Clustering

based on Loïc Cerf's slides (UFMG)

Marc Plantevit

March 11, 2021

UCBL – LIRIS – DM2L

Clustering Approaches

- Partition-based algorithms: build several partitions then assess them w.r.t. some criteria.
- Hierarchy-based algorithms: create a hierarchical decomposition of the objects w.r.t. some criteria.
- Density-based algorithms: based on the notions of density and connectivity.

Characteristics

- extensibility
- ability to handle different data types
- prior for parameter settings
- ability to handle noisy data and outliers

Outline

Outline

Inductive database vision

Querying a clustering:

$$\{X \in P \mid \mathcal{Q}(X, \mathcal{D})\}$$

where:

- $\mathcal{D}$ is a set of objects $\mathcal{O}$ associated with a similarity measure,
- P is $\{(C_1, \ldots, C_k) \in (2^{\mathcal{O}})^k \mid \begin{cases} \forall i = 1..k, C_i \neq \emptyset \\ \forall j \neq i, C_i \cap C_j \neq \emptyset \\ \cup_{l=1}^{k} C_l = \mathcal{O} \end{cases}\}$,
- $\mathcal{Q}$ is a function to optimize. It quantifies how similar are pairs of objects in a same cluster and how dissimilar are those in two different clusters

Inductive database vision

Querying a clustering with k-means:

$$\{X \in P \mid \mathcal{Q}(X, \mathcal{D})\}$$

where:

- $\mathcal{D}$ is a set of objects $\mathcal{O}$ associated with a similarity measure,
- P is $\{(C_1, \ldots, C_k) \in (2^{\mathcal{O}})^k \mid \begin{cases} \forall i = 1..k, C_i \neq \emptyset \\ \forall j \neq i, C_i \cap C_j \neq \emptyset \\ \cup_{l=1}^{k} C_l = \mathcal{O} \end{cases} \}$ where $k \in \mathbb{N} \setminus \{0\}$ is fixed,
- $\mathcal{Q}$ is the maximization of the sum, over all objects, of the similarities to the centers of the assigned clusters:

$$(C_1, \ldots, C_k) \mapsto \sum_{i=1}^{k} \sum_{o \in C_i} s(o, \frac{\sum_{o \in C_i}}{|C_i|})$$

Inductive database vision

Querying a clustering with k-means:

$$\{X \in P \mid \mathcal{Q}(X, \mathcal{D})\}$$

where:

- $\mathcal{D}$ is a set of objects $\mathcal{O}$ associated with a similarity measure,
- P is $\{(C_1, \ldots, C_k) \in (2^{\mathcal{O}})^k \mid \begin{cases} \forall i = 1..k, C_i \neq \emptyset \\ \forall j \neq i, C_i \cap C_j \neq \emptyset \\ \cup_{l=1}^{k} C_l = \mathcal{O} \end{cases} \}$ where $k \in \mathbb{N} \setminus \{0\}$ is fixed,
- $\mathcal{Q}$ is the maximization of the sum, over all objects, of the similarities to the centers of the assigned clusters:

$$(C_1, \ldots, C_k) \mapsto \sum_{i=1}^{k} \sum_{o \in C_i} s(o, \mu_i)$$

Exact algorithm

Input: $\mathcal{O}, \mathcal{D}, k \in \mathbb{N} \setminus \{0\}$
Output: the clustering of $\mathcal{O}$ maximizing f: the sum, over all objects, of the similarities to the centers of the assigned clusters

$\mathcal{C}_{\text{max}} \leftarrow \emptyset$
$f_{\text{max}} \leftarrow -\infty$
for all k-clustering $\mathcal{C}$ of $\mathcal{O}$ **do**
  if $f(\mathcal{C}, \mathcal{D}) > f_{\text{max}}$ **then**
    $f_{\text{max}} \leftarrow f(\mathcal{C}, \mathcal{D})$
    $\mathcal{C}_{\text{max}} \leftarrow \mathcal{C}$
  end if
end for
output$(\mathcal{C}_{\text{max}})$

Number of k-clusterings

Question

How many k-clusterings are enumerated?

Number of k-clusterings

Question

How many k-clusterings are enumerated? The Stirling number of the second kind, i. e., $\frac{1}{k!}\sum_{t=0}^{k}(-1)^t \binom{k}{t}(k-t)^n = O(k^n)$.

k-means principles

k-means is a greedy iterative approach that always converges to a local maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters.

k-means principles

k-means is a greedy iterative approach that always converges to a local maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters.

An iteration consists in two steps:

E Each object is assigned to the cluster whose center is the most similar (thus defining a clustering);

k-means principles

k-means is a greedy iterative approach that always converges to a local maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters.

An iteration consists in two steps:

- E Each object is assigned to the cluster whose center is the most similar (thus defining a clustering);
- M The center of each cluster is updated to the mean of the objects assigned to it.

k-means principles

k-means is a greedy iterative approach that always converges to a local maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters.

An iteration consists in two steps:

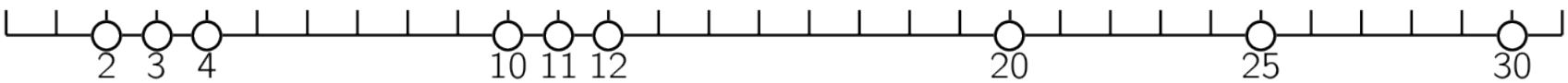
- E Each object is assigned to the cluster whose center is the most similar (thus defining a clustering);
- M The center of each cluster is updated to the mean of the objects assigned to it.

Initially, the centers of the clusters are randomly drawn. The procedure stops when, from an iteration to the next one, the centers of the clusters have not changed much (or at all).

2-means with $|\mathcal{A}| = 1$: illustration

2-means clustering of the objects in a one-dimensional space using the Euclidean distance.

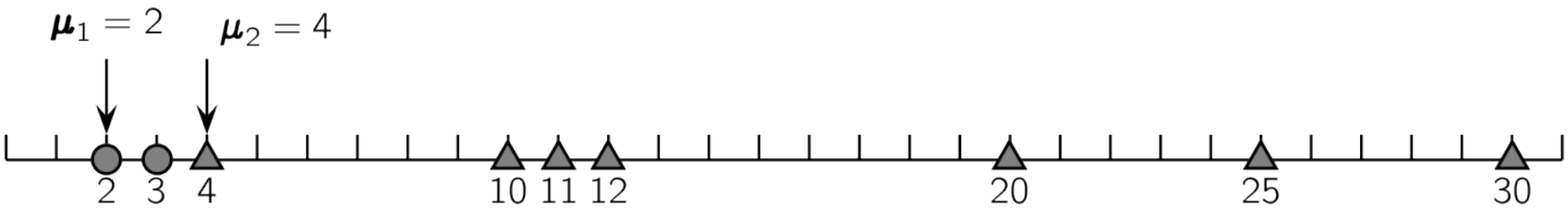
Dataset:



2-means with $|\mathcal{A}| = 1$: illustration

2-means clustering of the objects in a one-dimensional space using the Euclidean distance.

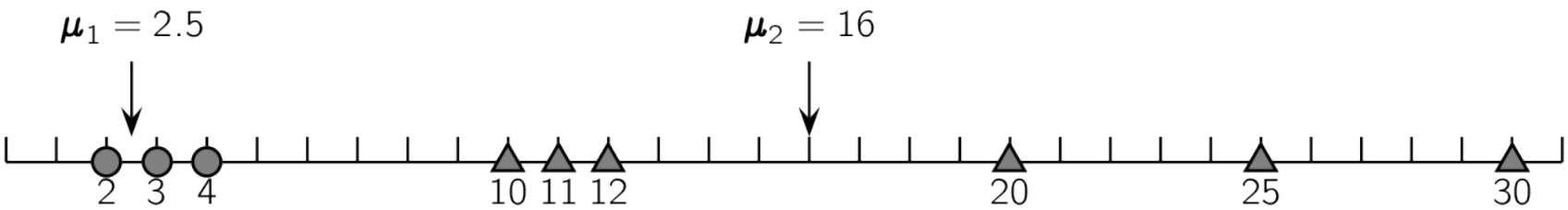
Iteration 1:



2-means with $|\mathcal{A}| = 1$: illustration

2-means clustering of the objects in a one-dimensional space using the Euclidean distance.

