

## Uitwerkingen

In de uitgewerkte voorbeelden worden vanwege de leesbaarheid afgeronde tussenresultaten gepresenteerd. De eindresultaten zijn echter altijd berekend zonder tussentijds afronden.

## Hoofdstuk 11

### Antwoord 11.1

Euclidische afstandsmatrix

0,000	88,786	45,978	780,700	779,461	905,252
88,786	0,000	113,759	703,785	706,086	835,818
45,978	113,759	0,000	815,833	816,267	943,640
780,700	703,785	815,833	0,000	62,209	182,151
779,461	706,086	816,267	62,209	0,000	142,411
905,252	835,818	943,640	182,151	142,411	0,000

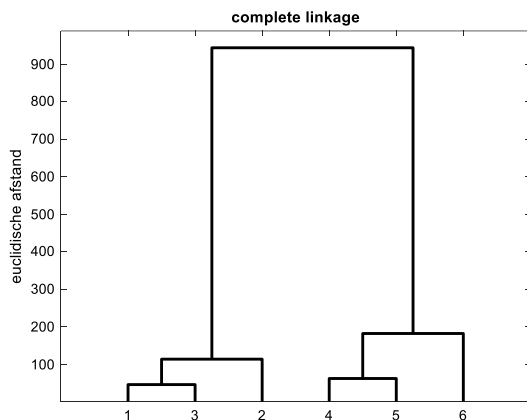
City-block-afstandsmatrix

0	135	70	1171	1153	1336
135	0	165	1066	1048	1231
70	165	0	1231	1213	1396
1171	1066	1231	0	98	221
1153	1048	1213	98	0	183
1336	1231	1396	221	183	0

Similariteitsmatrix

1,000	0,906	0,951	0,173	0,174	0,041
0,906	1,000	0,879	0,254	0,252	0,114
0,951	0,879	1,000	0,135	0,135	0,000
0,173	0,254	0,135	1,000	0,934	0,807
0,174	0,252	0,135	0,934	1,000	0,849
0,041	0,114	0,000	0,807	0,849	1,000

Het dendrogram op basis van de euclidische afstandsmatrix en de complete-linkage-methode:



### Opgave 11.2

Correctie antwoorden: In de afstandsmatrices  $D_3$ ,  $D_4$  en  $D_5$  moeten enkele correcties worden aangebracht die in de volgende tabel geel gemarkeerd zijn.

Voor de Fe-Ti-dataset uit tabel 11.1 is een *hiërarchische agglomeratieve* clustering uitgevoerd op basis van euclidische afstanden (tabel 11.2) met de average-linkage-methode. Hiervoor zijn stapsgewijs voor de gevormde clusters afstandsmatrices berekend die zijn weergegeven in de volgende tabel. De procedure start met de zes afzonderlijke objecten en de bijbehorende euclidische afstandsmatrix  $D_1$ .

De kleinste afstand in de afstandsmatrix  $D_1$  is de afstand tussen de objecten 5 en 6,  $d_{5,6} = 40$ . Deze objecten combineren daarom op een afstand 40 tot een nieuwe cluster (5,6). De afstanden van de objecten 1, 2, 3 en 4 tot de nieuwe cluster (5,6) in de eerste gereduceerde afstandsmatrix  $D_2$  worden bij de average-linkage-methode als volgt berekend:

$$d_{1,(5,6)} = d_{(5,6),1} = \text{gem}(d_{1,5}; d_{1,6}) = (315 + 335)/2 = 325,0$$

$$d_{2,(5,6)} = d_{(5,6),2} = \text{gem}(d_{2,5}; d_{2,6}) = (226 + 230)/2 = 228,0$$

$$d_{3,(5,6)} = d_{(5,6),3} = \text{gem}(d_{3,5}; d_{3,6}) = (258 + 274)/2 = 266,0$$

$$d_{4,(5,6)} = d_{(5,6),4} = \text{gem}(d_{4,5}; d_{4,6}) = (111 + 73)/2 = 92,0$$

De overige afstanden blijven ongewijzigd. Dit resulteert in de nieuwe afstandsmatrix  $D_2$  die een object minder heeft en dus een dimensie heeft van  $5 \times 5$ .

In  $D_2$  is de afstand tussen de objecten 1 en 3 het kleinst,  $d_{1,3} = 71$ . De objecten 1 en 3 combineren daarom op een afstand van 71 tot een nieuwe cluster (1,3). De afstanden van de objecten 2 en 4 en de cluster (5,6) tot de nieuwe cluster (1,3) in de tweede gereduceerde afstandsmatrix  $D_3$  zijn bij de single-linkage-methode:

$$d_{2,(1,3)} = d_{(1,3),2} = \text{gem}(d_{2,1}; d_{2,3}) = (151 + 81)/2 = 116,0$$

$$d_{4,(1,3)} = d_{(1,3),4} = \text{gem}(d_{4,1}; d_{4,3}) = (395 + 330)/2 = 362,5$$

$$d_{(5,6),(1,3)} = d_{(1,3),(5,6)} = \text{gem}(d_{(5,6),1}; d_{(5,6),3}) = (325 + 266)/2 = 295,5$$

De overige afstanden blijven ongewijzigd. Dit resulteert in de nieuwe gereduceerde  $4 \times 4$ -afstandsmatrix  $D_3$ .

