Mining Massive Datasets - Problem Set 2

Exercise 01

A = (0,0), B = (10,10), C = (21,21), D = (33,33), E = (5,27), F = (28,6)

1.1 Use centroid distance

	A	B	C	D	E	F
A	0.0	14.14	29.7	46.67	27.5	28.6
B		0.0	15.6	32.5	17.7	18.4
C			0.0	17.0	17.1	16.6
D				0.0	28.6	27.5
E					0.0	31.1
F						0.0

merge A and B

→ AB centroid (5,5)

	AB	C	D	E	F
AB	0.0	22.6	29.6	22.0	23.0
C		0.0	17.0	17.1	16.6
D			0.0	28.6	27.5
E				0.0	31.1
F					0.0

merge C and F

→ CF centroid (24.5, 13.5)

	AB	CF	D	E
AB	0.0	21.3	31.6	27.0
CF		0.0	21.3	23.7
D			0.0	28.6
E				0.0

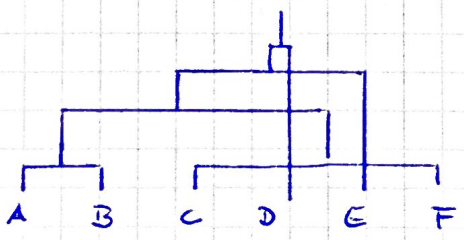
merge AB and CF

→ ABCF centroid (14.8, 9.3)

	ABCF	D	E
ABCF	0.0	29.9	20.2
D		0.0	28.6
E			0.0

merge ABCF and E

->



1.2. Single Linkage

- Schritt 1 s.o.

	AB	C	D	E	F
AB	0.0	14.1	32.5	17.7	18.4
C		0.0	17.0	17.1	16.6
D			0.0	28.6	27.5
E				0.0	31.1
F					0.0

Anm: als distance btw. AB and C choose minimum of d(AC) and d(BC)

usw.

merge AB and C

- Minimum in matrix

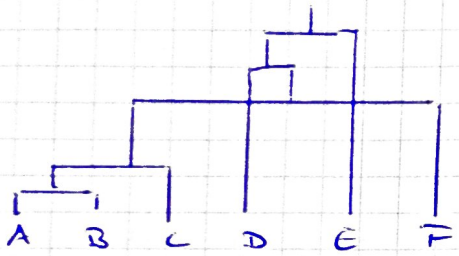
	ABC	D	E	F
ABC	0.0	17.0	31.1	16.6
D		0.0	28.6	17.5
E			0.0	31.1
F				0.0

merge ABC and F

	ABCF	D	E
ABCF	0.0	17.0	31.1
D		0.0	28.6
E			0.0

merge ABCF and D

→



1.3 Complete Linkage

- Schritt 1 s.o.

	AB	C	D	E	F
AB	0.0	29.7	46.7	27.5	28.6
C		0.0	42.0	31.1	16.6
D			0.0	48.6	27.5
E				0.0	31.1
F					0.0

Anm: als distance btw. AB and C choose maximum of d(AC) and d(BC)

merge C and F

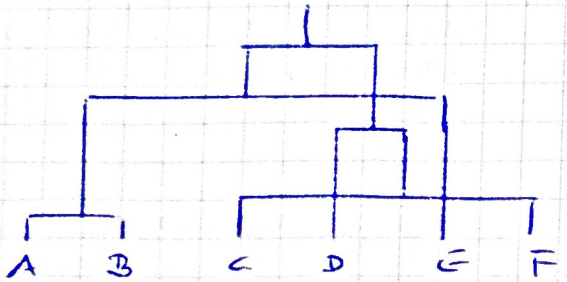
	AB	CF	D	E
AB	0.0	29.7	46.7	27.5
CF		0.0	17.0	31.1
D			0.0	31.1
E				0.0

merge CF and D

	AB	CFD	E
AB	0.0	46.7	27.5
CFD		0.0	31.1
E			0.0

merge AB and E

→



1.4 Average Linkage

- Schritt 1 s.o.

	AB	C	D	E	F
AB	0.0	22.7	39.6	22.6	23.5
C		0.0	17.0	17.1	16.6
D			0.0	28.6	27.5
E				0.0	31.1
F					0.0

Anm: as distance btw. AB and C choose avg(d(AC), d(BC))

merge C and F

	AB	CF	D	E
AB	0.0	22.1	39.6	22.6
CF		0.0	27.3	24.1
D			0.0	31.1
E				0.0

merge CF and D

	AB	CFD	E
AB	0.0	28.6	22.6
CFD		0.0	25.6
E			0.0

merge AB and E

-> dendrogram identical to the one from 1.3.