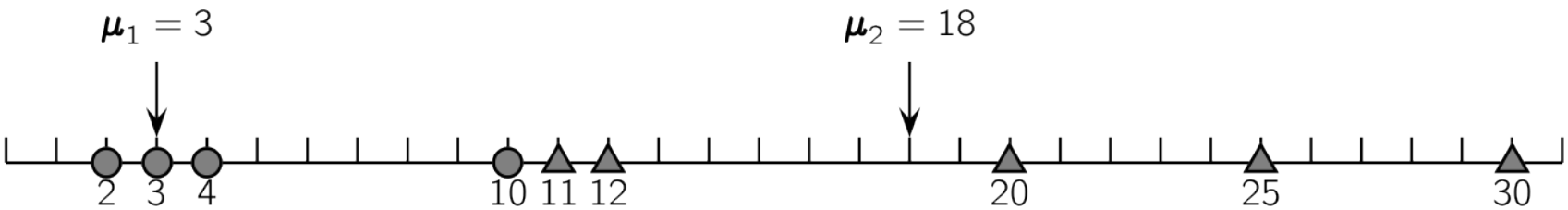
Iteration 2:



2-means with $|\mathcal{A}| = 1$: illustration

2-means clustering of the objects in a one-dimensional space using the Euclidean distance.

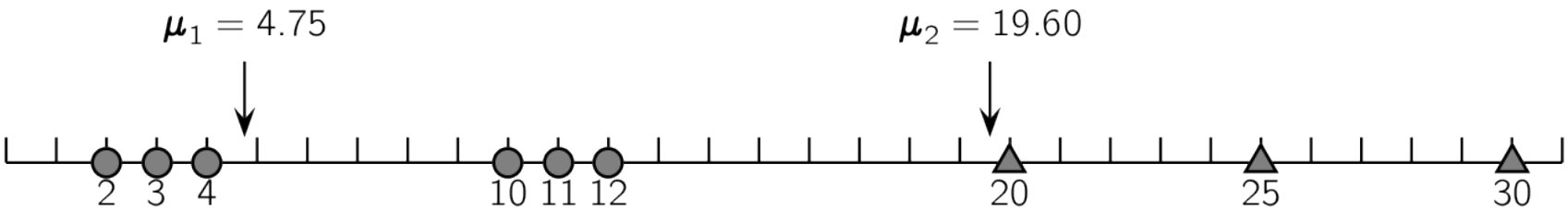
Iteration 3:



2-means with $|\mathcal{A}| = 1$: illustration

2-means clustering of the objects in a one-dimensional space using the Euclidean distance.

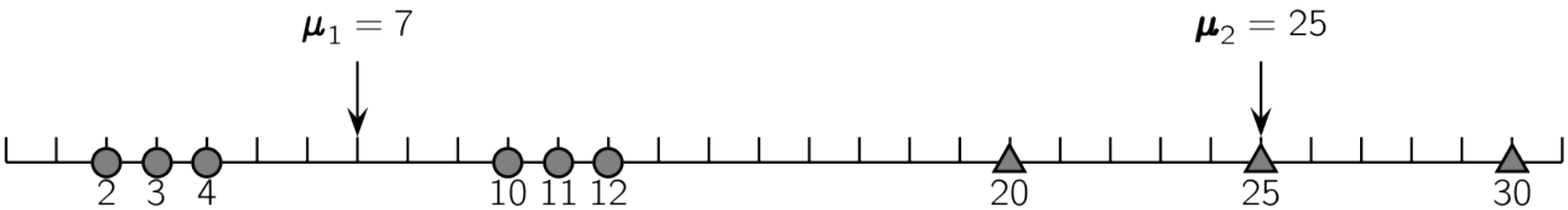
Iteration 4:



2-means with $|\mathcal{A}| = 1$: illustration

2-means clustering of the objects in a one-dimensional space using the Euclidean distance.

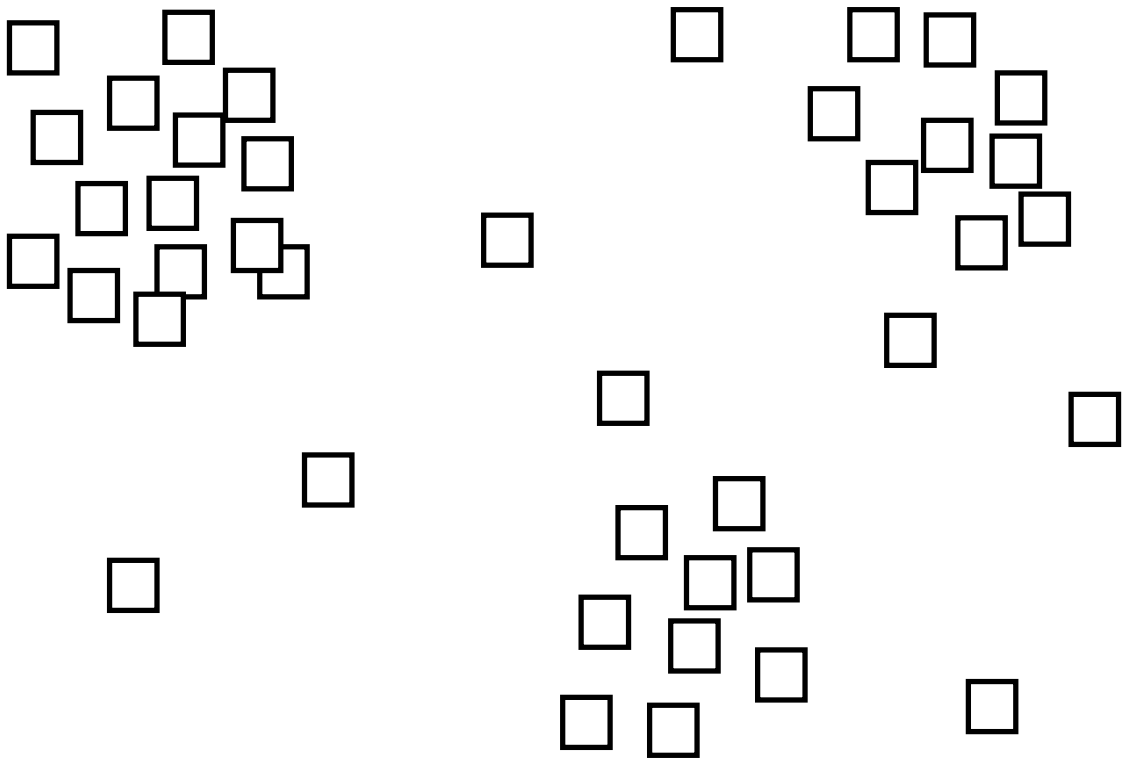
Iteration 5:



3-means with $|\mathcal{A}| = 2$: illustration

3-means clustering of the objects in a two-dimensional space using the Euclidean distance.

	x	y
o_1	91	70
o_2	129	91
o_3	359	243
o_4	322	254
o_5	100	104
o_6	464	113
o_7	342	297
o_8	410	65
o_9	334	329
$\vdots$	$\vdots$	$\vdots$



k-means algorithm

Input: $\mathcal{O}, \mathcal{D}, k \in \mathbb{N} \setminus \{0\}$
Output: a clustering of $\mathcal{O}$ *locally* maximizing the sum, over all objects, of the similarities to the centers of the assigned clusters
$(\mu_i)_{i=1..k} \leftarrow \text{random}(\mathcal{D})$
repeat
  $(C_i)_{i=1..k} \leftarrow \text{assign\_cluster}(\mathcal{O}, \mathcal{D}, (\mu_i)_{i=1..k})$
  $(c, (\mu_i)_{i=1..k}) \leftarrow \text{update\_centers}(\mathcal{D}, (C_i)_{i=1..k}, (\mu_i)_{i=1..k}))$
until c
output$((C_i)_{i=1..k})$

assign_cluster

Input: $\mathcal{O}, \mathcal{D}, (\mu_i)_{i=1..k} \in (\mathbb{R}^{|\mathcal{A}|})^k$
Output: $(C_i)_{i=1..k}$ the clustering of $\mathcal{O}$ such that $\forall i = 1..k, \forall j \neq i, \forall o \in C_i, s(o, \mu_i) \geq s(o, \mu_j)$
for all $o \in \mathcal{O}$ **do**
  $a \leftarrow \arg\max_{i=1..k} s(o, \mu_i)$
  $C_a \leftarrow C_a \cup \{o\}$
end for
return$((C_i)_{i=1..k})$

Complexity of assign_cluster

Question

Assuming the computation of a similarity is linear in the number of attributes $|\mathcal{A}|$, what is the complexity of **assign_cluster**?

Complexity of assign_cluster

Question

Assuming the computation of a similarity is linear in the number of attributes $|\mathcal{A}|$, what is the complexity of **assign_cluster**?
$O(k|\mathcal{O} \times \mathcal{A}|)$.

update_centers

Input: $\mathcal{D}, (C_i)_{i=1..k}$ a clustering of $\mathcal{O}, (\mu_i)_{i=1..k} \in (\mathbb{R}^{|\mathcal{A}|})^k$
Output: $c \in \{\text{false}, \text{true}\}$ indicating whether the convergence is reached, $(\mu'_i)_{i=1..k}) \in (\mathbb{R}^{|\mathcal{A}|})^k$ such that $\forall i = 1..k, \mu'_i = \frac{\sum_{o \in C_i} o}{|C_i|}$

$c \leftarrow$ true
for $i = 1 \rightarrow k$ **do**
  $\mu'_i \leftarrow \frac{\sum_{o \in C_i} o}{|C_i|}$
  if $\mu'_i \neq \mu_i$ **then**
    $c \leftarrow$ false
  end if
end for
return$(c, (\mu'_i)_{i=1..k})$

Complexity of k-means

Question

Assuming the computation of a similarity is linear in the number of attributes $|\mathcal{A}|$, what is the complexity of **assign_cluster**? $O(k|\mathcal{O} \times \mathcal{A}|)$.

