Clustering Algorithms

Johannes Blömer

WS 2015/16

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Introduction

Clustering techniques for data management and analysis that classify/group given set of objects into categories/subgroups or clusters

Clustering techniques for data management and analysis that classify/group given set of objects into categories/subgroups or clusters

Clusters homogeneous subgroups of objects such that similarity b/w objects in one subgroup is larger than similarity b/w objects from different subgroups

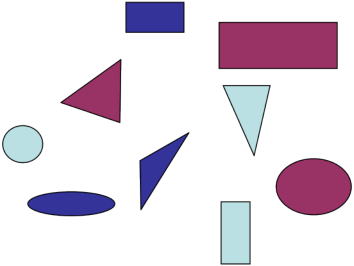
Clustering techniques for data management and analysis that classify/group given set of objects into categories/subgroups or clusters

Clusters homogeneous subgroups of objects such that similarity b/w objects in one subgroup is larger than similarity b/w objects from different subgroups

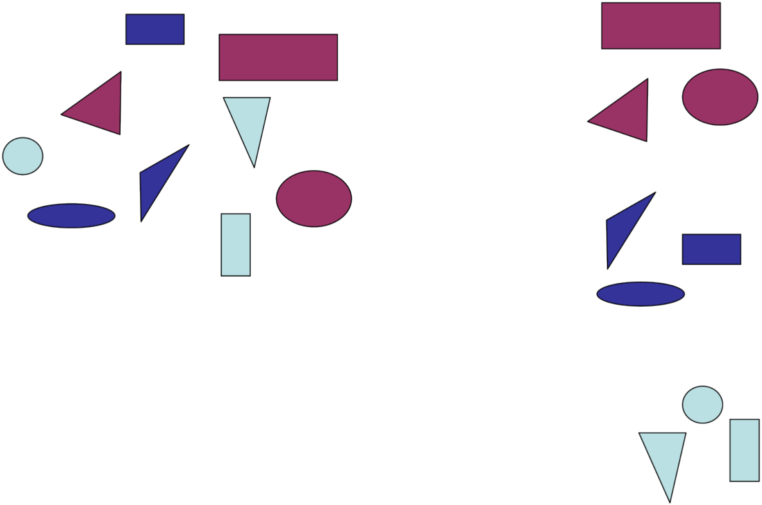
Goals

1. find structures in large set of objects/data
2. simplify large data sets

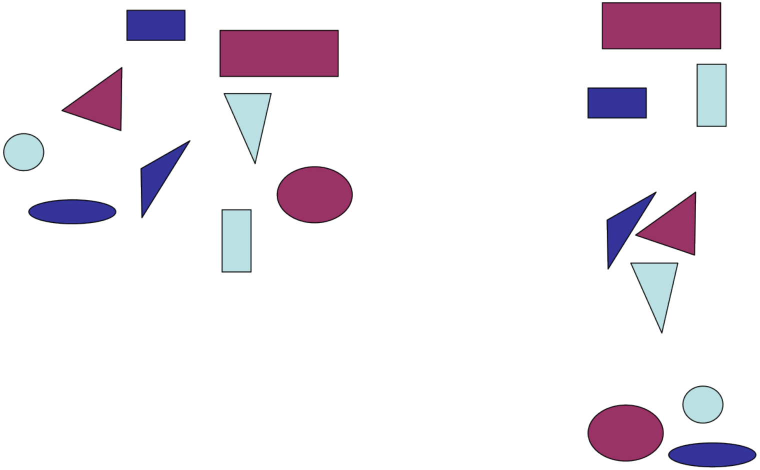
Example

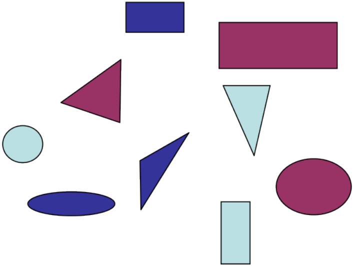


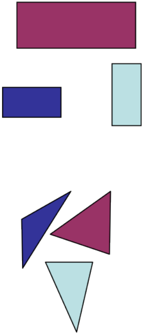
Example



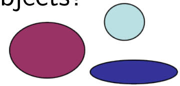
Example



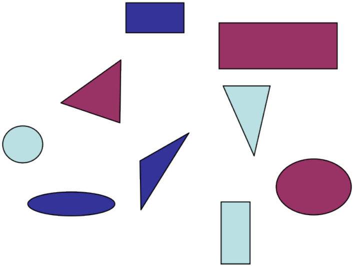


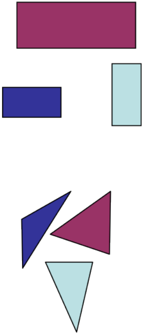


How do we measure similarity/dissimilarity of objects?



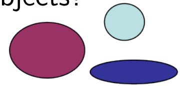
Example





How do we measure similarity/dissimilarity of objects?

How do we measure quality of clustering?



Application areas

1. information retrieval

Application areas

1. information retrieval
2. data mining

Application areas

1. information retrieval
2. data mining
3. machine learning

Application areas

1. information retrieval
2. data mining
3. machine learning
4. statistics

Application areas

1. information retrieval
2. data mining
3. machine learning
4. statistics
5. pattern recognition

Application areas

1. information retrieval
2. data mining
3. machine learning
4. statistics
5. pattern recognition
6. computer graphics

Application areas

1. information retrieval
2. data mining
3. machine learning
4. statistics
5. pattern recognition
6. computer graphics
7. data compression

Application areas

1. information retrieval
2. data mining
3. machine learning
4. statistics
5. pattern recognition
6. computer graphics
7. data compression
8. bioinformatics

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Application areas

1. information retrieval
2. data mining
3. machine learning
4. statistics
5. pattern recognition
6. computer graphics
7. data compression
8. bioinformatics
9. speech recognition.

Goals of this course

- different models for clustering

Goals of this course

- different models for clustering
