

KONPUTAZIO-ZIENTZIEN METODO MATEMATIKOAK 2011.ko maiatza

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

1. (Puntu 1) Gainbegiraturako sailkatzaileen ebaluazioa. Demagun N kasuz osaturako datubasea dugula, eta datubase horretatik abiatuz sailkapen-zuhaitz bat induzitu eta ebaluatu nahi dugula, honako bi metodoak erabiliz:
 - 1.1 Holdout. Azal ezazu zein diren eman beharreko urratsak, bai sailkapen-zuhaitza induitzeko eta baita ebaluatzeko. Nola kalkulatu da asmatze-tasa?
 - 1.2 10-fold cross-validation edo 10-geruzako balidazio gurutzatua. Azal ezazu zein diren eman beharreko urratsak, bai sailkapen-zuhaitza induitzeko eta baita ebaluatzeko. Nola kalkulatu da asmatze-tasa?

Aurreko bi atalak erantzuterakoan, grafikoren bat ere irudika ezazu metodoak ahalik eta argien azaltzeko. Metodoak azaldu ondoren, konpara itzazu bi ebaluazio metodoak. (Oharra: Ez eman $ID3$ edo $C4.5$ algoritmoen inolako azalpenik).

2. (2 puntu) Naive Bayes sailkatzailea. Teoria eta ariketa.
 - 2.1 (Puntu 1) Teoria.
 - Zein teorematan oinarritzen da Naive Bayes sailkatzailea?
 - Zein dira egiten diren hipotesiak ereduaren sortzerakoan?
 - Zenbat parametroren estimazioa egin behar da Naive Bayes ereduaren, n aldagai iragarle dituen problema batean klase-aldagaiak m balio posible hartzen baditu? Jar ezazu adibide bat 3 aldagai iragarle izanik klase-aldagaiak 2 balio posible ditueneko.
 - 2.2 (Puntu 1) Emandako datubasetik abiatuz, eraiki ezazu Naive Bayes ereduaren. Ondoren, erabil ezazu ereduaren $\mathbf{x}_{11} = (0, 0, 1)$ kasu berriaren C klase-aldagaiaren iragarpena egiteko. Probabilitateen estimazioak kalkulatzeko, ez ezazu Laplaceren zuzenketa (rule of succession) erabili.

Kasuak	X	Y	Z	C
$\mathbf{x}_1$	0	0	1	—
$\mathbf{x}_2$	1	0	1	+
$\mathbf{x}_3$	0	1	0	—
$\mathbf{x}_4$	1	0	0	—
$\mathbf{x}_5$	1	0	1	+
$\mathbf{x}_6$	0	0	1	+
$\mathbf{x}_7$	1	1	0	—
$\mathbf{x}_8$	0	0	0	—
$\mathbf{x}_9$	0	1	0	+
$\mathbf{x}_{10}$	1	1	1	+

3. (1.5 puntu) Tenisean jokatzeko zaletasuna duen kirolari batek 14 egunetako eguraldi-informazioa eta egun horietako bakoitzean hartutako erabakia (jokatu? bai ala ez) gorde ditu honako datubasean:

Eguna	X_1 (Zerua)	X_2 (Tenperatura)	X_3 (Hezetasuna)	X_4 (Haizea)	C Jokatu?
1	garbia	beroa	altua	ahula	ez
2	garbia	beroa	altua	zakarra	ez
3	hodeitsu	beroa	altua	ahula	bai
4	euritsu	epela	altua	ahula	bai
5	euritsu	hotza	normala	ahula	bai
6	euritsu	hotza	normala	zakarra	ez
7	hodeitsu	hotza	normala	zakarra	bai
8	garbia	epela	altua	ahula	ez
9	garbia	hotza	normala	ahula	bai
10	euritsu	epela	normala	ahula	bai
11	garbia	epela	normala	zakarra	bai
12	hodeitsu	epela	altua	zakarra	bai
13	hodeitsu	beroa	normala	ahula	bai
14	euritsu	epela	altua	zakarra	ez

Eguraldiari buruzko informazioa gordetzeko erabili dituen lau aldagai iragarleek honako informazioa ematen dute: zeruaren egoera (X_1), tenperatura (X_2), hezetasuna (X_3) eta haizea (X_4).

Aldagaien aukeraketa egokia egin nahi da, ondoren sailkapen-zuhaitz bat eraiki eta kasu berri bat sailkatu.

- 3.1 Emandako datubasetik abiatuta, aldagai-aukeraketa egin behar duzu “Filter” izeneko teknikari jarraituz. Horretarako, aldagai iragarleek C klase-aldagaiarekiko duten elkarrekiko informazioak kalkula itzazu: $I(X_1, C)$, $I(X_2, C)$, $I(X_3, C)$, $I(X_4, C)$. Informazio horren arabera, zein aldagai-aukeraketa egingo zenuke? Aldagai iragarle guztiak esanguratsuak direla esango al zenuke?
- 3.2 Aurreko ataleko kalkuluetan oinarrituz, aldagai-aukeraketan aukera itzazu bi aldagai iragarle esanguratsuenak. Ondoren, ID3 algoritmoa erabiliz, sailkapen-zuhaitz bat eraiki ezazu. Induzituko den sailkapen-zuhaitz hori bakarra al da?
- 3.3 Demagun kasu berri hau datorkigula: X_1 (Zerua)=euritsu, X_2 (Tenperatura)=hotza, X_3 (Hezetasuna)=altua, X_4 (Haizea)=zakarra. Kasu berri horrentzat zein izango da egingo den iragarpena, aurreko atalean eraikitako sailkapen-zuhaitza aplikatuz?
4. (0.5 puntu) Taulan dauden datuekin Forgy-ren algoritmoa erabili nahi dugu $k = 3$ klaseak aurkitzeko. Hasierako partiketa zoriz eginez, azken zutabean (klasea) dagoen partiketa atera da. Jarrai itzazu algoritmoaren pausuak amaitu arte.

Kasua	X	Y	Klasea
x_1	1	1	3
x_2	1	6	1
x_3	4	3	3
x_4	5	5	1
x_5	6	1	2
x_6	6	7	2