Question

What is the complexity of **update_centers**?

Complexity of k-means

Question

Assuming the computation of a similarity is linear in the number of attributes $|\mathcal{A}|$, what is the complexity of **assign_cluster**? $O(k|\mathcal{O} \times \mathcal{A}|)$.

Question

What is the complexity of **update_centers**? $O(|\mathcal{O} \times \mathcal{A}|)$.

Complexity of k-means

Question

Assuming the computation of a similarity is linear in the number of attributes $|\mathcal{A}|$, what is the complexity of **assign_cluster**? $O(k|\mathcal{O} \times \mathcal{A}|)$.

Question

What is the complexity of **update_centers**? $O(|\mathcal{O} \times \mathcal{A}|)$.

Question

What is the complexity of k-means if $t \in \mathbb{N}$ iterations are necessary to converge?

Complexity of k-means

Question

Assuming the computation of a similarity is linear in the number of attributes $|\mathcal{A}|$, what is the complexity of **assign_cluster**? $O(k|\mathcal{O} \times \mathcal{A}|)$.

Question

What is the complexity of **update_centers**? $O(|\mathcal{O} \times \mathcal{A}|)$.

Question

What is the complexity of k-means if $t \in \mathbb{N}$ iterations are necessary to converge? $O(tk|\mathcal{O} \times \mathcal{A}|)$.

Convergence

Worst-case scenarios require $2^{\Omega(|\mathcal{O}|)}$ iterations to converge but a smoothed analysis gives a polynomial complexity.

Convergence

Worst-case scenarios require $2^{\Omega(|\mathcal{O}|)}$ iterations to converge but a smoothed analysis gives a polynomial complexity.

The low complexity of k-means is its greatest advantage.

Limitations of k-means

- Convergence towards a *local* maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters;

Limitations of k-means

- Convergence towards a *local* maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters;
- Sensitivity to outliers (k-medoids replaces the means by medians);

Limitations of k-means

- Convergence towards a *local* maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters;
- Sensitivity to outliers (k-medoids replaces the means by medians);
- Tendency to produce equi-sized clusters;

Limitations of k-means

- Convergence towards a *local* maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters;
- Sensitivity to outliers (k-medoids replaces the means by medians);
- Tendency to produce equi-sized clusters;
- The number of clusters must be known beforehand.

Limitations of k-means

- Convergence towards a *local* maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters;
- Sensitivity to outliers (k-medoids replaces the means by medians);
- Tendency to produce equi-sized clusters;
- The number of clusters must be known beforehand.

The elbow method

Plot a measure of the quality of the k clusters (e. g., the sum, over all objects, of the similarities to the centers of the assigned clusters) when k increases. Choose k after a large drop of the growth.

The elbow method

Plot a measure of the quality of the k clusters (e. g., the sum, over all objects, of the similarities to the centers of the assigned clusters) when k increases. Choose k after a large drop of the growth.

More principled method exist and can be seen as variants (finding the best trade-off between quality and compression).

The elbow method

Plot a measure of the quality of the k clusters (e. g., the sum, over all objects, of the similarities to the centers of the assigned clusters) when k increases. Choose k after a large drop of the growth.

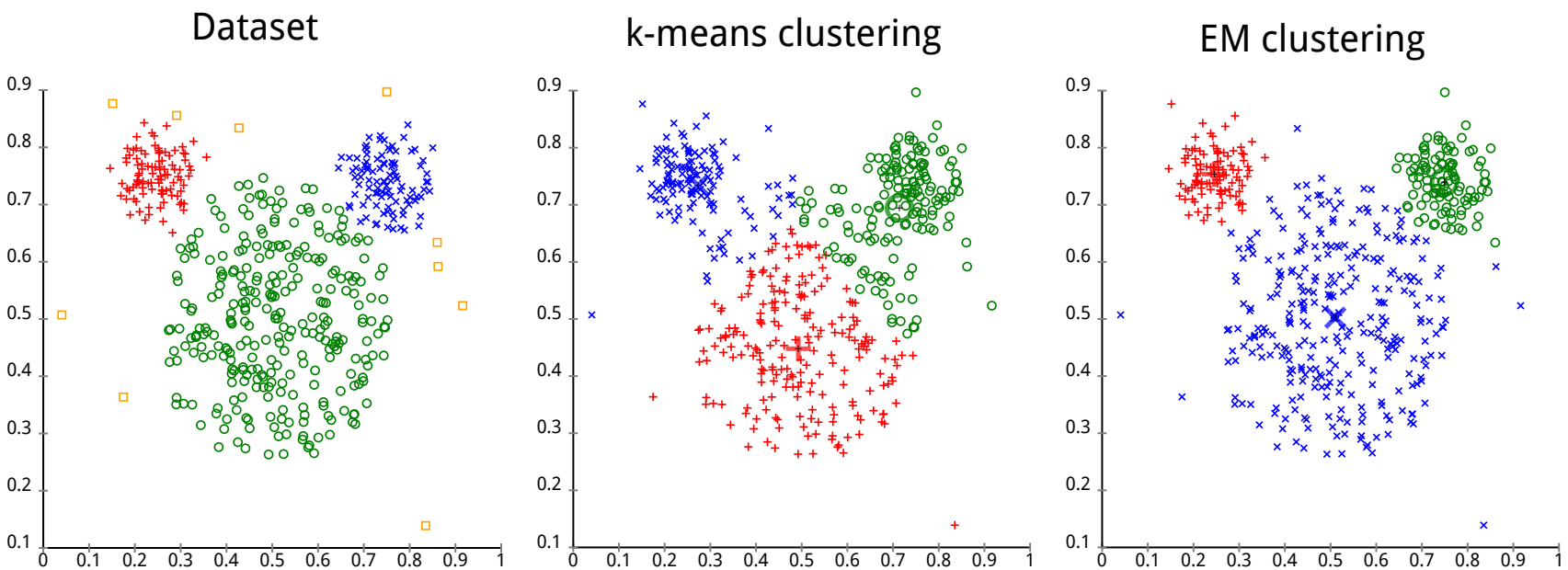
More principled method exist and can be seen as variants (finding the best trade-off between quality and compression).

If the quadratic time complexity of a hierarchical agglomeration is not prohibitive, the number of clusters can be determined from the dendrogram.

Limitations of k-means

- Convergence towards a *local* maximum of the sum, over all objects, of the similarities to the centers of the assigned clusters;
- Sensitivity to outliers (k-medoids replaces the means by medians);
- Tendency to produce equi-sized clusters;
- The number of clusters must be known beforehand.

Tendency to produce equi-sized clusters



Outline

EM assumptions

The dataset $\mathcal{D}$ is seen as a random sample from a $|\mathcal{A}|$-dimensional random variable O.

EM assumptions

The dataset $\mathcal{D}$ is seen as a random sample from a $|\mathcal{A}|$-dimensional random variable O.

This probability density function is given as a mixture model of the $k \in \mathbb{N} \setminus \{0\}$ clusters $(C_i)_{i=1..k}$:

$$f(o) = \sum_{i=1}^{k} f_i(o) P(C_i)$$

, where $P(C_i)$ is the probability to belong to the cluster C_i and f_i is the probability density function of this cluster whose type of distribution is chosen beforehand.

Maximum likelihood estimation

EM searches a parametrization θ of f (i. e., $(P(C_i)_{i=1..k}$ and the parametrization of the $(f_i)_{i=1..k}$) so that the likelihood that $\mathcal{D}$ is indeed a random sample of O is maximized:

$$\arg\max_{\theta} P(\mathcal{D}|\theta) \quad .$$

Maximum likelihood estimation

EM searches a parametrization θ of f (i. e., $(P(C_i)_{i=1..k}$ and the parametrization of the $(f_i)_{i=1..k}$) so that the likelihood that $\mathcal{D}$ is indeed a random sample of O is maximized:

$$\arg\max_{\theta} P(\mathcal{D}|\theta) \ .$$

Since the dataset is assumed to be a random sample from O (i. e., independent and identically distributed as O), the objective becomes the computation of:

$$\arg\max_{\theta} \prod_{o \in \mathcal{O}} f(o) \ .$$

Maximum likelihood estimation

EM searches a parametrization θ of f (i. e., $(P(C_i)_{i=1..k}$ and the parametrization of the $(f_i)_{i=1..k}$) so that the likelihood that $\mathcal{D}$ is indeed a random sample of O is maximized:

$$\arg\max_{\theta} P(\mathcal{D}|\theta) \ .$$

Since the dataset is assumed to be a random sample from O (i. e., independent and identically distributed as O), the objective becomes the computation of:

$$\arg\max_{\theta} \prod_{o \in \mathcal{O}} f(o) \ .$$

It usually is hard to analytically compute $\arg\max_\theta \prod_{o\in\mathcal{O}} f(o)$.

