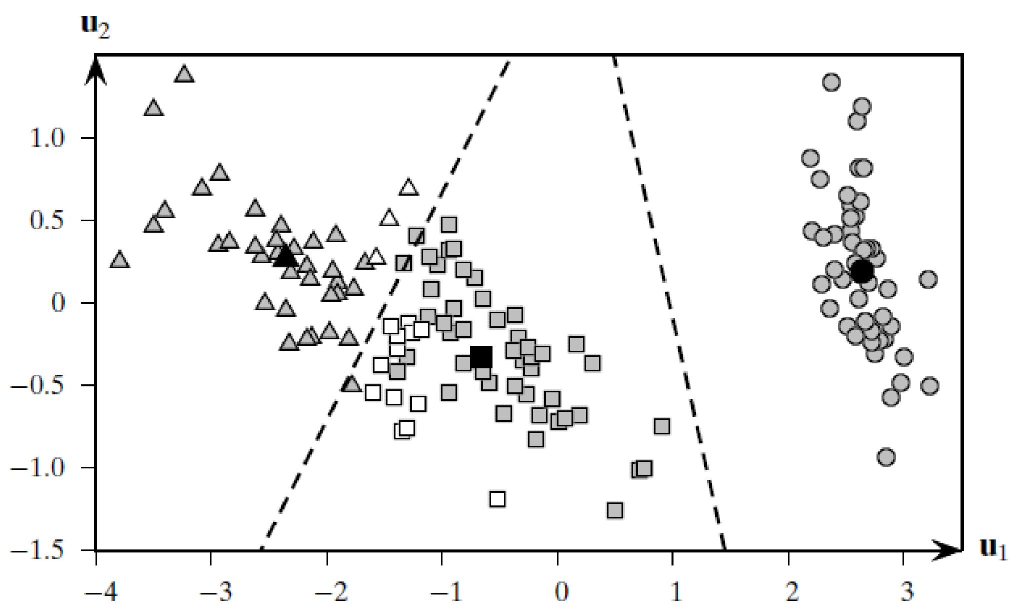
$$prec = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$FM = \sqrt{prec \cdot recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

Example 17.3

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3
C_1	0	47	14
C_2	50	0	0
C_3	0	3	36



Example 17.3

$$TP = \binom{47}{2} + \binom{14}{2} + \binom{50}{2} + \binom{3}{2} + \binom{36}{2}$$
$$= 1081 + 91 + 1225 + 3 + 630 = 3030$$

$FN = 645$ $FP = 766$ $TN = 6734$

$$Jaccard = \frac{3030}{3030 + 645 + 766} = \frac{3030}{4441} = 0.68$$
$$Rand = \frac{3030 + 6734}{11175} = \frac{9764}{11175} = 0.87$$
$$FM = \frac{3030}{\sqrt{3675 \cdot 3796}} = \frac{3030}{3735} = 0.81$$

Example 17.3

$TP = 2891$ $FN = 784$ $FP = 2380$ $TN = 5120$

	Jaccard	*Rand*	*FM*
(a) Good	0.682	0.873	0.811
(b) Bad	0.477	0.717	0.657

Correlation Measures

$$\mu_X = \frac{1}{N}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbf{X}(i,j) = \frac{1}{N}\mathbf{x}^T\mathbf{x} \qquad \mathbf{z}_X = \mathbf{x} - \mathbf{1}\cdot\mu_X$$

$$\Gamma = \frac{1}{N}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbf{X}(i,j)\cdot\mathbf{Y}(i,j) = \frac{1}{N}\mathbf{x}^T\mathbf{y}$$

$$\Gamma_n = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\big(\mathbf{X}(i,j)-\mu_X\big)\big(\cdot\mathbf{Y}(i,j)-\mu_Y\big)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\big(\mathbf{X}(i,j)-\mu_X\big)^2 \quad \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\big(\mathbf{Y}[i]-\mu_Y\big)^2}} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2\sigma_Y^2}}$$

Hubert Statistic

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\mathbf{X}(i,j)-\mu_X\right)^2 = \frac{1}{N}\mathbf{z}_x^T\mathbf{z}_x = \frac{1}{N}\|\mathbf{z}_x\|^2$$

$$\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\mathbf{Y}(i,j)-\mu_Y\right)^2 = \frac{1}{N}\mathbf{z}_y^T\mathbf{z}_y = \frac{1}{N}\left\|\mathbf{z}_y\right\|^2$$

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\mathbf{X}(i,j)-\mu_X\right)\left(\mathbf{Y}(i,j)-\mu_Y\right) = \frac{1}{N}\mathbf{z}_x^T\mathbf{z}_y$$

$$\Gamma_n = \frac{\mathbf{z}_x^T\mathbf{z}_y}{\|\mathbf{z}_x\|\cdot\left\|\mathbf{z}_y\right\|} = \cos\theta$$

Discretized Hubert Statistic

$$\mathbf{T}(i,j) = \begin{cases} 1 & \text{if } y_i = y_j, i \neq j \\ 0 & \text{otherwise} \end{cases} \qquad \mathbf{C}(i,j) = \begin{cases} 1 & \text{if } \hat{y}_i = \hat{y}_j, i \neq j \\ 0 & \text{otherwise} \end{cases}$$

$$\Gamma = \frac{1}{N}\mathbf{t}^T\mathbf{c} = \frac{TP}{N}$$

$$\Gamma_n = \frac{\mathbf{z}_t^T\mathbf{z}_c}{\|\mathbf{z}_t\| \cdot \|\mathbf{z}_c\|} = \cos\theta$$

$$\mu_T = \frac{\mathbf{t}^T\mathbf{t}}{N} = \frac{TP+FN}{N} \qquad \mu_C = \frac{\mathbf{c}^T\mathbf{c}}{N} = \frac{TP+FP}{N}$$

Normalized Discretized Hubert Statistic

$$\begin{aligned}
\mathbf{z}_t^T \mathbf{z}_c &= (\mathbf{t} - \mathbf{1} \cdot \mu_T)^T (\mathbf{c} - \mathbf{1} \cdot \mu_C) \\
&= \mathbf{t}^T \mathbf{c} - \mu_C \mathbf{t}^T \mathbf{1} - \mu_T \mathbf{c}^T \mathbf{1} + \mathbf{1}^T \mathbf{1} \mu_T \mu_C \\
&= \mathbf{t}^T \mathbf{c} - N \mu_C \mu_T - N \mu_T \mu_C + N \mu_T \mu_C \\
&= \mathbf{t}^T \mathbf{c} - N \mu_T \mu_C \\
&= TP - N \mu_T \mu_C
\end{aligned}$$

$$\|\mathbf{z}_t\|^2 = \mathbf{z}_t^T \mathbf{z}_t = \mathbf{t}^T \mathbf{t} - N \mu_T^2 = N \mu_T - N \mu_T^2 = N \mu_T (1 - \mu_T)$$

$$\|\mathbf{z}_c\|^2 = \mathbf{z}_c^T \mathbf{z}_c = \mathbf{c}^T \mathbf{c} - N \mu_C^2 = N \mu_C - N \mu_C^2 = N \mu_C (1 - \mu_C)$$

$$\Gamma_n = \frac{\frac{TP}{N} - \mu_T \mu_C}{\sqrt{\mu_T \mu_C (1 - \mu_T)(1 - \mu_C)}}$$

Example 17.4

$TP = 3030$ $FN = 645$ $FP = 766$ $TN = 6734$

$$\mu_T = \frac{TP + FN}{N} = \frac{3675}{11175} = 0.33$$
$$\mu_C = \frac{TP + FP}{N} = \frac{3796}{11175} = 0.34$$

$$\Gamma = \frac{3030}{11175} = 0.271$$

$$\Gamma_n = \frac{0.27 - 0.33 \cdot 0.34}{\sqrt{0.33 \cdot 0.34 \cdot (1 - 0.33) \cdot (1 - 0.34)}} = \frac{0.159}{0.222} = 0.717$$

$TP = 2891$ $FN = 784$ $FP = 2380$ $TN = 5120$

$\Gamma = 0.258$ $\Gamma_n = 0.442$

External Measures, Part 2

The End

Internal Measures

Internal Measures

- Internal evaluation measures have no relationships with the ground-truth partitioning
- To evaluate the measures we must utilize intracluster similarities or compactness
- Internal measure of the pairwise distances are the distance matrix also referred to as the proximity matrix

Internal Measures

$$\mathbf{W} = \left\{\delta(\mathbf{x}_i, \mathbf{x}_j)\right\}_{i,j=1}^{n} \qquad \delta(\mathbf{x}_i, \mathbf{x}_j) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2$$

$$W(S,R) = \sum_{\mathbf{x}_i \in S} \sum_{\mathbf{x}_j \in R} w_{ij}$$

$$W_{in} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, C_i) \qquad N_{in} = \sum_{i=1}^{k}\binom{n_i}{2} = \frac{1}{2}\sum_{i=1}^{k} n_i(n_i - 1)$$

$$W_{out} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1}\sum_{j>i} W(C_i, C_j) \qquad N_{out} = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i \cdot n_j = \frac{1}{2}\sum_{i=1}^{k}\sum_{\substack{j=1 \\ j \neq i}}^{k} n_i \cdot n_j$$

$$N = N_{in} + N_{out} = \binom{n}{2} = \frac{1}{2}n(n-1)$$

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

BetaCV Measure and C-index

$$BetaCV = \frac{W_{in}/N_{in}}{W_{out}/N_{out}} = \frac{N_{out}}{N_{in}} \cdot \frac{W_{in}}{W_{out}} = \frac{N_{out}}{N_{in}} \frac{\sum_{i=1}^{k} W(C_i, C_i)}{\sum_{i=1}^{k} W(C_i, \overline{C_i})}$$

$$Cindex = \frac{W_{in} - W_{\min}(N_{in})}{W_{\max}(N_{in}) - W_{\min}(N_{in})}$$

