DSS
Decision Sciences & Systems

Technische Universität München

Tutorial Business Analytics

Tutorial 9: Ensemble Methods and Clustering

Decision Sciences & Systems (DSS)
Department of Informatics
TU München

Tutorial 9 Business Analytics: Clustering and Ensemble Methods

Today’s Agenda

2. Clustering

2.1 Theory: Difference Between **Classification and Clustering** (Tutorial Video)

2.2 Theory: Partitional Clustering: **K-Means**
Practise: **Exercise 9.1**

2.3 Theory Probabilistic Clustering: **EM Algorithm**
Practise: **Exercise 9.2**

1. Ensemble Methods (Tutorial Video)

1.1 Theory: What are **Ensemble Methods**?

1.2 Theory: **Bagging**

1.3 Theory: **Boosting**

1.4 Theory: **Stacking**

Homework

- **Exercise 9.3**
- **Exercise 9.4**
- **Exercise 9.5**

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – A centroid-based technique

Idea: 2-step method to partition the instances into *k* clusters, $C_1 \ldots C_k$,with high intra-cluster similarity and high inter-cluster dissimilarity.

Method:

- [init] Initially choose *k* random centers or randomly pick *k* instances as initial centers
- [repeat]
  - **Step 1**: Assign instances to the closest cluster
  - **Step 2**: Update cluster center
- [until no change]

Step 1: How to Identify the Closest Cluster?

Distance function: Euclidean distance

$$d(p, c) = \sqrt{(x(p) - x(c))^2 + (y(p) - y(c))^2}$$

Step 2: How to Update Cluster Centre

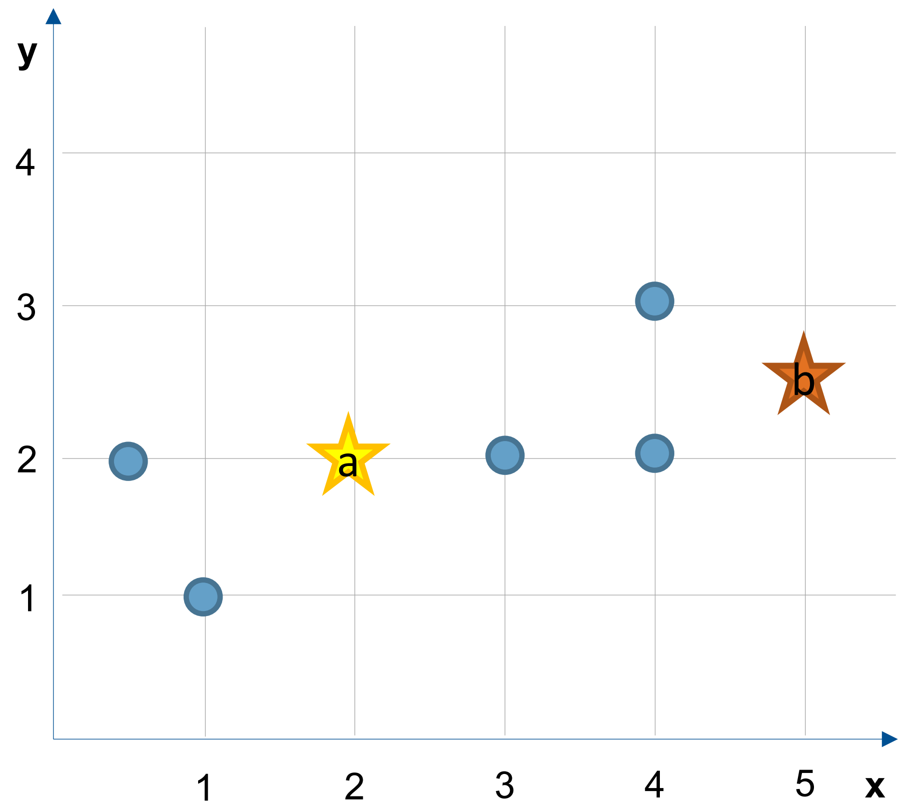
Calculate centroid:

$$x'(c_i) = \frac{\sum_{p \epsilon C_i} x(p)}{|c_i|} \text{ and } y'(c_i) = \frac{\sum_{p \in C_i} y(p)}{|c_i|}$$

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Example

Group the data into two clusters applying the k-Means algorithm and the Euclidean distance function



