LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Database Systems Group • Prof. Dr. Thomas Seidl

Exercise 4:
Sequential Patterns + Process Mining + kmeans

Knowledge Discovery in Databases I

SS 2016

Ex4.1: Sequential Pattern Mining

Sequence Database
<(ABBA)(BBACA)(CBAA)
(ACA)(BAAB)>

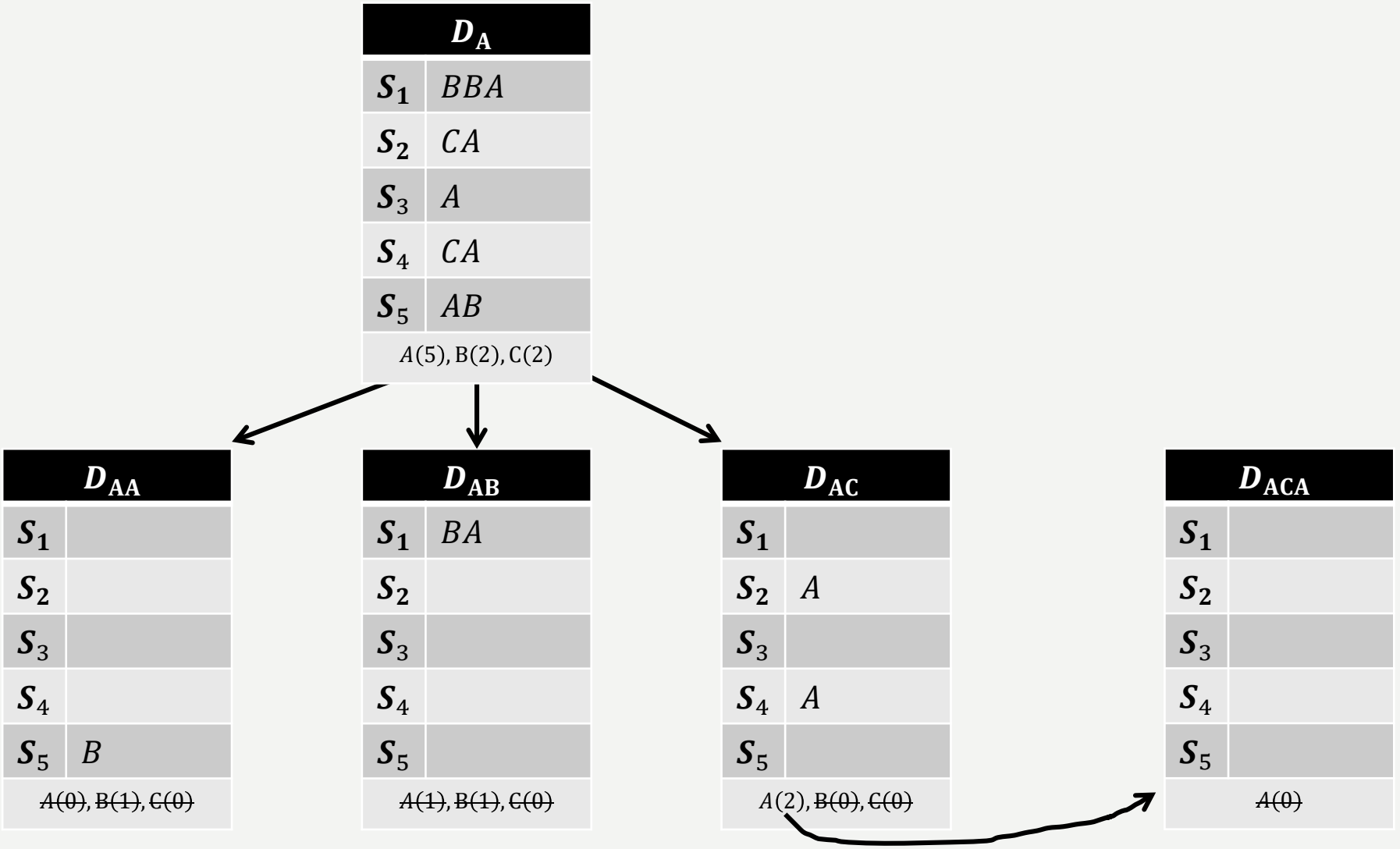
$D_\emptyset$	
S_1	$ABBA$
S_2	$BBACA$
S_3	$CBAA$
S_4	ACA
S_5	$BAAB$
$A(5), B(4), C(3)$	