EM principles

EM is a greedy iterative approach that always converges to a local maximum of $P(\mathcal{D}|\theta)$.

EM principles

EM is a greedy iterative approach that always converges to a local maximum of $P(\mathcal{D}|\theta)$.

An iteration consists in two steps:

- E Given θ, the posterior probabilities of each object to belong to each cluster is computed;

EM principles

EM is a greedy iterative approach that always converges to a local maximum of $P(\mathcal{D}|\theta)$.

An iteration consists in two steps:

- E Given θ, the posterior probabilities of each object to belong to each cluster is computed;
- M θ is updated to reflect these probabilities.

EM principles

EM is a greedy iterative approach that always converges to a local maximum of $P(\mathcal{D}|\theta)$.

An iteration consists in two steps:

- E Given θ, the posterior probabilities of each object to belong to each cluster is computed;
- M θ is updated to reflect these probabilities.

Initially, the parametrization of θ is randomly drawn and $\forall i = 1..k, P(C_i) = \frac{1}{k}$. The procedure stops when, from an iteration to the next one, the parametrization has not changed much (or at all).

Expectation step

Given θ, the posterior probability of an object $o \in \mathcal{O}$ to belong to a cluster C_i is:

$$\begin{aligned} P(C_i|o) &= \frac{P(C_i \wedge o)}{P(o)} \\ &= \frac{P(o|C_i)P(C_i)}{\sum_{a=1..k} P(o \wedge C_a)} \\ &= \frac{P(o|C_i)P(C_i)}{\sum_{a=1..k} P(o|C_a)P(C_a)} \\ &= \frac{f_i(o)P(C_i)}{\sum_{a=1..k} f_a(o)P(C_a)} \ . \end{aligned}$$

Maximization step (1/2)

The distribution of a cluster usually is assumed multivariate normal, thus parametrized with a *location* (the center of the cluster) and a *covariance matrix*.

Maximization step (1/2)

The distribution of a cluster usually is assumed multivariate normal, thus parametrized with a *location* (the center of the cluster) and a *covariance matrix*.

Given $(P(C_i|o))_{i=1..k,o\in\mathcal{O}}$, the location of the cluster C_i is updated to the weighted sample mean μ_i:

$$\frac{\sum_{o\in\mathcal{O}} P(C_i|o)o}{\sum_{o\in\mathcal{O}} P(C_i|o)} \ .$$

Maximization step (2/2)

Given $(P(C_i|o))_{i=1..k,o\in\mathcal{O}}$, the covariance of the cluster C_i between the random variables O_a and O_b is updated to the weighted sample covariance:

$$\frac{\sum_{o\in\mathcal{O}} P(C_i|o)(o_a - \mu_{i,a})(o_b - \mu_{i,b})}{\sum_{o\in\mathcal{O}} P(C_i|o)} .$$

Maximization step (2/2)

Given $(P(C_i|o))_{i=1..k,o\in\mathcal{O}}$, the covariance of the cluster C_i between the random variables O_a and O_b is updated to the weighted sample covariance:

$$\frac{\sum_{o\in\mathcal{O}} P(C_i|o)(o_a - \mu_{i,a})(o_b - \mu_{i,b})}{\sum_{o\in\mathcal{O}} P(C_i|o)} .$$

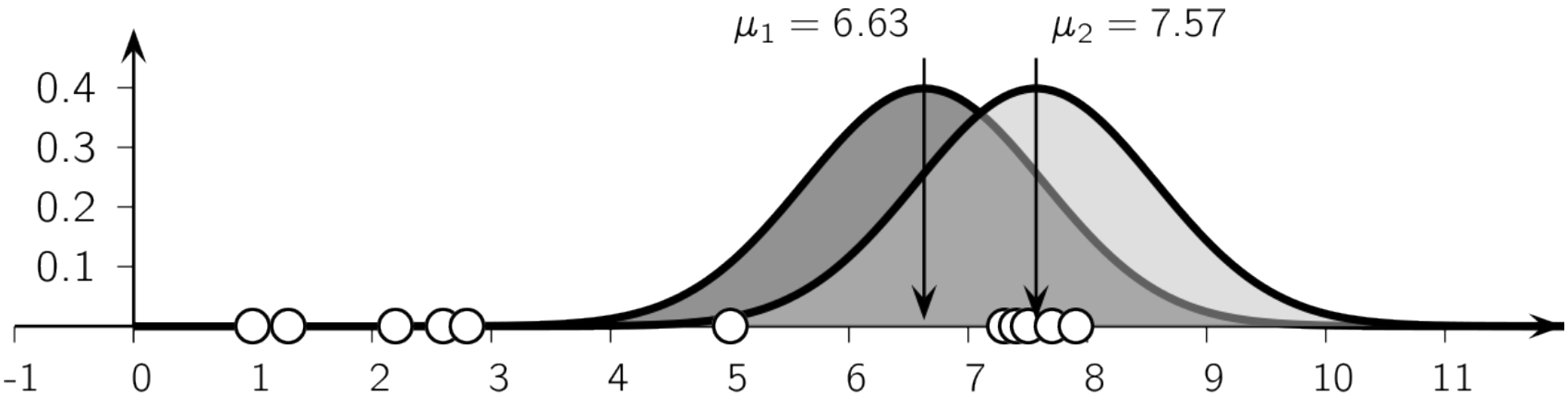
Given $(P(C_i|o))_{i=1..k,o\in\mathcal{O}}$, the prior probability of belonging to the cluster C_i is updated to:

$$\frac{\sum_{o\in\mathcal{O}} P(C_i|o)}{|\mathcal{O}|} .$$

EM with $|\mathcal{A}| = 1$ and $k = 2$: illustration

EM clustering of the objects in a one-dimensional space using the Euclidean distance.

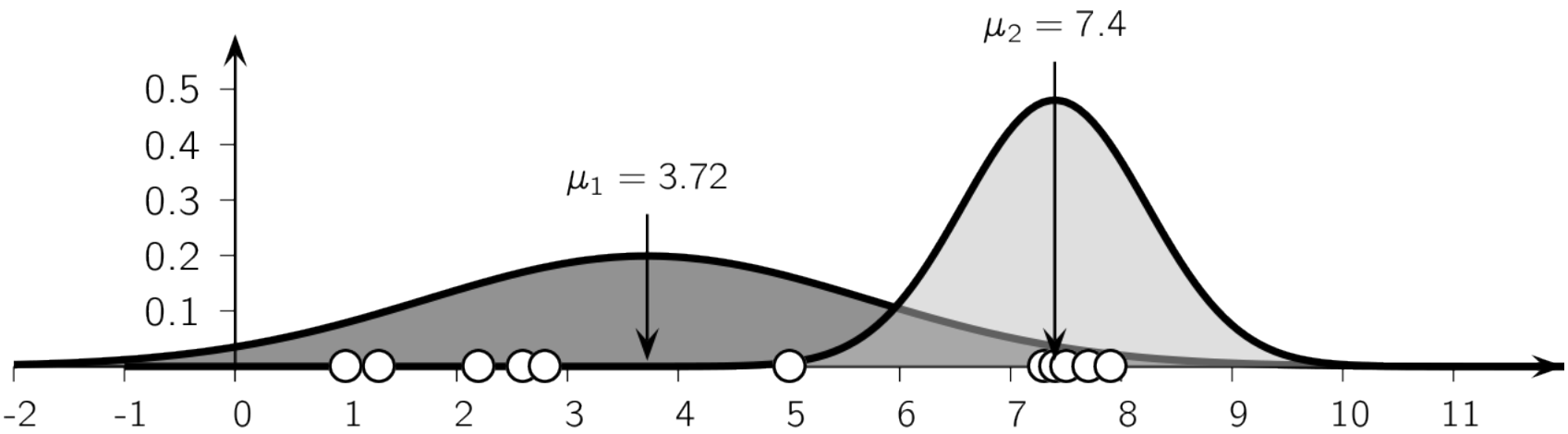
Dataset:



EM with $|\mathcal{A}| = 1$ and $k = 2$: illustration

EM clustering of the objects in a one-dimensional space using the Euclidean distance.