Dataset

p_i	x	y
1	4	3
2	0.5	2
3	3	2
4	4	2
5	1	1

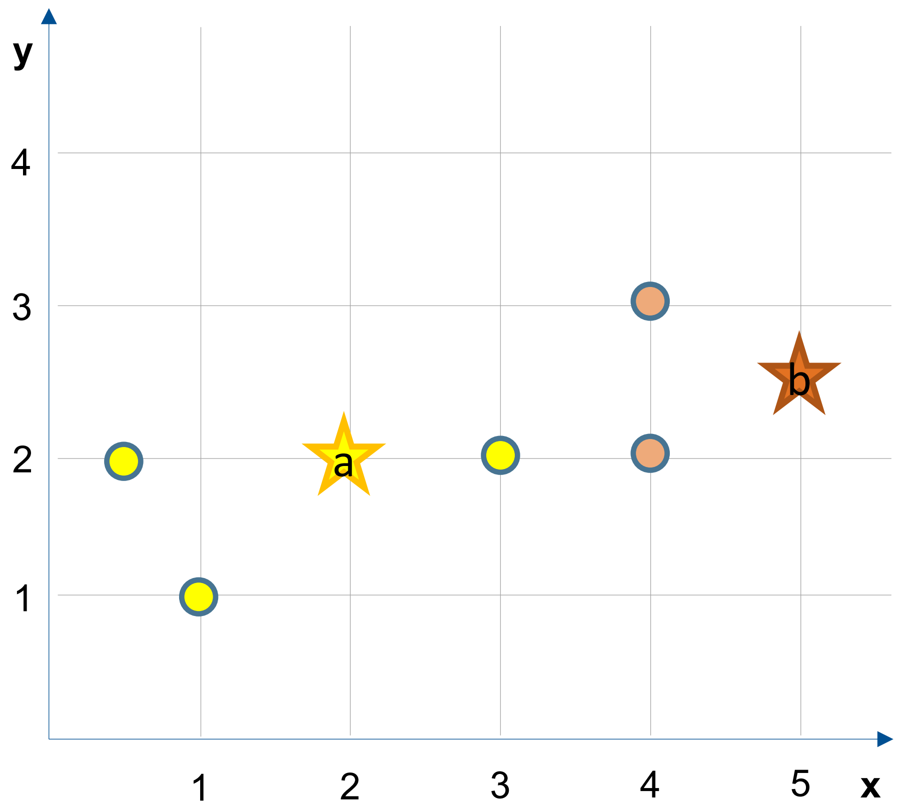
Centroids

c_i	x	y
a	2	2
b	5.0	2.5

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Example

Step 1: (Re)assign instances to the closest cluster



Dataset

p_i	x	y
1	4	3
2	0.5	2
3	3	2
4	4	2
5	1	1

Centroids

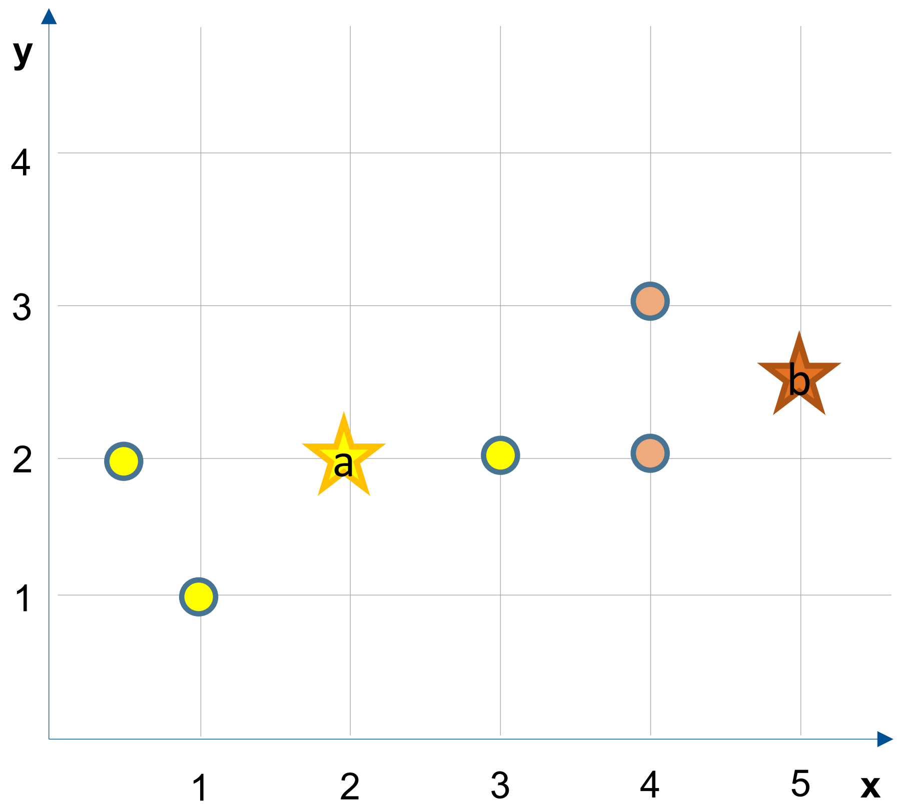
c_i	x	y
a	2	2
b	5.0	2.5

DSS
Decision Sciences & Systems

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Example

Step 2: Update cluster center



Dataset

p_i	x	y
1	4	3
2	0.5	2
3	3	2
4	4	2
5	1	1

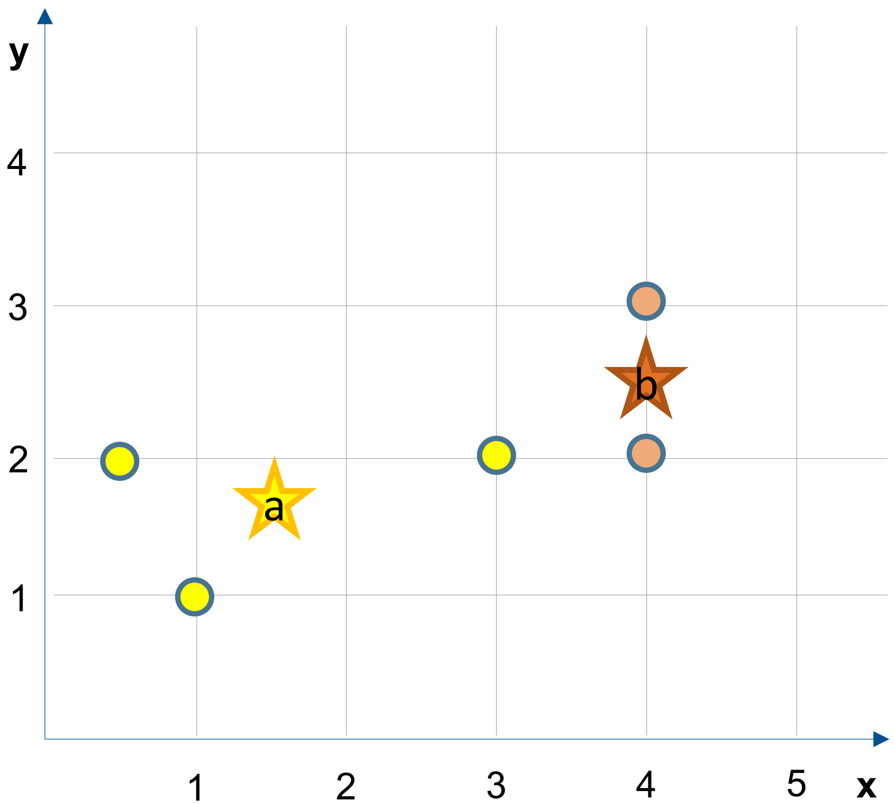
Centroids

c_i	x	y
a		
b		

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Example

Step 2: Update cluster center



Dataset

p_i	x	y
1	4	3
2	0.5	2
3	3	2
4	4	2
5	1	1

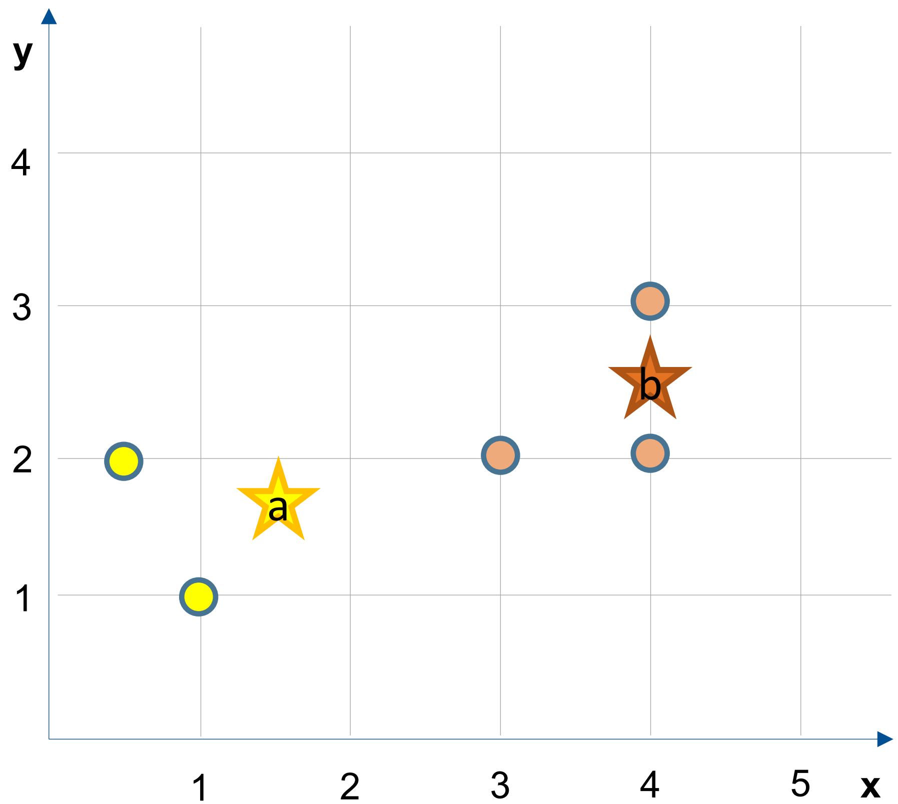
Centroids

c_i	x	y
a	1.5	1.67
b	4	2.5

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Example

Step 1: (Re)assign instances to the closest cluster



Dataset

p_i	x	y
1	4	3
2	0.5	2
3	3	2
4	4	2
5	1	1

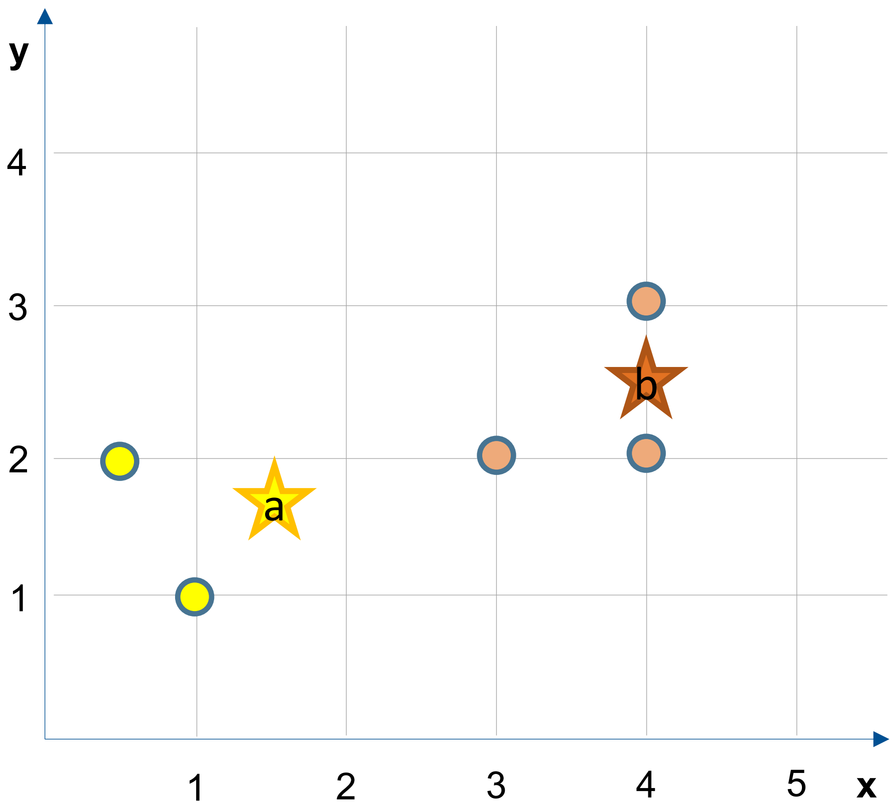
Centroids