- many important clustering heuristics, including agglomerative clustering, Lloyd's algorithm, and the EM algorithm

Goals of this course

- different models for clustering
- many important clustering heuristics, including agglomerative clustering, Lloyd's algorithm, and the EM algorithm
- the limitations of these heuristics

Goals of this course

- different models for clustering
- many important clustering heuristics, including agglomerative clustering, Lloyd's algorithm, and the EM algorithm
- the limitations of these heuristics
- improvements to these heuristics

Goals of this course

- different models for clustering
- many important clustering heuristics, including agglomerative clustering, Lloyd's algorithm, and the EM algorithm
- the limitations of these heuristics
- improvements to these heuristics
- various theoretical results about clustering, including NP-hardness results and approximation algorithms

Goals of this course

- different models for clustering
- many important clustering heuristics, including agglomerative clustering, Lloyd's algorithm, and the EM algorithm
- the limitations of these heuristics
- improvements to these heuristics
- various theoretical results about clustering, including NP-hardness results and approximation algorithms
- general techniques to improve the efficiency of heuristics and approximation algorithms, i.e. dimension reduction techniques.

Organization

Information about this course
http://www.cs.uni-paderborn.de/fachgebiete/ag-bloemer/lehre/2015/ws/clusteringalgorithms.html

Here you find

- announcements
- handouts
- slides
- literature

Organization

Information about this course
http://www.cs.uni-paderborn.de/fachgebiete/ag-bloemer/lehre/2015/ws/clusteringalgorithms.html

Here you find

- announcements
- handouts
- slides
- literature
- lecture notes (will be written and appear as course progresses)

Organization

Information about this course
http://www.cs.uni-paderborn.de/fachgebiete/ag-bloemer/lehre/2015/ws/clusteringalgorithms.html

Here you find

- announcements
- handouts
- slides
- literature
- lecture notes (will be written and appear as course progresses)

- There is only one tutorial, Thursday 13:00 -14:00.
- It starts next week.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Prerequisites

- design and analysis of algorithms
- basic complexity theory
- probability theory and stochastic
- some linear algebra

Objects

- objects described by d different features

Objects

- objects described by d different features
- features continuous or binary

Objects

- objects described by d different features
- features continuous or binary
- objects described as elements in $\mathbb{R}^d$ or $\{0,1\}^d$

