

Gainbegiratu Gabeko Sailkapena: Clustering

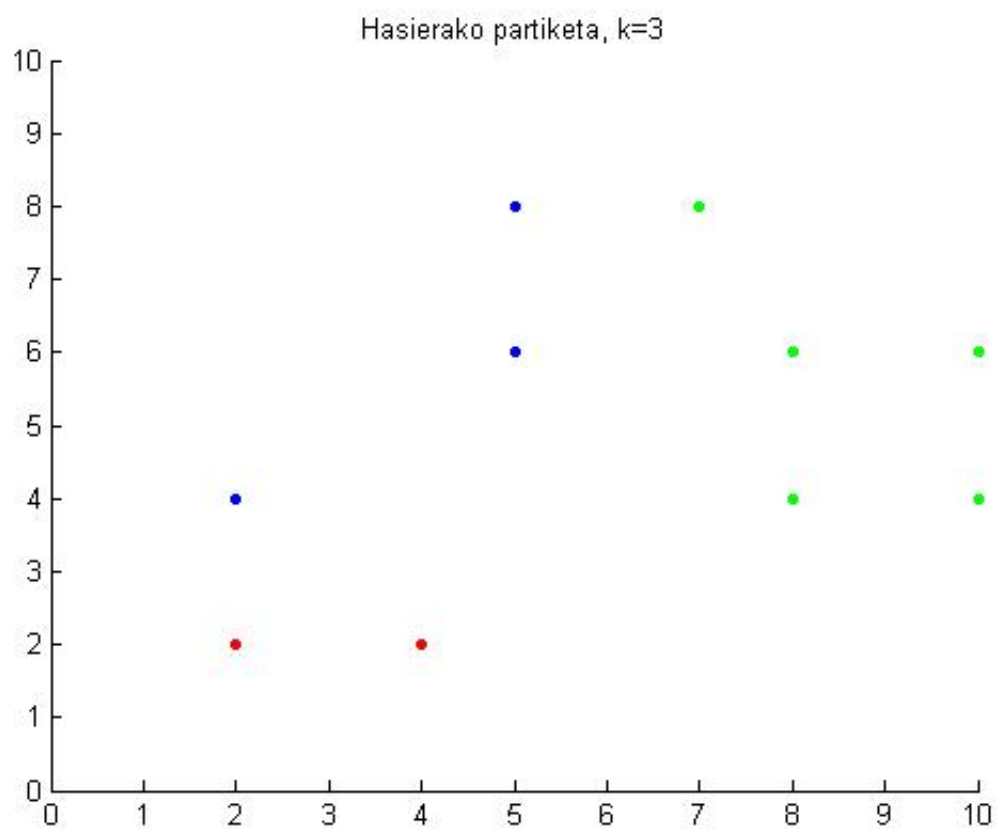
2011-2012 ikasurtea

1 Forgy-ren metodoa.

Datu base honekin egingo dugu:

