

3. Gainbegiraturako Sailkatzaileen Ebaluazioa Ariketak

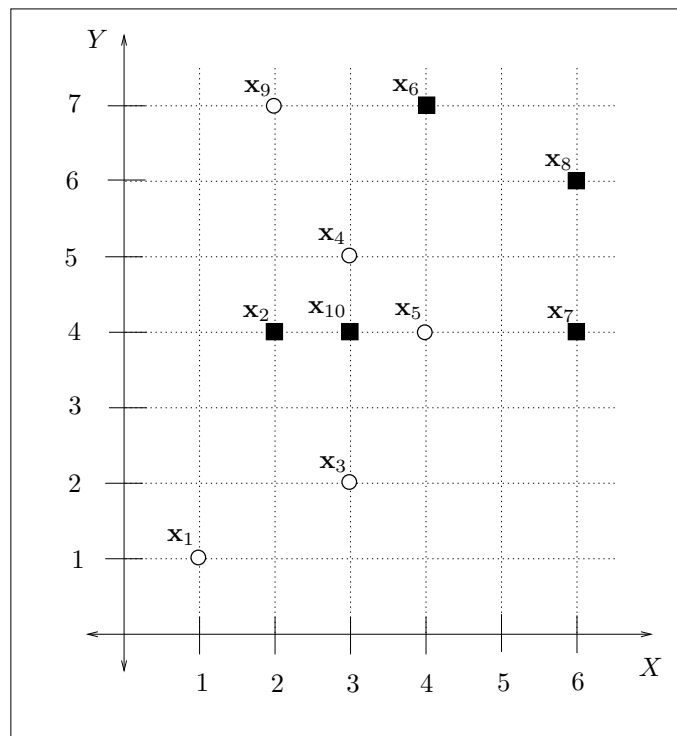
- Demagun bi dimentsioko 10 kasuz osatutako honako datubasea dugula. Demagun X eta Y aldagaien balioak dagoeneko normalizatuak izan direla, eta kasuen arteko distantzia Euklides-tarrak kalkulatu ditugula. Demagun baita klase positiboa 1 klasea dela (TPR, FPR, TNR, FPR koefizienteak eta ROC kurba estimatzeko).

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_1	1	1	0	x_1	0	3.1	2.2	4.4	4.2	6.7	5.8	7	6	3.6
x_2	2	4	1	x_2	3.1	0	2.2	1.4	2	3.6	4	4.4	3	1
x_3	3	2	0	x_3	2.2	2.2	0	3	2.2	5	3.6	5	5	2
x_4	3	5	0	x_4	4.4	1.4	3	0	1.4	2.2	3.1	3.1	2.2	1
x_5	4	4	0	x_5	4.2	2	2.2	1.4	0	3	2	2.8	3.6	1
x_6	4	7	1	x_6	6.7	3.6	5	2.2	3	0	3.6	2.2	2	3.1
x_7	6	4	1	x_7	5.8	4	3.6	3.1	2	3.6	0	2	5	3
x_8	6	6	1	x_8	7	4.4	5	3.1	2.8	2.2	2	0	4.1	3.6
x_9	2	7	0	x_9	6	3	5	2.2	3.6	2	5	4.1	0	3.1
x_{10}	3	4	1	x_{10}	3.6	1	2	1	1	3.1	3	3.6	3.1	0

Datubasea

Datubaseko kasuen arteko distantzia Euklides-tarrak

Datubasearen adierazpen grafikoa espazio Euklides-tarrean:



Datubase horretatik abiatuz lortutako K-NN sailkatzaileen ebaluazioa egin ezazu. K-NN sailkatzaileari berdinketa kasuak sortzen zaizkionean suposa ezazu ausaz datubasean lehenengo dagoen kasua aukeratuko dela. Honako estimazio-teknikak erabil itzazu sailkatzaileen ebaluazioa egiteko:

1. Holdout. Demagun ondoren zehazten diren entrenamendurako (%60) eta testerako (beste %40) kasuen azpimultzoak sortu direla zoriz.
 - 1.1 Entrenamendua: $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$. Testa: $\{\mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}\}$. Egin itzazu 1-NN, 2-NN eta 3-NN sailkatzaileen asmatze-tasen eta TPR, FNR, FPR eta TNR koefizienteen estimazioak, eta errorearen matrizeak osa itzazu. Irudika itzazu hiru sailkatzaileak ROC kurba batean. Zein da sailkatzailearik onena? Estimatu berri dituzun koefizienteetan oinarrituz, arrazonatu sailkatzaile bakoitzaren ezaugarriak.
 - 1.2 Entrenamendua: $\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9\}$. Testa: $\{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_{10}\}$. Aurreko azpiataleko galdera berberak erantzun itzazu.
 - 1.3 Entrenamendua: $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_8, \mathbf{x}_{10}\}$. Testa: $\{\mathbf{x}_1, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_9\}$. Aurreko azpiataleko galdera berberak erantzun itzazu.