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

Normalized Cut Measure

$$NC = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{vol(C_i)} = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{W(C_i, V)}$$

$$NC = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{W(C_i, C_i) + W(C_i, \overline{C_i})} = \sum_{i=1}^{k} \frac{1}{\frac{W(C_i, C_i)}{W(C_i, \overline{C_i})} + 1}$$

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

Modularity

$$Q = \sum_{i=1}^{k} \left(\frac{W(C_i, C_i)}{W(V, V)} - \left(\frac{W(C_i, V)}{W(V, V)} \right)^2 \right)$$

$$\begin{aligned} W(V, V) &= \sum_{i=1}^{k} W(C_i, V) \\ &= \sum_{i=1}^{k} W(C_i, C_i) + \sum_{i=1}^{k} W(C_i, \overline{C_i}) \\ &= 2(W_{in} + W_{out}) \end{aligned}$$

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

Dunn Index

$$Dunn = \frac{W_{out}^{\min}}{W_{in}^{\max}}$$

$$W_{out}^{\min} = \min_{i,j>i} \left\{ w_{ab} | \mathbf{x}_a \in C_i, \mathbf{x}_b \in C_j \right\}$$

$$W_{in}^{\max} = \max_{i} \left\{ w_{ab} | \mathbf{x}_a, \mathbf{x}_b \in C_i \right\}$$

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

Davies-Bouldin Index

$$\mu_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j \qquad \sigma_{\mu_i} = \sqrt{\frac{\sum_{\mathbf{x}_j \in C_i} \delta(\mathbf{x}_j, \mu_i)^2}{n_i}} = \sqrt{var(C_i)}$$

$$DB_{ij} = \frac{\sigma_{\mu_i} + \sigma_{\mu_j}}{\delta(\mu_i, \mu_j)}$$

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \{DB_{ij}\}$$

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

Silhouette Coefficient

$$s_i = \frac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\left\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\right\}}$$

$$\mu_{in}(\mathbf{x}_i) = \frac{\sum_{\mathbf{x}_j \in C_{\hat{y}_i}, j \neq i} \delta(\mathbf{x}_i, \mathbf{x}_j)}{n_{\hat{y}_i} - 1}$$

$$\mu_{out}^{\min}(\mathbf{x}_i) = \min_{j \neq \hat{y}_i} \left\{ \frac{\sum_{\mathbf{y} \in C_j} \delta(\mathbf{x}_i, \mathbf{y})}{n_j} \right\}$$

$$SC = \frac{1}{n} \sum_{i=1}^{n} s_i$$

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

Hubert Statistic

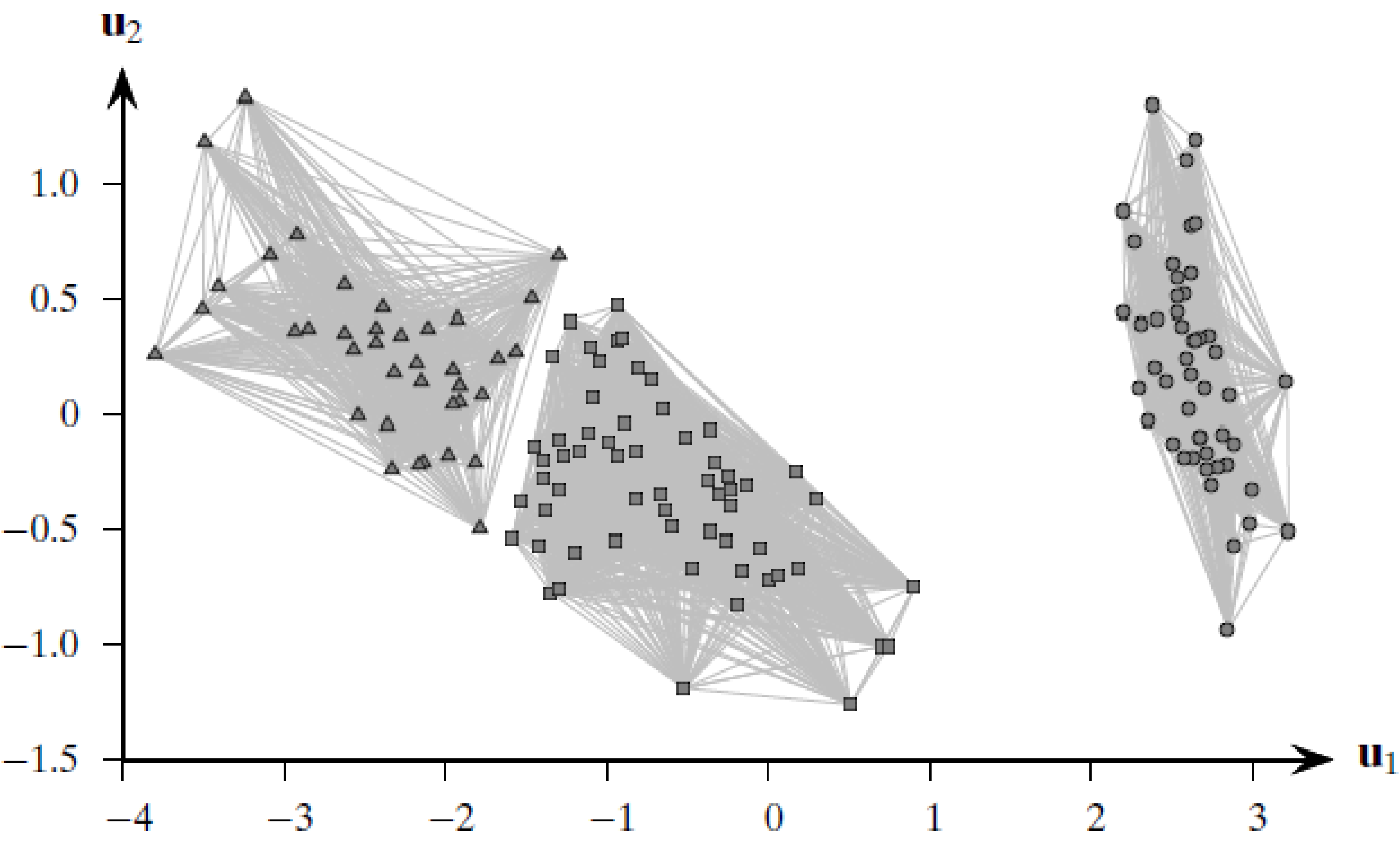
- The Hubert statistic can be used as an internal measure when X = W

$$\mathbf{Y} = \left\{ \delta(\mu_{\hat{y}_i}, \mu_{\hat{y}_j}) \right\}_{i,j=1}^{n}$$

- If W and Y are symmetric Γ and Γ_n can be computer over their upper elements

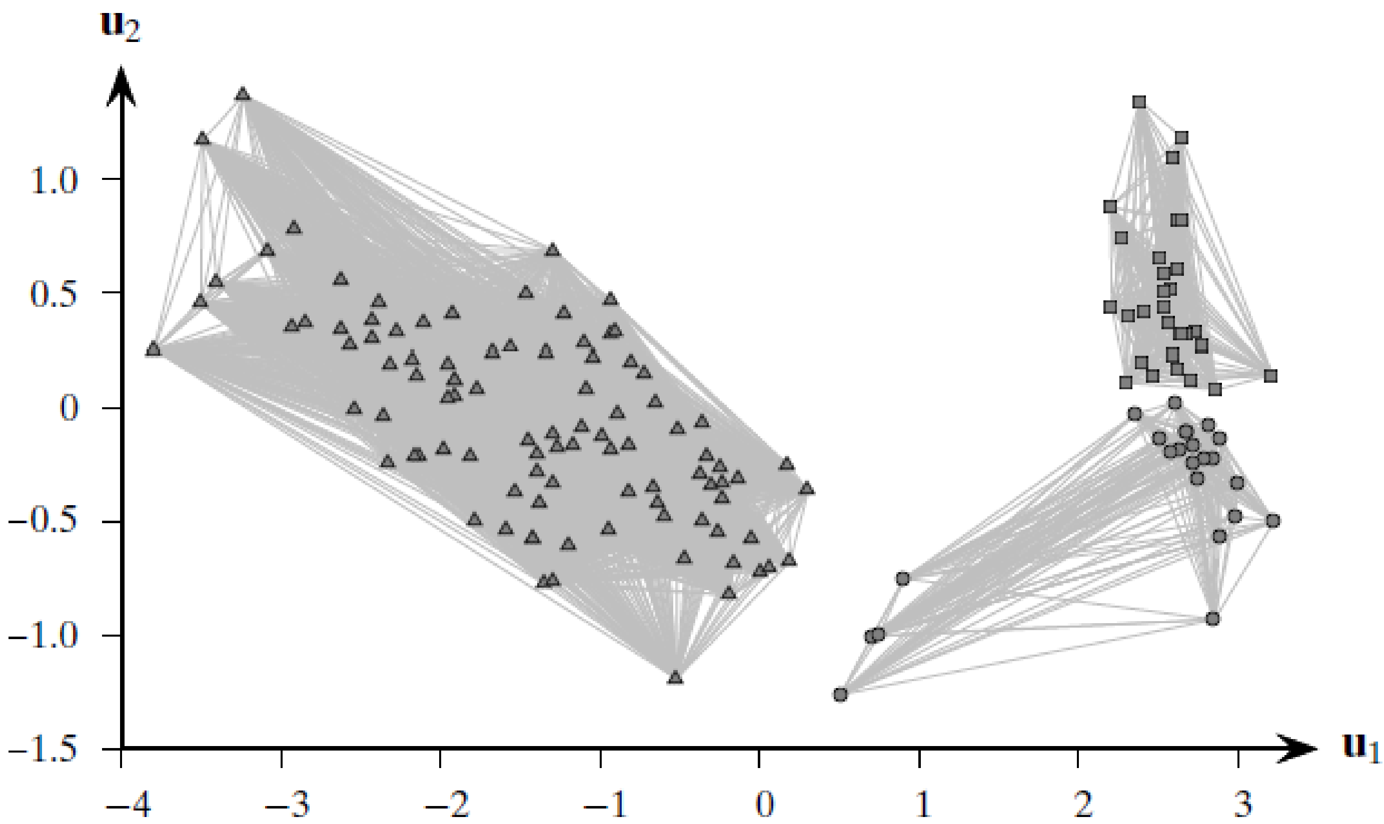
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 17.6