D_A	
S_1	BBA
S_2	CA
S_3	A
S_4	CA
S_5	AB
$A(5), B(2), C(2)$	

D_B	
S_1	BA
S_2	$BACA$
S_3	AA
S_4	
S_5	AAB
$A(4), B(3), \sout{C(1)}$	

D_C	
S_1	
S_2	A
S_3	BAA
S_4	A
S_5	
$A(3), \sout{B(1)}, \sout{C(0)}$	

Ex4.1: Sequential Pattern Mining



Ex4.1: Sequential Pattern Mining

D_B	
S_1	BA
S_2	$BACA$
S_3	AA
S_4	
S_5	AAB
$A(4), B(3), \sout{C(1)}$	

D_{BAA}	
S_1	
S_2	
S_3	
S_4	
S_5	
$\sout{A(0)}$	

D_{BA}	
S_1	
S_2	A
S_3	A
S_4	
S_5	AB
$A(3), \sout{B(1)}, \sout{C(1)}$	

D_{BB}	
S_1	A
S_2	AA
S_3	
S_4	
S_5	
$A(2), \sout{B(0)}, \sout{C(0)}$	

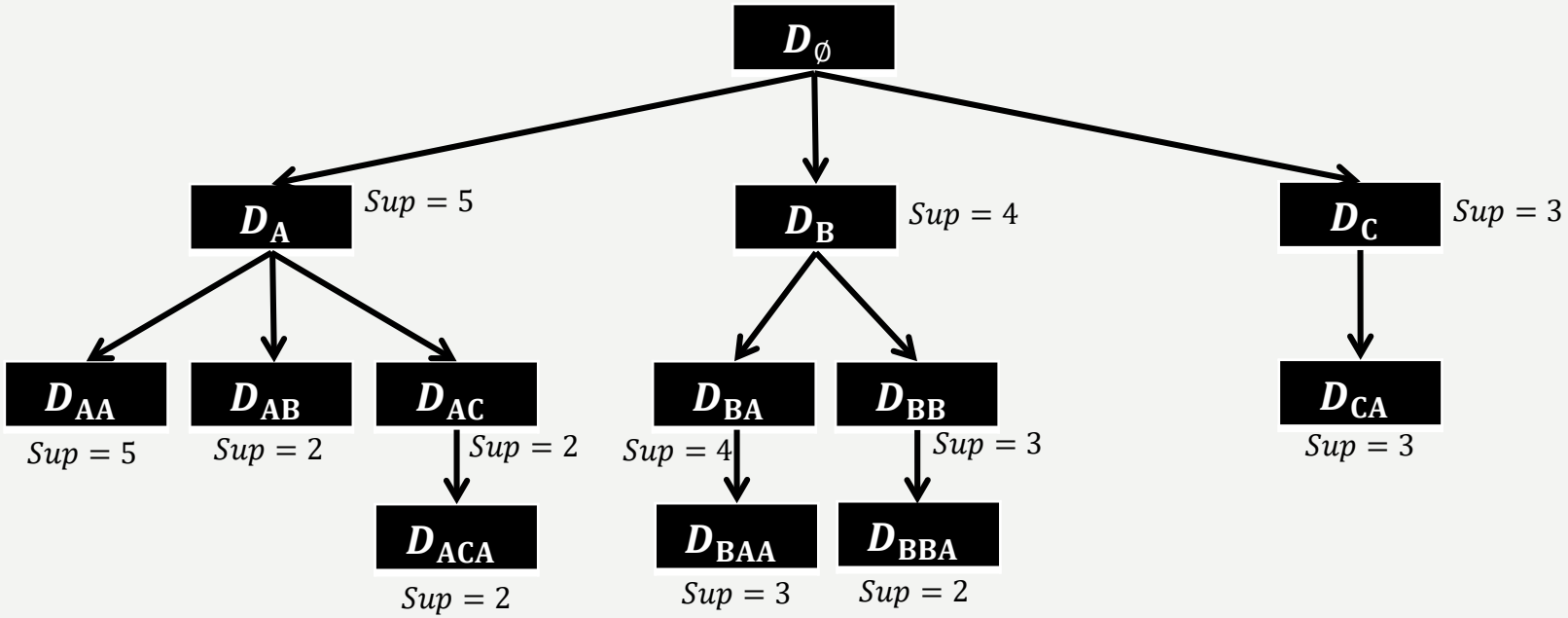
D_{BBA}	
S_1	
S_2	A
S_3	
S_4	
S_5	
$\sout{A(1)}$	