In  $D_3$  is de afstand tussen object 4 en cluster (5,6) het kleinst,  $d_{4,(5,6)} = 92$ . Deze combineren daarom in het dendrogram op een afstand van 92 tot een nieuwe cluster (4,(5,6)). De afstanden van object 2 en de cluster (1,3) tot de nieuwe cluster (4,(5,6)) in de derde gereduceerde afstandsmatrix  $D_4$  zijn:

$$d_{2,(4,(5,6))} = d_{(4,(5,6)),2} = \text{gem}(d_{2,4}; d_{2,(5,6)}) = (274 + 228)/2 = 251,0$$

$$d_{(1,3),(4,(5,6))} = d_{(4,(5,6)),(1,3)} = \text{gem}(d_{(1,3),4}; d_{(1,3),(5,6)}) = (362,5 + 295,5)/2 = 329,0$$

De afstand  $d_{2,(1,3)}$  blijft ongewijzigd. Dit resulteert in de nieuwe gereduceerde  $3 \times 3$ -afstandsmatrix  $D_4$ .

In  $D_4$  is de afstand tussen object 2 en cluster (1,3) het kleinst  $d_{2,(1,3)} = 116$ . Deze combineren in het dendrogram op een afstand van 116 tot een nieuwe cluster (2,(1,3)). De afstand tussen de clusters (2,(1,3)) en (4,(5,6)) in de vierde gereduceerde afstandsmatrix  $D_5$  is:

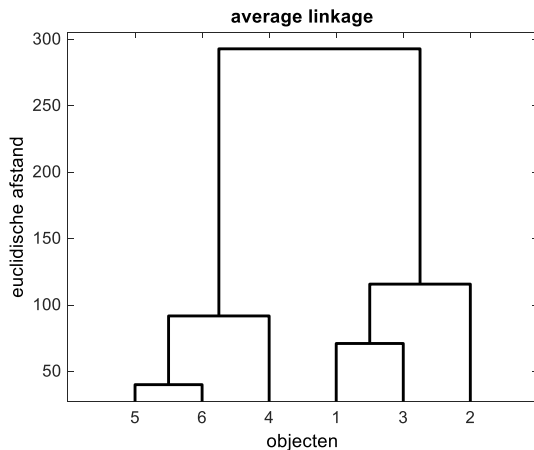
$$d_{(2,(1,3)),(4,(5,6))} = d_{(4,(5,6)),(2,(1,3))} = \text{gem}(d_{2,(4,(5,6))}; d_{(1,3),(4,(5,6))}) = (251,0 + 329,0)/2 = 290,0$$

Dit resulteert in de nieuwe gereduceerde  $2 \times 2$ -afstandsmatrix  $D_5$ . De laatste twee clusters (2,(1,3)) en (4,(5,6)) combineren op een afstand van 290,0 tot één cluster.

Datamatrices voor clustering van de Fe-Ti-dataset volgens de average-linkage-methode:

<b>D<sub>1</sub></b>	1	2	3	4	5	6
1	0	151	71	395	315	335
2	151	0	81	274	226	230
3	71	81	0	330	258	274
4	395	274	330	0	111	73
5	315	226	258	111	0	40
6	335	230	274	73	40	0
<b>D<sub>2</sub></b>	1	2	3	4	(5,6)	
1	0	151	71	395	325,0	
2	151	0	81	274	228,0	
3	71	81	0	330	266,0	
4	395	274	330	0	92,0	
(5,6)	325,0	228,0	266,0	92,0	0	
<b>D<sub>3</sub></b>	(1,3)	2	4	(5,6)		
(1,3)	0,0	116,0	362,5	295,5		
2	116,0	0,0	274,0	228,0		
4	362,5	274,0	0,0	92,0		
(5,6)	295,5	228,0	92,0	0,0		
<b>D<sub>4</sub></b>	(1,3)	2	(4,(5,6))			
(1,3)	0,0	116,0	329,0			
2	116,0	0,0	251,0			
(4,(5,6))	329,0	251,0	0,0			
<b>D<sub>5</sub></b>	(2,(1,3))	(4,(5,6))				
(2,(1,3))	0,0	290,0				
(4,(5,6))	290,0	0,0				

De hiërarchische structuur van de gevormde clusters is afgebeeld in het volgende *dendrogram* of *boomdiagram*



De clustering van de objecten komt overeen met die in afbeelding 11.4 waarbij de objecten zijn geclusterd volgens de single-linkage-methode. Alleen de niveaus waarop de objecten clusteren verschillen.