

2. gaia. K -NN sailkatzaileak

Irakasgaia: Datu Meatzaritza
Espezialitatea: Konputazioa, hirugarren ikasmaila
Titulazioa: Informatika Ingeniaritzako Gradua
Konputazio Zientzia eta Adimen Artifiziala saila
Universidad del País Vasco-Euskal Herriko Unibertsitatea

2.1 Sarrera

Gai honetan K -NN (K -Nearest Neighbour) izenez ezaguna den sailkapenerako paradigma ikasiko dugu. Paradigma honen arabera, kasu berri bat sailkatzerakoan bere hurbileneko K auzokideen klaseak aztertu eta sarrien agertzen den klasea egokituko zaio kasu berriari. Hortaz, oso ideia sinplean eta intuitiboan oinarritzen da paradigma, eta inplementazioa erraza denez, sailkapeneterako oso paradigma erabilia da.

Hasteko, oinarritzko K -NN algoritmoa eta algoritmoaren zenbait aldaera azalduko dira gai honetan. Ondoren, kasuen aukeraketarako zenbait metodo aztertuko dira.

2.2 Oinarritzko K -NN algoritmoa

Gai honetan erabiliko den notazioa honakoa da (ikus 1. Irudia):

- D fitxategi bat da eta bertan N kasu daude. Kasu horietako bakoitza n aldagai iragarleen bidez, $X_1, \dots, X_n$, eta iragarri beharreko aldagaiaz, hau da C klase-aldagaiaz karakterizaturik dago.
- N kasuak honela adierazten dira,

$$\begin{aligned} (\mathbf{x}_1, c_1), \dots, (\mathbf{x}_N, c_N) & \quad \text{non} \\ \mathbf{x}_i &= (x_{i,1} \dots x_{i,n}) & i = 1, \dots, N \\ c_i &\in \{c^1, \dots, c^m\} & i = 1, \dots, N \end{aligned}$$

$c^1, \dots, c^m$ dira C klase-aldagaiaren m balio posibleak.

- Sailkatu nahi den kasu berria $\mathbf{x} = (x_1, \dots, x_n)$ moduan adierazten da.

D		X_1	$\dots$	X_j	$\dots$	X_n	C
$(\mathbf{x}_1, c_1)$	1	x_{11}	$\dots$	x_{1j}	$\dots$	x_{1n}	c_1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$(\mathbf{x}_i, c_i)$	i	x_{i1}	$\dots$	x_{ij}	$\dots$	x_{in}	c_i
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$(\mathbf{x}_N, c_N)$	N	x_{N1}	$\dots$	x_{Nj}	$\dots$	x_{Nn}	c_N
$\mathbf{x}$		x_1	$\dots$	x_j	$\dots$	x_n	?

Irudia 1: K -NN paradigmarako notazioa

2. Irudian oinarritzko K -NN sailkatzaileko sasikodea agertzen da. Bertan ikus daitekeen bezala, sailkatu nahi den $\mathbf{x}$ kasu berriaren eta dagoeneko sailkatuta dauden

kasu guztien arteko distantziak kalkulatzeko dira. Sailkatu nahi den $\mathbf{x}$ aldagaitik hurbilen dauden sailkatutako K kasuak aukeratu ondoren $D_{\mathbf{x}}^K$, K kasu horietarako C klase-aldagaia aztertu eta sarrien agertzen den klasea egokituko zaio sailkatu behar den $\mathbf{x}$ kasu berriari.