Kasua	X	Y
1	2	2
2	2	4
3	4	2
4	5	6
5	5	8
6	7	8
7	8	4
8	8	6
9	10	4
10	10	6

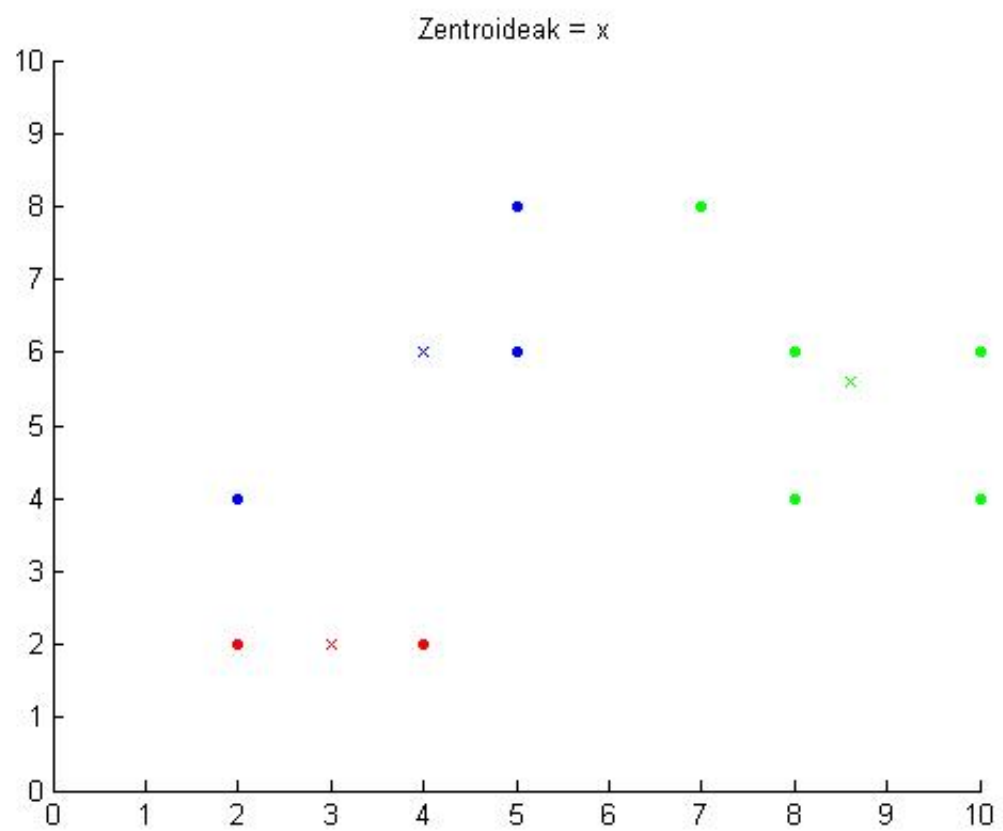
Taula 1: Datu basea



Irudia 1: Datu basearen grafikoa

Lehenengo pausuan aukeratu dugu partiketa hau:

$P_1 = \{1, 3\}$, (gorriak) $P_2 = \{2, 4, 5\}$ (urdinak) eta $P_3 = \{6, 7, 8, 9, 10\}$ (berdeak)



Irudia 2: Datuak eta zentroideak

Orain konglomeratu bakoitzaren zentroidea kalkulatu dugu:
 $Z_1 = [3, 2]$, $Z_2 = [4, 6]$, $Z_3 = [8.6, 5.6]$
eta grafikoan ikusiko ditugu.

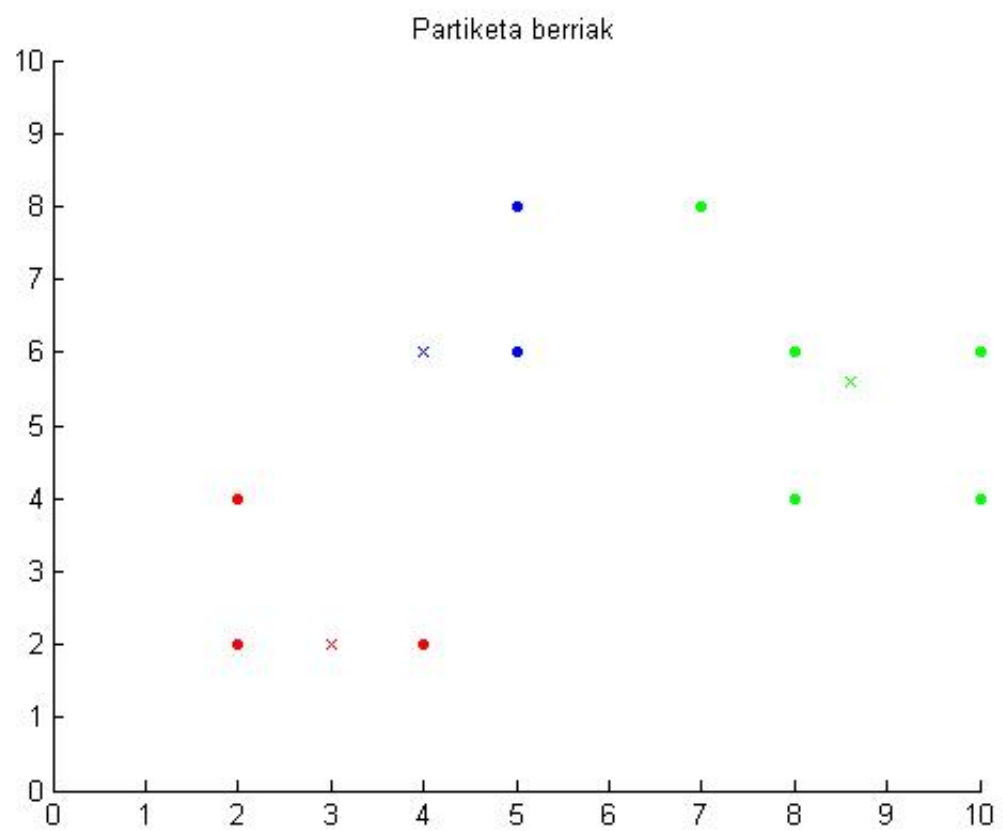
	1	2	3	4	5	6	7	8	9	10
Z_1	1.00	2.24	1.00	4.47	6.32	7.21	5.39	6.40	7.28	8.06
Z_2	4.47	2.83	4.00	1.00	2.24	3.61	4.47	4.00	6.32	6.00
Z_3	7.52	6.79	5.84	3.62	4.33	2.88	1.71	0.72	2.13	1.46

