

KONPUTAZIO-ZIENTZIEN METODO MATEMATIKOAK

2010.eko ekaina

(Azterketa honetan gehienez 5 puntu lortuko dira. Praktikan ere gehienez 5. Biak gaitzitu behar dira irakasgaia bere osotasunean gaitzitzeko.)

1. (0.5 puntu) Ondorengo taulan entrenamendurako 8 kasuz osatutako datubasea agertzen da.

Kasua	X	Y	Klasea
$\mathbf{x}_1$	2	0	1
$\mathbf{x}_2$	4	2	1
$\mathbf{x}_3$	1	1	1
$\mathbf{x}_4$	2	4	2
$\mathbf{x}_5$	1	2	1
$\mathbf{x}_6$	2	3	2
$\mathbf{x}_7$	3	4	2
$\mathbf{x}_8$	3	6	2
$\mathbf{x}_9$	4	3	?

Distantzia Euklidestarrak kalkulatu eta eman ezazu $\mathbf{x}_9$ kasu berriaren klasearen iragarpen bat honako sailkatzaileak erabiliz:

1.1 Oinarrizko 1-NN sailkatzailea

1.2 3-NN batzbesteko distantziarekin

2. (0.5 puntu) Ondorengo taulan entrenamendurako 8 kasuz osatutako datubasea agertzen da. Hart-en kondentsazioa erabiliz murriztu ezazu datubasea, $k=1$.

Kasua	X	Y	Klasea
$\mathbf{x}_1$	2	0	0
$\mathbf{x}_2$	4	2	1
$\mathbf{x}_3$	1	1	0
$\mathbf{x}_4$	2	4	1
$\mathbf{x}_5$	1	2	0
$\mathbf{x}_6$	2	3	1
$\mathbf{x}_7$	3	4	0
$\mathbf{x}_8$	3	6	1

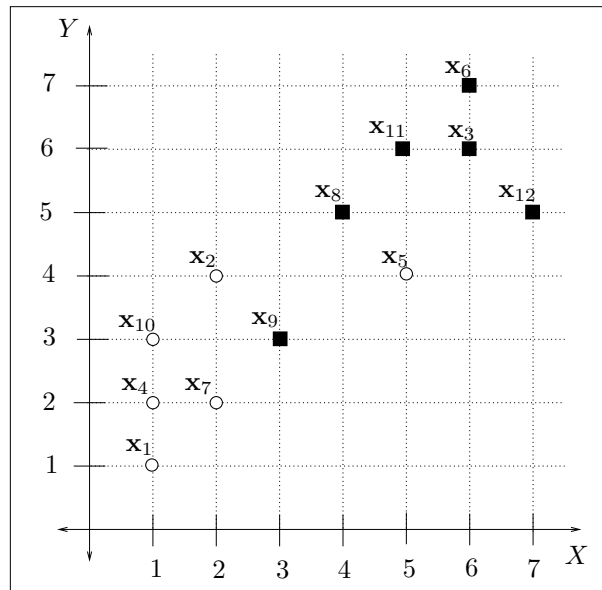
3. (0.5 puntu) Demagun bi dimentsioko 12 kasuz osatutako honako datubasea dugula.

Kasua	X	Y	Klasea		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
x_1	1	1	0	x_1	0	3.1	7.0	1.0	5.0	7.8	1.4	5.0	2.8	2.0	6.4	7.2
x_2	2	4	0	x_2	3.1	0	4.4	2.2	3.0	5.0	2.0	2.2	1.4	1.4	3.6	5.0
x_3	6	6	1	x_3	7.0	4.4	0	6.4	2.2	1.0	5.6	2.2	4.2	5.8	1.0	1.4
x_4	1	2	0	x_4	1.0	2.2	6.4	0	4.4	7.0	1.0	4.2	2.2	1.0	5.6	6.7
x_5	5	4	0	x_5	5.0	3.0	2.2	4.4	0	3.1	3.6	1.4	2.2	4.1	2.0	2.2
x_6	6	7	1	x_6	7.8	5.0	1.0	7.0	3.1	0	6.4	2.8	5.0	6.4	1.4	2.2
x_7	2	2	0	x_7	1.4	2.0	5.6	1.0	3.6	6.4	0	3.6	1.4	1.4	5.0	5.8
x_8	4	5	1	x_8	5.0	2.2	2.2	4.2	1.4	2.8	3.6	0	2.2	3.6	1.4	3.0
x_9	3	3	1	x_9	2.8	1.4	4.2	2.2	2.2	5.0	1.4	2.2	0	2.0	3.6	4.4
x_{10}	1	3	0	x_{10}	2.0	1.4	5.8	1.0	4.1	6.4	1.4	3.6	2.0	0	5.0	6.3
x_{11}	5	6	1	x_{11}	6.4	3.6	1.0	5.6	2.0	1.4	5.0	1.4	3.6	5.0	0	2.2
x_{12}	7	5	1	x_{12}	7.2	5.0	1.4	6.7	2.2	2.2	5.8	3.0	4.4	6.3	2.2	0