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

Example 17.6



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

Example 17.6

$n_1 = 61$ $\quad n_2 = 50$ $\quad n_3 = 39$

$$N_{in} = \binom{61}{2} + \binom{50}{2} + \binom{31}{2} = 1830 + 1225 + 741 = 3796$$

$$N_{out} = 61 \cdot 50 + 61 \cdot 39 + 50 \cdot 39 = 3050 + 2379 + 1950 = 7379$$

$$N = N_{in} + N_{out} = 3796 + 7379 = 11175$$

$$\begin{array}{c|ccc}
W & C_1 & C_2 & C_3 \\
\hline
C_1 & 3265.69 & 10402.30 & 4418.62 \\
C_2 & 10402.30 & 1523.10 & 9792.45 \\
C_3 & 4418.62 & 9792.45 & 1252.36
\end{array}$$

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

Example 17.6

$$W_{in} = \frac{1}{2}(3265.69 + 1523.10 + 1252.36) = 3020.57$$

$$W_{out} = (10402.30 + 4418.62 + 9792.45) = 24613.37$$

$$BetaCV = \frac{N_{out} \cdot W_{in}}{N_{in} \cdot W_{out}} = \frac{7379 \times 3020.57}{3796 \times 24613.37} = 0.239$$

$$W_{\min}(N_{in}) = 2535.96 \qquad W_{\max}(N_{in}) = 16889.57$$

$$Cindex = \frac{W_{in} - W_{\min}(N_{in})}{W_{\max}(N_{in}) - W_{\min}(N_{in})} = \frac{3020.57 - 2535.96}{16889.57 - 2535.96} = \frac{484.61}{14535.61} = 0.0338$$

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

Example 17.6

$$W(C_1, \overline{C_1}) = 10402.30 + 4418.62 = 14820.91$$

$$W(C_2, \overline{C_2}) = 10402.30 + 9792.45 = 20194.75$$

$$W(C_3, \overline{C_3}) = 4418.62 + 9792.45 = 14211.07$$

$$W(C_1, V) = 3265.69 + W(C_1, \overline{C_1}) = 18086.61$$

$$W(C_2, V) = 1523.10 + W(C_2, \overline{C_2}) = 21717.85$$

$$W(C_3, V) = 1252.36 + W(C_3, \overline{C_3}) = 15463.43$$

$$W(V, V) = W(C_1, V) + W(C_2, V) + W(C_3, V) = 55267.89$$

$$NC = \frac{14820.91}{18086.61} + \frac{20194.75}{21717.85} + \frac{14211.07}{15463.43} = 0.819 + 0.93 + 0.919 = 2.67$$

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

Example 17.6

$$Q = \left(\frac{3265.69}{55267.89} - \left(\frac{18086.61}{55267.89}\right)^2\right) + \left(\frac{1523.10}{55267.89} - \left(\frac{21717.85}{55267.89}\right)^2\right)$$
$$+ \left(\frac{1252.36}{55267.89} - \left(\frac{15463.43}{55267.89}\right)^2\right)$$
$$= -0.048 - 0.1269 - 0.0556 = -0.2305$$

$W^{\min}$	C_1	C_2	C_3
C_1	0	1.62	0.198
C_2	1.62	0	3.49
C_3	0.198	3.49	0

$W^{\max}$	C_1	C_2	C_3
C_1	2.50	4.85	4.81
C_2	4.85	2.33	7.06
C_3	4.81	7.06	2.55

$$Dunn = \frac{W_{out}^{\min}}{W_{in}^{\max}} = \frac{0.198}{2.55} = 0.078$$

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

Example 17.6

$$\mu_1 = \begin{pmatrix} -0.664 \\ -0.33 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 2.64 \\ 0.19 \end{pmatrix} \qquad \mu_3 = \begin{pmatrix} -2.35 \\ 0.27 \end{pmatrix}$$

$$\sigma_{\mu_1} = 0.723 \qquad \sigma_{\mu_2} = 0.512 \qquad \sigma_{\mu_3} = 0.695$$

$$\begin{pmatrix} \begin{array}{c|ccc} DB_{ij} & C_1 & C_2 & C_3 \\ \hline C_1 & - & 0.369 & 0.794 \\ C_2 & 0.369 & - & 0.242 \\ C_3 & 0.794 & 0.242 & - \end{array} \end{pmatrix}$$

$$DB = \frac{1}{3}(0.794 + 0.369 + 0.794) = 0.652$$

$$s_1 = \frac{1.902 - 0.701}{\max\{1.902, 0.701\}} = \frac{1.201}{1.902} = 0.632 \qquad SC = 0.598$$

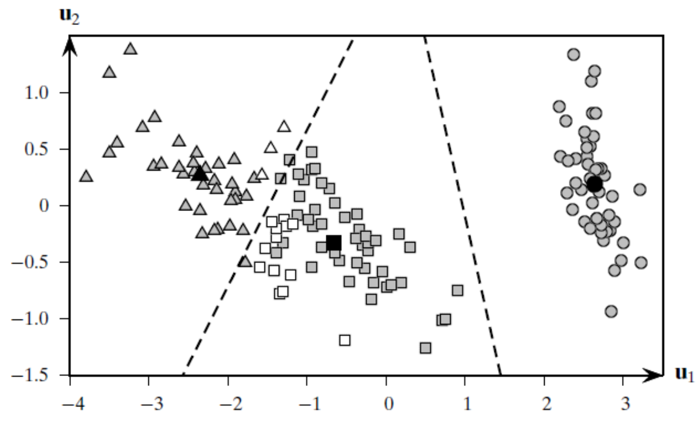
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

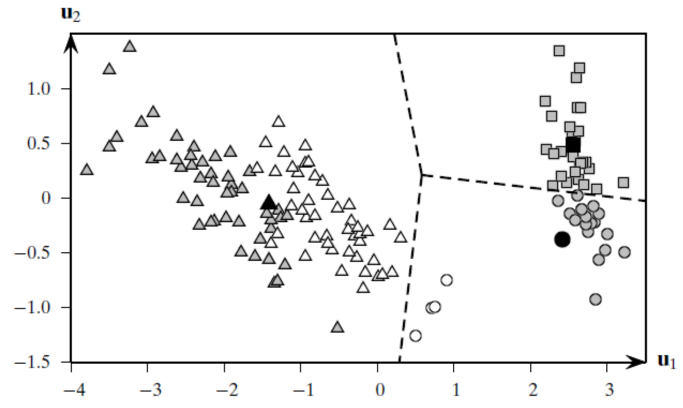
Example 17.6

$$\Gamma = \frac{\mathbf{w}^T\mathbf{y}}{N} = \frac{91545.85}{11175} = 8.19$$

$$\Gamma_n = \frac{\mathbf{z}_w^T\mathbf{z}_y}{\|\mathbf{x}_w\| \cdot \|\mathbf{z}_y\|} = 0.918$$

	Lower better				Higher better				
	BetaCV	*Cindex*	*Q*	*DB*	*NC*	*Dunn*	*SC*	Γ	Γ_n
(a) Good	0.24	0.034	−0.23	0.65	2.67	0.08	0.60	8.19	0.92
(b) Bad	0.33	0.08	−0.20	1.11	2.56	0.03	0.55	7.32	0.83