HASIERA

Sarrera: $D = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_N, c_N)\}$

$\mathbf{x} = (x_1, \dots, x_n)$ sailkatu beharreko kasu berria

Sailkatutako (x_i, c_i) objektu guztietarako

$d_i = d(\mathbf{x}_i, \mathbf{x})$ kalkulatu

Ordenatu $d_i (i = 1, \dots, N)$ txikienetik handienara

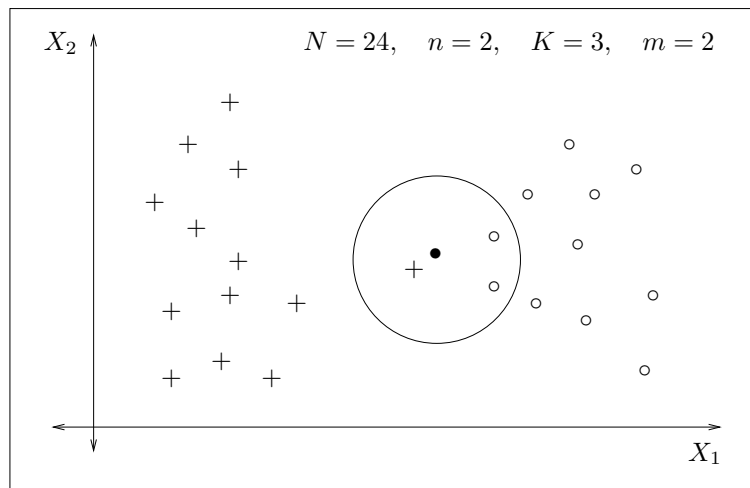
$\mathbf{x}$ kasutik hurbilen dauden sailkatutako K kasuak hartu, $D_{\mathbf{x}}^K$

$\mathbf{x}$ kasuari $D_{\mathbf{x}}^K$ multzoan sarrien agertzen den klasea esleitu

AMAIERA

Irudia 2: K -NN sailkatzaileko sasikodea

Aurrekoaren adibide grafiko bat 3. Irudian ikus daiteke.



Irudia 3: Oinarrizko K -NN algoritmoaren aplikaziorako adibidea

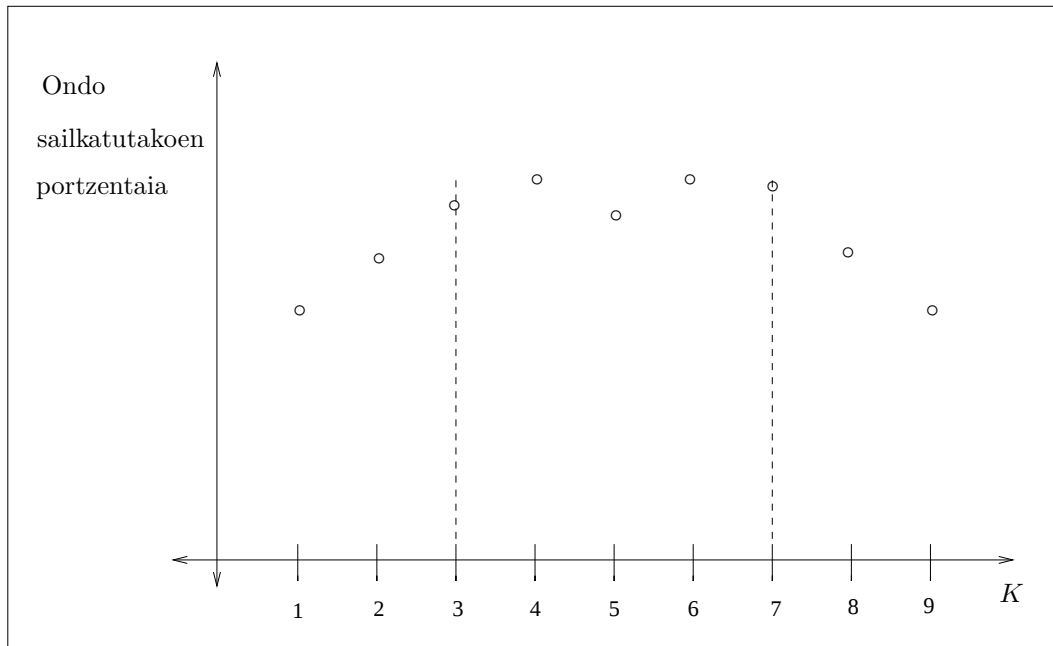
3. Irudian ikus daitekeen bezala, sailkatutako $N = 24$ kasu ditugu bi klasetan sailkatuak ($m = 2$). Aldagai iragarleak X_1 eta X_2 dira, eta $K = 3$ aukeratu da. Sailkatu beharreko $\mathbf{x}$ kasu berriari (irudian $\bullet$ moduan adierazia) hurbilen geratzen zaizkion sailkatutako 3 kasuak aztertuz, bi kasu o klasekoak direla ikusten da. Hortaz, 3-NN sailkatzaileak o klasea aukeratzeko du kasu berriari esleitzeko. Esan behar da $\mathbf{x}$ kasu berriak hurbilen geratzen den sailkatutako kasua $+$ klasekoa dela, eta ondorioz, sailkapena burutzeko 1-NN sailkatzailea aukeratu izan bagenu, $\mathbf{x}$ kasuari $+$ klasea esleitzeko geniokeela.

