

5. Sailkatzaile Bayestarrak.

Ariketak

1. Futboleko munduko txapelketako finalaren atsedenaldean A eta B taldeen arteko partidari buruzko apustua egin nahi da. Horretarako, bi parametrotan oinarritzen den sailkapen paradigma bat eraiki nahi da. Parametroak honakoak dira: atsedenaldean partidaren emaitza zeinen aldekoa den eta txapelketa osoan zehar jasotako golak. A eta B futbol taldeek txapelketa irabazteko a priori probabilitate berbera dute.

Txapelketaren irabazlea A taldea bada, orduan atsedenalde garairako irabazten joateko %70-eko probabilitatea du. Txapelketaren irabazlea B taldea bada, orduan %60-ko probabilitatez atsedenalde garairako irabazten joango da.

Txapelketaren irabazlea A taldea bada, %80-ko probabilitatez B taldeak baino gol gutxiago jaso du txapelketan zehar. Irabazlea B taldea bada, %50-eko probabilitatez A taldeak baino gol gutxiago jaso du.

Atsedenalde garaian A taldea B taldeari irabazten ari zaio, eta B taldeak A taldeak baino gol gutxiago jaso du txapelketan zehar. Futbol partidaren emaitzari buruzko iragarpena egin ezazu, zein apustu egitea komeni zaigun erabaki ahal izateko. Erabilitako paradigma justifika ezazu.

2. Korreo elektronikoa zaborra diren mezuak jasotzeko a priori probabilitatea %50-ekoa da. Sailkapenerako eredu bat eraiki nahi da mezuak zaborra diren ala ez erabakitzeko. Horretarako, mezuan bi ezaugarri aztertuko ditugu: mezua bidaltzen duenaren helbide elektronikoa arriskutsuak diren helbideen zerrendakoa den, eta mezuan karaktere susmagarriak agertzen diren.

Mezua zaborra baldin bada, orduan helbide arriskutsu batetik bidalia izan da %80-ko probabilitatez. Zaborra ez bada, helbide arriskutsu batetik bidalia izan da %10-eko probabilitatez.

Mezua zaborra baldin bada, bertan karaktere susmagarriak daude %70-eko probabilitatez. Zaborra ez bada, karaktere susmagarriak izateko probabilitatea %40-ra jaisten da.

Helbide arriskutsu batetik bidalia izan den mezu bat jaso dugu. Bertan ez dago karaktere susmagarriarik. Zein izango da mezu horrentzat egingo dugun diagnostikoa? Zaborra da? Erabilitako paradigma justifika ezazu.

3. Pertsona batek gaixotasun jakin bat baduen erabaki ahal izateko bi proba egin zaizkio. Gauza jakina da gaixotasun hori izateko %1-eko probabilitatea existitzen dela populazio jakin batentzako.

Lehen probaren emaitza + edo - izan daiteke. Pertsona gaixo badago, lehen proban + emaitza ateratzeko %95-eko probabilitatea dago. Pertsona gaixo ez badago, lehen proban + emaitza ateratzeko %2-ko probabilitatea dago.

Bigarren probaren emaitza posibleak “altua”, “normala” eta “bajua” dira. Pertsona gaixo badago, orduan bigarren probaren emaitza “altua” da %85-etan. Pertsona gaixo badago, bigarren probaren emaitza “bajua” da %5-etan. Pertsona gaixo ez badago, orduan bigarren probaren emaitza “altua” da %20-tan. Pertsona gaixo ez badago, bigarren probaren emaitza “bajua” da %10-etan.

Demagun pertsona bati bi proba horiek egin dizkiotela. Lehen probaren emaitza + izan da, eta bigarren probarena “normala”. Eman ezazu pertsona horrentzako diagnostikoa. Erabilitako paradigma justifika ezazu.

4. Motor mota batek matxura elektrikoa du 10^{-3} -ko probabilitatez. Matxura mekanikoa 10^{-5} -eko probabilitatez izaten da. Bi matxura motak ez dira aldi berean gertatuko.

Matxura elektrikoa denean %95-etan argi gorria pizten da. Matxura mekanikoa denean %99-tan pizten da argi gorria. Motorrak matxurarik ez duenean milioi batetik behin pizten da argi gorria.

Matxurarik ez dagoenean motorraren tenperatura altua da %17-ko probabilitatez, eta bajua %5-eko probabilitatez. Matxura elektrikoa dagoenean motorraren tenperatura altua da %90-eko probabilitatez, eta bajua %1-eko probabilitatez. Matxura mekanikoa dagoenean motorraren tenperatura altua da %10-etan, eta bajua %40-tan.