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

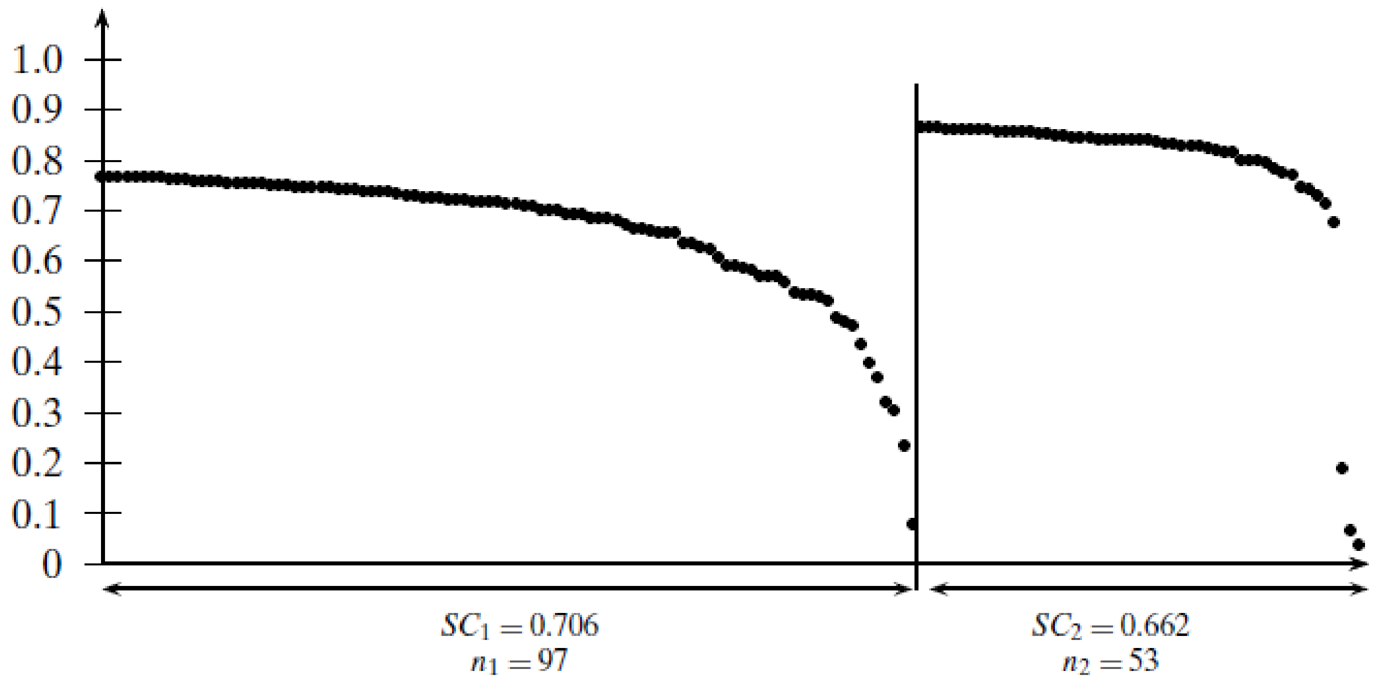
Internal Measures

The End

Relative Measures

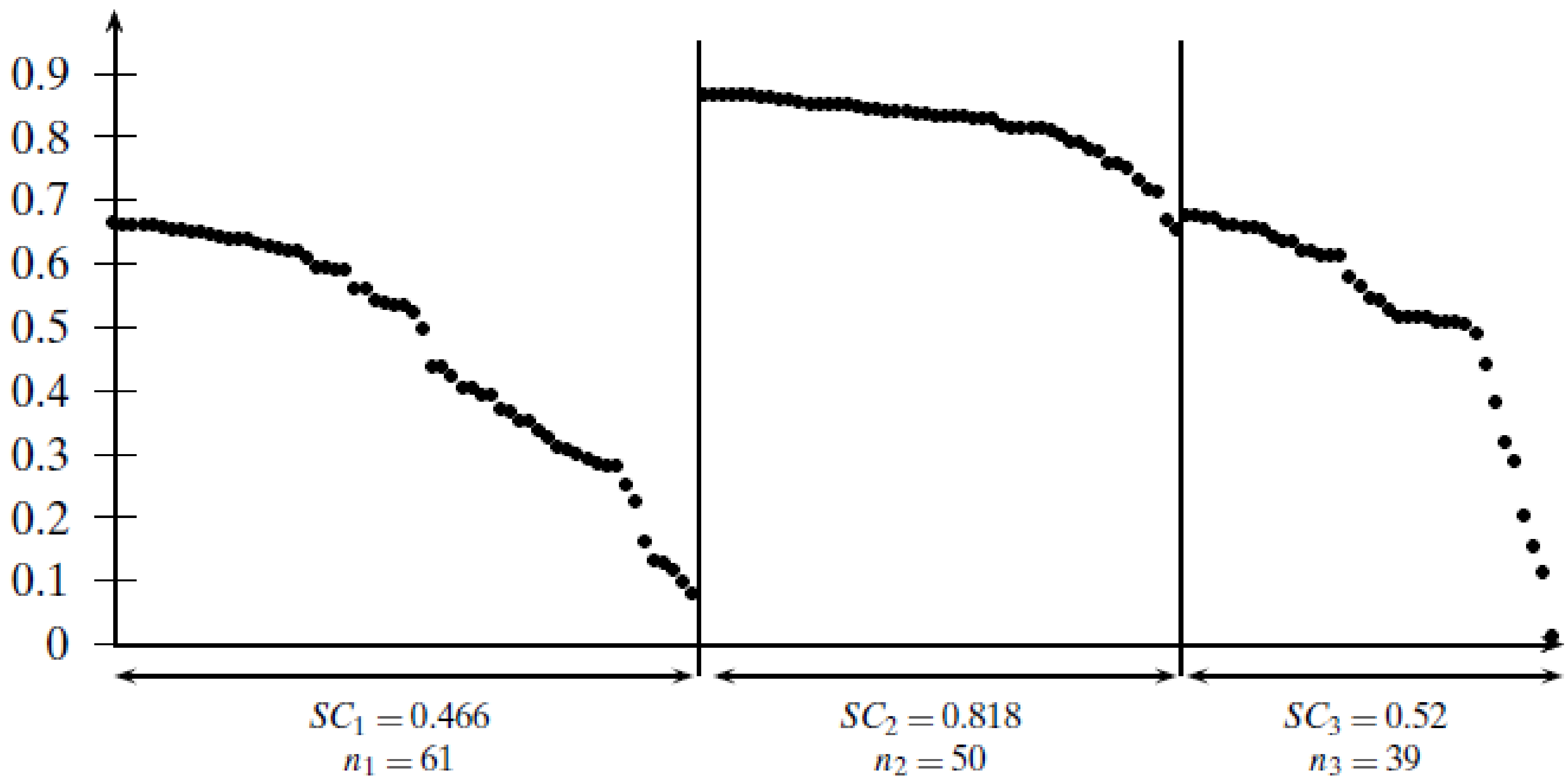
Silhouette Coefficient

$$SC_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} s_j$$



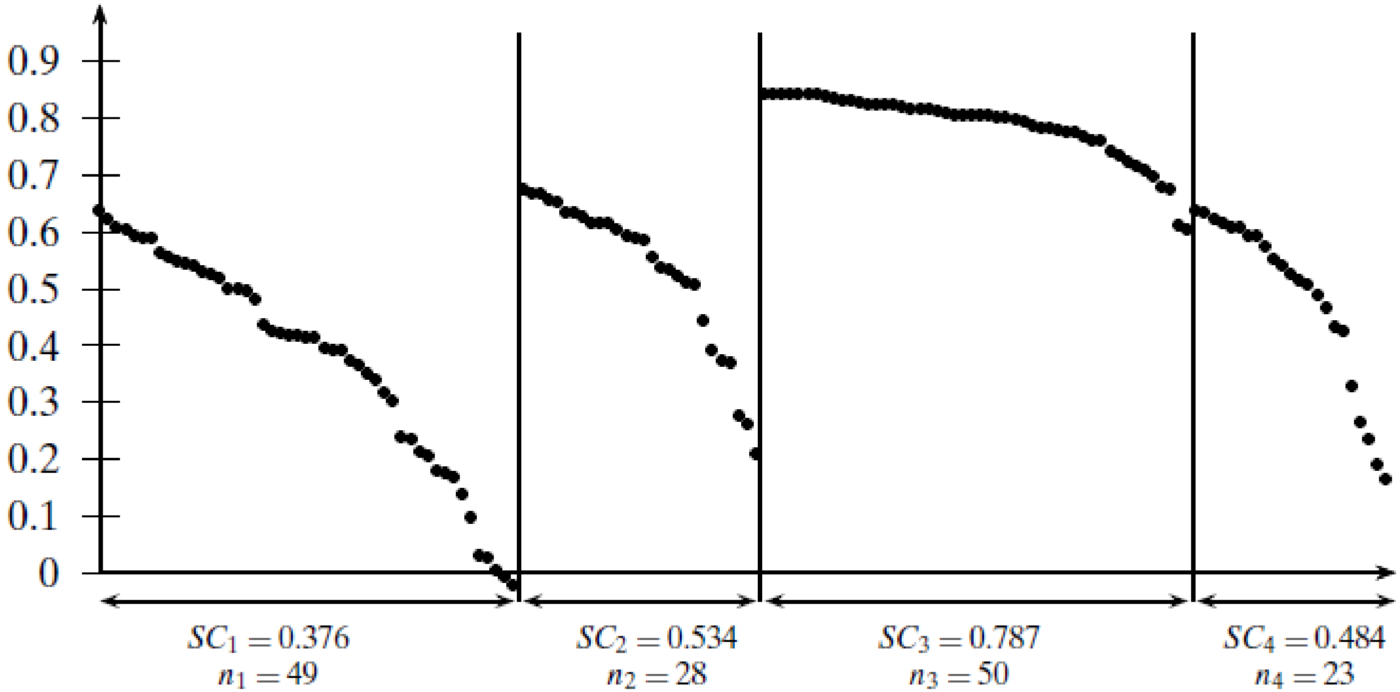
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Silhouette Coefficient



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

Silhouette Coefficient



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

Calinski-Harabasz Index

$$\mathbf{S} = n\boldsymbol{\Sigma} = \sum_{j=1}^{n} \left(\mathbf{x}_j - \mu\right)\left(\mathbf{x}_j - \mu\right)^T$$

$$\mathbf{S}_W = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \left(\mathbf{x}_j - \mu_i\right)\left(\mathbf{x}_j - \mu_i\right)^T$$