Datubasea

Distantzia Euklidestarrek

Datubasearen adierazpen grafikoa espazio Euklidestarrean:



Datubase horretatik abiatuz lortutako 1-NN sailkatzailearen ebaluazioa egin ezazu 3-fold cross-validation edo 3-geruzako balidazio gurutzatua erabiliz, geruzak honakoak izanik:

$$\{x_1, x_2, x_3, x_4\}, \quad \{x_5, x_6, x_7, x_8\}, \quad \{x_9, x_{10}, x_{11}, x_{12}\}$$

Auzokide hurbilena kalkulatzekoan berdinketa gertatzen denean datubasean lehena agertzen den kasua aukera ezazu. Zein da sailkatzailearen asmatze-tasa?

4. (2 puntu) Izan bedi 10 kasuz osatutako datubasea. Datubaseko kasu bakoitza lau aldagai iragarleren bidez emana dator: X , Y , Z eta T . Taulako azken zutabea adierazten da kasu bakoitzaren klasea: $C = 0$ edo $C = 1$.

Kasuak	X	Y	Z	T	C
$\mathbf{x}_1$	1	1	1	1	0
$\mathbf{x}_2$	2	2	1	2	0
$\mathbf{x}_3$	4	2	3	2	0
$\mathbf{x}_4$	3	1	3	1	0
$\mathbf{x}_5$	4	3	5	1	1
$\mathbf{x}_6$	3	2	3	2	0
$\mathbf{x}_7$	1	3	3	1	1
$\mathbf{x}_8$	2	3	2	1	1
$\mathbf{x}_9$	2	1	5	2	0
$\mathbf{x}_{10}$	4	2	1	2	0

Aldagai iragarleen eta C klase-aldagaiaren elkarrekiko informazioak eza-gunak dira: $I(X, C) = 0.1303$, $I(Y, C) = 0.8813$, $I(Z, C) = 0.3568$ eta $I(T, C) = 0.3958$. C klase-aldagaiaren baldintzapean bi aldagai iragarleren artean dagoen elkarrekiko informazio kantitatea ere ezaguna da: $I(X, Y|C) = 0.2897$, $I(X, Z|C) = 1.0897$, $I(X, T|C) = 0.4042$, $I(Y, Z|C) = 0.1387$, $I(Y, T|C) = 0.3287$, $I(Z, T|C) = 0.0532$.

- 4.1 Eman ezazu elkarrekiko informazio-kantitatearen definizioa. Adieraz ezazu nola kalkulatu den $I(Y, C) = 0.8813$. Besteekin konparatuta, zergatik uste duzu dela hain altua?
- 4.2 Emandako datubasetik abiatuz, aplika ezazu ID3 algoritmoa sailkapen-zuhaitz bat indultzeko. Ondoren, erabil ezazu induzitutako zuhaitza $\mathbf{x}_{11} = (4, 1, 2, 1)$ kasu berriaren klase-aldagaiaren iragarpena egiteko.
- 4.3 Emandako datubasetik abiatuz, eraiki ezazu Naive Bayes eredua. Ondoren, erabil ezazu eredua $\mathbf{x}_{11} = (4, 1, 2, 1)$ kasu berriaren klase-aldagaiaren iragarpena egiteko.
- 4.4 Emandako datubasetik abiatuz, eraiki ezazu TAN eredua. Ondoren, erabil ezazu eredua $\mathbf{x}_{11} = (4, 1, 2, 1)$ kasu berriaren klase-aldagaiaren iragarpena egiteko.
5. (0.5 puntu) Demagun 3. ariketan emandako datubasea dugula eta demagun 12 kasu horien klasea (taulako azken zutabea) ez dela ezaguna. Hasierako partiketan kasu bakoitza konglomeratu batean dago. Esan ezazu pausu batean zein konglomeratu bilduko diren distantziaren arabera.
6. (1 puntu) Idatz ezazu algoritmo genetiko sinplearen sasikodea. Adieraz ezazu algoritmoaren parametro bakoitzaren erabilera.