c_i	x	y
a	1.5	1.67
b	4	2.5

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Example

Step 2: Update cluster center



Dataset

p_i	x	y
1	4	3
2	0.5	2
3	3	2
4	4	2
5	1	1

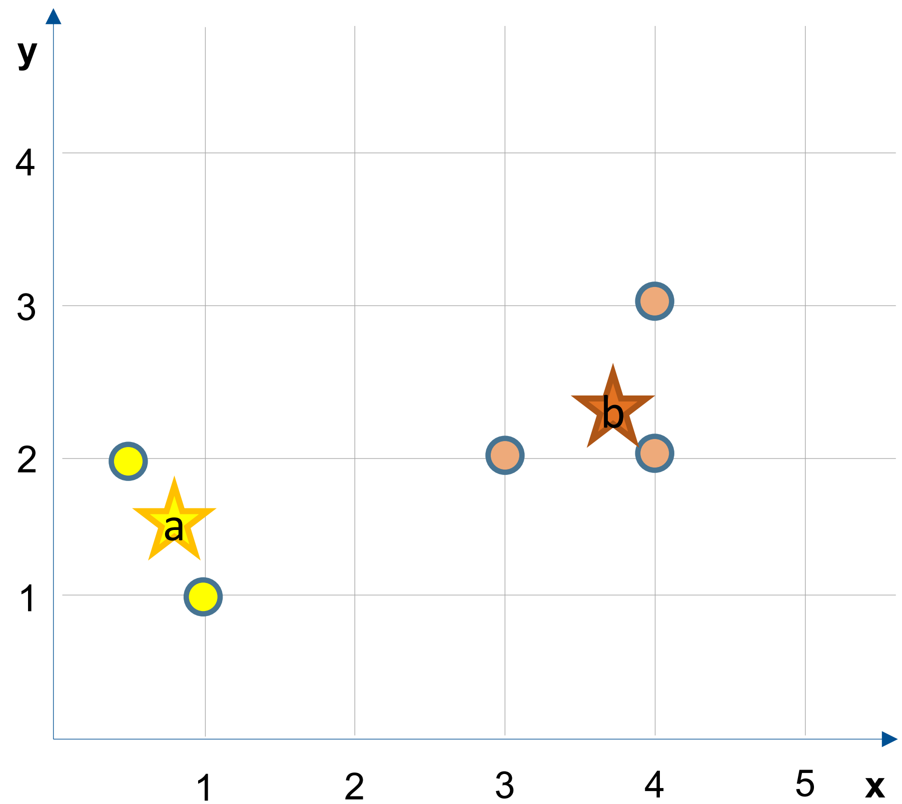
Centroids

c_i	x	y
a		
b		

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Example

Step 2: Update cluster center



Dataset

p_i	x	y
1	4	3
2	0.5	2
3	3	2
4	4	2
5	1	1

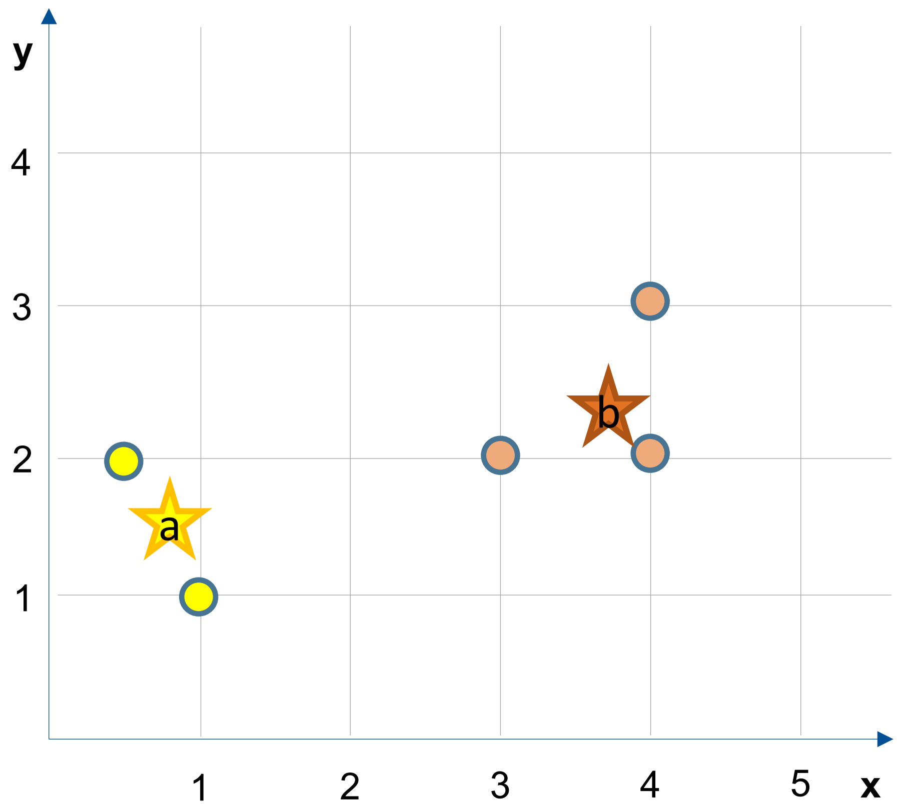
Centroids

c_i	x	y
a	0.75	1.5
b	3.67	2.33

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Example

Step 1: (Re)assign instances to the closest cluster – No Reassignment: Algorithm terminates



Dataset

p_i	x	y
1	4	3
2	0.5	2
3	3	2
4	4	2
5	1	1

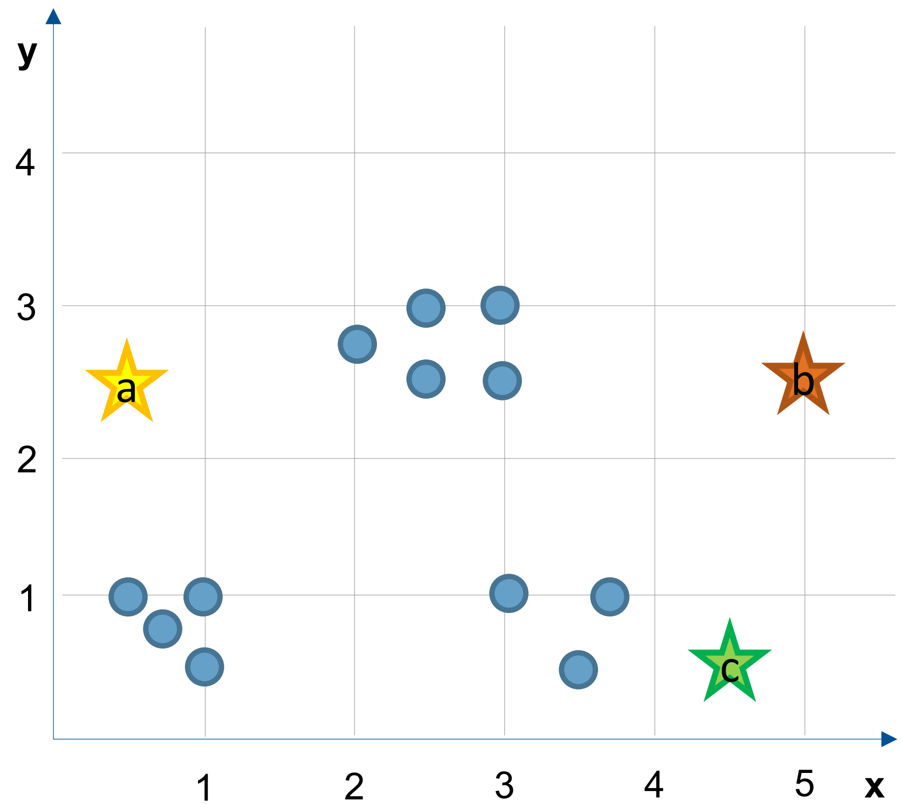
Centroids

c_i	x	y
a	0.75	1.5
b	3.67	2.33

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Exercise 9.1

Group the data into three clusters applying the k-Means algorithm and the Euclidean distance function



Dataset

p_i	x	y
1	2.5	3
2	3	3
3	2	2.75
4	2.5	2.5
5	3	2.5
6	0.5	1
7	1	1
8	3	1
9	3.75	1
10	0.75	0.75
11	1	0.5
12	3.5	0.5