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Objects

- objects described by d different features
- features continuous or binary
- objects described as elements in $\mathbb{R}^d$ or $\{0,1\}^d$
- objects from $M \subseteq \mathbb{R}^d$ or $M \subseteq \{0,1\}^d$

Definition 1.1

$D : M \times M \to \mathbb{R}$ *is called a distance function, if for all* $x, y, z \in M$

- $D(x, y) = D(y, x)$ *(symmetry)*
- $D(x, y) \geq 0$ *(positivity),*

Definition 1.1

$D : M \times M \to \mathbb{R}$ is called a distance function, if for all $x, y, z \in M$

- $D(x, y) = D(y, x)$ *(symmetry)*
- $D(x, y) \geq 0$ *(positivity),*

D is called a metric, if in addition,

- $D(x, y) = 0 \Leftrightarrow x = y$ *(reflexivity)*
- $D(x, z) \leq D(x, y) + D(y, z)$ *(triangle inequality)*

Example 1.2 (Euclidean distance)

$M = \mathbb{R}^d$,

$$D_{l_2}(x, y) = \|x - y\|_2 = \Big(\sum_{i=1}^{d} |x_i - y_i|^2\Big)^{\frac{1}{2}},$$

where $x = (x_1, \ldots, x_d)$ *and* $y = (y_1, \ldots, y_d)$.

Example 1.3 (Squared Euclidean distance)

$M = \mathbb{R}^d$,

$$D_{l_2^2}(x, y) = \|x - y\|_2 = \sum_{i=1}^{d} |x_i - y_i|^2,$$

where $x = (x_1, \ldots, x_d)$ *and* $y = (y_1, \ldots, y_d)$.

Example 1.4 (Minkowski distances, l_p-norms)

$M = \mathbb{R}^d$, $p \geq 1$,

$$D_{l_p}(x, y) = \|x - y\|_p = \Big(\sum_{i=1}^{d} |x_i - y_i|^p\Big)^{\frac{1}{p}}.$$

Example 1.4 (Minkowski distances, l_p-norms)

$M = \mathbb{R}^d$, $p \geq 1$,

$$D_{l_p}(x,y) = \|x-y\|_p = \Big(\sum_{i=1}^{d} |x_i - y_i|^p\Big)^{\frac{1}{p}}.$$

Example 1.5 (maximum distance)

$M = \mathbb{R}^d$,

$$D_{l_\infty}(x,y) = \|x-y\|_\infty = \max_{1\leq i\leq d} |x_i - y_i|.$$

Example 1.6 (Pearson correlation)

$M = \mathbb{R}^d$,

$$D_{Pearson}(x,y) = \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{d}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{d}(x_i - \bar{x})^2 \sum_{i=1}^{d}(y_i - \bar{y})^2}}\right),$$

where $\bar{x} = \frac{1}{d}\sum x_i$ *and* $\bar{y} = \frac{1}{d}\sum y_i$.

Example 1.7 (Mahalanobis divergence)

$A \in \mathbb{R}^{d \times d}$ *positive definite, i.e.* $x^T A x > 0$ *for* $x \neq 0$, $M = \mathbb{R}^d$,

$$D_A(x, y) = (x - y)^T A (x - y)$$

Example 1.7 (Mahalanobis divergence)

$A \in \mathbb{R}^{d \times d}$ *positive definite, i.e.* $x^T A x > 0$ *for* $x \neq 0, M = \mathbb{R}^d$,

$$D_A(x, y) = (x - y)^T A (x - y)$$

Example 1.8 (Itakura-Saito divergence)

$M = \mathbb{R}^d_{>0}$,

$$D_{IS}(x, y) = \sum \frac{x_i}{y_i} - \ln\left(\frac{x_i}{y_i}\right) - 1.$$

Example 1.9 (Kullback-Leibler divergence)

$M = S^d := \{x \in \mathbb{R}^d : \forall i : x_i \geq 0, \sum x_i = 1\}$,

$$D_{KLD}(x, y) = \sum x_i \ln(x_i / y_i),$$

where by definition $0 \cdot ln(0) = 0$.