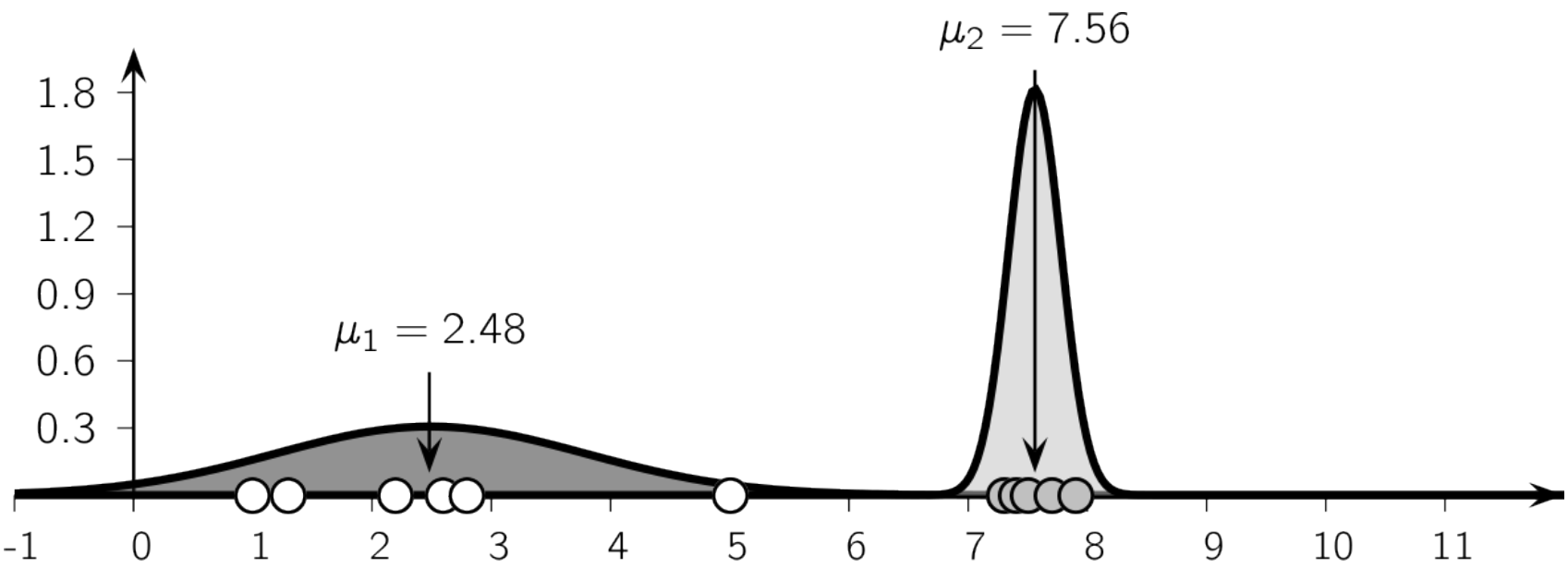
Iteration 1:



EM with $|\mathcal{A}| = 1$ and $k = 2$: illustration

EM clustering of the objects in a one-dimensional space using the Euclidean distance.

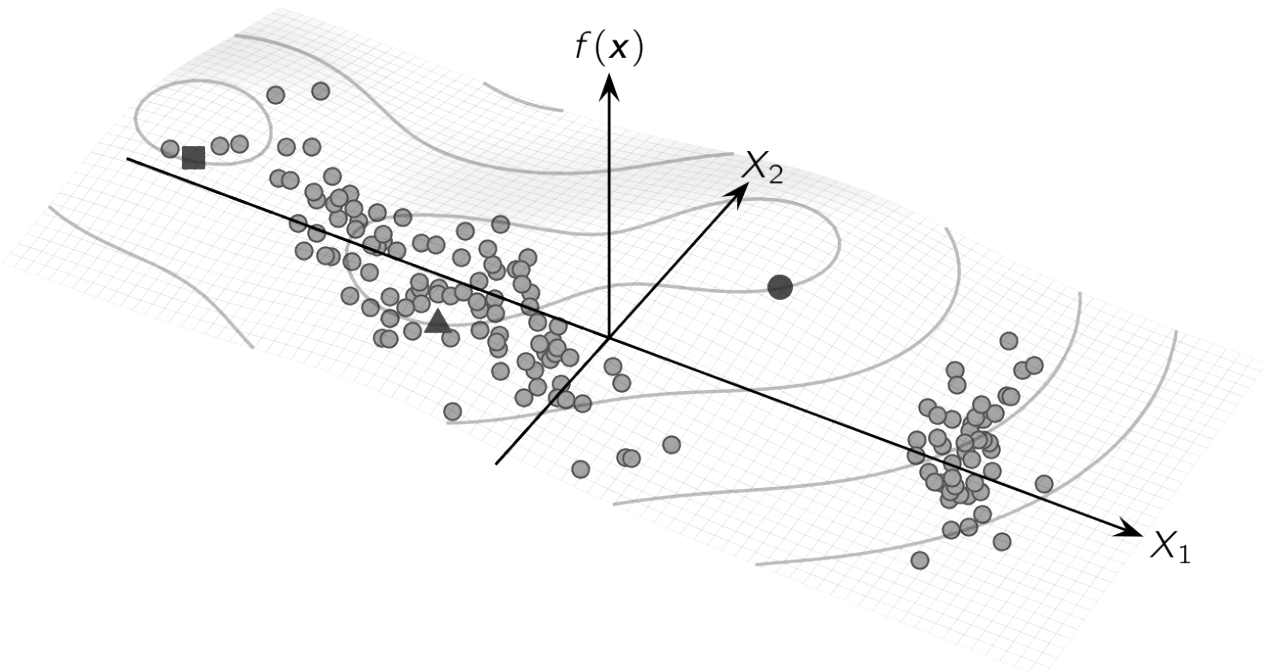
Iteration 5:



EM with $|\mathcal{A}| = 2$ and $k = 3$

EM clustering of the objects in a two-dimensional space using the Euclidean distance.

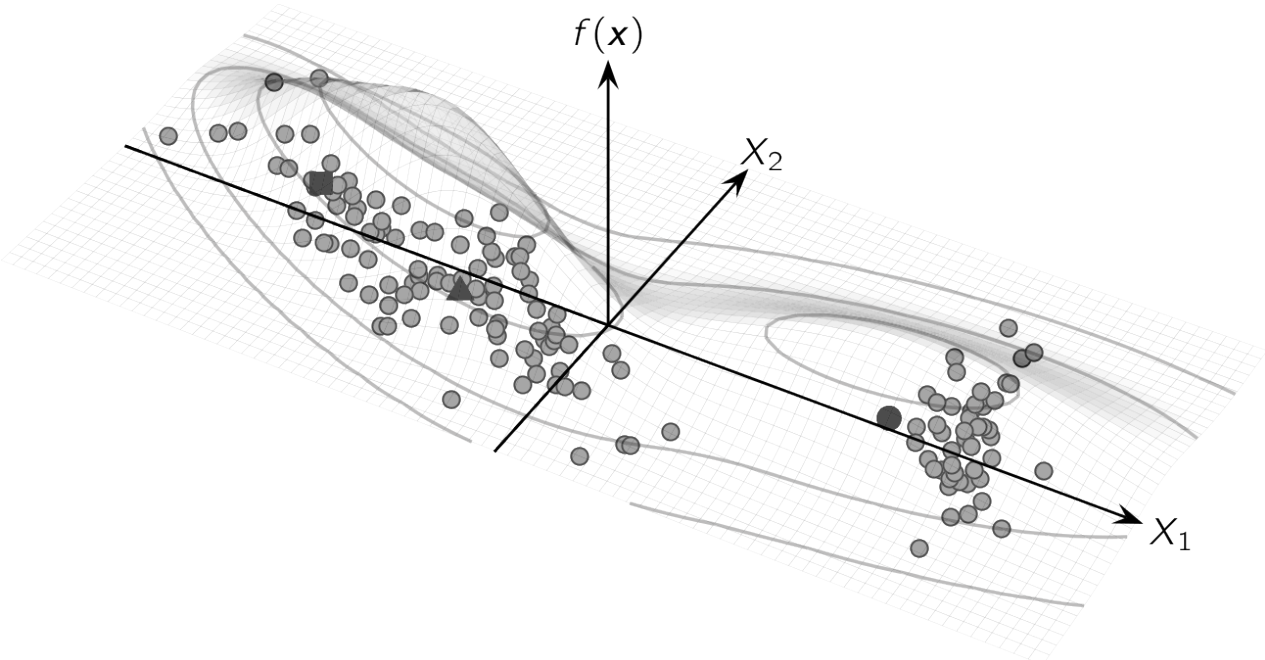
Dataset:



EM with $|\mathcal{A}| = 2$ and $k = 3$

EM clustering of the objects in a two-dimensional space using the Euclidean distance.