Ex4.1: Sequential Pattern Mining

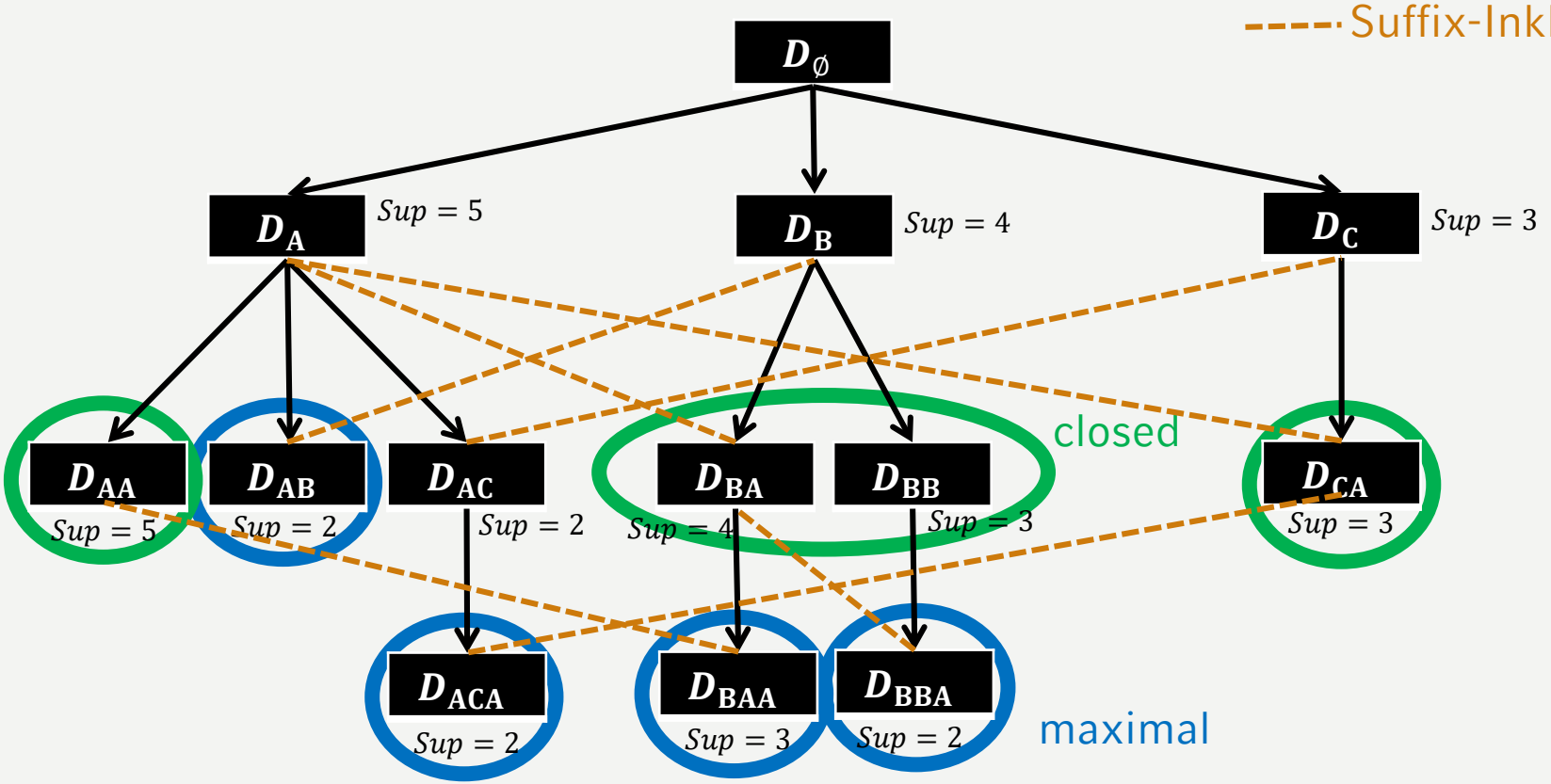
D_C	
S_1	
S_2	A
S_3	BAA
S_4	A
S_5	
$A(3), \sout{B(1)}, \sout{C(0)}$	

D_{CA}	
S_1	
S_2	
S_3	A
S_4	
S_5	
$\sout{A(1)}$	

Ex4.1: Sequential Pattern Mining



Frequent patterns:
$A, AA, AB, AC, ACA, B, BA, BAA, BB, BBA, C, CA$



Frequent patterns:

$A, AA, AB, AC, ACA, B, BA, BAA, BB, BBA, C, CA$

Mining for consecutive patterns

1) Large frequent patterns can start later. Consider all starting points when projecting in the first step:

$$ABAAABCD$$
$$\Rightarrow AAABCD \wedge CD \in D_{AB}$$

2) Do not skip objects. Next one is important for later projections!

D_A	
S_1	BBA
S_2	CA
S_3	A
S_4	CA
S_5	AB
$A(2), B(1), C(2)$	

→

D_{AA}	
S_1	
S_2	
S_3	
S_4	
S_5	B
~~$A(0), B(1), C(0)$~~	

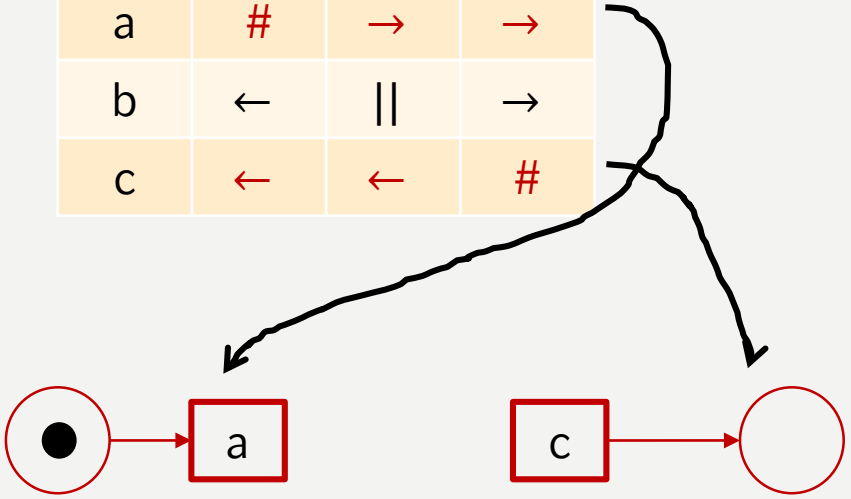
3) Pruning strategy has to be changed

Ex4.2: α-Algorithm

Log $L = \{ac, abc, abbc, abbbc\}$

	a	b	c
a	#	→	→
b	←	||	→
c	←	←	#

Obviously:

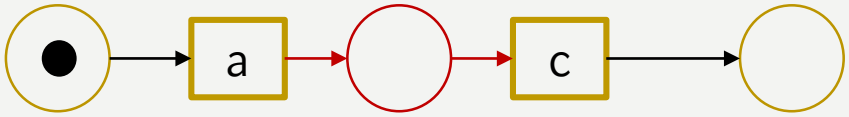


Ex4.2: α-Algorithm

Log $L = \{ac, abc, abbc, abbbc\}$

	a	b	c
a	#	→	→
b	←	||	→
c	←	←	#

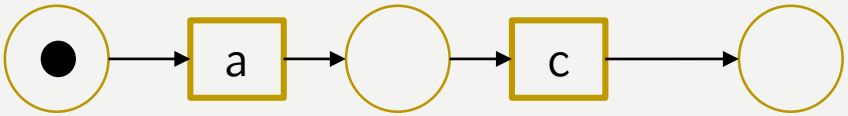
Due to $A = \{a\}, B = \{c\}$



Ex4.2: α-Algorithm

Log $L = \{ac, abc, abbc, abbbc\}$

	a	b	c
a	#	→	→
b	←	||	→
c	←	←	#



Other possibilities for a place p and two sets $A \to p \to B$?

$b \notin A, B$ because $b \# b$

Observation: || on diagonal (= 1-loop) impossible for α–Algorithm

Ex4.3: k-means

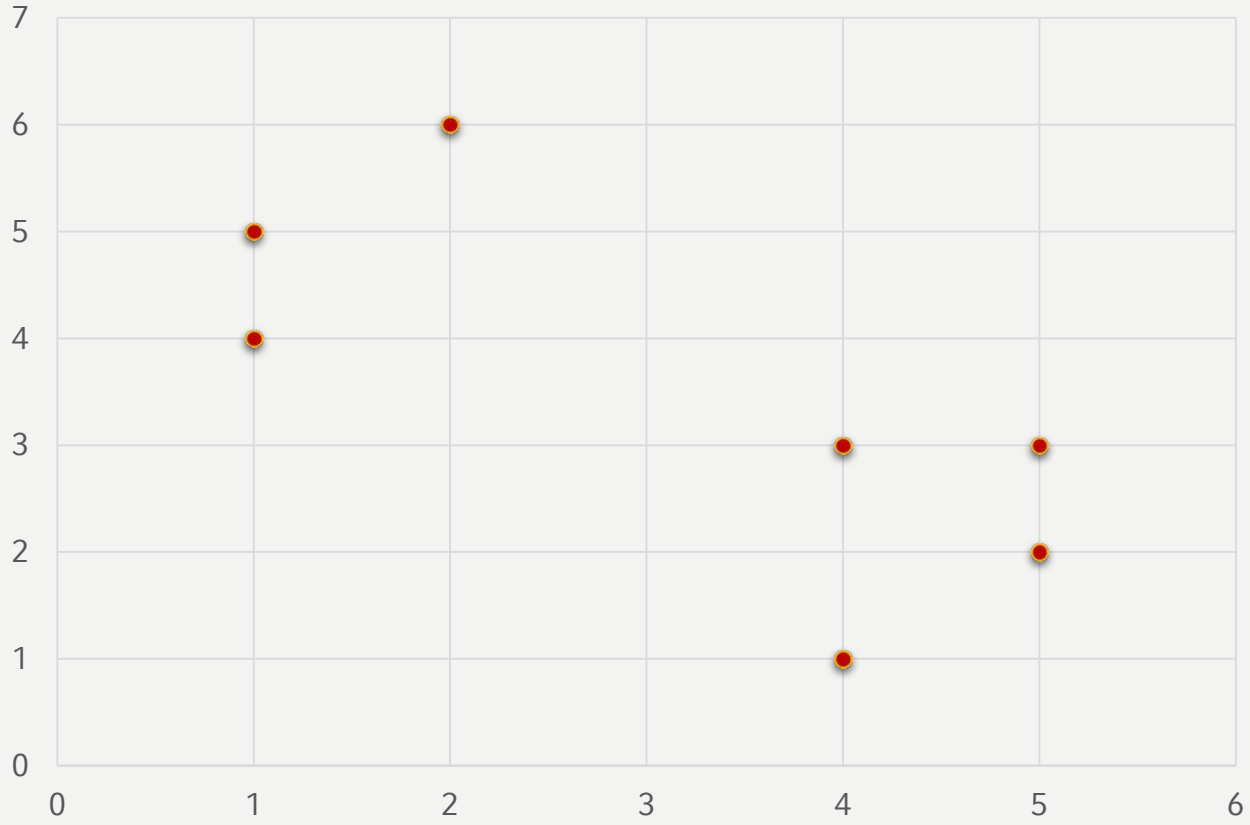
Mapping from sequences to $\mathbb{R}^2$:

„Bag-of-Words-Model“ defines classes (→ vector components) and stores support of these classes in a vector