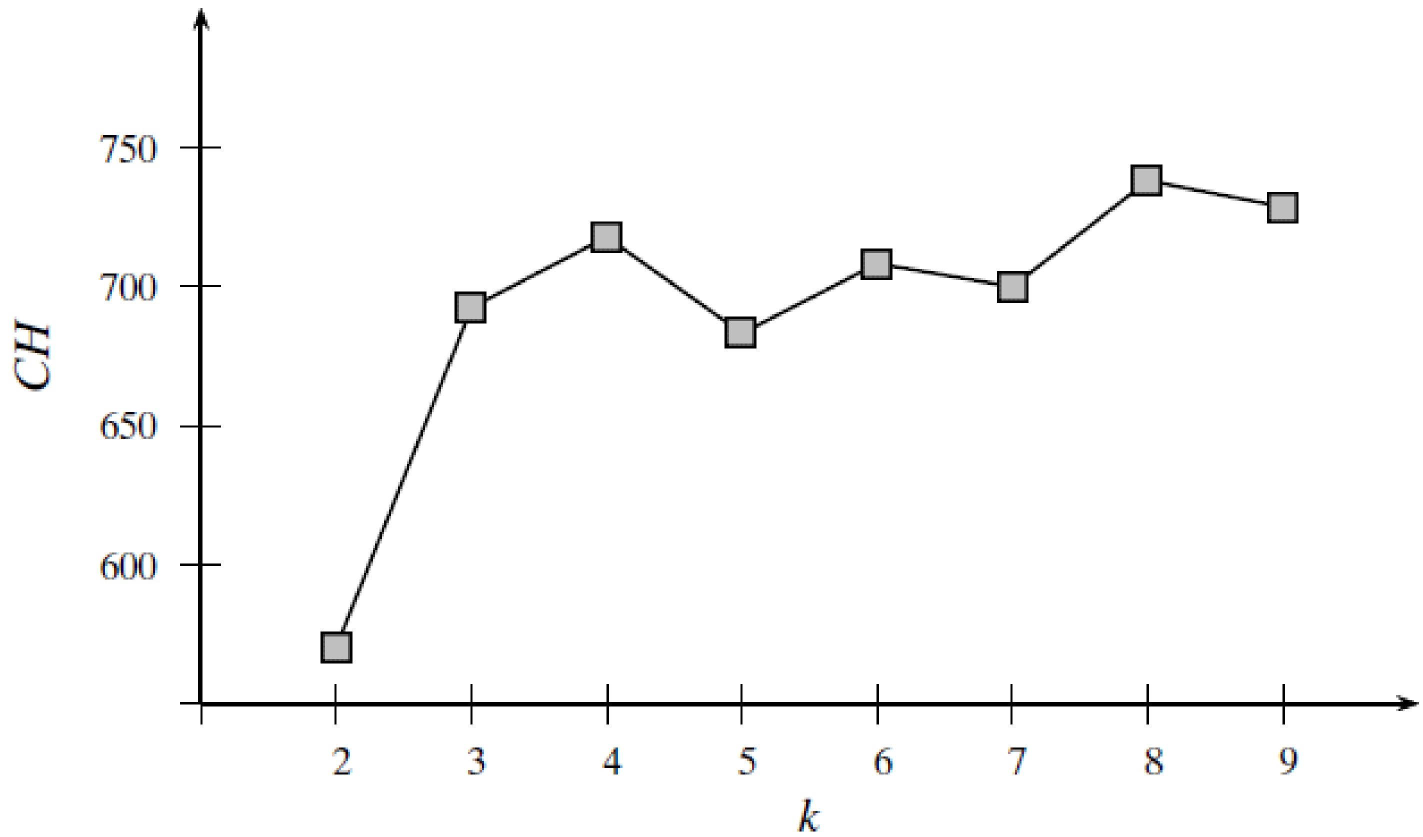
$$\mathbf{S}_B = \sum_{i=1}^{k} n_i \left(\mu_i - \mu\right)\left(\mu_i - \mu\right)^T$$

$$CH(k) = \frac{tr(\mathbf{S}_B)/(k-1)}{tr(\mathbf{S}_W)/(n-k)} = \frac{n-k}{k-1} \cdot \frac{tr(\mathbf{S}_B)}{tr(\mathbf{S}_W)}$$

$$\Delta(k) = \Big(CH(k+1) - CH(k)\Big) - \Big(CH(k) - CH(k-1)\Big)$$

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

Calinski-Harabasz Index



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

Example 17.8

$$\mathbf{S}_W = \begin{pmatrix} 39.14 & -13.62 \\ -13.62 & 24.73 \end{pmatrix}$$

$$\mathbf{S}_B = \begin{pmatrix} 590.36 & 13.62 \\ 13.62 & 11.36 \end{pmatrix}$$

$$CH(3) = \frac{(150-3)}{(3-1)} \cdot \frac{(590.36+11.36)}{(39.14+24.73)} = (147/2) \cdot \frac{601.72}{63.87} = 73.5 \cdot 9.42 = 692.4$$

k	2	3	4	5	6	7	8	9
$CH(k)$	570.25	692.40	717.79	683.14	708.26	700.17	738.05	728.63
$\Delta(k)$	–	−96.78	−60.03	59.78	−33.22	45.97	−47.30	–

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

Gap Statistic

$$\mu_W(k) = \frac{1}{t}\sum_{i=1}^{t} \log W_{in}^k(\mathbf{R}_i)$$

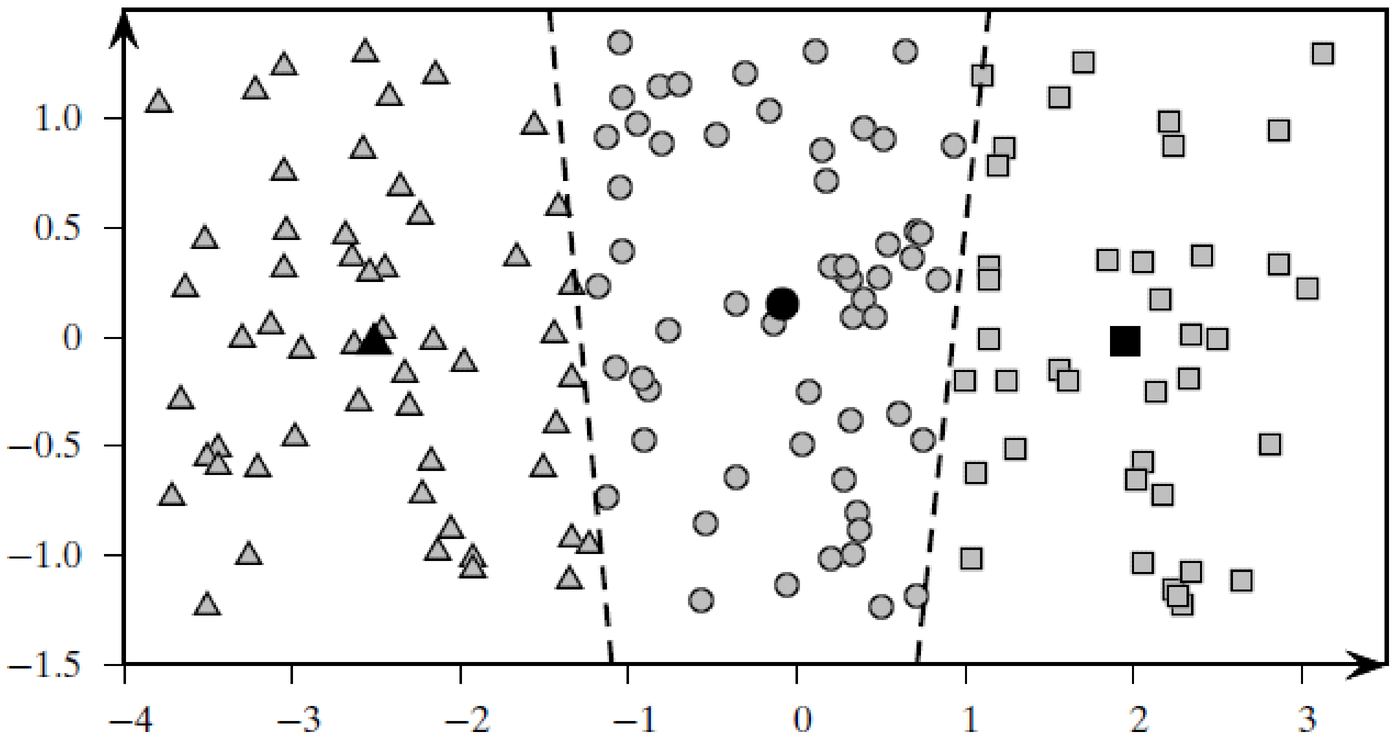
$$\sigma_W(k) = \sqrt{\frac{1}{t}\sum_{i=1}^{t}\Big(\log W_{in}^k(\mathbf{R}_i) - \mu_W(k)\Big)^2}$$

$$gap(k) = \mu_W(k) - \log W_{in}^k(\mathbf{D})$$

$$k^* = \arg\min_k \Big\{ gap(k) \geq gap(k+1) - \sigma_W(k+1) \Big\}$$

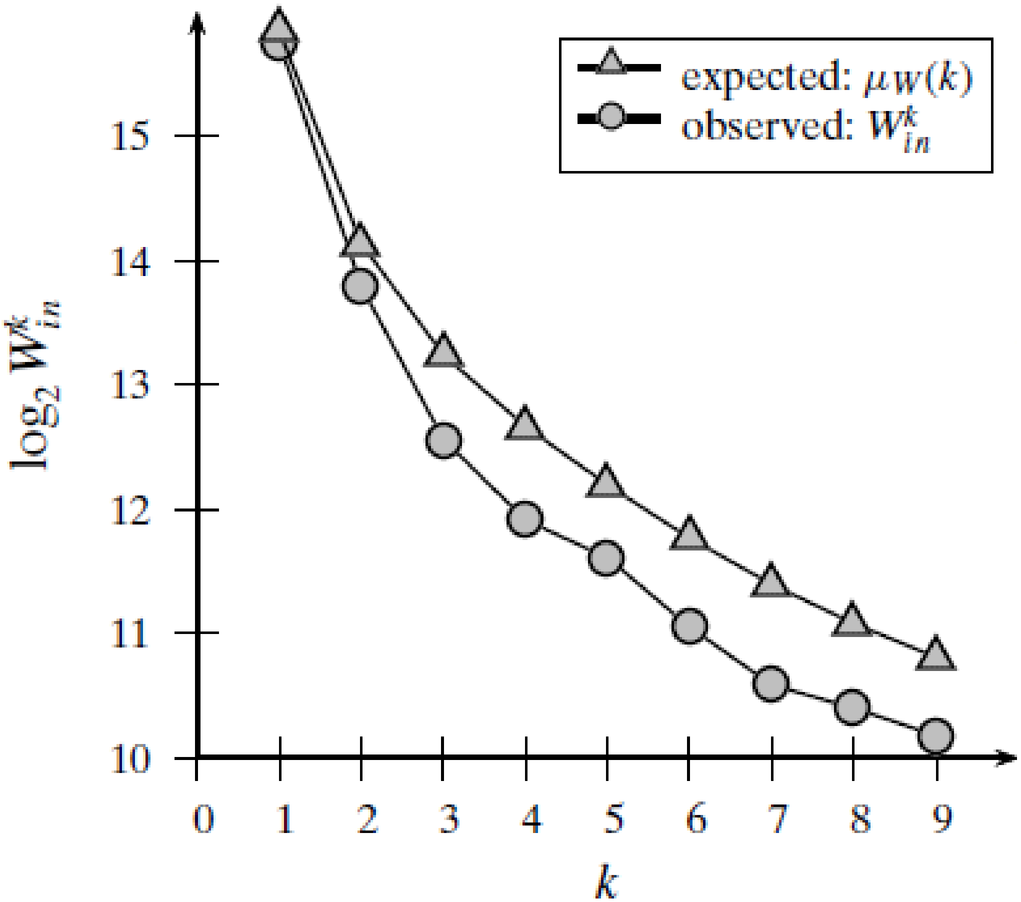
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Gap Statistic