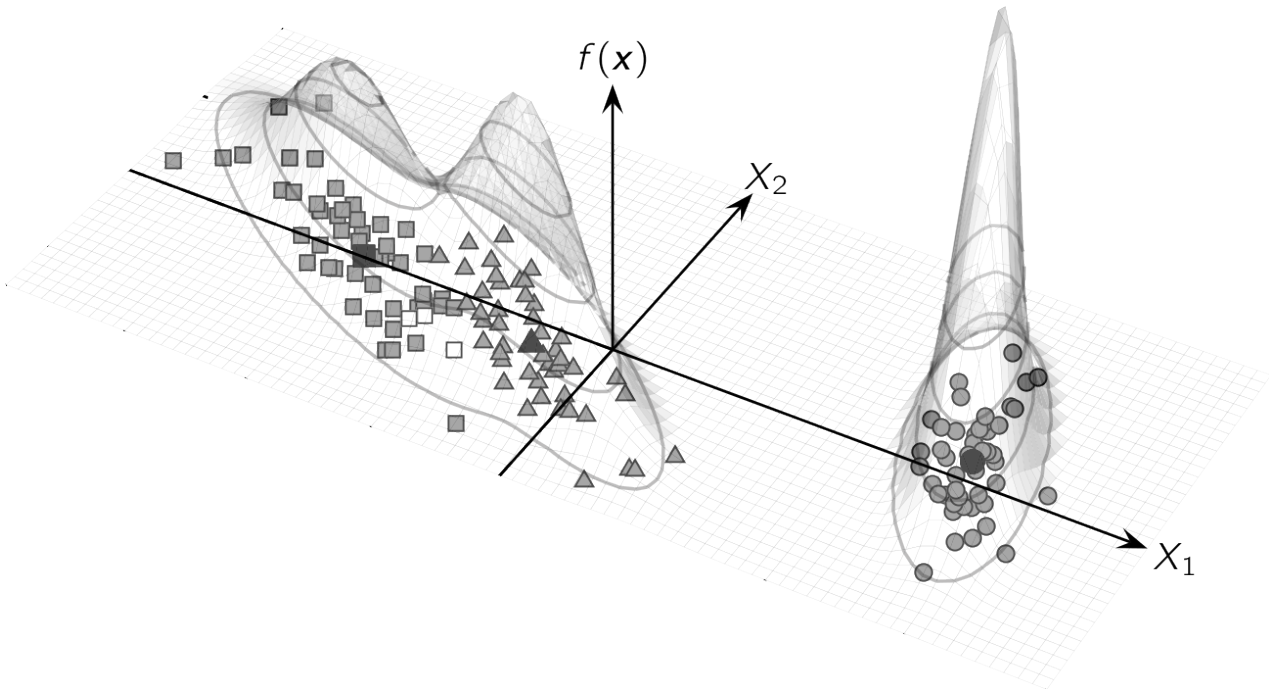
Iteration 1:



EM with $|\mathcal{A}| = 2$ and $k = 3$

EM clustering of the objects in a two-dimensional space using the Euclidean distance.

Iteration 36:



EM algorithm with mixture of Gaussians

Input: $\mathcal{O}, \mathcal{D}, k \in \mathbb{N} \setminus \{0\}$
Output: a fuzzy clustering of $\mathcal{O}$ corresponding to posterior probabilities of a *locally* maximized likelihood of a mixture of Gaussians

$(\mu_i)_{i=1..k} \leftarrow \text{random}(\mathcal{D})$
$(\Sigma_i)_{i=1..k} \leftarrow (I, \ldots, I)$
$(P(C_i))_{i=1..k} \leftarrow (\frac{1}{k}, \ldots, \frac{1}{k})$
repeat
  $(P(C_i|o))_{i=1..k, o \in \mathcal{O}} \leftarrow$
  $\text{expectation}(\mathcal{O}, \mathcal{D}, (\mu_i)_{i=1..k}, (\Sigma_i)_{i=1..k}, (P(C_i))_{i=1..k})$
  $(c, (\mu_i)_{i=1..k}, (\Sigma_i)_{i=1..k}, (P(C_i))_{i=1..k}) \leftarrow$
  $\text{maximization}(\mathcal{D}, (P(C_i|o))_{i=1..k, o \in \mathcal{O}}, (\mu_i)_{i=1..k})$
until c
output$((P(C_i|o))_{i=1..k, o \in \mathcal{O}})$

expectation

Input: $\mathcal{O}, \mathcal{D}, (\mu_i)_{i=1..k} \in (\mathbb{R}^{|\mathcal{A}|})^k, (\Sigma_i)_{i=1..k} \in (\mathbb{R}_+^{|\mathcal{A}|\times|\mathcal{A}|})^k, (P(C_i))_{i=1..k} \in [0,1]^k$
Output: $(P(C_i|o))_{i=1..k, o\in\mathcal{O}}$ the fuzzy assignment of the objects in $\mathcal{O}$ to the clusters given by the mixture of Gaussians parametrized with $(\mu_i)_{i=1..k}, (\Sigma_i)_{i=1..k}, (P(C_i))_{i=1..k}$
for all $o \in \mathcal{O}$ **do**
  for $i = 1 \to k$ **do**
    $P(C_i|o) \leftarrow \frac{f_i(o)P(C_i)}{\sum_{a=1}^{k} f_a(o)P(C_a)}$
  end for
end for
return$((P(C_i|o))_{i=1..k, o\in\mathcal{O}})$

expectation

Input: $\mathcal{O}, \mathcal{D}, (\mu_i)_{i=1..k} \in (\mathbb{R}^{|\mathcal{A}|})^k, (\Sigma_i)_{i=1..k} \in (\mathbb{R}_+^{|\mathcal{A}|\times|\mathcal{A}|})^k, (P(C_i))_{i=1..k} \in [0,1]^k$
Output: $(P(C_i|o))_{i=1..k, o\in\mathcal{O}}$ the fuzzy assignment of the objects in $\mathcal{O}$ to the clusters given by the mixture of Gaussians parametrized with $(\mu_i)_{i=1..k}, (\Sigma_i)_{i=1..k}, (P(C_i))_{i=1..k}$
for all $o \in \mathcal{O}$ **do**
  for $i = 1 \to k$ **do**

$$P(C_i|o) \leftarrow \frac{(\det(\Sigma_i) e^{(o-\mu_i)^T \Sigma_i^{-1} (o-\mu_i)})^{\frac{-1}{2}} P(C_i)}{\sum_{a=1}^{k} (\det(\Sigma_a) e^{(o-\mu_a)^T \Sigma_a^{-1} (o-\mu_a)})^{\frac{-1}{2}} P(C_a)}$$

  end for
end for
return$((P(C_i|o))_{i=1..k, o\in\mathcal{O}})$

Complexity of expectation

Question

What is the complexity of computing $(\det(\Sigma_i))_{i=1..k}$?

Complexity of expectation

Question

What is the complexity of computing $(\det(\Sigma_i))_{i=1..k}$? $kO(|\mathcal{A}|^3)$.

Complexity of expectation

Question

What is the complexity of computing $(\det(\Sigma_i))_{i=1..k}$? $kO(|\mathcal{A}|^3)$.

Question

What is the complexity of computing $(\Sigma_i^{-1})_{i=1..k}$ once $(\det(\Sigma_i))_{i=1..k}$ is known?

Complexity of expectation

Question

What is the complexity of computing $(\det(\Sigma_i))_{i=1..k}$? $kO(|\mathcal{A}|^3)$.

Question

What is the complexity of computing $(\Sigma_i^{-1})_{i=1..k}$ once $(\det(\Sigma_i))_{i=1..k}$ is known? $O(k|\mathcal{A}|^2)$.

Complexity of expectation

Question

What is the complexity of computing $(\det(\Sigma_i))_{i=1..k}$? $kO(|\mathcal{A}|^3)$.

Question

What is the complexity of computing $(\Sigma_i^{-1})_{i=1..k}$ once $(\det(\Sigma_i))_{i=1..k}$ is known? $O(k|\mathcal{A}|^2)$.

Question

What is the complexity of computing one Mahalanobis distance, $(o, o') \mapsto (o - o')^T \Sigma_i^{-1} (o - o')$, once Σ_i^{-1} is known?

Complexity of expectation

Question

What is the complexity of computing $(\det(\Sigma_i))_{i=1..k}$? $kO(|\mathcal{A}|^3)$.

Question

What is the complexity of computing $(\Sigma_i^{-1})_{i=1..k}$ once $(\det(\Sigma_i))_{i=1..k}$ is known? $O(k|\mathcal{A}|^2)$.

Question

What is the complexity of computing one Mahalanobis distance, $(o, o') \mapsto (o - o')^T \Sigma_i^{-1} (o - o')$, once Σ_i^{-1} is known? $O(|\mathcal{A}|^2)$.