Argi gorria piztu da, eta motorraren tenperatura bajua da. Motorrak matxurarik du? Egin ezazu diagnostikoa, eta justifika ezazu erabilitako paradigma.

5. Bihotzekoa lehen aldiz jasan duen pertsona batek berriro ere bihotzeko arazoren bat izateko a priori probabilitatea %30-ekoa da. Bi faktorek dute eragina berriro ere bihotzeko arazoren bat izango duen auresateko eraiki dugun ereduan: egunerokotasunean kirol lasaia egiteak eta adinak. Adinak banaketa normalari jarraitzen diola suposatuko dugu.

Berriro ere bihotzeko arazorik izan ez dutenen artean %80-ak egunerokotasunean kirol lasaia egiten du. Berriro ere bihotzeko arazoren bat izan dutenen artean aldiz %10-ak besterik ez du egunerokotasunean kirol lasaia egiten.

Berriro ere bihotzeko arazorik izan ez dutenen adinak banaketa normalari jarraitzen dio, 60 urteko itxaropen matematikoarekin eta 10 urteko desbideratze estandarra. Berriro ere bihotzeko arazoren bat izan dutenen adinak aldiz banaketa normalari jarraitzen dio, 75 urteko itxaropen matematikoarekin eta 15 urteko desbideratze estandarra.

Bihotzekoa izan zuen 65 urteko pertsona batek egunerokotasunean kirol lasaia egin du. Berriro ere bihotzeko arazoren bat izango duela esango zenuke? Egin ezazu bere diagnostikoa, eta erabakia hartzeko erabili duzun paradigma justifika ezazu.

6. Gibeiko zirrosia izateko a priori probabilitatea 0.13koa da. Gaixotasunaren garapenean bi faktorek eragiten dute: adinak eta edari alkoholduen kontsumoak. Edari alkoholduen kontsumoa hiru mailatan neurtzen da: “bajua”, “normala” eta “altua”. Adina aldiz aldagai jarraia da, eta banaketa normalari jarraitzen diola suposatuko dugu.

Gibeiko zirrosia ez dutenen artean alkoholduen edarien kontsumoa “bajua” da %60-ko probabilitatez, eta “altua” %10-eko probabilitatez. Gibeiko zirrosia dutenen artean aldiz alkoholduen edarien kontsumoa “bajua” da %8-ko probabilitatez eta “altua” %70-eko probabilitatez.

Adinari dagokionez, gibeiko zirrosia ez dutenentzat 60 urteko itxaropen matematikoa eta 10 urteko desbideratze estandarra dituen banaketa normalari jarraitzen dio. Gibeiko zirrosia dutenen adinak 50 urteko itxaropen matematikoa eta 5 urteko desbideratze estandarra dituen banaketa normalari jarraitzen dio.

65 urte dituen eta alkoholduen edariak kantitate “normalean” hartzen dituen norbaiti diagnostikoa egin nahi diogu. Gibealeko zirrosia duela esango genuke? Zein probabilitatez? Erabakia hartzeko erabili duzun paradigma justifika ezazu.

7. 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).

Emandako datubasetik abiatuta aldagai iragarleen eta C klase-aldagaiaren arteko honako elkarrekiko informazioak estimatu dira:

$$I(X_1, C) = 0.2467, I(X_2, C) = 0.0292, I(X_3, C) = 0.1518, I(X_4, C) = 0.0481$$

C klase-aldagaiaren baldintzapean dauden elkarrekiko informazioak ere estimatu dira:

$$I(X_1, X_2|C) = 0.4196, \quad I(X_1, X_3|C) = 0.2228, \quad I(X_1, X_4|C) = 0.3118,$$

$$I(X_2, X_3|C) = 0.4196, \quad I(X_2, X_4|C) = 0.1689, \quad I(X_3, X_4|C) = 0.0611$$

- Egiazta ezazu $I(X_1, X_2|C) = 0.4196$ balioaren estimazioa.
- Kalkula ezazu Naïve Bayes eredua.
- Kalkula ezazu TAN algoritmoa aplikatuz lortzen den sailkapen-eredua.
- Kalkula ezazu sailkapen-eredua k -DB algoritmoa erabiliz, $k = 2$ baliorako.