Centroids

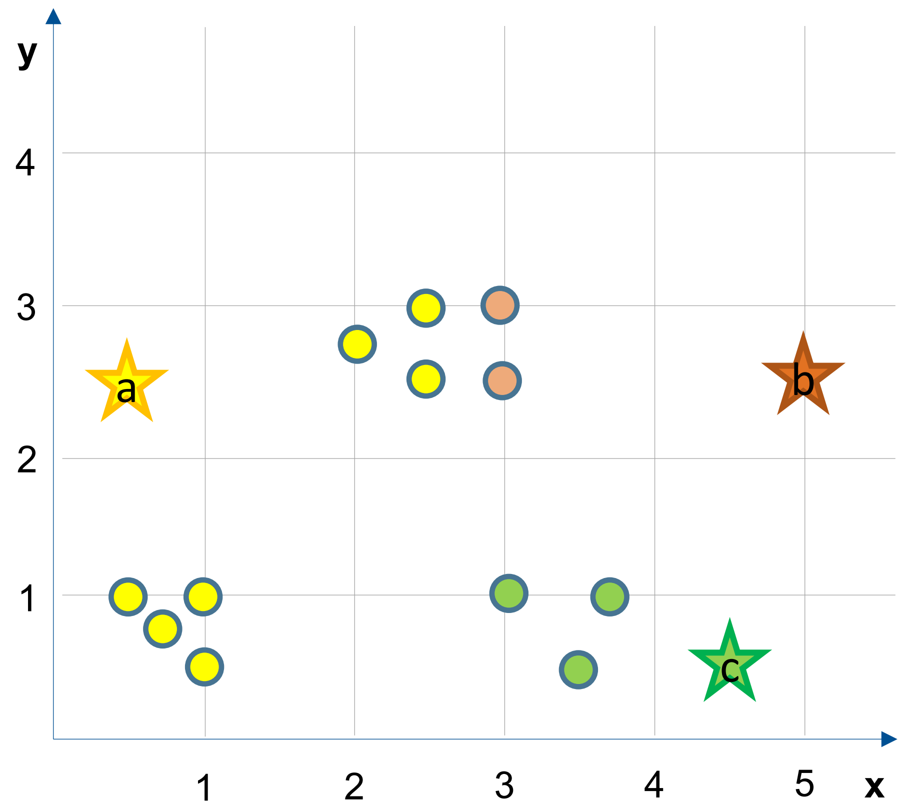
c_i	x	y
a	0.5	2.5
b	5.0	2.5
c	4.5	0.5

DSS Decision Sciences & Systems

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Exercise 9.1

Solution Step 1: (Re)assign instances to the closest cluster



Dataset

p_i	x	y
1	2.5	3
2	3	3
3	2	2.75
4	2.5	2.5
5	3	2.5
6	0.5	1
7	1	1
8	3	1
9	3.75	1
10	0.75	0.75
11	1	0.5
12	3.5	0.5

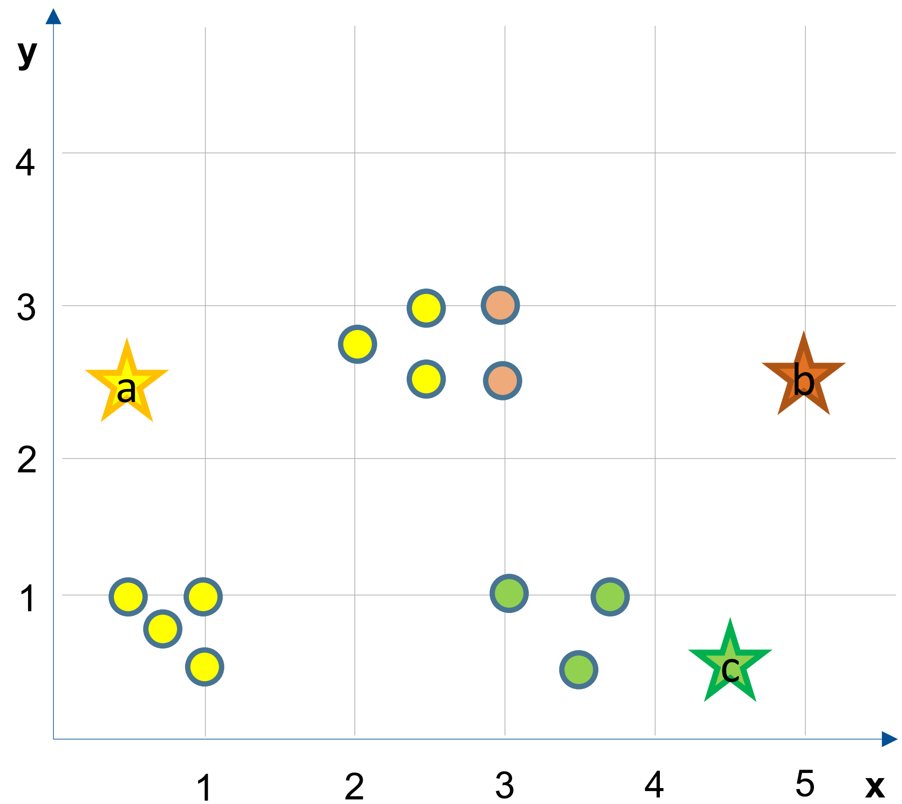
Centroids

c_i	x	y
a	0.5	2.5
b	5.0	2.5
c	4.5	0.5

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Exercise 9.1

Solution Step 2: Update cluster center



Dataset

p_i	x	y
1	2.5	3
2	3	3
3	2	2.75
4	2.5	2.5
5	3	2.5
6	0.5	1
7	1	1
8	3	1
9	3.75	1
10	0.75	0.75
11	1	0.5
12	3.5	0.5

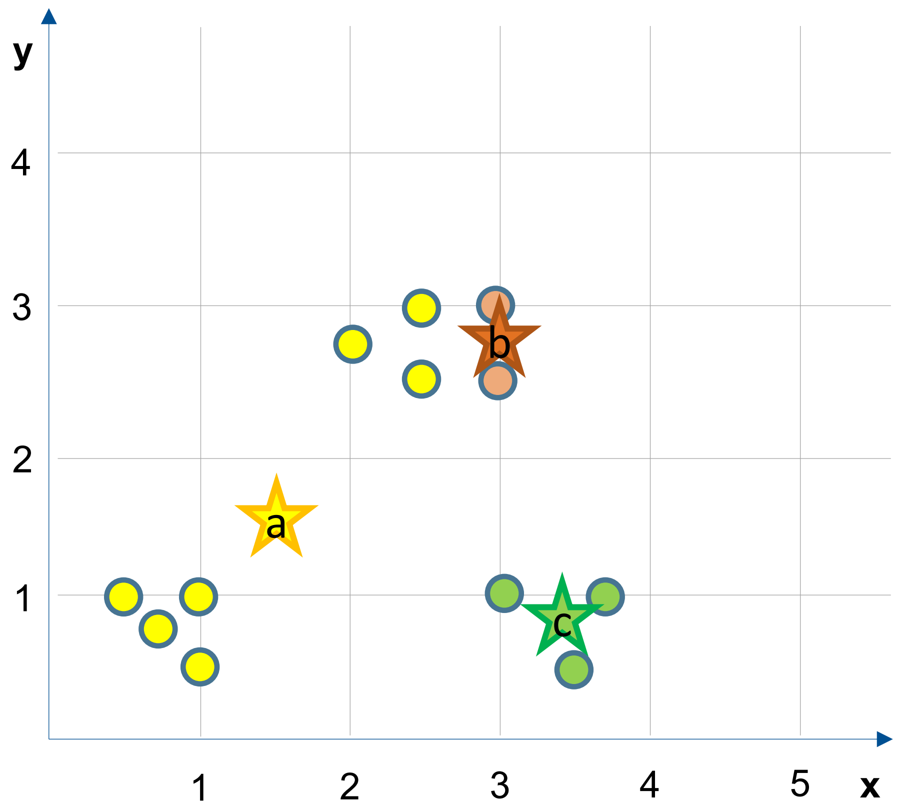
Centroids

c_i	x	y
a		
b		
c		

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Exercise 9.1

Solution Step 2: Update cluster center



Dataset

p_i	x	y
1	2.5	3
2	3	3
3	2	2.75
4	2.5	2.5
5	3	2.5
6	0.5	1
7	1	1
8	3	1
9	3.75	1
10	0.75	0.75
11	1	0.5
12	3.5	0.5

