

6. Sailkatzaileen Konbinaketa Ariketak

- Demagun bi dimentsioko 10 kasuz osatutako datubasea dugula, eta $\mathbf{x}_{11}$ kasu berria sailkatu nahi dugula. Demagun X eta Y aldagaien balioak dagoeneko normalizatuak izan direla, eta kasuen arteko distantzia Euklidentarrak kalkulatu ditugula.

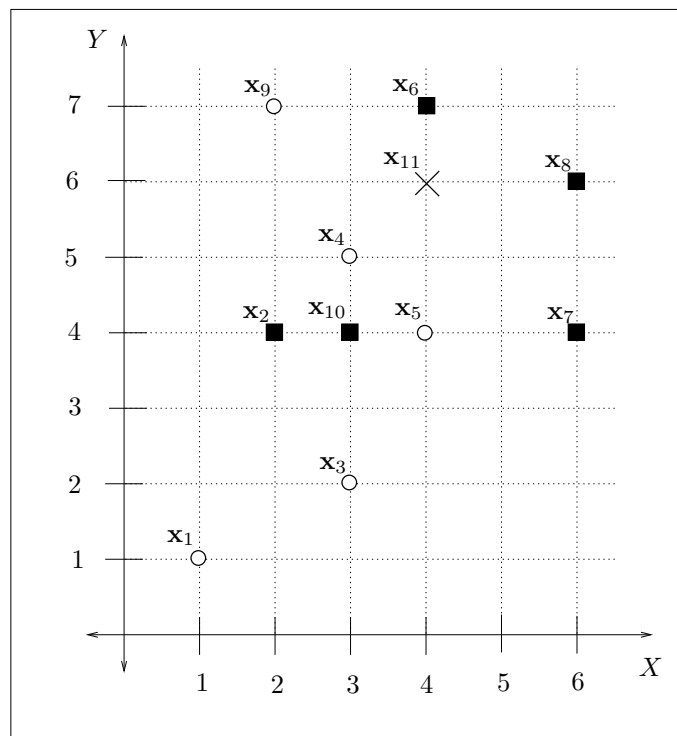
Kasua	X	Y	Klasea
$\mathbf{x}_1$	1	1	0
$\mathbf{x}_2$	2	4	1
$\mathbf{x}_3$	3	2	0
$\mathbf{x}_4$	3	5	0
$\mathbf{x}_5$	4	4	0
$\mathbf{x}_6$	4	7	1
$\mathbf{x}_7$	6	4	1
$\mathbf{x}_8$	6	6	1
$\mathbf{x}_9$	2	7	0
$\mathbf{x}_{10}$	3	4	1
$\mathbf{x}_{11}$	4	6	—

Datubasea

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$
$\mathbf{x}_1$	0	3.1	2.2	4.4	4.2	6.7	5.8	7	6	3.6
$\mathbf{x}_2$	3.1	0	2.2	1.4	2	3.6	4	4.4	3	1
$\mathbf{x}_3$	2.2	2.2	0	3	2.2	5	3.6	5	5	2
$\mathbf{x}_4$	4.4	1.4	3	0	1.4	2.2	3.1	3.1	2.2	1
$\mathbf{x}_5$	4.2	2	2.2	1.4	0	3	2	2.8	3.6	1
$\mathbf{x}_6$	6.7	3.6	5	2.2	3	0	3.6	2.2	2	3.1
$\mathbf{x}_7$	5.8	4	3.6	3.1	2	3.6	0	2	5	3
$\mathbf{x}_8$	7	4.4	5	3.1	2.8	2.2	2	0	4.1	3.6
$\mathbf{x}_9$	6	3	5	2.2	3.6	2	5	4.1	0	3.1
$\mathbf{x}_{10}$	3.6	1	2	1	1	3.1	3	3.6	3.1	0
$\mathbf{x}_{11}$	5.8	2.8	4.1	1.4	2	1	2.8	2	2.2	2.2

Datubaseko kasuen arteko distantzia Euklidentarrak

Datubasearen adierazpen grafikoa espazio Euklidentarrean:



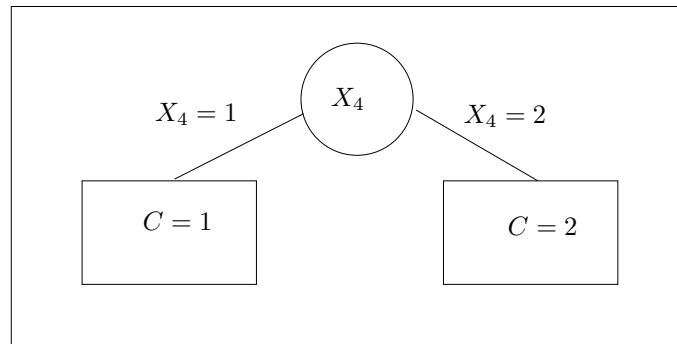
Eman ezazu $\mathbf{x}_{11}$ kasu berriarentzako klasearen iragarpena oinarritzko 5-NN sailkatzailearekin eta naïve Bayesekin hibrido bat eraikiz. Probabilitateen estimazioak egiteko Laplaceren zuzenketa (rule of succession) erabil ezazu.

2. Izan bedi 20 kasu dituen datubasea, non 4 aldagai iragarle eta C klase-aldagai hauek ditugun: X_1 : “Analisi Matematikoa”, X_2 : “Kalkulua”, X_3 : “Optimizazioa”, X_4 : “Probabilitatea eta Estatistika”, C : “Datu-Analisia” irakasgaia lehen deialdian gaitutuko den, bai ala ez

$X_i, i = 1, \dots, 4$ aldagai iragarleak 1 balioa badu, orduan irakasgai hori lehen deialdian gaitutua izan zela esan nahi du, eta 2 balioa badu, orduan deialdi bat baino gehiago behar izan zela irakasgaia gaitutzeko.

Kasuak	X_1	X_2	X_3	X_4	C
$\mathbf{x}_1$	1	2	1	1	1
$\mathbf{x}_2$	2	2	1	1	1
$\mathbf{x}_3$	1	2	1	1	1
$\mathbf{x}_4$	2	2	1	1	1
$\mathbf{x}_5$	1	1	1	1	1
$\mathbf{x}_6$	2	2	1	1	1
$\mathbf{x}_7$	1	1	1	1	1
$\mathbf{x}_8$	2	2	1	1	1
$\mathbf{x}_9$	1	1	2	2	1
$\mathbf{x}_{10}$	2	1	2	2	1
$\mathbf{x}_{11}$	1	2	1	2	2
$\mathbf{x}_{12}$	2	2	1	2	2
$\mathbf{x}_{13}$	1	2	2	2	2
$\mathbf{x}_{14}$	2	2	2	2	2
$\mathbf{x}_{15}$	1	2	1	2	2
$\mathbf{x}_{16}$	2	2	2	2	2
$\mathbf{x}_{17}$	1	2	1	2	2
$\mathbf{x}_{18}$	2	1	2	1	2
$\mathbf{x}_{19}$	1	1	2	1	2
$\mathbf{x}_{20}$	2	2	2	1	2

Datubasetik abiatuz sailkapen-zuhaitz bat indutitu da, eta inausketa-prozesua burutu da inausketarako irizpide bat erabiliz. Honakoa da inausketa egin ondoren geratu den zuhaitza:



Zuhaitz horretatik abiatuz hibrido bat eraiki nahi da, hostoetan naïve Bayes lokal bat eraikiz. Egin ezazu datubasean ez dauden ondoko bi kasuetarako C klase-aldagaiaren iragarpena:

$$\mathbf{x}_{21} = (1, 2, 2, 1) \quad , \quad \mathbf{x}_{22} = (1, 1, 1, 2)$$

Probabilitateen estimazioak egiteko Laplaceren zuzenketa (rule of succession) erabil ezazu.

3. Demagun bi dimentsioko 10 kasuz osatutako datubasea dugula, eta $\mathbf{x}_{11}$ kasu berria sailkatu nahi dugula. Demagun X eta Y aldagaien balioak dagoeneko normalizatuak izan direla, eta kasuen arteko distantzia Euklidentarrak kalkulatu ditugula.