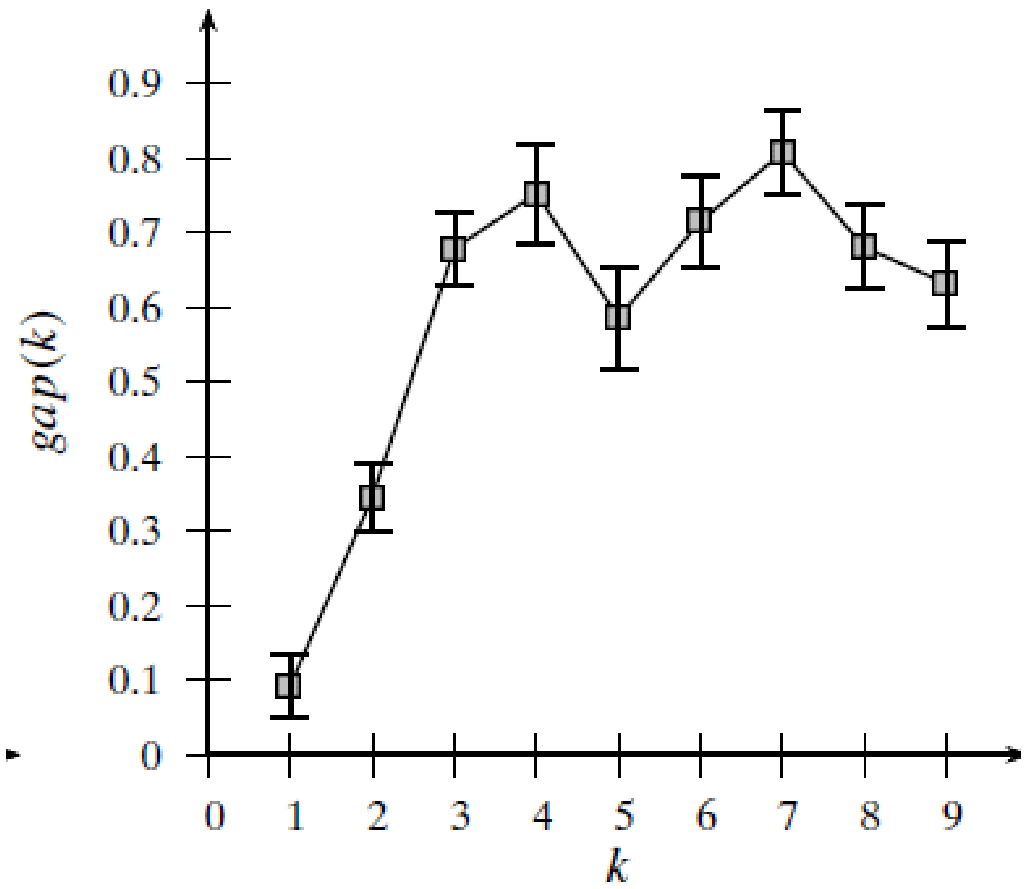
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 17.9



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

Example 17.9



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

Example 17.9

k	$gap(k)$	$\sigma_W(k)$	$gap(k) - \sigma_W(k)$
1	0.093	0.0456	0.047
2	0.346	0.0486	0.297
3	0.679	0.0529	0.626
4	0.753	0.0701	0.682
5	0.586	0.0711	0.515
6	0.715	0.0654	0.650
7	0.808	0.0611	0.746
8	0.680	0.0597	0.620
9	0.632	0.0606	0.571

$$gap(4) = 0.753 > gap(5) - \sigma_W(5) = 0.515$$

$$gap(3) = 0.679 > gap(4) - 2\sigma_W(4) = 0.753 - 2 \cdot 0.0701 = 0.613$$

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

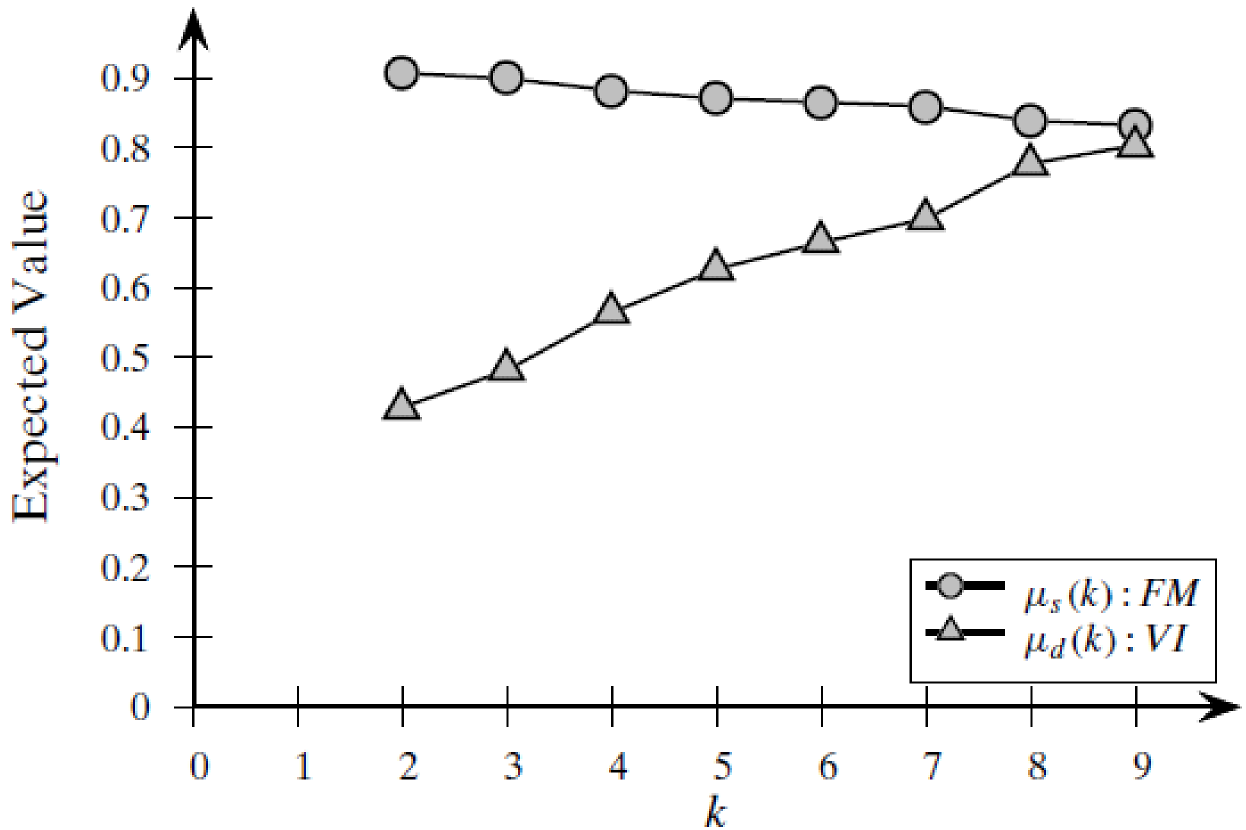
Cluster Stability

ClusteringStability $(A, t, k^{\max}, \mathbf{D})$:

$n \leftarrow |\mathbf{D}|$
`// Generate t samples`
for $i = 1, 2, \ldots, t$ **do**
  $\mathbf{D}_i \leftarrow$ sample n points from $\mathbf{D}$ with replacement
`// Generate clusterings for different values of k`
for $i = 1, 2, \ldots, t$ **do**
  for $k = 2, 3, \ldots, k^{\max}$ **do**
    $\mathcal{C}_k(\mathbf{D}_i) \leftarrow$ cluster $\mathbf{D}_i$ into k clusters using algorithm A
`// Compute mean difference between clusterings for each k`
foreach *pair* $\mathbf{D}_i, \mathbf{D}_j$ *with* $j > i$ **do**
  $\mathbf{D}_{ij} \leftarrow \mathbf{D}_i \cap \mathbf{D}_j$ `// create common dataset using Eq.` (17.30)
  for $k = 2, 3, \ldots, k^{\max}$ **do**
    $d_{ij}(k) \leftarrow d\big(\mathcal{C}_k(\mathbf{D}_i), \mathcal{C}_k(\mathbf{D}_j), \mathbf{D}_{ij}\big)$ `// distance between clusterings`
for $k = 2, 3, \ldots, k^{\max}$ **do**
  $\mu_d(k) \leftarrow \frac{2}{t(t-1)} \sum_{i=1}^{t} \sum_{j>i} d_{ij}(k)$ `// expected pairwise distance`
`// Choose best k`
$k^* \leftarrow \arg\min_k \{\mu_d(k)\}$