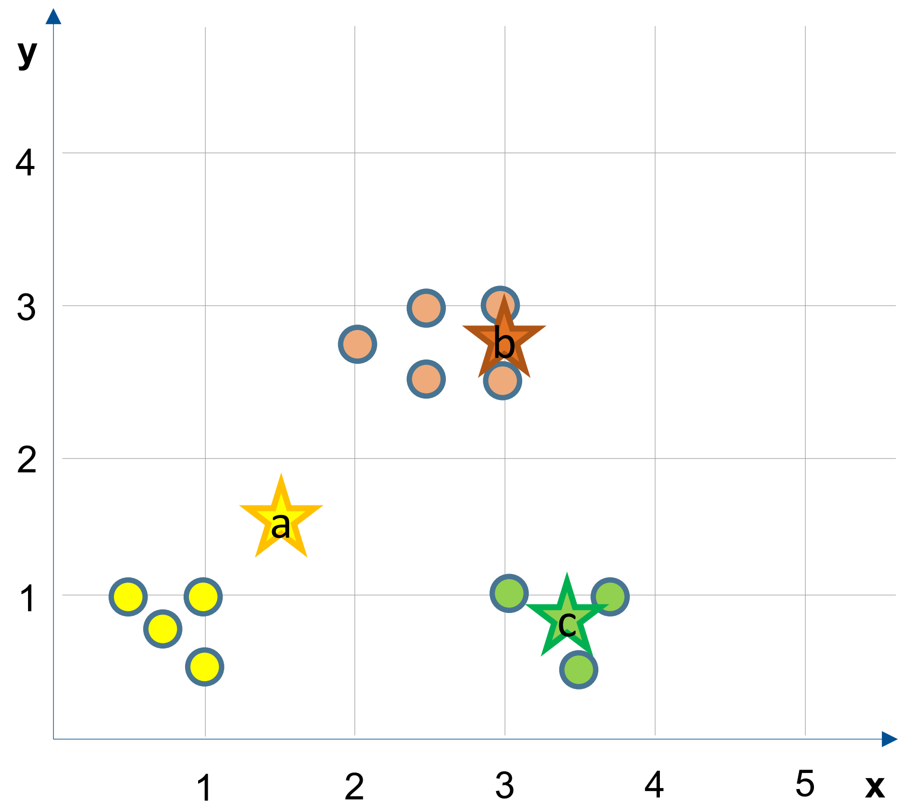
Centroids

c_i	x	y
a	1.46	1.64
b	3.00	2.75
c	3.42	0.83

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Exercise 9.1

Solution Step 1: (Re)assign instances to the closest cluster



Dataset

p_i	x	y
1	2.5	3
2	3	3
3	2	2.75
4	2.5	2.5
5	3	2.5
6	0.5	1
7	1	1
8	3	1
9	3.75	1
10	0.75	0.75
11	1	0.5
12	3.5	0.5

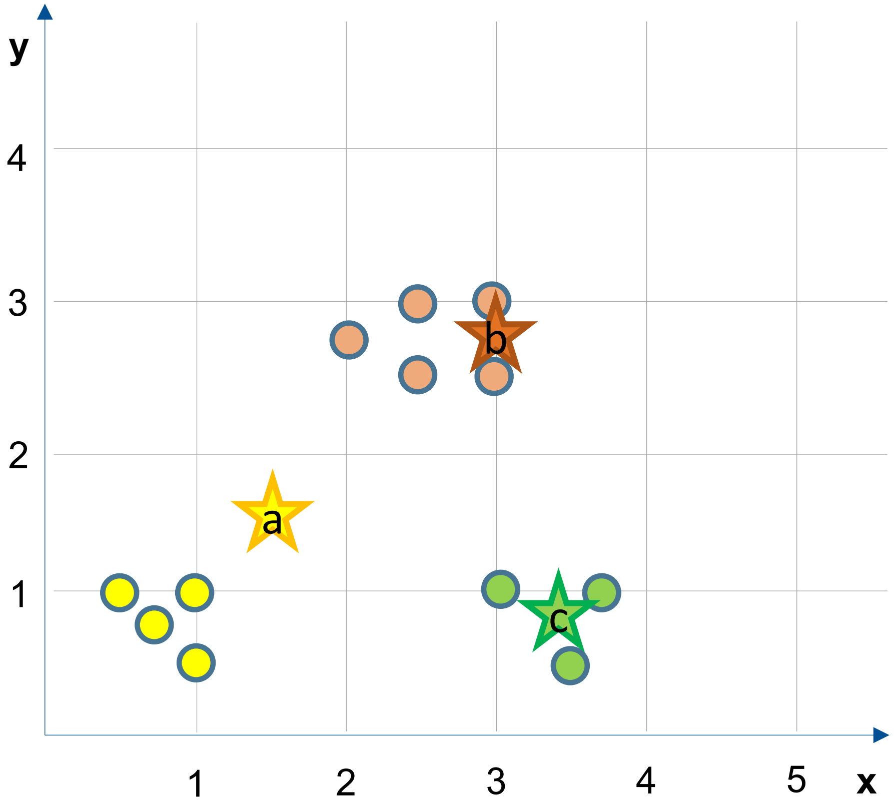
Centroids c_i

c_i	x	y
a	1.46	1.64
b	3.00	2.75
c	3.42	0.83

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Exercise 9.1

Solution Step 2: Update cluster center



Dataset

p_i	x	y
1	2.5	3
2	3	3
3	2	2.75
4	2.5	2.5
5	3	2.5
6	0.5	1
7	1	1
8	3	1
9	3.75	1
10	0.75	0.75
11	1	0.5
12	3.5	0.5

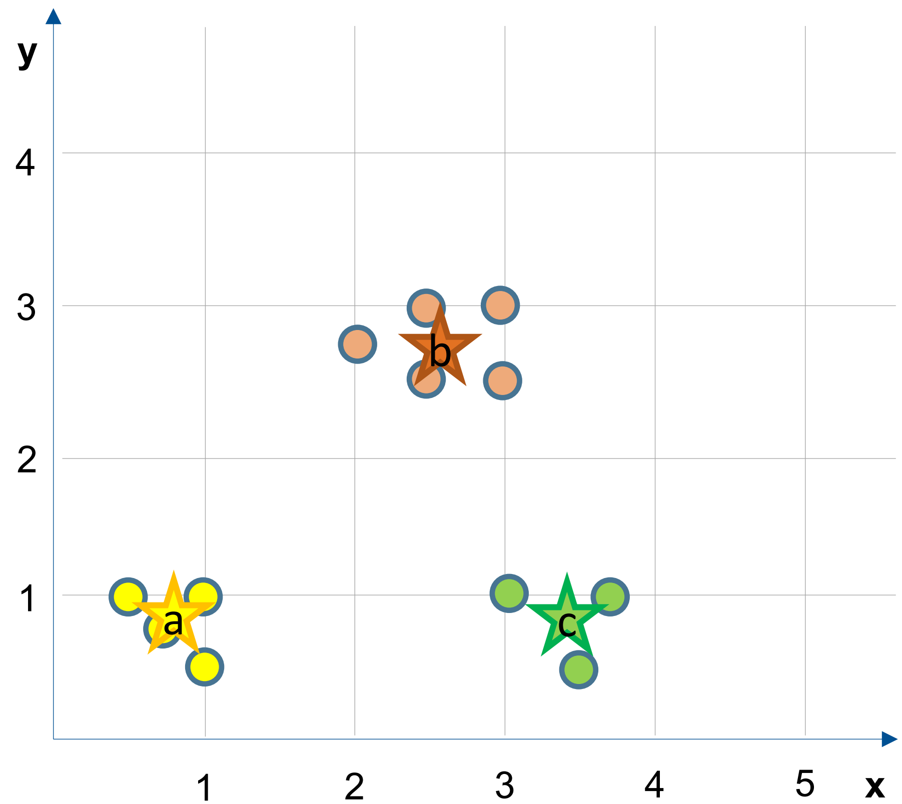
Centroids c_i

c_i	x	y
a		
b		
c		

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Exercise 9.1

Solution Step 2: Update cluster center



Dataset

p_i	x	y
1	2.5	3
2	3	3
3	2	2.75
4	2.5	2.5
5	3	2.5
6	0.5	1
7	1	1
8	3	1
9	3.75	1
10	0.75	0.75
11	1	0.5
12	3.5	0.5