8. 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 gaindituko den, bai ala ez

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

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

Emandako datubasetik abiatuta aldagai iragarleen eta C klase-aldagaiaren arteko honako elkarrekiko informazioak estimatu dira:

$$I(X_1, C) = 0, \quad I(X_2, C) = 0.0348, \quad I(X_3, C) = 0.1245, \quad I(X_4, C) = 0.1911$$

C klase-aldagaiaren baldintzapean dauden elkarrekiko informazioen estimazioak:

$$I(X_1, X_2|C) = 0.0623, \quad I(X_1, X_3|C) = 0.0623, \quad I(X_1, X_4|C) = 0.0174,$$

$$I(X_2, X_3|C) = 0.2464, \quad I(X_2, X_4|C) = 0.3842, \quad I(X_3, X_4|C) = 0.5016.$$

- Egiazta ezazu $I(X_1, X_2|C) = 0.0623$ balioaren estimazioa.
- Kalkula ezazu Naïve Bayes eredua.
- Kalkula ezazu TAN algoritmoa aplikatuz lortzen den sailkapen-eredua.
- Kalkula ezazu sailkapen-eredua k -DB algoritmoa erabiliz, $k = 2$ baliorako.

9. Emandako datubasetik abiatuta elkarrekiko informazio hauek estimatu dira:
 $I(X_1, C) = 0.4812$, $I(X_2, C) = 0.2075$, $I(X_3, C) = 0.6258$, $I(X_4, C) = 0.4591$
 C klase-aldagaiaren baldintzapean dauden elkarrekiko informazioen estimazioak:
 $I(X_1, X_2|C) = 0.0409$, $I(X_1, X_3|C) = 0.4779$, $I(X_1, X_4|C) = 0.7704$,
 $I(X_2, X_3|C) = 0.2075$, $I(X_2, X_4|C) = 0.1038$, $I(X_3, X_4|C) = 0.5409$.

Kasuak	X_1	X_2	X_3	X_4	C
$\mathbf{x}_1$	1	2	1	2	1
$\mathbf{x}_2$	2	2	1	1	1
$\mathbf{x}_3$	1	1	1	2	1
$\mathbf{x}_4$	2	1	2	1	1
$\mathbf{x}_5$	1	2	2	3	2
$\mathbf{x}_6$	1	2	2	3	2
$\mathbf{x}_7$	2	2	2	2	2
$\mathbf{x}_8$	3	2	3	2	2
$\mathbf{x}_9$	3	2	3	2	3
$\mathbf{x}_{10}$	2	2	3	2	3
$\mathbf{x}_{11}$	3	1	2	1	3
$\mathbf{x}_{12}$	3	2	2	1	3

- Egiazta ezazu $I(X_2, X_3|C) = 0.2075$ balioaren estimazioa.
 - Kalkula ezazu Naïve Bayes eredua.
 - Kalkula ezazu TAN algoritmoa aplikatuz lortzen den sailkapen-eredua.
 - Kalkula ezazu sailkapen-eredua k -DB algoritmoa erabiliz, $k = 2$ baliorako.
10. Emandako datubasetik abiatuta honako elkarrekiko informazioak estimatu dira:
 $I(X_1, C) = 0.7245$, $I(X_2, C) = 0.0490$, $I(X_3, C) = 0.6099$, $I(X_4, C) = 0$
 C klase-aldagaiaren baldintzapean dauden elkarrekiko informazioen estimazioak:
 $I(X_1, X_2|C) = 0.4464$, $I(X_1, X_3|C) = 0.1610$, $I(X_1, X_4|C) = 0.5219$,
 $I(X_2, X_3|C) = 0.1610$, $I(X_2, X_4|C) = 0.5219$, $I(X_3, X_4|C) = 0.0365$

Kasuak	X_1	X_2	X_3	X_4	C
$\mathbf{x}_1$	1	1	1	1	1
$\mathbf{x}_2$	1	2	1	1	1
$\mathbf{x}_3$	2	3	1	2	1
$\mathbf{x}_4$	4	1	1	1	1
$\mathbf{x}_5$	4	2	2	1	1
$\mathbf{x}_6$	3	3	2	1	2
$\mathbf{x}_7$	3	1	2	1	2
$\mathbf{x}_8$	3	2	2	1	2
$\mathbf{x}_9$	2	3	2	2	2
$\mathbf{x}_{10}$	2	1	2	1	2

- Egiazta ezazu $I(X_1, X_4|C) = 0.5219$ balioaren estimazioa.
- Kalkula ezazu Naïve Bayes eredua.
- Kalkula ezazu TAN algoritmoa aplikatuz lortzen den sailkapen-eredua.
- Kalkula ezazu sailkapen-eredua k -DB algoritmoa erabiliz, $k = 2$ baliorako.