Aurreko hiru azpiataletan egin dituzun ebaluazioetan oinarrituz, Holdout estimazio-metodoa hiru aldiz errepikatua erabiliz gero, zein da sailkatzaile bakoitzaren asmatze-tasa? Zein da sailkatzailearik onena? Ohartzen al zara zorizkotasunaren eraginaz? Lortu dituzun emaitzetan oinarrituz arrazona ezazu. Ausa tartean denean errepikatzea ezinbestekoa da.

2. k-fold cross-validation edo k-geruzako balidazio gurutzatua. 1-NN sailkatzailearen balidazio gurutzatua egin ezazu ondorengo hiru geruza kopuruak erabiliz:
 - 2.1 Datubasea 2 geruzatan banatuz. Demagun $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ eta $\{\mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}\}$ direla 2 geruzak. Zein da sailkatzailearen asmatze-tasa? Konturatzen al zara balidazio gurutzatua 2 geruzatan egitea holdout 2 aldiz egitearen berdina dela, entrenamendurako eta testerako kasuen %50 hartuta eta entrenamendua eta testa elkartrukatuta?
 - 2.2 Datubasea 5 geruzatan banatuz. Demagun 5 geruzak honakoak direla: $\{\mathbf{x}_1, \mathbf{x}_2\}$, $\{\mathbf{x}_3, \mathbf{x}_4\}$, $\{\mathbf{x}_5, \mathbf{x}_6\}$, $\{\mathbf{x}_7, \mathbf{x}_8\}$, $\{\mathbf{x}_9, \mathbf{x}_{10}\}$. Zein da sailkatzailearen asmatze-tasa?
 - 2.3 Datubasea 10 geruzatan banatuz: leave one out. Zein da sailkatzailearen asmatze-tasa?

Ohar zaitez geruza-kopuruak duen eraginaz asmatze-tasa estimatzerakoan balidazio gurutzatua k-geruzatan egiten denean.

3. Bootstrap bidezko estimazio-metodoa. 3-NN sailkatzailearen ebaluazioa egin ezazu. Suposa ezazu datubasetik kasuak birjarritz entrenamendurako honako hiru laginak sortu direla zoriz:

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

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

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

Zein da sailkatzailearen asmatze-tasa?

Bukatzeko, konpara itzazu ebaluaziorako estimazio-teknika ezberdinak erabiliz lortu dituzun asmatze-tasak.

Azken gogoeta:

Ez dago sailkatzaile unibertsalik!

Ez dago ebaluazio-teknika unibertsalik!

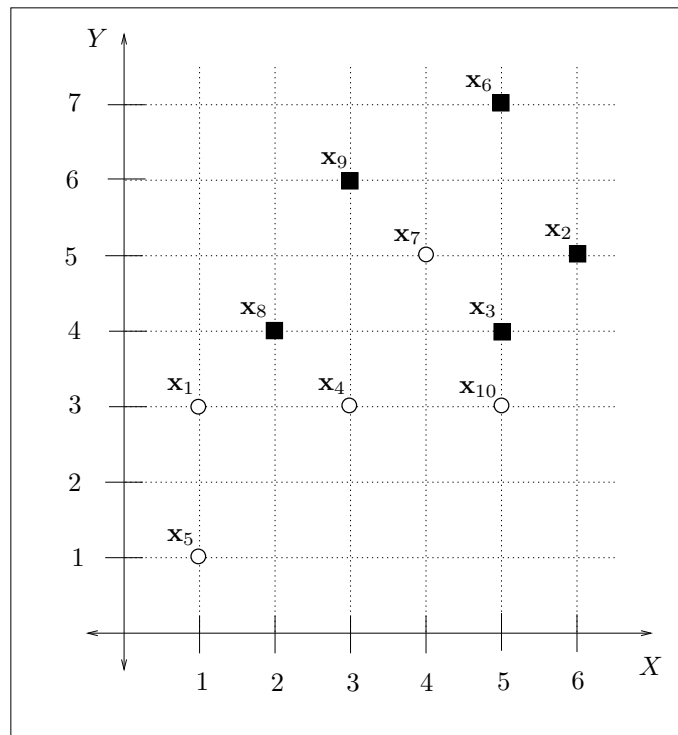
2. Aurreko ataleko galdera berberak erantzun itzazu ondorengo beste database honetarako.