Complexity of expectation

Question

What is the complexity of computing $(\det(\Sigma_i))_{i=1..k}$? $kO(|\mathcal{A}|^3)$.

Question

What is the complexity of computing $(\Sigma_i^{-1})_{i=1..k}$ once $(\det(\Sigma_i))_{i=1..k}$ is known? $O(k|\mathcal{A}|^2)$.

Question

What is the complexity of computing one Mahalanobis distance, $(o, o') \mapsto (o - o')^T \Sigma_i^{-1} (o - o')$, once Σ_i^{-1} is known? $O(|\mathcal{A}|^2)$.

Question

What is the complexity of **expectation**?

Complexity of expectation

Question

What is the complexity of computing $(\det(\Sigma_i))_{i=1..k}$? $kO(|\mathcal{A}|^3)$.

Question

What is the complexity of computing $(\Sigma_i^{-1})_{i=1..k}$ once $(\det(\Sigma_i))_{i=1..k}$ is known? $O(k|\mathcal{A}|^2)$.

Question

What is the complexity of computing one Mahalanobis distance, $(o, o') \mapsto (o - o')^T \Sigma_i^{-1} (o - o')$, once Σ_i^{-1} is known? $O(|\mathcal{A}|^2)$.

Question

What is the complexity of **expectation**? $O(k|\mathcal{A}|^2(|\mathcal{O}| + |\mathcal{A}|))$.

maximization

Input: $\mathcal{D}, (P(C_i|o))_{i=1..k, o\in\mathcal{O}} \in [0,1]^{k|\mathcal{O}|}, (\mu_i)_{i=1..k} \in (\mathbb{R}^{|\mathcal{A}|})^k$
Output: $c \in \{\text{false}, \text{true}\}$ indicating whether the convergence is reached, the new parametrization of the mixture of Gaussians

$c \leftarrow$ true
for $i = 1 \rightarrow k$ **do**
  $\mu_i' \leftarrow \frac{\sum_{o\in\mathcal{O}} P(C_i|i)o}{\sum_{o\in\mathcal{O}} P(C_i|i)}$
  if $\mu_i' \neq \mu_i$ **then**
    $c \leftarrow$ false
  end if
  $\Sigma_i' \leftarrow \frac{\sum_{o\in\mathcal{O}} P(C_i|o)(o-\mu_i')(o-\mu_i')^T}{\sum_{o\in\mathcal{O}} P(C_i|o)}$
  $P(C_i)' \leftarrow \frac{\sum_{o\in\mathcal{O}} P(C_i|o)}{|\mathcal{O}|}$
end for
return$(c, (\mu_i')_{i=1..k}, (\Sigma_i')_{i=1..k}, (P(C_i)')_{i=1..k})$

Complexity of EM

Question

What is the complexity of **expectation**? $O(k|\mathcal{A}|^2(|\mathcal{O}| + |\mathcal{A}|))$.

Question

What is the complexity of **maximization**?

Complexity of EM

Question

What is the complexity of **expectation**? $O(k|\mathcal{A}|^2(|\mathcal{O}| + |\mathcal{A}|))$.

Question

What is the complexity of **maximization**? $O(k|\mathcal{O}||\mathcal{A}|^2)$.

Complexity of EM

Question

What is the complexity of **expectation**? $O(k|\mathcal{A}|^2(|\mathcal{O}| + |\mathcal{A}|))$.

Question

What is the complexity of **maximization**? $O(k|\mathcal{O}||\mathcal{A}|^2)$.

Question

What is the complexity of EM if $t \in \mathbb{N}$ iterations are necessary to converge?

Complexity of EM

Question

What is the complexity of **expectation**? $O(k|\mathcal{A}|^2(|\mathcal{O}| + |\mathcal{A}|))$.

Question

What is the complexity of **maximization**? $O(k|\mathcal{O}||\mathcal{A}|^2)$.

Question

What is the complexity of EM if $t \in \mathbb{N}$ iterations are necessary to converge? $O(tk|\mathcal{A}|^2(|\mathcal{O}| + |\mathcal{A}|))$.

Diagonal covariance matrix

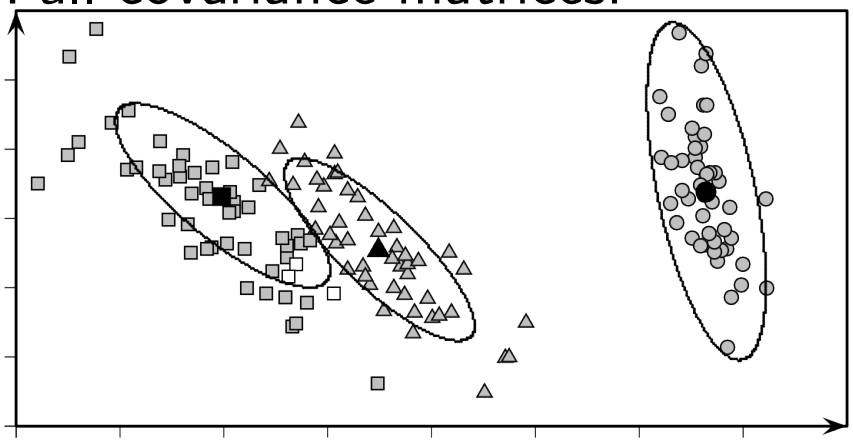
A lower complexity is obtained by assuming all attributes independent, i. e., all covariance matrices diagonal. The operations involving such a matrix become linear in $|\mathcal{A}|$ and the total time complexity of EM becomes $O(tk|\mathcal{O} \times \mathcal{A}|)$.

Diagonal covariance matrix

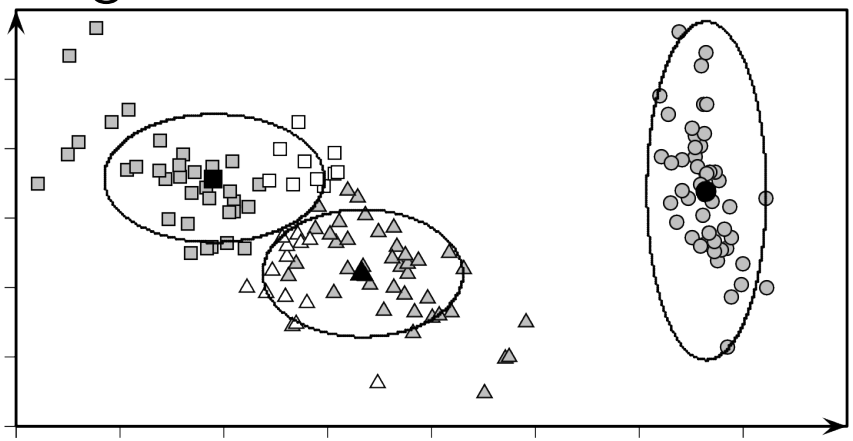
A lower complexity is obtained by assuming all attributes independent, i. e., all covariance matrices diagonal. The operations involving such a matrix become linear in $|\mathcal{A}|$ and the total time complexity of EM becomes $O(tk|\mathcal{O} \times \mathcal{A}|)$.

However, if the attributes are not really independent, the obtained fuzzy clustering become much worse:

Full covariance matrices:



Diagonal covariance matrices:



k-means as specialization of EM

k-means is EM with f_i chosen as follows:

$$\begin{cases} 1 \text{ if } C_i = \arg\max_{a=1..k} s(o, \mu_a) \\ 0 \text{ otherwise} \end{cases} .$$

Outline

Hierarchical Clustering

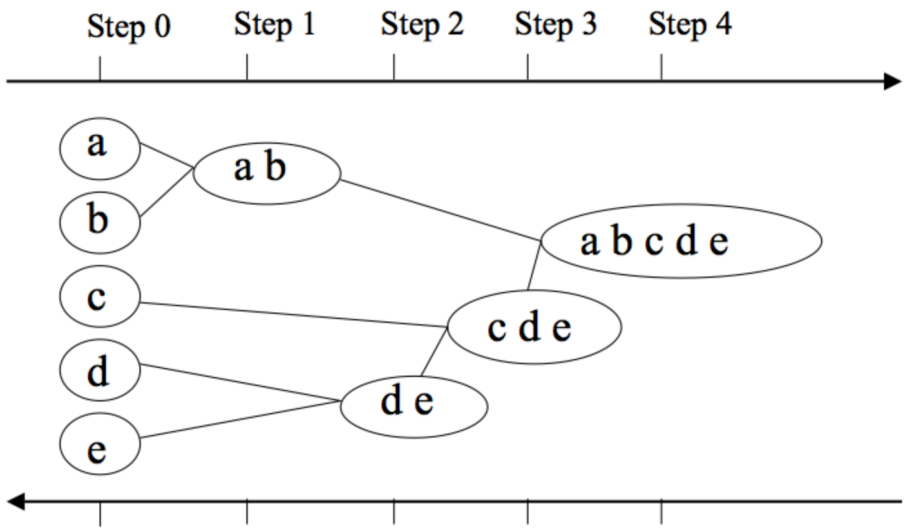
- Build a hierarchy of clusters (not an unique partition);
- The number of clusters k is not required as input;
- Use a distance matrix as clustering criteria
- An early-termination condition can be used (ex. nb clusters).