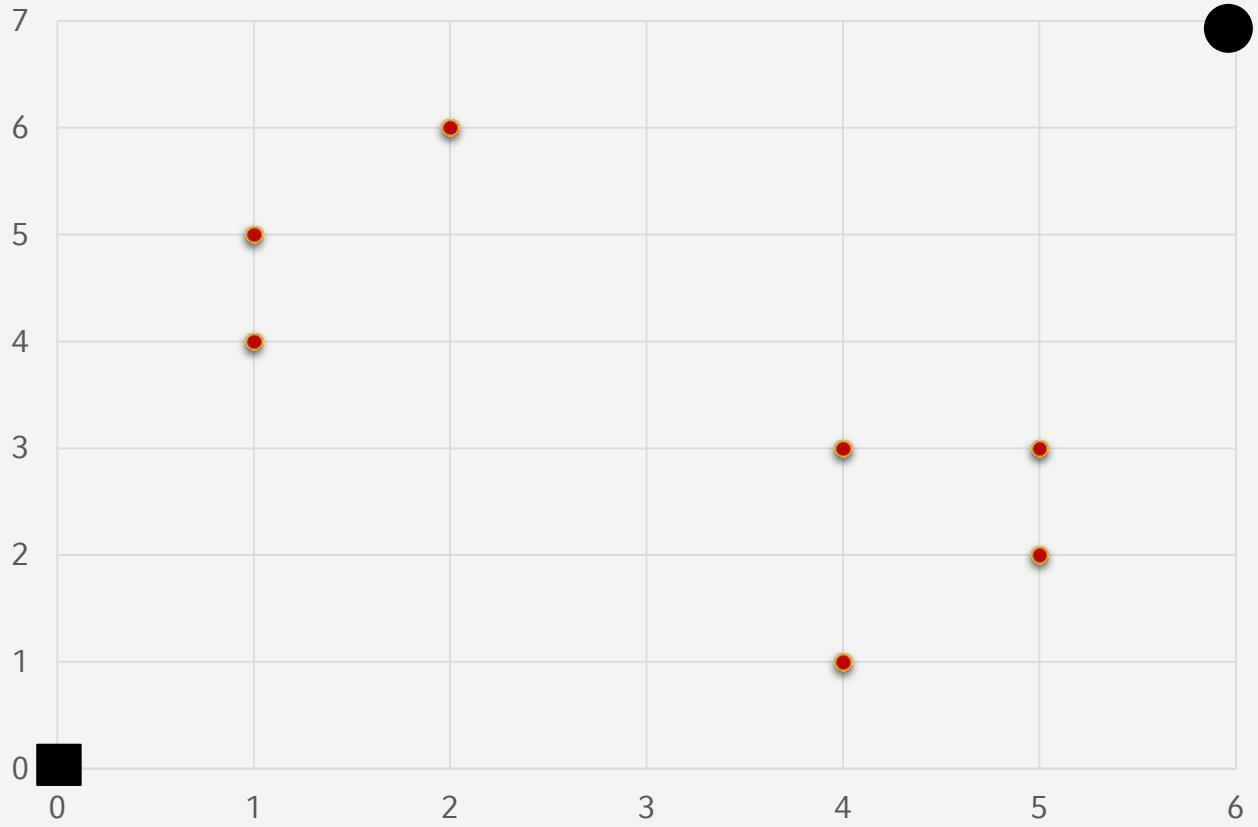
In the sequence case, we lose order information

day	sequence	Vector repr.
monday	bbabb	(1,4)
tuesday	aabaa	(4,1)
wednesday	ababaaa	(5,2)
thursday	baababaa	(5,3)
friday	bbaabaa	(4,3)
saturday	babbbb	(1,5)
sunday	bbbabbba	(2,6)