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_1	1	3	0	x_1	0	5.3	4.1	2	2	5.6	3.6	1.4	3.6	4
x_2	6	5	1	x_2	5.3	0	1.4	3.6	6.4	2.2	2	4.1	3.1	2.2
x_3	5	4	1	x_3	4.1	1.4	0	2.2	5	3	1.4	3	2.8	1
x_4	3	3	0	x_4	2	3.6	2.2	0	2.8	4.4	2.2	1.4	3	2
x_5	1	1	0	x_5	2	6.4	5	2.8	0	7.2	5	3.1	5.3	4.4
x_6	5	7	1	x_6	5.6	2.2	3	4.4	7.2	0	2.2	4.2	2.2	4
x_7	4	5	0	x_7	3.6	2	1.4	2.2	5	2.2	0	2.2	1.4	2.2
x_8	2	4	1	x_8	1.4	4.1	3	1.4	3.1	4.2	2.2	0	2.2	3.1
x_9	3	6	1	x_9	3.6	3.1	2.8	3	5.3	2.2	1.4	2.2	0	3.6
x_{10}	5	3	0	x_{10}	4	2.2	1	2	4.4	4	2.2	3.1	3.6	0

Datubasea

Datubaseko kasuen arteko distantzia Euklidestarrak

Datubasearen adierazpen grafikoa espazio Euklidestarrean:



3. Har dezagun “Congressional Voting Records Data Set” datubasea. Datubaseari buruzko zehaztasunak honako helbidean aurkituko dituzu:

<http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records>

Weka aplikazioa erabiliz, 1-NN, 2-NN, 3-NN eta 4-NN sailkatzaileak sortu ditugu. Auzokide hurbilenak aurkitzeko distantzia Euklidebarra erabili da. Erabili den k -NN algoritmoaren aldaera, k -NN auzokideen pisaketarekin izan da, eta pisua distantziarekiko alderantzizko proportzionala definitu da.

Eraikitako lau sailkatzaile horien ebaluaziorako, 10 geruzako balidazio gurutzatua erabili da. Ikus behean weka aplikazioaren irteera. Kontuan izan behar da erroreen matrizea osatzera-koan, weka aplikazioak errenkadetan klase errealeari buruzko informazioa kokatzen duela, eta zutabeetan iragarritako klaseari buruzkoa.

Erabaki ezazu lau sailkatzaileen artean zein den onena, bai asmatze-tasari erreparatuz, eta baita Roc kurba interpretatuz.

1-NN === Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.914	0.06	0.961	0.914	0.937	0.974	democrat
	0.94	0.086	0.873	0.94	0.905	0.974	republican
Weighted Avg.	0.924	0.07	0.927	0.924	0.925	0.974	

=== Confusion Matrix ===

```

a   b <-- classified as
244 23 | a = democrat
10 158 | b = republican

```

2-NN === Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.921	0.06	0.961	0.921	0.941	0.978	democrat
	0.94	0.079	0.883	0.94	0.911	0.978	republican
Weighted Avg.	0.929	0.067	0.931	0.929	0.929	0.978	

=== Confusion Matrix ===

```

a   b <-- classified as
246 21 | a = democrat
10 158 | b = republican

```

3-NN === Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.918	0.054	0.965	0.918	0.94	0.981	democrat
	0.946	0.082	0.878	0.946	0.911	0.981	republican
Weighted Avg.	0.929	0.065	0.931	0.929	0.929	0.981	

=== Confusion Matrix ===

```

a   b <-- classified as
245 22 | a = democrat
9 159 | b = republican

```

4-NN === Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.918	0.048	0.968	0.918	0.942	0.983	democrat
	0.952	0.082	0.879	0.952	0.914	0.983	republican
Weighted Avg.	0.931	0.061	0.934	0.931	0.931	0.983	

=== Confusion Matrix ===

```

a   b <-- classified as
245 22 | a = democrat
8 160 | b = republican

```