Taula 2: Distantzia eukldestarrak zentroidekiko

Hurrengo pausuan kalkulatu ditugu puntu guztien distantziak zentroidekiko:

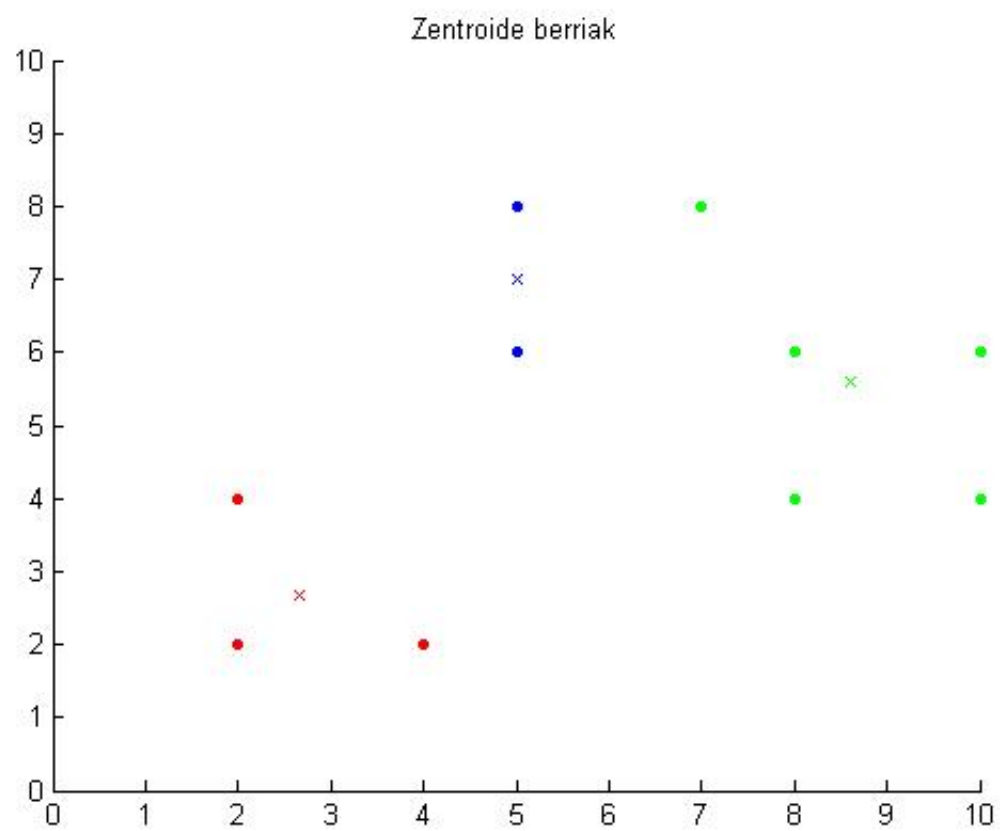
Konglomeratu berriak:

$P_1 = \{1, 2, 3\}$, (gorriak) $P_2 = \{4, 5\}$ (urdinak) eta $P_3 = \{6, 7, 8, 9, 10\}$ (berdeak)



Irudia 3: Partiketa berriak

Irudian ikusten dira partiketa berriak.



Irudia 4: Partiketa eta zentroide berriak

Orain konglomeratu bakoitzaren zentroidea kalkulatu dugu:
 $Z_1 = [2.67, 2.67]$, $Z_2 = [5.0, 7.0]$, $Z_3 = [8.6, 5.6]$

Berririo kalkulatuako ditugu elementu guztien distantziak zentroidekiko.