K -NN paradigma gainontzeko sailkapen-paradigmekin konparatuz gero, ez-ohikoa dela esan daiteke, gainontzeko paradigmekin kasu berri baten sailkapena bi urratsetan burutzen den bitartean (eredu sailkatzailearen *indukzioa* eta ondoren kasu berriaren gain aplikatzea edo *dedukzioa*), K -NN paradigmak bi urratsak bakar batean elkartzen dira; *transindukzio* izena eman ohi zaio urrats bakar horri. K -NN paradigmaren eredu

espliziturik existitzen ez delako gertatzen da hori.

Bi klase edo gehiagoren artean *berdinketa* gertatzen denerako *erregla heuristiko* bat izatea komeni da klaseen artean erabakitzeke. Honakoak dira erregela heuristiko batzuren adibideak: hurbileneko auzokidea duen klasea aukeratzea, batazbesteko distantzia txikiena duen klasea aukeratzea, eta abar.

K balioaren aukeraketa ere garrantzia duen kontua da. Enpirikoki egiazta daiteke ondo sailkatutako kasuen portzentaia ez dela K balioarekiko monotonoa (ikus 4. Irudia). Oro har K balioa 3 eta 7 balioen artean aukeratzea komeni dela esan ohi da.



Irudia 4: Ondo sailkatutakoen portzentaia K balioaren funtzioan. Erlazio ez monotonorako adibidea

2.3 Oinarrizko algoritmoaren aldaerak

Atal honetan oinarrizko algoritmoaren inguruko zenbait aldaera azalduko ditugu.

2.3.1 K-NN bermegabeak baztertuz

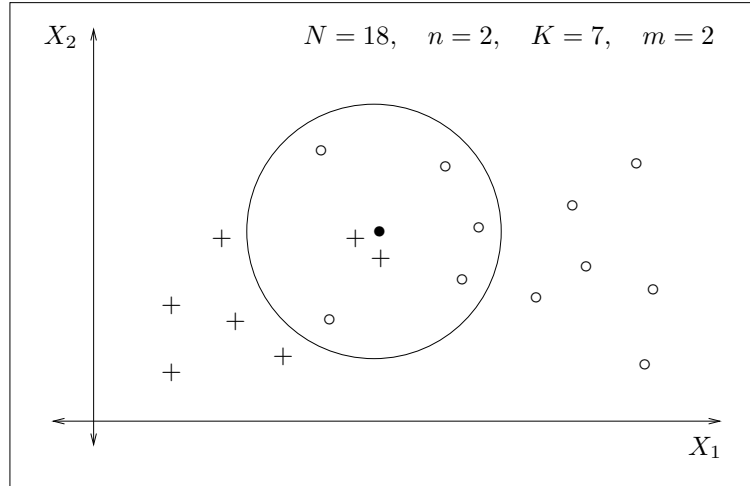
K -NN algoritmoa bermegabeak baztertuz aplikatzen denean kasu berri bat sailkatu ahal izateko nahikoa berme izan behar da. Hori dela eta, sailkatu beharreko kasuren bat sailkatu gabe gera daiteke, kasu horri esleituko zaion klasea zuzena izateko berme nahikorik ez badago.

