

7. Aldagai Aukeraketa.

Ariketak

1. Izan bedi lau aldagai iragarleren bidez eta C klase-aldagaiaren bidez adierazitako 12 kasu dituen honako datubasea:

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

Aldagaien aukeraketa egokia egin nahi da, horietan oinarrituz zenbait sailkatzaile eraikitzeko. “Filter” izeneko teknika erabili nahi da.

- 1.1 Emandako datubasetik abiatuta, aldagai iragarleek C klase-aldagaiarekiko duten elkarrekiko informazio hauek estimatu dira: $I(X_1, C) = 0.4812$, $I(X_2, C) = 0.2075$, $I(X_3, C) = 0.6258$, $I(X_4, C) = 0.4591$. Informazio horren arabera, zein aldagai iragarle aukeratuko zenituzke? Zein ordenatan? Denak esanguratsuak direla esango al zenuke? Har itzazu hiru aukerarik onenak, eta eraiki ezazu 1-NN sailkatzailea. Ondoren, 4 geruzako balidazio gurutzatua erabiliz, asmatze-tasaren estimazioa kalkula ezazu, eta esan hiruren artetik zein den onena. Egiaztatu ezazu aldagai iragarle gehiago kontuan hartzeagatik ez dela beti sailkatzaile hobea lortuko. Ba al dago aldagai iragarle erredundanterik? Nola neurtuko zenuke informazio hori?
- 1.2 Aldagai iragarleak binaka hartuta, C klase-aldagaiaren baldintzapeko elkarrekiko informazioen estimazio hauek kalkulatu dira: $I(X_1, X_2|C) = 0.0409$, $I(X_1, X_3|C) = 0.4779$, $I(X_1, X_4|C) = 0.7704$, $I(X_2, X_3|C) = 0.2075$, $I(X_2, X_4|C) = 0.1038$, $I(X_3, X_4|C) = 0.5409$. Har itzazu, orain, elkarrekin hartuta informazio gehien ematen duten bi aldagai iragarleak. Eraiki ezazu 1-NN sailkatzailea, eta 4 geruzako balidazio gurutzatua erabiliz asmatze-tasaren estimazioa kalkula ezazu. Aldagai iragarleak banaka hartuta baino hobea atera al zaizu?
- 1.3 Lehenengo atalean egin dituzun hiru aldagai aukeraketak eta bigarren atalean egin duzuna kontuan hartuz, eraiki itzazu lau sailkapen-zuhaitzak.
- 1.4 Lehenengo atalean egin dituzun hiru aldagai aukeraketak eta bigarren atalean egin duzuna kontuan hartuz, eraiki itzazu lau Naive Bayes ereduak.
- 1.5 Har ezazu honako kasu berria: $\mathbf{x} = (2, 1, 2, 3)$. Egin ezazu kasu berri horrentzako C klase-aldagaiaren iragarpena 1-NN sailkatzailea erabiliz, eraiki dituzun lau sailkapen-zuhaitzak erabiliz eta eraiki dituzun lau Naive Bayes ereduak erabiliz. Naive Bayes ereduak erabiltzean, probabilitateen estimazioak kalkulatzekoan Laplaceren zuzenketa erabil ezazu.