	1	2	3	4	5	6	7	8	9	10
Z_1	0.94	1.49	1.49	4.07	5.82	6.87	5.50	6.29	7.45	8.06
Z_2	5.83	4.24	5.10	1.00	1.00	2.24	4.24	3.16	5.83	5.10
Z_3	7.52	6.79	5.84	3.62	4.33	2.88	1.71	0.72	2.13	1.46

Taula 3: Distantzia euklidentarrak zentroidekiko

Konglomeratuak ez dira aldatzen:

$P_1 = \{1, 2, 3\}$, (gorriak) $P_2 = \{4, 5\}$ (urdinak) eta $P_3 = \{6, 7, 8, 9, 10\}$ (berdeak)

	1	2	3	4	5	6	7	8	9	10
1	0	2.00	2.00	5.00	6.71	7.81	6.32	7.21	8.25	8.94
2	2.00	0	2.83	3.61	5.00	6.40	6.00	6.32	8.00	8.25
3	2.00	2.83	0	4.12	6.08	6.71	4.47	5.66	6.32	7.21
4	5.00	3.61	4.12	0	2.00	2.83	3.61	3.00	5.39	5.00
5	6.71	5.00	6.08	2.00	0	2.00	5.00	3.61	6.40	5.39
6	7.81	6.40	6.71	2.83	2.00	0	4.12	2.24	5.00	3.61
7	6.32	6.00	4.47	3.61	5.00	4.12	0	2.00	2.00	2.83
8	7.21	6.32	5.66	3.00	3.61	2.24	2.00	0	2.83	2.00
9	8.25	8.00	6.32	5.39	6.40	5.00	2.00	2.83	0	2.00
10	8.94	8.25	7.21	5.00	5.39	3.61	2.83	2.00	2.00	0

Taula 4: Distantzia matrize

2 McQueen-en k batezbesteko metodoa.

Datu baseko lehenengo k elementuekin hiru konglomeratu egingo dugu.

$P_1 = \{1\}$, (gorriak) $P_2 = \{2\}$ (urdinak) eta $P_3 = \{3\}$ (berdeak)

Fitxategian dauden ordenean, gelditzen diren 7 elementuak esleitu zentroide hurbilen duten konglomeratuari.

$P_1 = \{1\}$, (gorriak) $P_2 = \{2, 4, 5, 6\}$ (urdinak) eta $P_3 = \{3, 7, 8, 9, 10\}$ (berdeak)

Zentroideak birkalkulatuta:

$Z_1 = [2, 2]$, $Z_2 = [2.67, 2.67]$, $Z_3 = [6.25, 6.50]$

3 Clustering hierarkikoa

Bi mota dadude:

Clustering gorakor hierarkikoa

Clustering beherakor hierarkikoa

3.1 Clustering gorakor hierarkikoa.

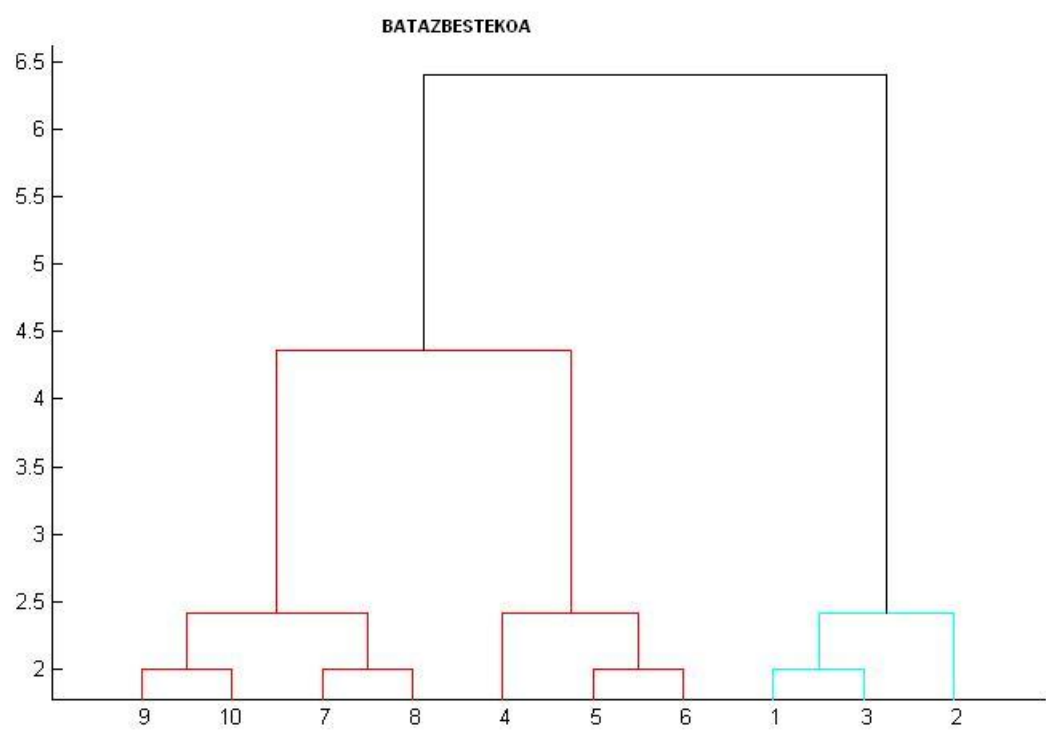
1. Kalkulatu distantzia matrizea.
 2. Elementu bakoitza cluster bat da.
Repeat
 - 3 Bildu hurbilen dauden bi clusterrak.
 4. Eguneratu distantzia matrizea.
- Errepikatu 3. eta 4. pausuak cluster bat bakarrik geratu arte.