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

Example 17.6



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

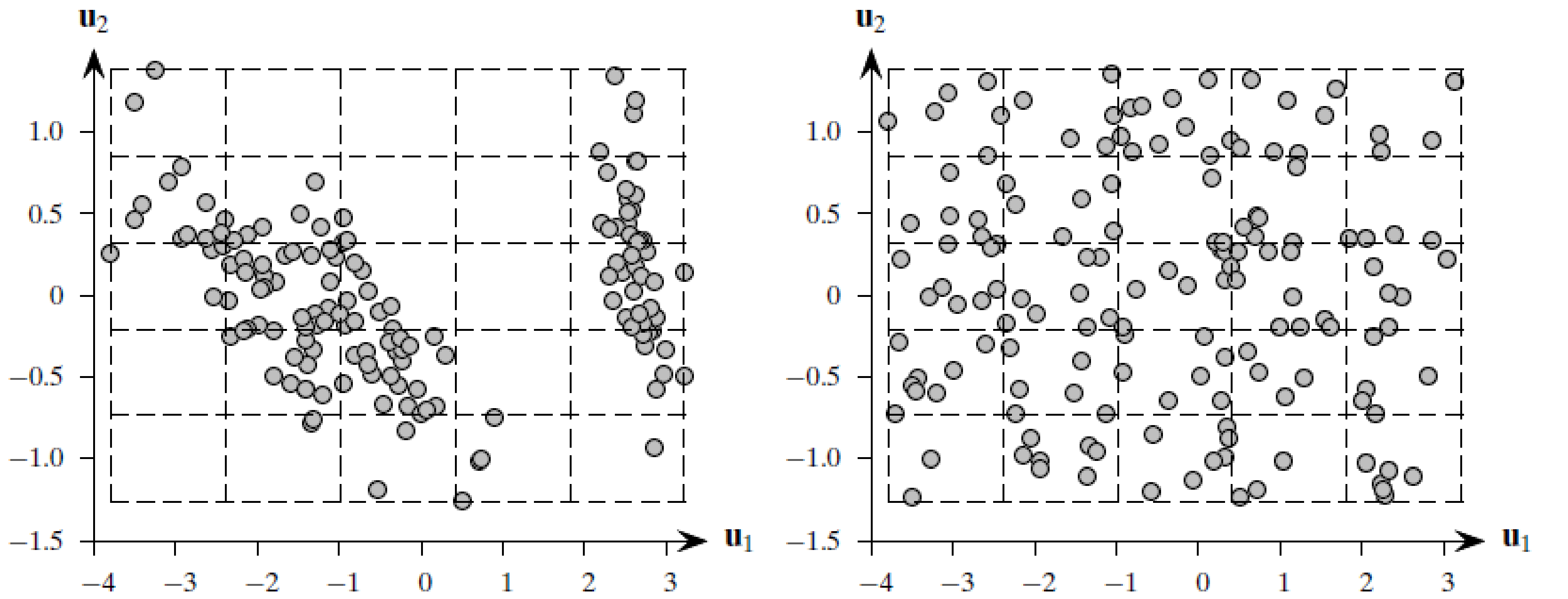
Spatial Histogram

$$f(\mathbf{i}) = P(\mathbf{x}_j \in \text{cell } \mathbf{i}) = \frac{\left|\{\mathbf{x}_j \in \text{cell } \mathbf{i}\}\right|}{n}$$

$$KL(f \mid g_j) = \sum_{\mathbf{i}} f(\mathbf{i}) \log\left(\frac{f(\mathbf{i})}{g_j(\mathbf{i})}\right)$$

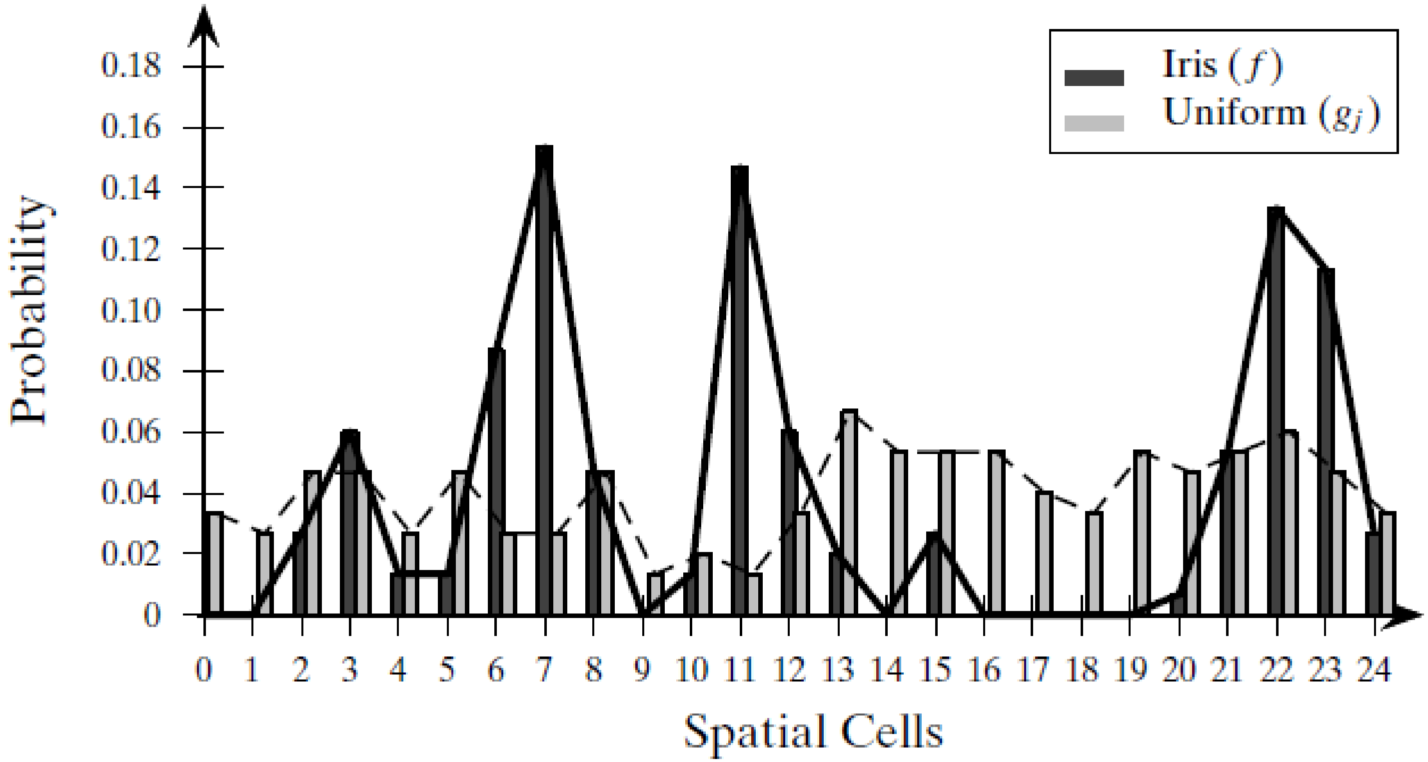
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 17.11



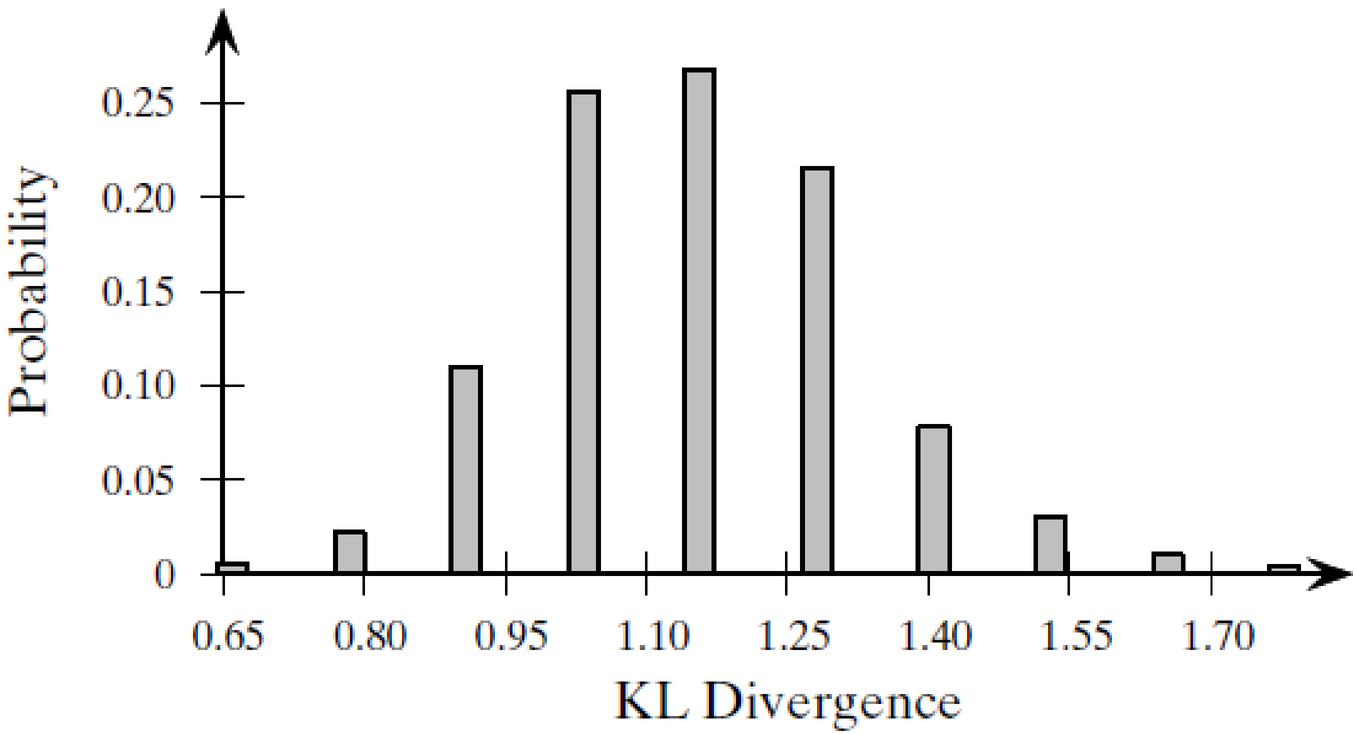
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.*