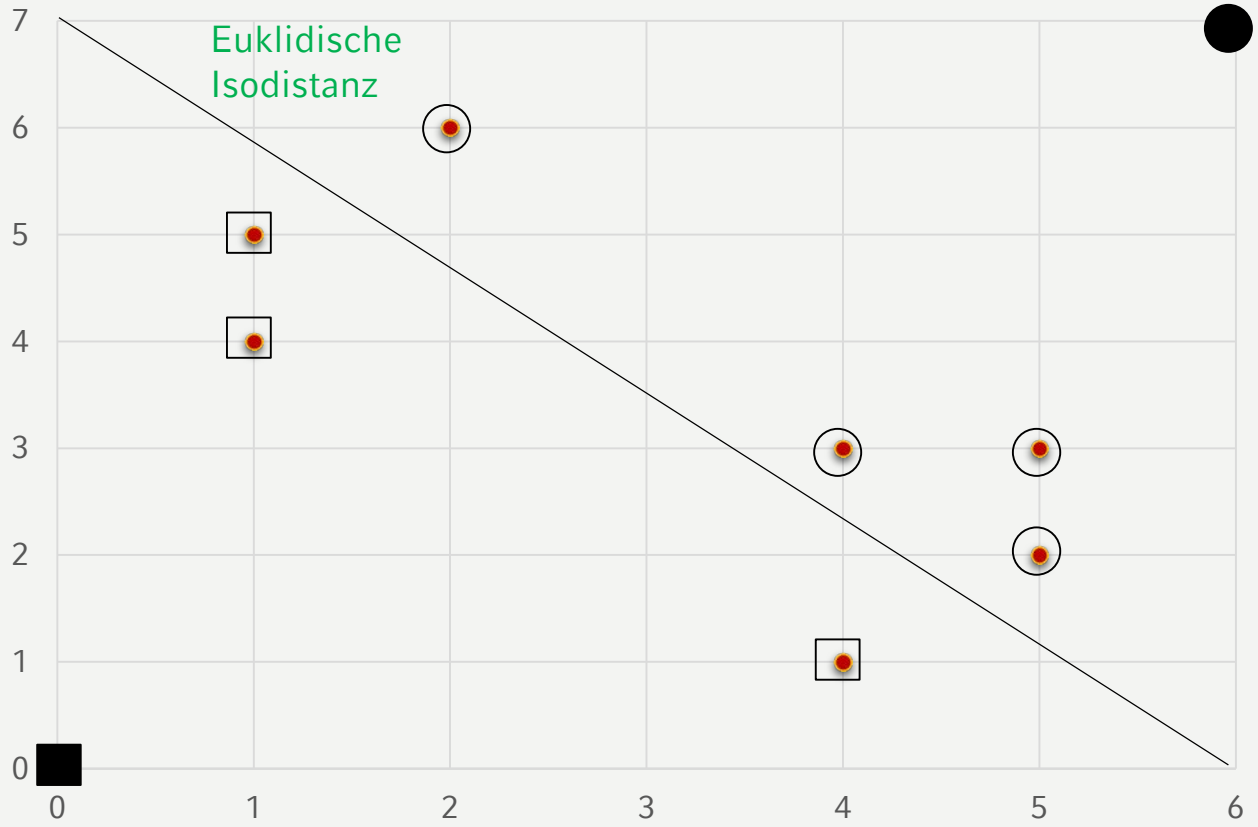
Ex4.3: k-means



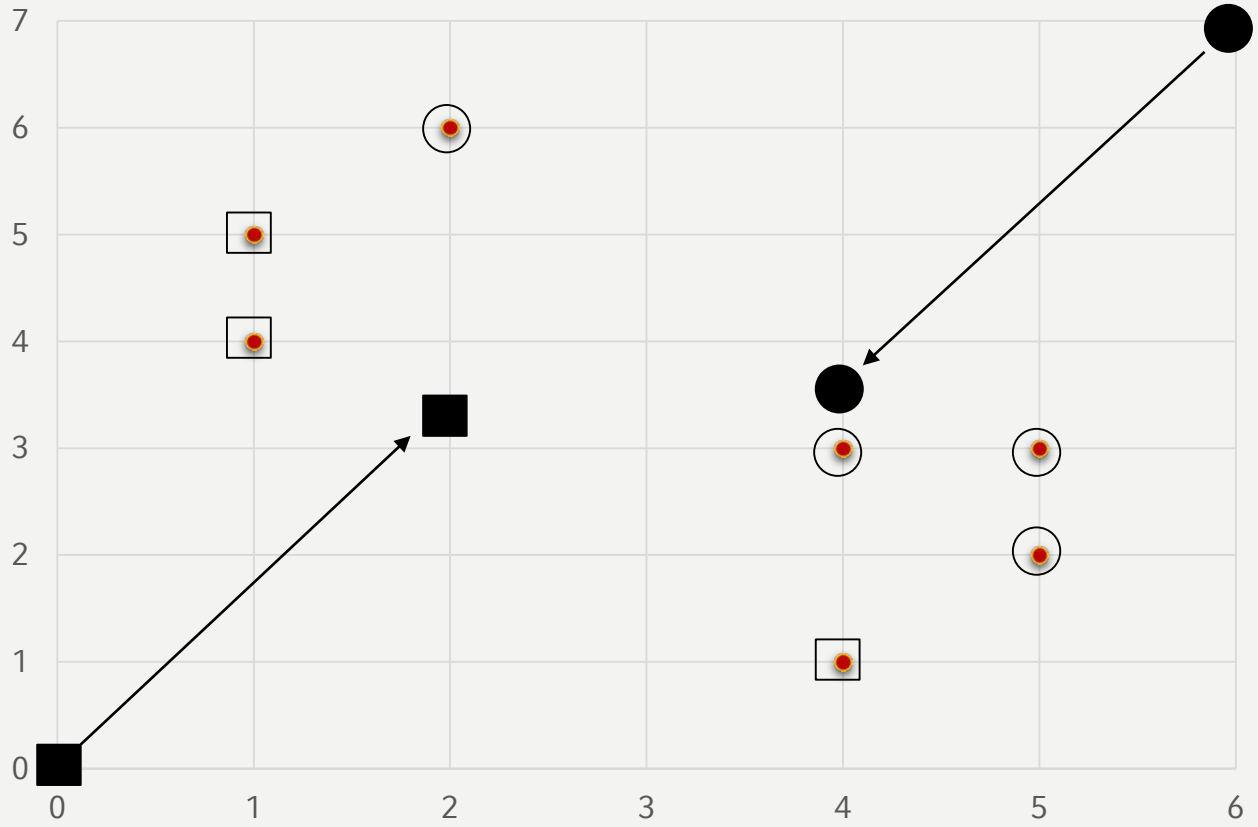
Ex4.3: k-means