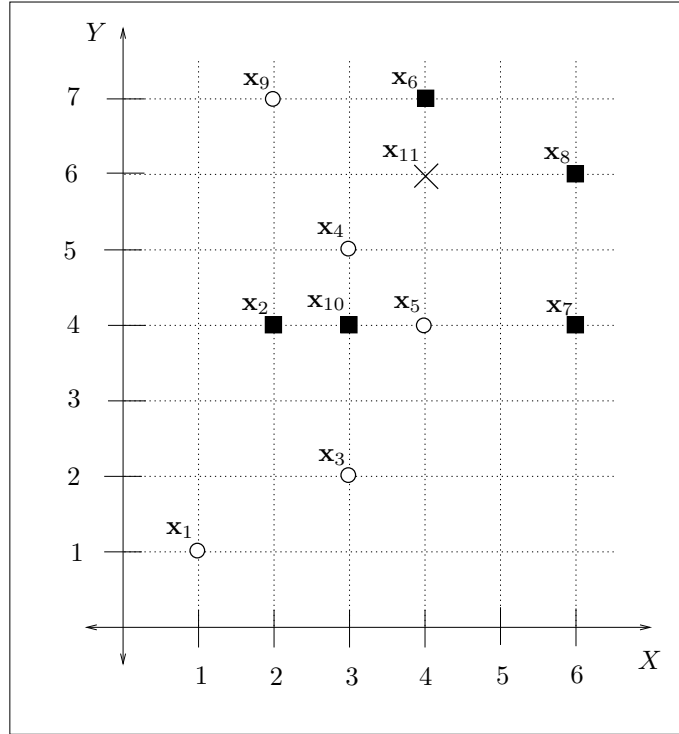
Kasua	X	Y	Klasea
$\mathbf{x}_1$	1	1	0
$\mathbf{x}_2$	2	4	1
$\mathbf{x}_3$	3	2	0
$\mathbf{x}_4$	3	5	0
$\mathbf{x}_5$	4	4	0
$\mathbf{x}_6$	4	7	1
$\mathbf{x}_7$	6	4	1
$\mathbf{x}_8$	6	6	1
$\mathbf{x}_9$	2	7	0
$\mathbf{x}_{10}$	3	4	1
$\mathbf{x}_{11}$	4	6	—

Datubasea

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$
$\mathbf{x}_1$	0	3.1	2.2	4.4	4.2	6.7	5.8	7	6	3.6
$\mathbf{x}_2$	3.1	0	2.2	1.4	2	3.6	4	4.4	3	1
$\mathbf{x}_3$	2.2	2.2	0	3	2.2	5	3.6	5	5	2
$\mathbf{x}_4$	4.4	1.4	3	0	1.4	2.2	3.1	3.1	2.2	1
$\mathbf{x}_5$	4.2	2	2.2	1.4	0	3	2	2.8	3.6	1
$\mathbf{x}_6$	6.7	3.6	5	2.2	3	0	3.6	2.2	2	3.1
$\mathbf{x}_7$	5.8	4	3.6	3.1	2	3.6	0	2	5	3
$\mathbf{x}_8$	7	4.4	5	3.1	2.8	2.2	2	0	4.1	3.6
$\mathbf{x}_9$	6	3	5	2.2	3.6	2	5	4.1	0	3.1
$\mathbf{x}_{10}$	3.6	1	2	1	1	3.1	3	3.6	3.1	0
$\mathbf{x}_{11}$	5.8	2.8	4.1	1.4	2	1	2.8	2	2.2	2.2

Datubaseko kasuen arteko distantzia Euklidentarrak

Datubasearen adierazpen grafikoa espazio Euklidentarrean:



Demagun datubasean bootstrap aplikatuz (kasuak birjarritz) honako hiru laginak sortu direla zoriz:

$$D_1 = \{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_5, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_{10}, \mathbf{x}_{10}\}$$

$$D_2 = \{\mathbf{x}_1, \mathbf{x}_1, \mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_6, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_8, \mathbf{x}_9\}$$

$$D_3 = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_4, \mathbf{x}_4, \mathbf{x}_4, \mathbf{x}_6, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}\}$$

- 3.1 Eman ezazu $\mathbf{x}_{11}$ kasu berriarentzako klasearen iragarpena oinarritzko sailkatzaile moduan 3-NN sailkatzailea erabiliz eta bagging metodoaren bidez konbinatuz.
- 3.2 Eman ezazu $\mathbf{x}_{11}$ kasu berriarentzako klasearen iragarpena oinarritzko sailkatzaile moduan naïve Bayes erabiliz eta bagging bidez konbinatuz. Probabilitateen estimazioak egiteko Laplaceren zuzenketa (rule of succession) erabil ezazu.