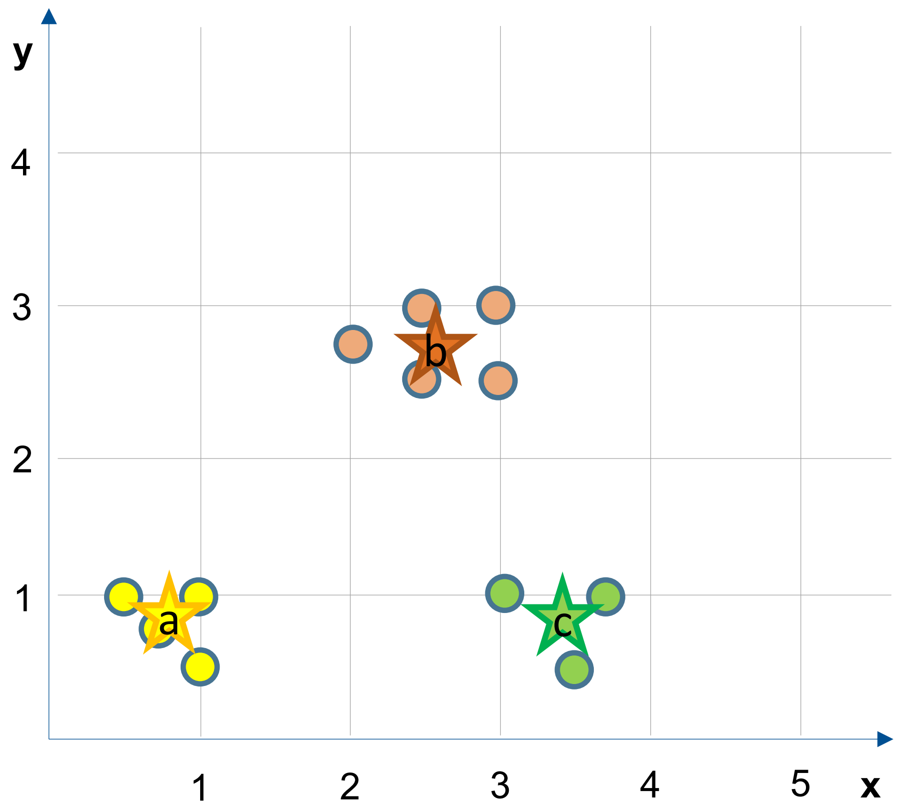
Centroids c_i

c_i	x	y
a	0.81	0.81
b	2.60	2.75
c	3.42	0.83

Tutorial 9 Business Analytics: Clustering

2.2 Partitional Clustering: K-Means – Exercise 9.1

Solution Step 1: (Re)assign instances to the closest cluster – No Reassignment: Algorithm terminates



Dataset

p_i	x	y
1	2.5	3
2	3	3
3	2	2.75
4	2.5	2.5
5	3	2.5
6	0.5	1
7	1	1
8	3	1
9	3.75	1
10	0.75	0.75
11	1	0.5
12	3.5	0.5

Centroids c_i

c_i	x	y
a	0.81	0.81
b	2.60	2.75
c	3.42	0.83

Tutorial 9 Business Analytics: Clustering

2.3 Probabilistic Clustering: Expectation-Maximization (EM) – Fuzzy Clustering

Why Fuzzy Clustering?

So far: Each object p in our data set can be assigned to one of clusters only

However: Sometimes, a fuzzy or flexible cluster assignment is realistic

Idea: 2-step method to calculate cluster assignment probabilities (1) & estimate distr. parameters (2)

Method:

- [init] Start with guesses for cluster centers and define k
- [repeat]
  - **Expectation step**: calculate likelihoods for instance p belonging to distribution (=cluster) A
  - **Maximization step**: optimize the distribution parameters based on the instance likelihoods
- [until convergence/no changes]

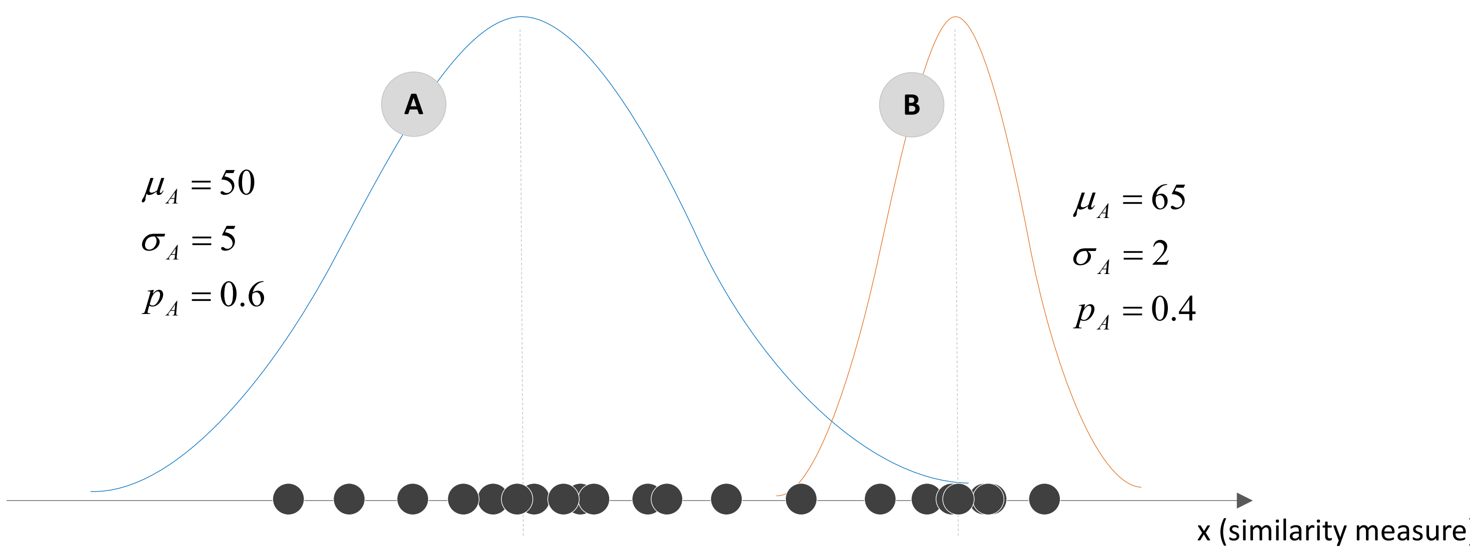
Tutorial 9 Business Analytics: Clustering

2.3 Probabilistic Clustering: Expectation-Maximization (EM) – Fuzzy Clustering

- Model the **various clusters** as **probability distributions**
- Identifying the best distribution for one cluster: **Maximum Likelihood estimates**
- Maximum-Likelihood estimates for the **Standard Normal Distribution** are

$$\mu = \bar{x} \text{ and } s^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

Example: single numeric attribute with a two cluster mixture model (k=2)



Tutorial 9 Business Analytics: Clustering

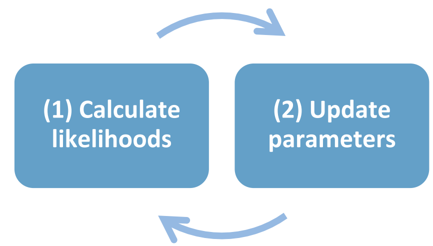
2.3 Probabilistic Clustering: Expectation-Maximization (EM) – Fuzzy Clustering

Initialization

- Assume **equal probabilities p_i** for each distribution/cluster

Initial guess for parameters

- **μ =** define one instance as the cluster mean
- **σ =** set standard deviation of each distribution to **1**



(1) Expectation Step (given k = A and B)

$$f(x,\mu_A,\sigma_A) = \frac{1}{\sigma_A \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu_A)^2}{2\cdot\sigma_A^2}}$$

$$\Pr[x] = f(x,\mu_A,\sigma_A)\cdot p_A + f(x,\mu_B,\sigma_B)\cdot p_B$$

Given instance x, the probability it belongs to cluster A is:

$$\Pr[A|x] = \frac{f(x,\mu_A,\sigma_A)\cdot\ p_A}{\Pr[x]}$$

Pr[A|x] serves as the weight w_i in the maximization step.