3.2 Nola elkartu bi cluster?

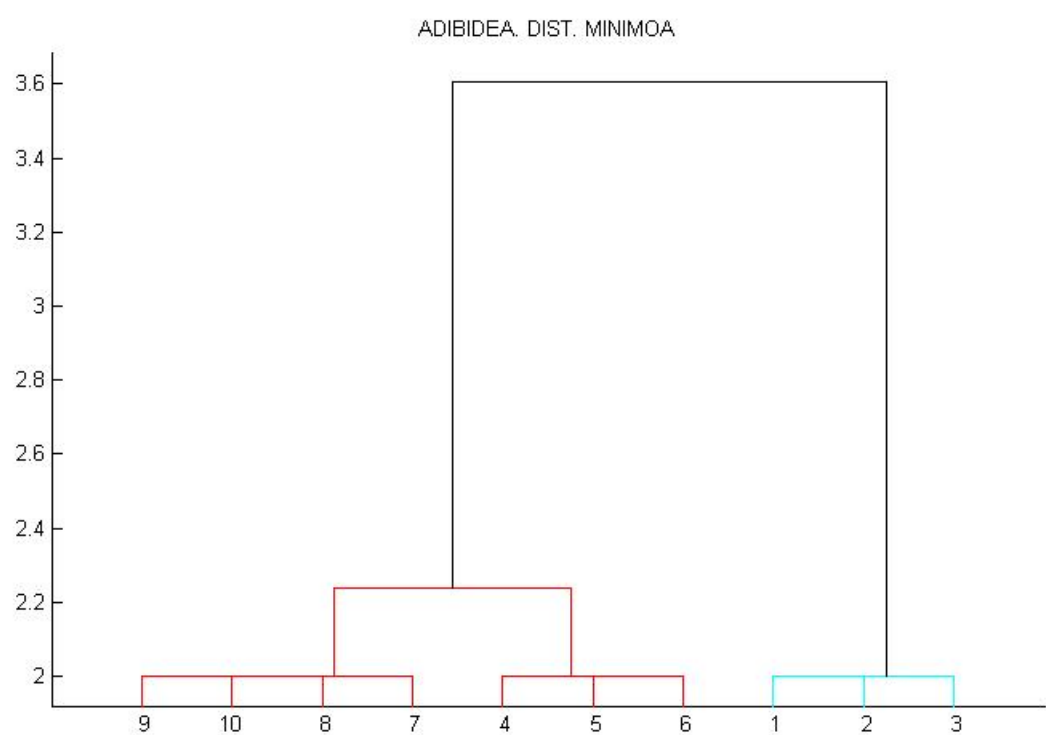
3.2.1 Konglomeratuen arteko batazbesteko lotura.(AVERAGE)

3.2.2 Lotura erraza. (MIN)

Pausu bakoitzean bi puntu hurbilen duten bi konglomeratuak elkartzen dira



Irudia 5: Batazbestekoa



Irudia 6: Lotura erraza