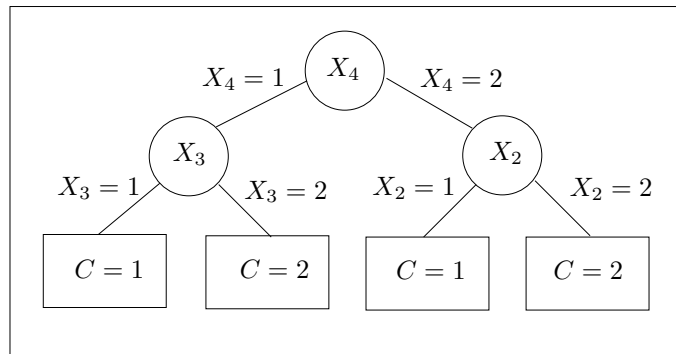
4. Izan bedi 20 kasu dituen datubasea, non 4 aldagai iragarle eta C klase-aldagai hauek ditugun: X_1 : “Analisi Matematikoa”, X_2 : “Kalkulua”, X_3 : “Optimizazioa”, X_4 : “Probabilitatea eta Estatistika”, C : “Datu-Analisia” irakasgaia lehen deialdian gaitutuko den, bai ala ez

$X_i, i = 1, \dots, 4$ aldagai iragarleak 1 balioa badu, orduan irakasgai hori lehen deialdian gaitutua izan zela esan nahi du, eta 2 balioa badu, orduan deialdi bat baino gehiago behar izan zela irakasgaia gaitutzeko.

Kasuak	X_1	X_2	X_3	X_4	C
$\mathbf{x}_1$	1	2	1	1	1
$\mathbf{x}_2$	2	2	1	1	1
$\mathbf{x}_3$	1	2	1	1	1
$\mathbf{x}_4$	2	2	1	1	1
$\mathbf{x}_5$	1	1	1	1	1
$\mathbf{x}_6$	2	2	1	1	1
$\mathbf{x}_7$	1	1	1	1	1
$\mathbf{x}_8$	2	2	1	1	1
$\mathbf{x}_9$	1	1	2	2	1
$\mathbf{x}_{10}$	2	1	2	2	1
$\mathbf{x}_{11}$	1	2	1	2	2
$\mathbf{x}_{12}$	2	2	1	2	2
$\mathbf{x}_{13}$	1	2	2	2	2
$\mathbf{x}_{14}$	2	2	2	2	2
$\mathbf{x}_{15}$	1	2	1	2	2
$\mathbf{x}_{16}$	2	2	2	2	2
$\mathbf{x}_{17}$	1	2	1	2	2
$\mathbf{x}_{18}$	2	1	2	1	2
$\mathbf{x}_{19}$	1	1	2	1	2
$\mathbf{x}_{20}$	2	2	2	1	2

Datubasetik abiatuz honako oinarritzko 5 sailkatzaileak induzitu dira:

- k -NN, $k = 3$ izanik.
- Naïve Bayes: $p(C|X_1, X_2, X_3, X_4) \propto p(C) \cdot p(X_1|C) \cdot p(X_2|C) \cdot p(X_3|C) \cdot p(X_4|C)$
- TAN: $p(C|X_1, X_2, X_3, X_4) \propto p(C) \cdot p(X_1|C) \cdot p(X_2|X_1, C) \cdot p(X_3|X_4, C) \cdot p(X_4|X_2, C)$
- 2-DB: $p(C|X_1, X_2, X_3, X_4) \propto p(C) \cdot p(X_4|C) \cdot p(X_3|X_4, C) \cdot p(X_2|X_3, X_4, C) \cdot p(X_1|X_2, X_3, C)$
- Sailkapen-zuhaitza:



Egin ezazu datubasean ez dauden ondoko hiru kasuetarako C klase-aldagairen iragarpena, eta konbina itzazu sailkatzaileen iragarpenak gehiengoaren aldeko bozketa erabiliz. Probabilitateen estimazioak egiteko Laplaceren zuzenketa (rule of succession) erabil ezazu.

$$\mathbf{x}_{21} = (1, 2, 2, 1) \quad , \quad \mathbf{x}_{22} = (2, 1, 1, 1) \quad , \quad \mathbf{x}_{23} = (1, 1, 1, 2)$$