(2) Maximization Step

Optimize distribution/cluster parameters based on the instance weights w_i (= the likelihoods):

$$\mu_A = \frac{w_1x_{1+}\ w_2x_2 + \cdots +\ w_nx_n}{w_1 + w_2 + \cdots + w_n}$$

$$\sigma_A = \sqrt{\frac{w_1(x_1-\mu_A)^2 +\ + \cdots +\ w_n(x_n-\mu_A)^2}{w_1 + w_2 + \cdots + w_n}}$$

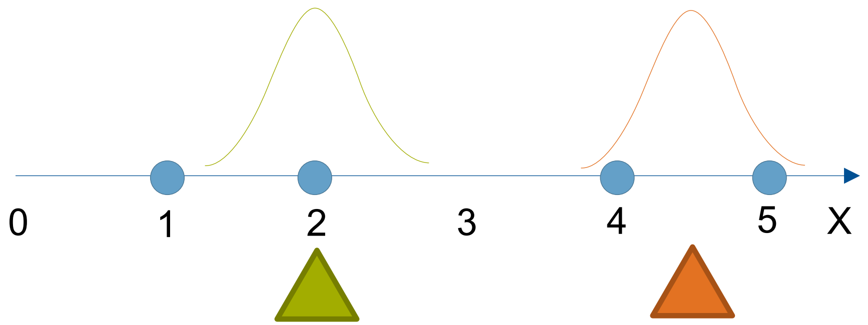
$$p_A = \frac{\sum w_A}{\sum w_A + \sum w_B}\ ; \qquad p_B = 1 - p_A$$

Tutorial 9 Business Analytics: Clustering

2.3 Probabilistic Clustering: Expectation-Maximization – Example

Given k=2, perform EM algorithm with the following instances

Instance	1	2	3	4
Value	1	2	4	5

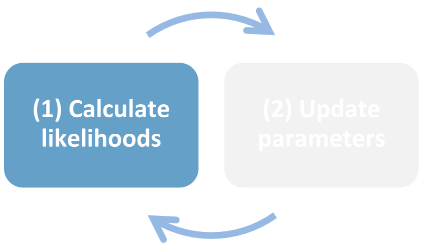


Initialisation

μ_A	2
σ_A	1.00
p_A	50%
μ_B	4.5
σ_B	1.00
p_B	50%

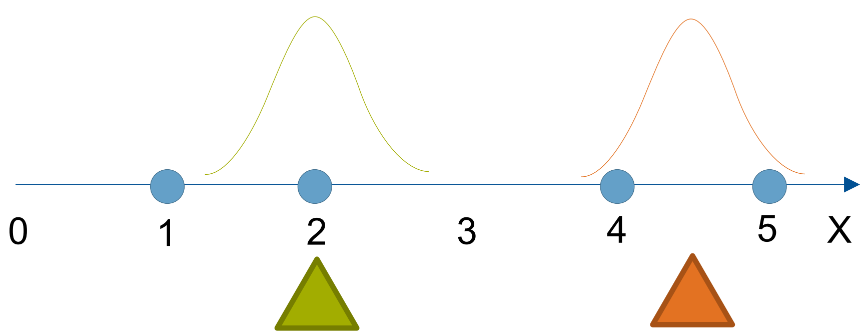
Tutorial 9 Business Analytics: Clustering

2.3 Probabilistic Clustering: Expectation-Maximization – Example



phase 1, step 1: Calculate cluster probabilities

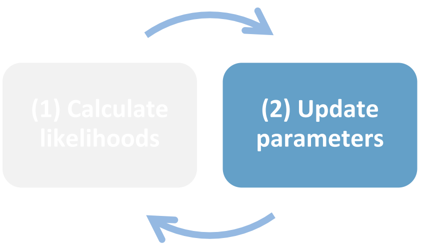
Instance	1	2	3	4
Value	1	2	4	5



μ_A	2
σ_A	1.00
p_A	50%
μ_B	4.5
σ_B	1.00
p_B	50%

	1	2	3	4
Pr[A|x]	99.64%	95.79%	13.30%	1.24%
Pr[B|x]	0.36%	4.21%	86.70%	98.76%

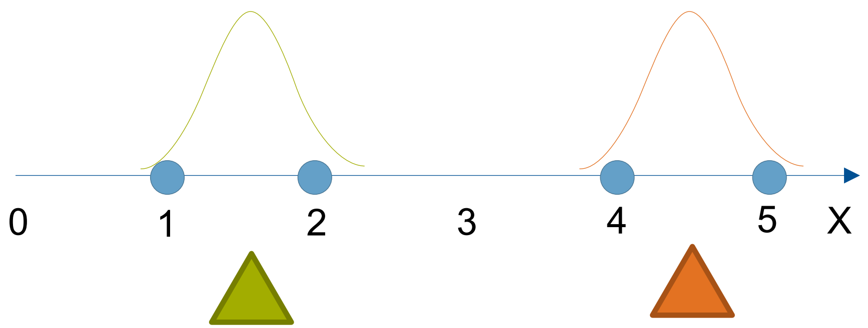
Tutorial 9 Business Analytics: Clustering



2.3 Probabilistic Clustering: Expectation-Maximization – Example

phase 1, step 2: Update distribution parameters

Instance	1	2	3	4
Value	1	2	4	5



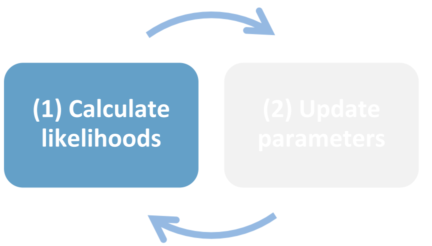
μ_A	1.67
σ_A	0.82
p_A	52%
μ_B	4.47
σ_B	0.64
p_B	48%

	1	2	3	4
Pr[A|x]	99.64%	95.79%	13.30%	1.24%
Pr[B|x]	0.36%	4.21%	86.70%	98.76%

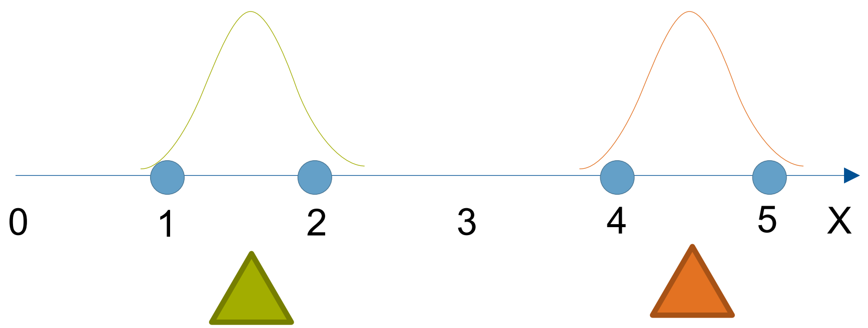
Tutorial 9 Business Analytics: Clustering

2.3 Probabilistic Clustering: Expectation-Maximization – Example

phase 2, step 1: Calculate cluster probabilities



Instance	1	2	3	4
Value	1	2	4	5

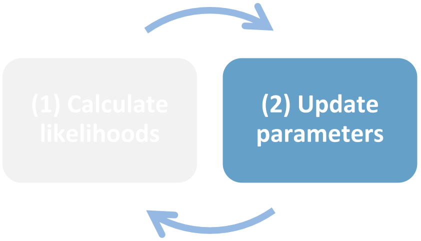


μ_A	1.67
σ_A	0.82
p_A	52%
μ_B	4.47
σ_B	0.64
p_B	48%

	1	2	3	4
Pr[A|x]	100.00%	99.93%	1.95%	0.03%
Pr[B|x]	0.00%	0.07%	98.05%	99.97%

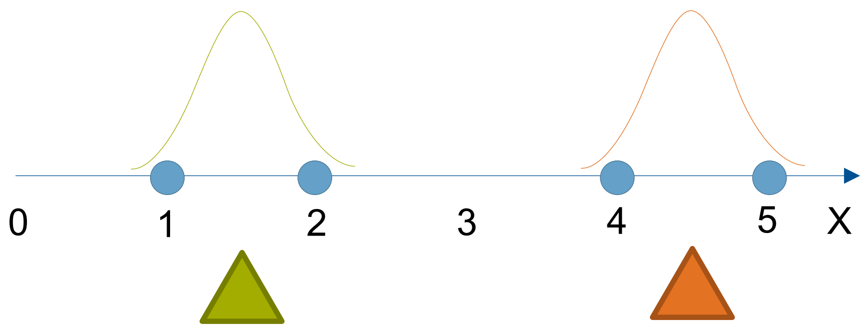
Tutorial 9 Business Analytics: Clustering

2.3 Probabilistic Clustering: Expectation-Maximization – Example



phase 2, step 2: Update distribution parameters

Instance	1	2	3	4
Value	1	2	4	5



μ_A	1.52
σ_A	0.56
p_A	50%
μ_B	4.50
σ_B	0.50
p_B	50%

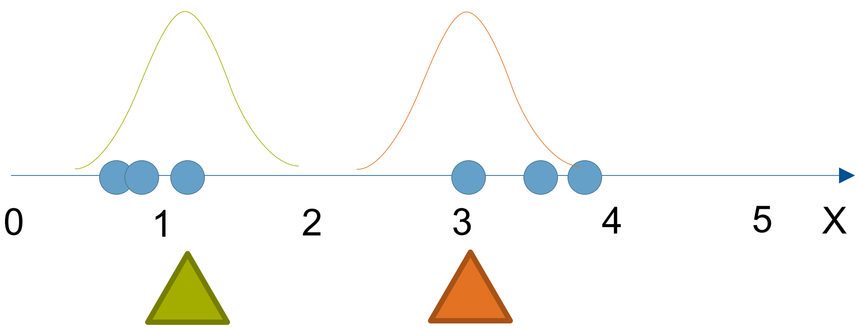
	1	2	3	4
Pr[A|x]	100.00%	99.93%	1.95%	0.03%
Pr[B|x]	0.00%	0.07%	98.05%	99.97%