Hona hemen sailkapena berme nahikoarekin egiteko bi adibide:

- Sailkatu beharreko kasu berriari esleituko zaion klaseak *aurretik ezarritako atalaseak* eskatu adina botu jaso beharko ditu. $K = 10$ dela suposatuz eta $m = 2$ izanik, atalasea 6 izan daiteke.
- Esleitu beharreko klaserako *gehiengo absolutu*-ren bat ezartzea. Adibidez, $K = 20$ eta $m = 4$ badira erabaki daiteke kasu berriari klase bat esleitu ahal izango zaiola sarrien agertu den klasearen eta hurrengoaren arteko diferentzia gutxienez 3koa bada.

2.3.2 K-NN batazbesteko distantziarekin

K -NN batazbesteko distantziarekin erabiltzen denean kasu berri bati batazbesteko distantziarik txikieneko klasea esleitzen zaio. Horrela, 5. Irudiko adibidean, sailkatu beharreko kasutik hurbilen dauden 7 kasuetatik 5 kasu $\circ$ klasekoak badira ere, $+$ klasea esleituko zaio, $+$ klaseko bi kasuetara dagoen batazbesteko distantzia $\circ$ klaseko 5 kasuetara dagoen batazbesteko distantzia baino txikiagoa delako.



Irudia 5: K-NN batazbesteko distantziarekin erabiltzen denerako adibidea

2.3.3 K-NN distantzia minimoarekin

K -NN distantzia minimoarekin erabiltzen denean, hasteko klase bakoitzeko kasu bat aukeratu behar da, klaseko ordezkari lanak egingo dituen, normalean klaseko elementu guztien barizentrutik hurbilen dagoen kasua. Urrats honetan, gorde beharreko kasuen fitxategiaren dimentsioa N izatetik m izatera murrizten da. Ondoren, kasu berriari hurbilen dagoen ordezkariaren klasea esleituko zaio.

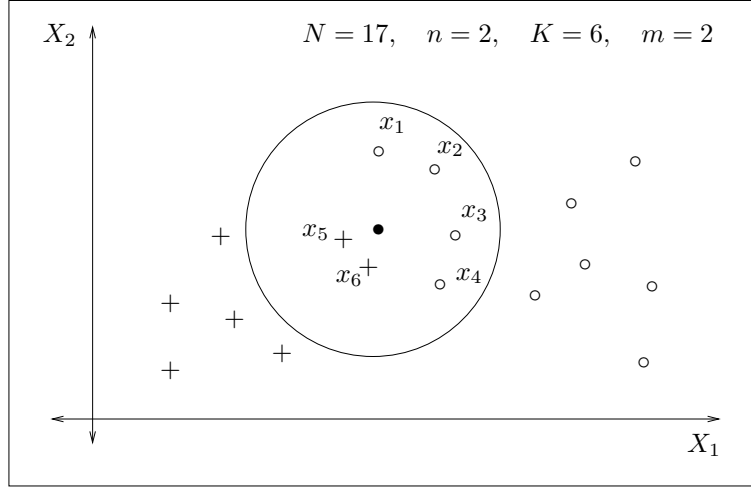
Prozedura hori 1-NN algoritmo baten moduan ikus daiteke, klase bakoitzeko kasu bat duen m kasuko multzo bati aplikatuta. Kostu konputazionala K -NN orokorrarena baino txikiagoa da, baina bere eraginkortasuna klase barruko homogenotasunaren mende dago; zenbat eta homogenotasun handiagoa egon klase barruan orduan eta eraginkorragoa izango da prozedura.

2.3.4 K-NN aukeratutako auzokideen pisaketarekin

Algoritmo honetan sailkatutako kasuen artetik sailkatu beharreko $\mathbf{x}$ kasu berrira hurbilen egoteagatik aukeratutako K kasuak ez dira berdin kontatzen. Aukeratutako kasu bakoitzari pisu bat egokituko zaio, pisu hori sailkatu beharreko kasu berrira dagoen distantziaren mende egongo delarik.

Adibide moduan, demagun aukeratutako kasuei esleituko diegun pisua kasutik sailkatu beharreko $\mathbf{x}$ kasura dagoen distantziaren alderantzizko proportzionala erabiltzen dugula. 6. Irudian agertzen dira $K = 6$ izanik aukeratuak izan diren $x_i, i = 1, \dots, 6$ kasuak, eta honako taulan agertzen dira 6 kasu horietatik sailkatu beharreko $\mathbf{x}$ kasura dauden $d(\mathbf{x}_i, \mathbf{x})$ distantziak, eta distantzia horiekiko modu alderantzizko proportzionalan kalkulaturako w_i pisuak.

x_i	$d(\mathbf{x}_i, \mathbf{x})$	w_i
$\mathbf{x}_1$	2	0.5
$\mathbf{x}_2$	2	0.5
$\mathbf{x}_3$	2	0.5
$\mathbf{x}_4$	2	0.5
$\mathbf{x}_5$	0.7	1/0.7
$\mathbf{x}_6$	0.8	1/0.8



Irudia 6: K -NN aukeratutako kasuen pisaketarekin erabiltzen denerako adibidea

Aukeratutako 6 kasuen artean, $\circ$ klasekoak diren x_1, x_2, x_3 eta x_4 kasuak sailkatu beharreko kasutik 2 distantziara daudenez, kasu horietako bakoitzari $w_i = 1/2 = 0,5$ pisua esleituko zaio. x_5 eta x_6 kasuak aldiz sailkatu beharreko kasura 0,7 eta 0,8 distantziara daude, eta $w_5 = 1/0,7$ eta $w_6 = 1/0,8$ pisuak esleituko zaizkie. $\circ$ klasearen eta $+$ klasearen artean aukeratzeko, pisuen batura egin:

$$\begin{aligned} \circ \text{ klasea} &\rightarrow w_1 + w_2 + w_3 + w_4 = 0,5 + 0,5 + 0,5 + 0,5 = 2 \\ + \text{ klasea} &\rightarrow w_5 + w_6 = \frac{1}{0,7} + \frac{1}{0,8} = 2,678 \end{aligned}$$

K -NN aukeratutako kasuen pisaketarekin erabiliz gero sailkatu beharreko kasu berriari $+$ klasea esleituko zaio, aukeratutako kasuen pisuen batura handiagoa delako.

2.3.5 K -NN aldagaien pisaketarekin

Orain arte ikusitako K -NN algoritmoaren aldaera guztietan, sailkatu beharreko $\mathbf{x}$ kasu berriaren eta sailkatutako kasuen D fitxategiko $\mathbf{x}_r, r = 1, \dots, N$ kasu bakoitzaren arteko distantzia kalkulatzeko $X_1, \dots, X_n$ aldagaiei pisu berbera eman izan zaie. Distantzia hori kalkulatzeko distantzia euklidearra erabiliz gero, adibidez, $\mathbf{x}$ kasuaren eta $\mathbf{x}_r$ kasuaren arteko $d(\mathbf{x}, \mathbf{x}_r)$ distantzia honela kalkulatu da:

$$d(\mathbf{x}, \mathbf{x}_r) = \sum_{j=1}^n (x_j, x_{rj})^2$$

Aldagai guztiei garrantzia berbera ematea zenbait kasutan ez da egokia gertatzen K -NN algoritmoa erabiltzerakoan, gerta baitaiteke aldagairen batek garrantziarik ez

izatea C klase-aldagairako. Horrelakoetan aldagai bakoitza modu egokian ponderatuko duen distantzia bat erabiltzea komeniko da. Honako hau, adibidez, $\mathbf{x}$ kasu berriaren eta $\mathbf{x}_r$ kasuaren arteko distantzia ponderatua kalkulatzeko erabiliko da.

$$d(\mathbf{x}, \mathbf{x}_r) = \sum_{j=1}^n w_j (x_j, x_{rj})^2$$

Distantzia horretan X_j aldagaiak w_j pisua du egokituta. Pisu hori aurretik finkatu egin behar da.