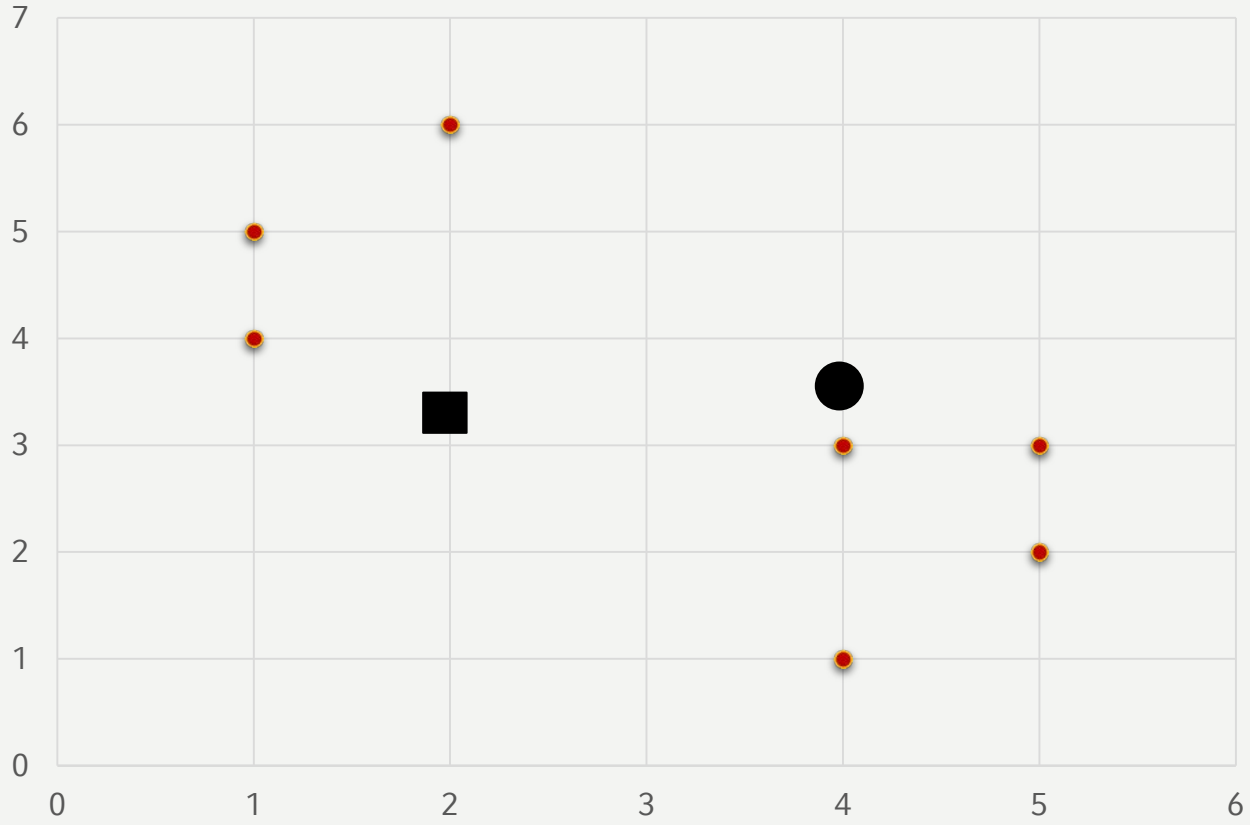
Ex4.3: k-means



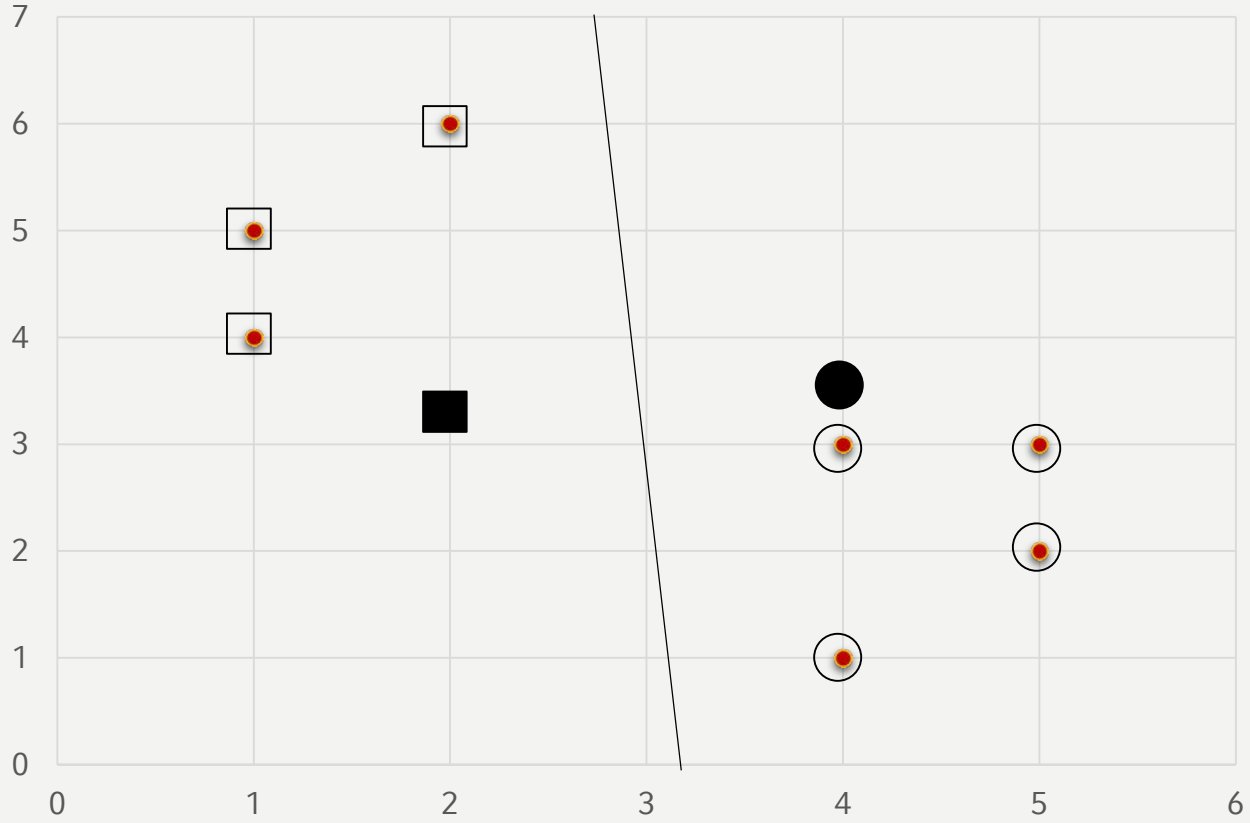
Ex4.3: k-means