Example 17.11



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

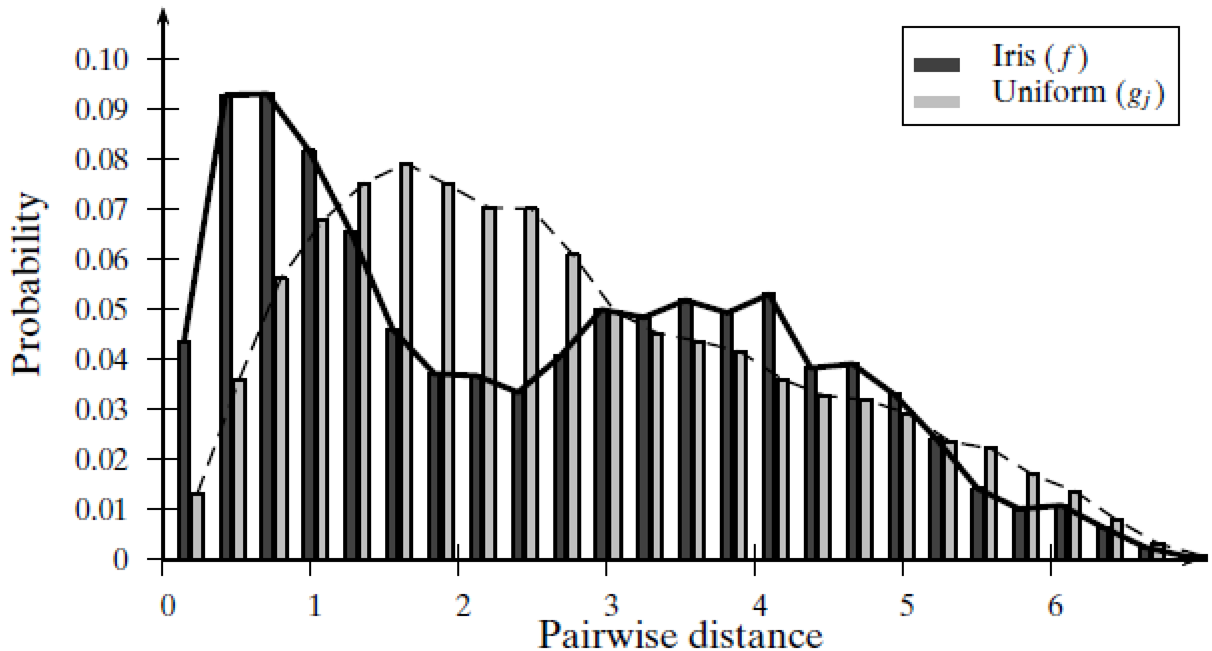
Example 17.11



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

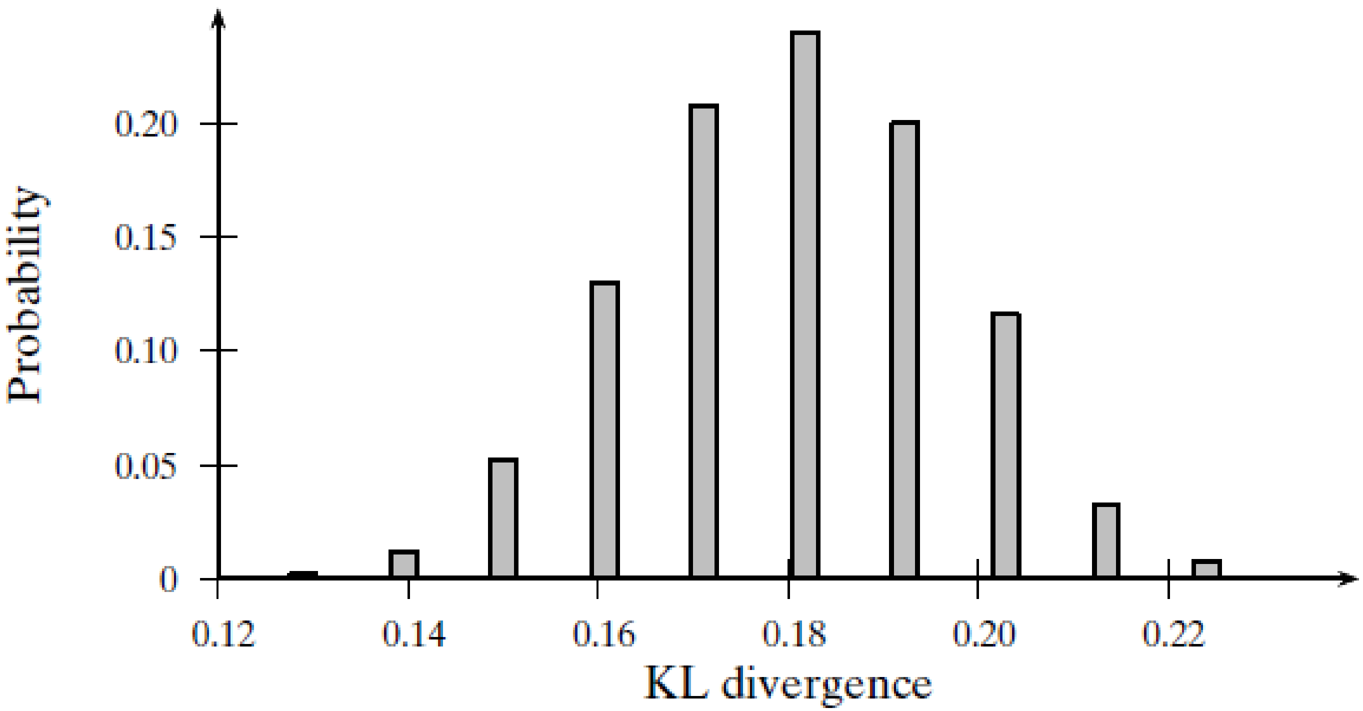
Distance Distributions

$$f(i) = P(w_{pq} \in \text{ bin } i \mid \mathbf{x}_p, \mathbf{x}_q \in \mathbf{D}, p < q) = \frac{\left|\{w_{pq} \in \text{ bin } i\}\right|}{n(n-1)/2}$$



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

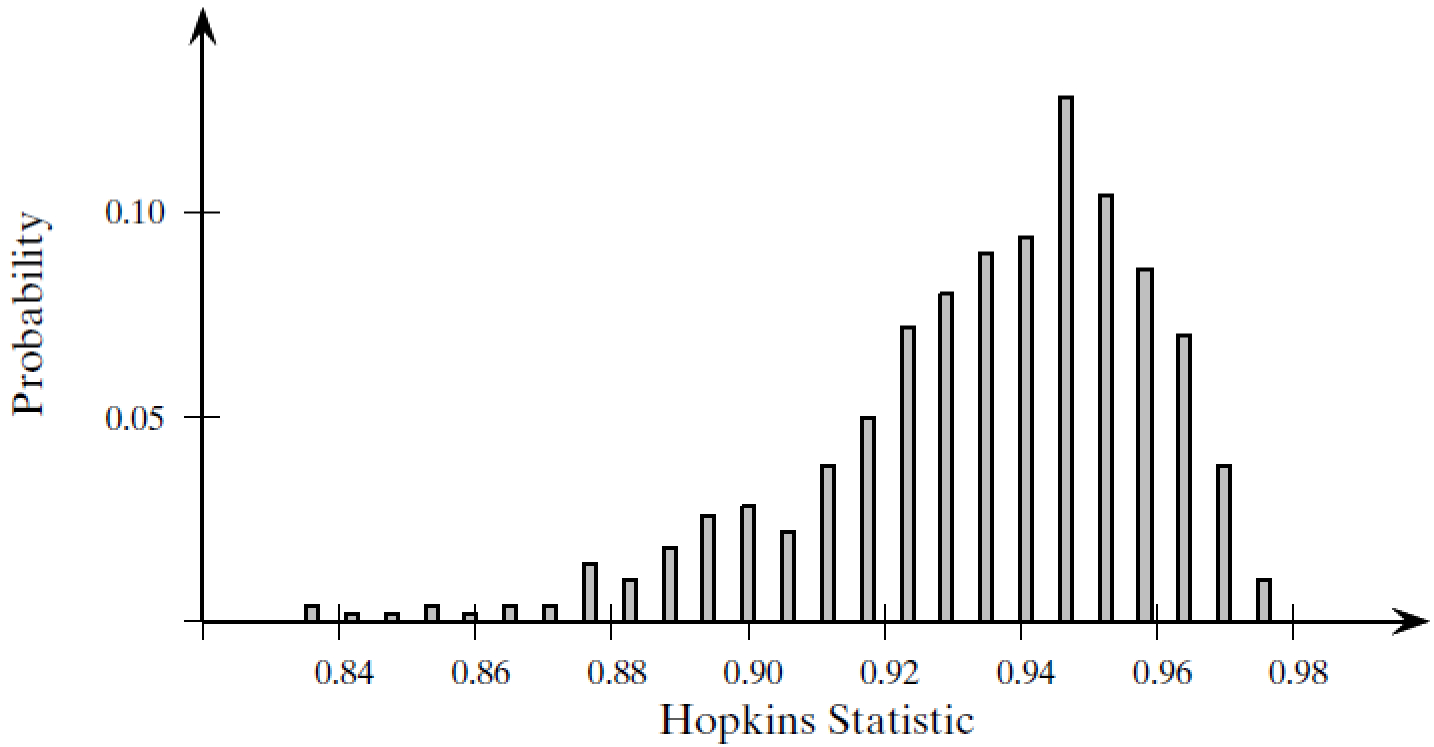
Distance Distributions



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

Hopkins Statistic

$$\delta_{\min}(\mathbf{x}_j) = \min_{\mathbf{x}_i \in \mathbf{D}, \mathbf{x}_i \neq \mathbf{x}_j} \left\{ \delta(\mathbf{x}_j, \mathbf{x}_i) \right\} \qquad HS_i = \frac{\sum_{\mathbf{y}_j \in \mathbf{R}_i} \left(\delta_{\min}(\mathbf{y}_j)\right)^d}{\sum_{\mathbf{y}_j \in \mathbf{R}_i} \left(\delta_{\min}(\mathbf{y}_j)\right)^d + \sum_{\mathbf{x}_j \in \mathbf{D}_i} \left(\delta_{\min}(\mathbf{x}_j)\right)^d}$$



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

The End