Ikus dezagun adibide bat. Demagun ondorengo taulako datuak ditugula:

X_1	X_2	C
0	0	1
0	0	1
0	0	1
1	0	1
1	0	1
1	1	1
0	1	0
0	1	0
0	1	0
1	1	0
1	1	0
1	0	0

Taula aztertuz ikus daiteke X_1 aldagaia ez dela esanguratsua C aldagairako, X_1 aldagaiak 0 balioa hartzen duen sei aldietatik hirutan C aldagaiak 1 balioa hartzen duelako eta beste hirutan 0 balioa; X_1 aldagaiak 1 balioa hartzen duen sei aldiak aztertzen baditugu, hirutan C aldagaiak 1 balioa hartzen du eta beste hirutan 0 balioa. X_2 aldagaiarekin aldiz, egoera zeharo kontrakoa da; X_2 aldagaiak 0 balioa hartzen duen sei aldietatik bost alditan C aldagaiak 1 balioa hartzen du, eta behin hartzen du 0 balioa. X_2 aldagaiak 1 balioa hartzen duen sei aldietatik bost alditan C aldagaiak 0 balioa hartzen du eta behin bakarrik hartzen du 1 balioa. Hortaz, distantziak kalkulatzeko nahiko nabarmen ikusten da X_2 aldagaiaren ekarpenari pisu handiagoa eman behar zaiola X_1 aldagaiaren ekarpenari baino, X_2 aldagaiaren balioak garrantzia handia duelako C aldagaiaren baliorako, X_1 aldagaiarenak aldiz ez.

X_i aldagai bakoitzari egokituko zaion w_i pisua kalkulatzeko modu bat X_i aldagaiaren eta C klase-aldagaiaren $I(X_i, C)$ elkarrekiko informazioaren neurria erabiltzea da, $i = 1 \dots n$. Neurria honela definitzen da:

$$I(X_i, C) = \sum_{x_i, c} p_{(X_i, C)}(x_i, c) \log_2 \frac{p_{(X_i, C)}(x_i, c)}{p_{X_i}(x_i) \cdot p_C(c)}$$

Bi aldagaien arteko elkarrekiko informazioaren neurria honela interpretatzen da: aldagai baten balioa ezagutzeak zer neurritan txikitzen duen beste aldagaiaren balioari buruz dugun ziurgabetasuna; zenbat eta handiagoa izan bi aldagaien arteko elkarrekiko informazioaren neurria, orduan eta mendekotasun handiagoa dago beren artean.

Atal honetan elkarrekiko informazioaren neurria erabiliko dugu X_i aldagaiari egokitu beharreko w_i pisua kalkulatzeko. Pisu hori X_i aldagaiaren eta C aldagaiaren arteko

$I(X_i, C)$ elkarrekiko informazioaren neurriarekiko proportzionala izango da. Taulako datuekin jarraituz, $I(X_1, C)$ eta $I(X_2, C)$ kalkulatu ditugu.

$$\begin{aligned}
I(X_1, C) &= p_{(X_1, C)}(0, 0) \log_2 \frac{p_{(X_1, C)}(0, 0)}{p_{X_1}(0) \cdot p_C(0)} + p_{(X_1, C)}(0, 1) \log_2 \frac{p_{(X_1, C)}(0, 1)}{p_{X_1}(0) \cdot p_C(1)} + \\
&\quad p_{(X_1, C)}(1, 0) \log_2 \frac{p_{(X_1, C)}(1, 0)}{p_{X_1}(1) \cdot p_C(0)} + p_{(X_1, C)}(1, 1) \log_2 \frac{p_{(X_1, C)}(1, 1)}{p_{X_1}(1) \cdot p_C(1)} = \\
&\quad \frac{3}{12} \log_2 \frac{\frac{3}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{3}{12} \log_2 \frac{\frac{3}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{3}{12} \log_2 \frac{\frac{3}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{3}{12} \log_2 \frac{\frac{3}{12}}{\frac{6}{12} \cdot \frac{6}{12}} = 0 \\
I(X_2, C) &= p_{(X_2, C)}(0, 0) \log_2 \frac{p_{(X_2, C)}(0, 0)}{p_{X_2}(0) \cdot p_C(0)} + p_{(X_2, C)}(0, 1) \log_2 \frac{p_{(X_2, C)}(0, 1)}{p_{X_2}(0) \cdot p_C(1)} + \\
&\quad p_{(X_2, C)}(1, 0) \log_2 \frac{p_{(X_2, C)}(1, 0)}{p_{X_2}(1) \cdot p_C(0)} + p_{(X_2, C)}(1, 1) \log_2 \frac{p_{(X_2, C)}(1, 1)}{p_{X_2}(1) \cdot p_C(1)} = \\
&\quad \frac{1}{12} \log_2 \frac{\frac{1}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{5}{12} \log_2 \frac{\frac{5}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{5}{12} \log_2 \frac{\frac{5}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{1}{12} \log_2 \frac{\frac{1}{12}}{\frac{6}{12} \cdot \frac{6}{12}}
\end{aligned}$$

2.4 Kasuen hasierako fitxategia murriztea

K -NN sailkatzaileak duen kostu konputazionalak kasuen fitxategian dagoen kasu kopuruarekiko, N -rekiko alegia, mendekotasun handia du. Mendekotasun hori arintzeko asmoz diseinatu diren teknikak ikusiko ditugu atal honetan. Kasuen hasierako fitxategia murrizteko metodoak bi motatakoak dira: *edizio-teknikak* eta *kondentsazio-teknikak*.

Edizio-tekniketan, hasierako fitxategitik kasu batzuk ezabatuz lortzen da kasuen azpimultzoa. Kondentsazio-tekniketan aldiz, zenbait kasu aukeratuz lortzen da kasuen azpimultzoa hasierako fitxategitik.

2.4.1 Wilson-en edizioa

Wilson-ek proposatutako edizioaren arabera, kasuen hasierako fitxategiko kasu bakoitza aztertuko da. Egiaztatu nahi dena da ea fitxategitik kasua kenduz eta gainontzeko kasuak erabiliz K -NN sailkatzaile bat aplikatuz, kendutako kasuari egokitzen zaion klasea bere benetako klasea ote den. Hori hala ez bada, kasua ezabatua izango da. Prozedura errepikatuko da fitxategiko kasu bakoitzerako. Neurri batean prozedura honek *leave-one-out* balidazioarekin badu antzekotasunik.

Wilson-en edizio errepikakorra ere kontsidera daiteke, eta prozedura gelditu segidako bi aukeraketatan aldaketarik gertatzen ez bada.

2.4.2 Hart-en kondentsazioa

Hart-en kondentsazioak arraro moduan har daitezkeen kasuen aukeraketa bat egiten du.

Horretarako, kasu bakoitzerako eta kasuak fitxategian gordeta dauden ordenari jarraituz, K -NN sailkatzaile bat eraikiko da aztergai den kasuaren aurretik dauden kasuekin bakarrik. Kasuak klase moduan egokituta duen balioa eta K -NN sailkatzaileak egokitu liokena desberdinak badira orduan kasua aukeratua izango da. Baina kasuaren benetako klasea eta K -NN sailkatzaileak egokitu lioken klasea berdinak izatekotan, kasua ez da aukeratua izango.

Garbi dago fitxategian kasuak agertzen direneko ordenaren mendekoa dela Hart-en kondentsazioaren metodoa.

Erreferentziak

1. D. Aha, D. Kibler, M.K. Albert (1991). Instance-based learning algorithms, *Machine Learning* **6**, 37-66
2. S. Cost, S. Salzberg (1993). A weighted nearest neighbour algorithm for learning with symbolic features, *Machine Learning* **10(1)**, 57-78
3. B.V. Dasarathy (1991). *Nearest Neighbour (NN) Norms: NN Pattern Recognition Techniques*, IEEE Computer Society Press
4. E. Fix, J.L.Hodges Jr. (1951). Discriminatory analysis, nonparametric discrimination. *Project 21-49-004, Rept. 4*, USAF School of Aviation Medicine, Randolph Field
5. P.E. Hart (1968). The condensed nearest neighbour rule, *IEEE Transactions on Information Theory*, **IT-14**, 515-516
6. D.L. Wilson (1972). Asymptotic properties of nearest neighbour rules using edited data. *IEEE Transactions on Systems, Man and Cybernetics*, Vol 2, 408-421