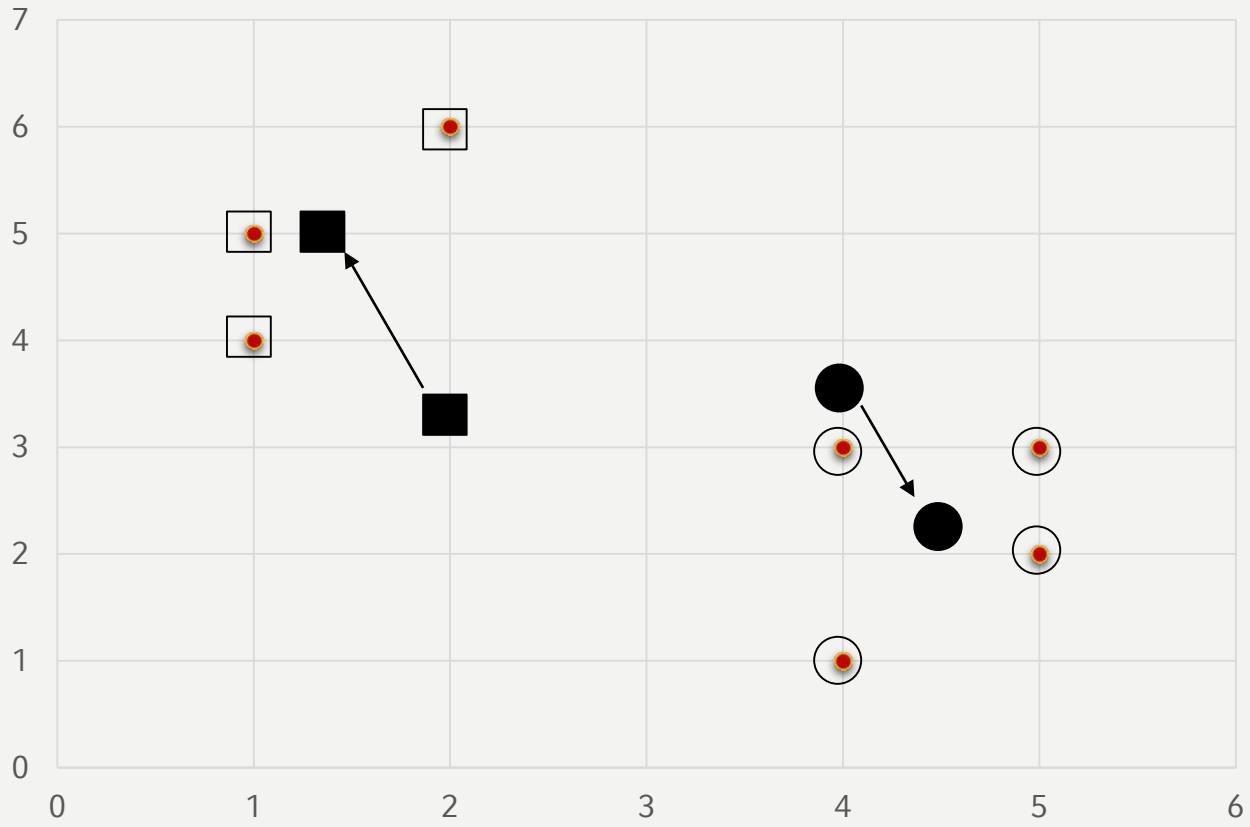
Ex4.3: k-means



Ex4.3: k-means



Ex4.3: k-means



Ex4.3: k-means