Example 1.9 (Kullback-Leibler divergence)

$M = S^d := \{x \in \mathbb{R}^d : \forall i : x_i \geq 0, \sum x_i = 1\}$,

$$D_{KLD}(x, y) = \sum x_i \ln(x_i / y_i),$$

where by definition $0 \cdot ln(0) = 0$.

Example 1.10 (generalized KLD)

$M = \mathbb{R}^d_{\geq 0}$,

$$D_{KLD}(x, y) = \sum x_i \ln(x_i / y_i) - (x_i - y_i),$$

Definition 1.11

$S : M \times M \to \mathbb{R}$ is called a similarity function, if for all $x, y, z \in M$

- $S(x, y) = S(y, x)$ *(symmetry)*
- $0 \leq S(x, y) \leq 1$ *(positivity),*

S is called a metric, if in addition,

- $S(x, y) = 1 \Leftrightarrow x = y$ *(reflexivity)*
- $S(x, y)S(y, z) \leq \big(S(x, y) + S(y, z)\big)S(x, z)$ *(triangle inequality)*

Example 1.12 (Cosine similarity)

$M = \mathbb{R}^d$,

$$S_{CS}(x,y) = \frac{x^T y}{\|x\|\|y\|} \quad or$$

$$\bar{S}_{CS}(x,y) = \frac{1 + S_{CS}(x,y)}{2}$$

Similarity for binary features

Let $x, y \in \{0,1\}^d$, then

$$n_{b\bar{b}}(x,y) := \left|\{1 \le i \le d : x_i = b, y_i = \bar{b}\}\right|$$

and for $w \in \mathbb{R}_{\ge 0}$

$$S_w(x,y) := \frac{n_{00}(x,y) + n_{11}(x,y)}{n_{00}(x,y) + n_{11}(x,y) + w\big(n_{01}(x,y) + n_{10}(x,y)\big)}.$$

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Let $x, y \in \{0,1\}^d$, then

$$n_{b\bar{b}}(x,y) := \left|\{1 \le i \le d : x_i = b, y_i = \bar{b}\}\right|$$

and for $w \in \mathbb{R}_{\ge 0}$

$$S_w(x,y) := \frac{n_{00}(x,y) + n_{11}(x,y)}{n_{00}(x,y) + n_{11}(x,y) + w\big(n_{01}(x,y) + n_{10}(x,y)\big)}.$$

Popular: $w = 1, 2, \frac{1}{2}$.

Similarity for binary features

Let $x, y \in \{0,1\}^d$, then

$$n_{b\bar{b}}(x,y) := \left|\{1 \le i \le d : x_i = b, y_i = \bar{b}\}\right|$$

and for $w \in \mathbb{R}_{\ge 0}$

$$S_w(x,y) := \frac{n_{00}(x,y) + n_{11}(x,y)}{n_{00}(x,y) + n_{11}(x,y) + w\big(n_{01}(x,y) + n_{10}(x,y)\big)}.$$

Popular: $w = 1, 2, \frac{1}{2}$.

Example 1.13 (matching coefficient)

$$w = 1, S_{mc}(x,y) = \frac{n_{00}(x,y) + n_{11}(x,y)}{d}.$$

Similarity for binary features

$$\bar{S}_w(x, y) := \frac{n_{11}(x, y)}{n_{11}(x, y) + w\big(n_{01}(x, y) + n_{10}(x, y)\big)}$$

Similarity for binary features

$$\bar{S}_w(x,y) := \frac{n_{11}(x,y)}{n_{11}(x,y) + w\big(n_{01}(x,y) + n_{10}(x,y)\big)}$$

Popular: $w = 1, 2, \frac{1}{2}$.

Similarity for binary features

$$\bar{S}_w(x,y) := \frac{n_{11}(x,y)}{n_{11}(x,y) + w\big(n_{01}(x,y) + n_{10}(x,y)\big)}$$

Popular: $w = 1, 2, \frac{1}{2}$.

Example 1.14 (Jaccard coefficient)

$$w = 1, S_{Jacard}(x,y) = \frac{n_{11}(x,y)}{n_{11}(x,y) + n_{01}(x,y) + n_{10}(x,y)}.$$

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft