8. Sailkapen gainbegiratu gabea: Clustering
Ariketak

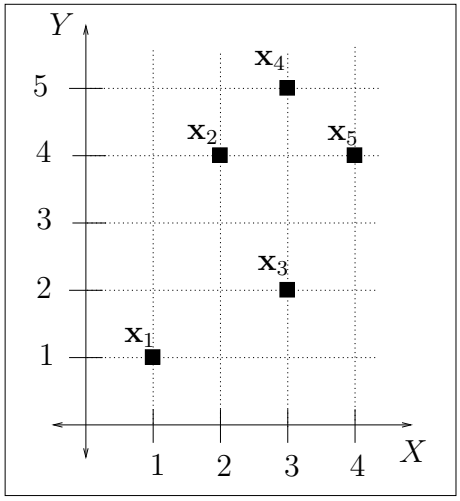
1. Demagun bi dimentsioko 5 kasuz osatutako datubasea dugula. Kasuak ez daude etiketatuta, eta beren arteko distantzia Euklidestarrak ezagunak dira.

Kasua	X	Y
$\mathbf{x}_1$	1	1
$\mathbf{x}_2$	2	4
$\mathbf{x}_3$	3	2
$\mathbf{x}_4$	3	5
$\mathbf{x}_5$	4	4

Datubasea

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{x}_1$	0	3.16	2.24	4.47	4.24
$\mathbf{x}_2$		0	2.24	1.41	2.00
$\mathbf{x}_3$			0	3.00	2.24
$\mathbf{x}_4$				0	1.41
$\mathbf{x}_5$					0

Distantzien matrizea



Erabil ezazu k**-means partiziozko clustering algoritmoa** honako kasuetan:

1.1 Cluster kopurua $k = 2$ suposatuz. Hasierako haziak: $\mathbf{x}_1 \in Cl_1$ eta $\mathbf{x}_2 \in Cl_2$

1.2 Cluster kopurua $k = 2$ suposatuz. Hasierako haziak: $\mathbf{x}_4 \in Cl_1$ eta $\mathbf{x}_5 \in Cl_2$

1.3 $k = 3$ suposatuz. Hasierako haziak: $\mathbf{x}_2 \in Cl_1$, $\mathbf{x}_4 \in Cl_2$ eta $\mathbf{x}_5 \in Cl_3$

Erabil ezazu **behetik gorako clustering algoritmo hierarkikoa**, cluster arteko lotura hauetarako:

1.4 Lotura distantzia minimoarekin ("single linkage")

1.5 Lotura distantzia maximoarekin ("complete linkage")

Demagun clustering-erako bi algoritmo erabiliz honako bi partizioak lortu direla:

$$P_1 = \{Cl_1, Cl_2\} = \{\{\mathbf{x}_1\}, \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}\}$$

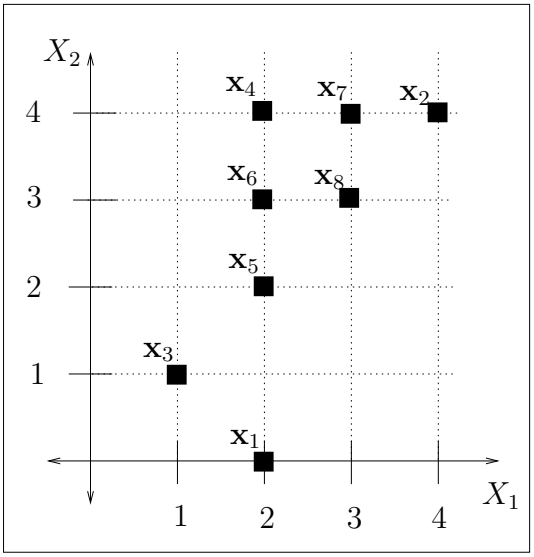
$$P_2 = \{Cl_1, Cl_2\} = \{\{\mathbf{x}_1, \mathbf{x}_3\}, \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5\}\}$$

1.6 Partizio bakoitzerako **Clusterren kohesioa** kalkula ezazu SSE (Sum of Squared Error) neurria erabiliz. Zeinek du kohesiorik altuena?

2. Demagun bi dimentsioko 8 kasuz osatutako datubasea dugula. Kasuak ez daude etiketatuta, eta beren arteko distantzia Euklidestarrak ezagunak dira.

Kasua	X_1	X_2
$\mathbf{x}_1$	2	0
$\mathbf{x}_2$	4	4
$\mathbf{x}_3$	1	1
$\mathbf{x}_4$	2	4
$\mathbf{x}_5$	2	2
$\mathbf{x}_6$	2	3
$\mathbf{x}_7$	3	4
$\mathbf{x}_8$	3	3

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$
$\mathbf{x}_1$	0	4.47	1.41	4.00	2.00	3.00	4.12	3.16
$\mathbf{x}_2$		0	4.24	2.00	2.83	2.24	1.00	1.41
$\mathbf{x}_3$			0	3.16	1.41	2.24	3.61	2.83
$\mathbf{x}_4$				0	2.00	1.00	1.00	1.41
$\mathbf{x}_5$					0	1.00	2.24	1.41
$\mathbf{x}_6$						0	1.41	1.00
$\mathbf{x}_7$							0	1.00
$\mathbf{x}_8$								0



Erabil ezazu k**-means partiziozko clustering algoritmoa** honako kasuetan:

2.1 Cluster kopurua $k = 2$ suposatuz. Hasierako haziak: $\mathbf{x}_1 \in Cl_1$ eta $\mathbf{x}_2 \in Cl_2$

2.2 Cluster kopurua $k = 2$ suposatuz. Hasierako haziak: $\mathbf{x}_2 \in Cl_1$ eta $\mathbf{x}_7 \in Cl_2$

2.3 $k = 3$ suposatuz. Hasierako haziak: $\mathbf{x}_2 \in Cl_1$, $\mathbf{x}_4 \in Cl_2$ eta $\mathbf{x}_7 \in Cl_3$

Erabil ezazu **behetik gorako clustering algoritmo hierarkikoa**, cluster arteko lotura hauetarako:

2.4 Lotura distantzia minimoarekin ("single linkage")

2.5 Lotura distantzia maximoarekin ("complete linkage")

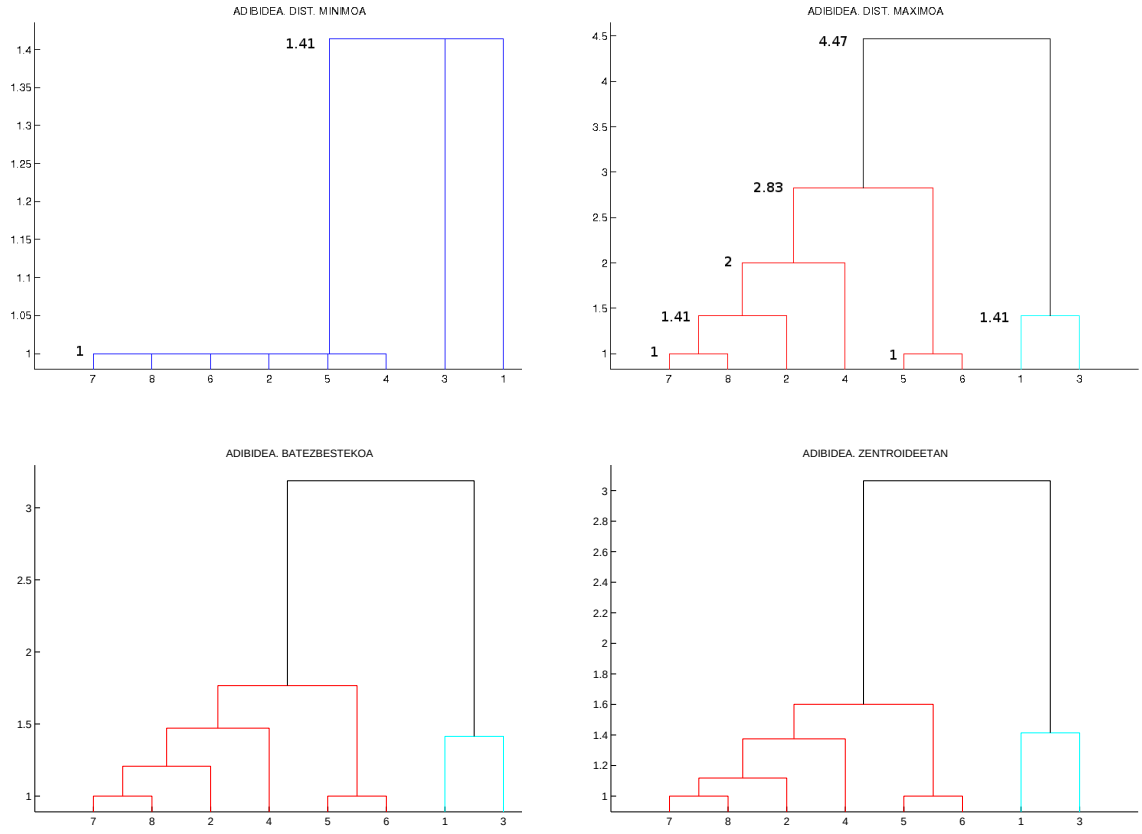
Demagun clustering-erako bi algoritmo erabiliz honako bi partizioak lortu direla:

$$P_1 = \{Cl_1, Cl_2\} = \{\{\mathbf{x}_1, \mathbf{x}_3\}, \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8\}\}$$

$$P_2 = \{Cl_1, Cl_2\} = \{\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_5\}, \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8\}\}$$

2.6 Partizio bakoitzerako **Clusterren kohesioa** kalkula ezazu SSE (Sum of Squared Error) neurria erabiliz. Zeinek du kohesiorik altuena?

3. Izan bedi 8 kasuz osatutako $D = \{1, 2, 3, 4, 5, 6, 7, 8\}$ datubasea. Indizedun clustering algoritmo hierarkiko bat aplikatu dugu, cluster arteko lotura desberdinetarako: distantzia minimoarekin (single linkage), distantzia maximoarekin (complete linkage), batazbesteko distantziarekin (average linkage), distantzia zentroidean (centroid linkage). Hauek dira lortutako lau dendrogramak:



Dendrograma bakoitzetik abiatuz,

3.1 Proposa ezazu, arrazoituz, datubasearen partiketa bat.

3.2 Cluster arteko loturak distantzia minimoa eta maximoa erabiliz lortutako bi dendrogrametarako (lehenengo biak) idatz ezazu dagokien ultrametriken matrizea.

Oharra: Lau dendrogramak 2. ariketako datubaseari dagozkio. Ikus bertan kasuen arteko distantzia euklidestarrak, eta konparatu dendrogrametatik lortutako ultrametriken matrizeekin.