Algorithm

Input: a sample of m objets $x_1, \ldots, x_m$.

1. The algorithm begins with m clusters (1 cluster = 1 object);
2. Merge the 2 clusters that are the closest.
3. End If it remains only one cluster.
4. Go to step 2.

Output: a dendogram



A hierarchy that can be split at a given level to form a partition.

- the hierarchy: a tree called dendogram
- the leaves = the objects

Distance between clusters

- Distance between the centers (centroid method)
- Minimal distance among the pairs composed of objects from the two clusters (Single Link Method):

$$d(i,j) = min_{x \in C_i; y \in C_j} d(x,y)$$

- Maximal distance among the pairs composed of objects from the two clusters (Complete Link Method):

$$d(i,j) = max_{x \in C_i; y \in C_j} d(x,y)$$

- Average distance among the pairs composed of objects from the two clusters (Average Linkage Method):

$$d(i,j) = avg_{x \in C_i; y \in C_j} d(x,y)$$

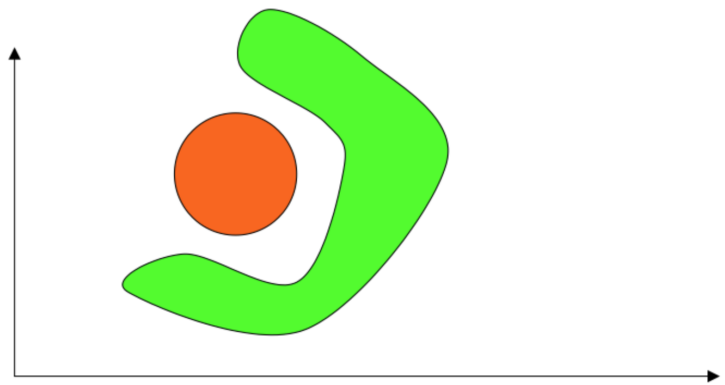
Pros:

- Conceptually simple.
- Theoretical properties well-known.

Cons:

- The clustering is definitive: erroneous decisions are impossible to modify later.
- Non-extensible method for large collections of objects ($\theta(n^2)$)

Outline

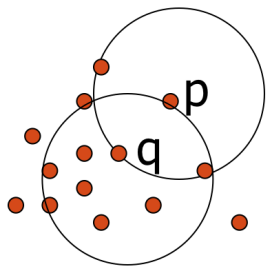


- For this kind of problem, the use of similarity (or distance) measures is less efficient than the use of neighborhood density

Density-based clustering

- Clusters are seen as dense regions separated by regions that are much less denser (noise)
- Two parameters:
  - Eps: The maximum radius of the neighborhood
  - MinPts: Minimum number of points within the Eps-neighborhood of a point.
- Neighborhood: $V_{Eps}(p) = \{q \in D \mid dist(p, q) \leq Eps\}$
- A point p is directly density-accessible from q w.r.t. Eps, MinPts if

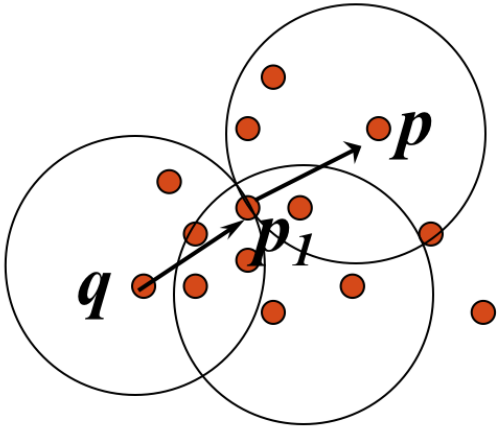
$P \in V_{Eps}(q)$ and $|V_{Eps}(q)| \geq MinPts$

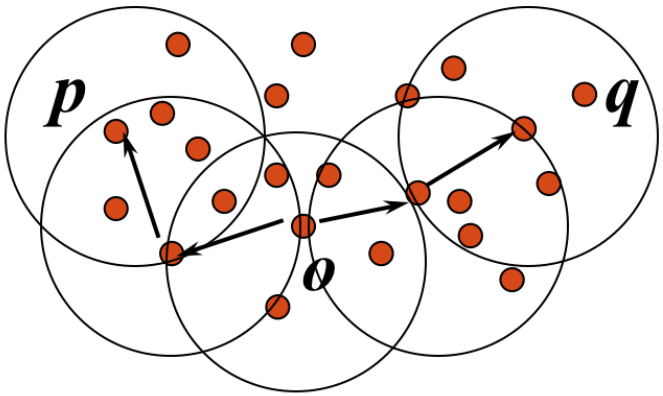


MinPts = 5

Eps = 1 cm

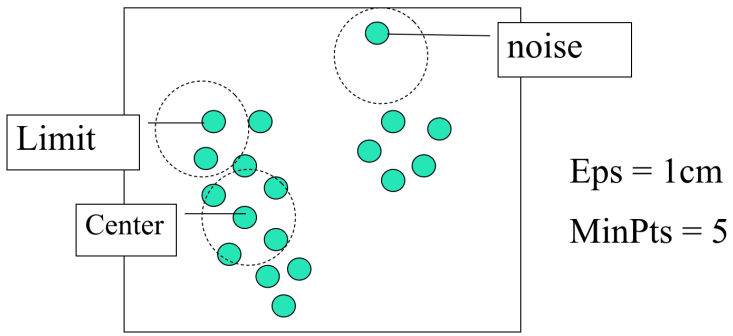
- Accessibility: p is accessible from q w.r.t. Eps, MinPts if there exists $p_1, \ldots, p_n$ such that $p_1 = q$, $p_n = p$ and p_{i+1} is directly accessible from p_i.
- Connexity: p is connected to q w.r.t. Eps and MinPts if there exists a point o such that p and q are accessible from o.





DBSCAN: Density Based Spatial Clustering of Applications with Noise

- A cluster is the maximal set of connected points
- Cluster shapes are not necessary convex



DBSCAN Algorithm

- Choose p
- Retrieve all poinst that accessible from p (w.r.t. Eps and MinPts)
- If p is a center, then a cluster is created.
- If p is a limit, then there is not accessible point from p, **Skip to another point**
- Repeat until it remains no point.

Outline

Summary

- k-means iteratively assigns each object to the cluster whose center is the most similar and recompute these centers as the mean of the objects they were assigned;

Summary

- k-means iteratively assigns each object to the cluster whose center is the most similar and recompute these centers as the mean of the objects they were assigned;
- Its strongest advantage is its low complexity;

Summary

- k-means iteratively assigns each object to the cluster whose center is the most similar and recompute these centers as the mean of the objects they were assigned;
- Its strongest advantage is its low complexity;
- Its worst drawback is its tendency to discover equi-sized clusters;

Summary

- k-means iteratively assigns each object to the cluster whose center is the most similar and recompute these centers as the mean of the objects they were assigned;
- Its strongest advantage is its low complexity;
- Its worst drawback is its tendency to discover equi-sized clusters;
- k-means actually is a specialization of a whole class of algorithms called EM;

Summary

- k-means iteratively assigns each object to the cluster whose center is the most similar and recompute these centers as the mean of the objects they were assigned;
- Its strongest advantage is its low complexity;
- Its worst drawback is its tendency to discover equi-sized clusters;
- k-means actually is a specialization of a whole class of algorithms called EM;
- They treat the dataset as a random sample of a multivariate random variable whose pdf is given as a mixture model;

Summary

- k-means iteratively assigns each object to the cluster whose center is the most similar and recompute these centers as the mean of the objects they were assigned;
- Its strongest advantage is its low complexity;
- Its worst drawback is its tendency to discover equi-sized clusters;
- k-means actually is a specialization of a whole class of algorithms called EM;
- They treat the dataset as a random sample of a multivariate random variable whose pdf is given as a mixture model;
- They *locally* maximize the likelihood, i. e., the probability of observing the dataset given the parametrization of the mixture model;

Summary

- k-means iteratively assigns each object to the cluster whose center is the most similar and recompute these centers as the mean of the objects they were assigned;
- Its strongest advantage is its low complexity;
- Its worst drawback is its tendency to discover equi-sized clusters;
- k-means actually is a specialization of a whole class of algorithms called EM;
- They treat the dataset as a random sample of a multivariate random variable whose pdf is given as a mixture model;
- They *locally* maximize the likelihood, i. e., the probability of observing the dataset given the parametrization of the mixture model;
- They iteratively compute the expectation of the likelihood and update the parametrization so that this expectation is maximized.

The end.