Tutorial 9 Business Analytics: Clustering

2.3 Probabilistic Clustering: Expectation-Maximization – Exercise 9.2

Given k=2, perform EM algorithm with the following instances

Instance	1	2	3	4	5	6
Value	0.76	0.86	1.12	3.05	3.51	3.75

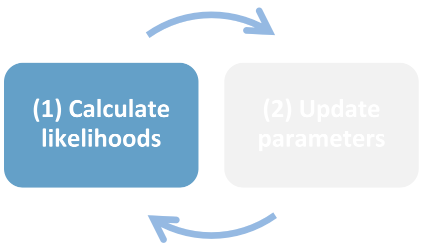


Initialisation

μ_A	1.12
σ_A	1.00
p_A	50%
μ_B	3.05
σ_B	1.00
p_B	50%

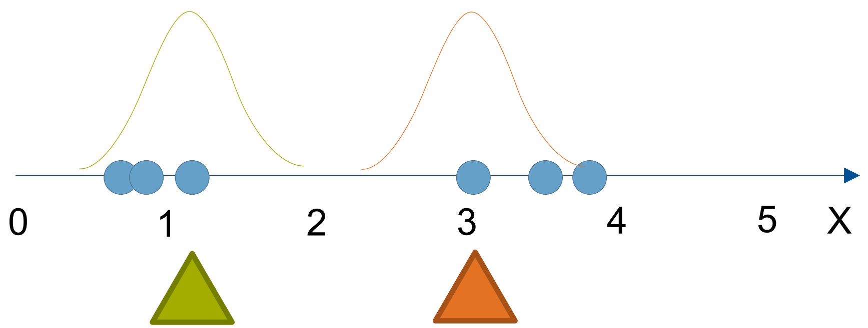
Tutorial 9 Business Analytics: Clustering

2.3 Probabilistic Clustering: Expectation-Maximization – Exercise 9.2



Solution phase 1, step 1: Calculate cluster probabilities

Instance	1	2	3	4	5	6
Value	0.76	0.86	1.12	3.05	3.51	3.75



μ_A	1.12
σ_A	1.00
p_A	50%
μ_B	3.05
σ_B	1.00
p_B	50%

	1	2	3	4	5	6
Pr[A|x]	92.81%	91.41%	86.56%	13.44%	6.01%	3.87%
Pr[B|x]	7.19%	8.59%	13.44%	86.56%	93.99%	96.13%

DSS
Decision Sciences & Systems

Tutorial 9 Business Analytics: Clustering

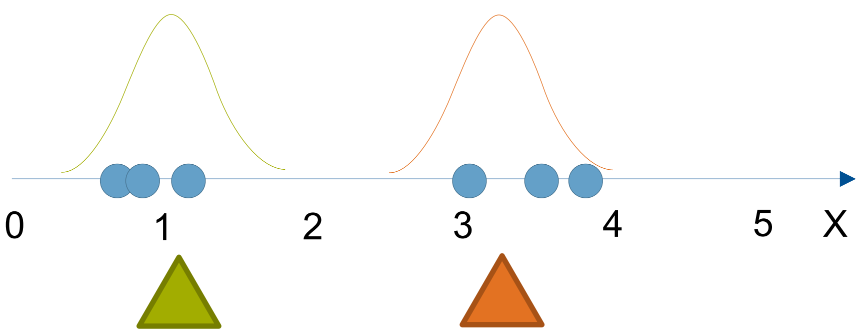
2.3 Probabilistic Clustering: Expectation-Maximization – Exercise 9.2

(1) Calculate likelihoods

(2) Update parameters

Solution phase 1, step 2: Update distribution parameters

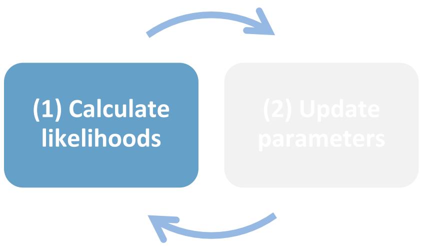
Instance	1	2	3	4	5	6
Value	0.76	0.86	1.12	3.05	3.51	3.75



μ_A	1.10
σ_A	0.66
p_A	49%
μ_B	3.21
σ_B	0.78
p_B	51%

	1	2	3	4	5	6
Pr[A|x]	92.81%	91.41%	86.56%	13.44%	6.01%	3.87%
Pr[B|x]	7.19%	8.59%	13.44%	86.56%	93.99%	96.13%

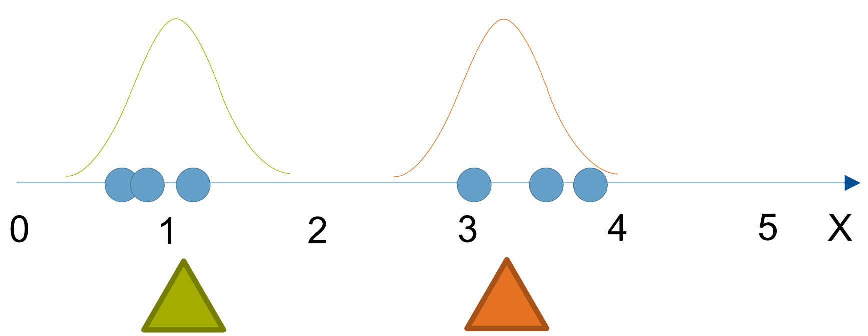
Tutorial 9 Business Analytics: Clustering



2.3 Probabilistic Clustering: Expectation-Maximization – Exercise 9.2

Solution phase 2, step 1: Calculate cluster probabilities

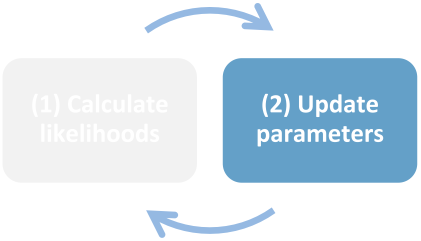
Instance	1	2	3	4	5	6
Value	0.76	0.86	1.12	3.05	3.51	3.75



μ_A	1.10
σ_A	0.66
p_A	49%
μ_B	3.21
σ_B	0.78
p_B	51%

	1	2	3	4	5	6
Pr[A|x]	99.25%	98.97%	97.55%	1.49%	0.16%	0.05%
Pr[B|x]	0.75%	1.03%	2.45%	98.51%	99.84%	99.95%

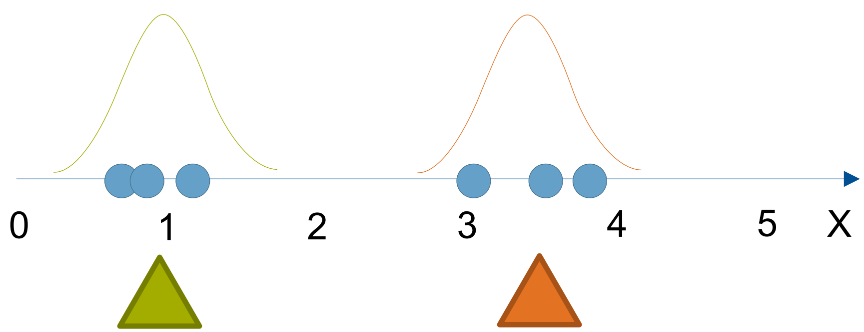
Tutorial 9 Business Analytics: Clustering



2.3 Probabilistic Clustering: Expectation-Maximization – Exercise 9.2

Solution phase 2, step 2: Update distribution parameters

Instance	1	2	3	4	5	6
Value	0.76	0.86	1.12	3.05	3.51	3.75



μ_A	0.92
σ_A	0.22
p_A	49%
μ_B	3.40
σ_B	0.41
p_B	51%

	1	2	3	4	5	6
Pr[A|x]	99.25%	98.97%	97.55%	1.49%	0.16%	0.05%
Pr[B|x]	0.75%	1.03%	2.45%	98.51%	99.84%	99.95%