

2. K-NN sailkatzaileak

Irakasgaia: Konputazio-Zientzien Metodo Matematikoak

Titulazioa: **Informatika Ingeniaria**

Konputazio Zientzia eta Adimen Artifiziala saila

Universidad del País Vasco - Euskal Herriko Unibertsitatea

- **K-NN**: (*K-Nearest Neighbour*)
- Kasu berri bat sailkatzerakoan bere **hurbileneko K auzokideen klaseen artean sarrien agertzen den klasea egokituko zaio**.
- Ideia simple eta intuitiboa. Implementazio erraza
- Ez dago eredu espliziturik. Distantzien kalkuluan oinarritzen da (ikus <http://es.wikipedia.org/wiki/Distancia>)

Notazioa

D datubasea, $X_1, \dots, X_n$ aldagai iragarleak, C klase-aldagaia, N kasu, $\mathbf{x}$ kasu berria.

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

Oinarrizko K -NN algoritmoa

Algoritmoaren sasikodea

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

ERREPIKATU sailkatutako (x_i, c_i) guztietarako

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

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

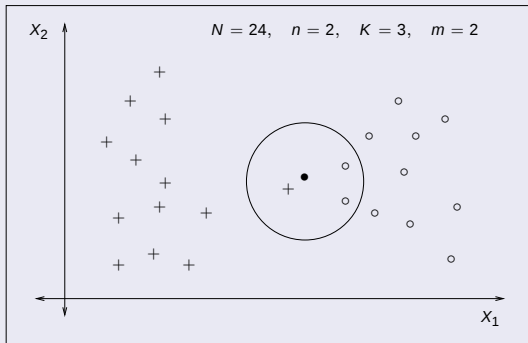
$\mathbf{x}$ -tik hurbilen dauden K kasuekin geratu, $D_{\mathbf{x}}^K$

$\mathbf{x}$ -ri $D_{\mathbf{x}}^K$ -ko sarrieneko klasea esleitu

AMAIERA

Oinarrizko K -NN algoritmoa

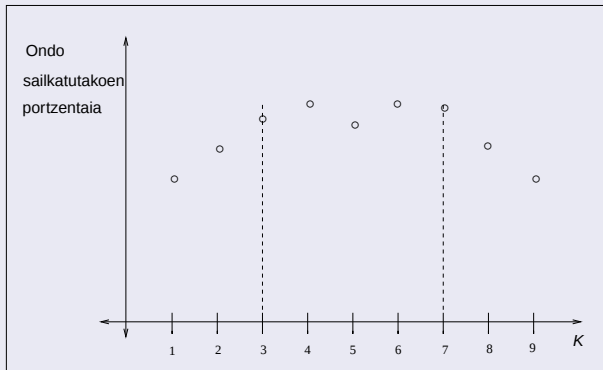
Adibide grafikoa: 3-NN, 24 kasu, 2 klase, 2 aldagai iragarle



3 auzokide hurbilenen artetik bi $\circ$ klasekoak, bat $+$ klasekoa
3-NN sailkatzaile honek $\circ$ iragarriko du $\bullet$ kasuarentzat
($K=1$ balitz?)

Oinarrizko K -NN algoritmoa

K parametroaren aukeraketa



Erlazioa ez da monotonoa. K bakoitia aukeratzea komeni da

K-NN algoritmoaren aldaerak

- 1 K-NN bermegabeak baztertuz
- 2 K-NN batazbesteko distantziarekin
- 3 K-NN distantzia minimoarekin
- 4 K-NN aukeratutako auzokideen pisaketarekin
- 5 K-NN aldagaien pisaketarekin

1. Aldaera: K-NN bermegabeak baztertuz

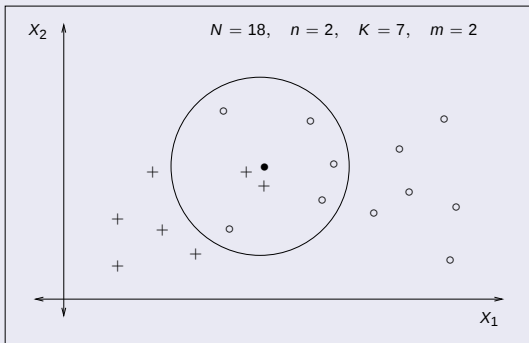
K-NN bermegabeak baztertuz

- Kasu berri bat sailkatzeko **nahikoa berme** izan behar da
- Ez izatekotan, **kasua sailkatu gabe utz daiteke**
- Aurretik **atalase bat** ezartzen da
- Gehiengo absolutua

2. Aldaera: K-NN batzbesteko distantziarekin

Batzbesteko distantziarik txikieneko klasea esleitu

Adibidea. 7 auzokide hurbilenak hartuz



- kasu berriari + klasea esleituko zaio

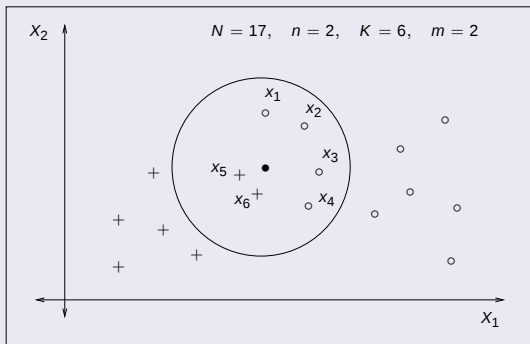
3. Aldaera: K-NN distantzia minimoarekin

- Klase bakoitzerako ordezkari bat aukeratu (adb. klaseko barizentrutik hurbilen dagoen kasua)
- Gorde beharreko kasuen fitxategiaren dimentsioa N -tik m -ra murriztuko da
- 1-NN algoritmoa aplikatu murriztutako fitxategiari
- Eraginkortasuna klase barruko homogenotasunaren mende dago
- Klase bateko kasuak multzo edo cluster batean baino gehiagotan banatuta badaude, aldaera ez da egokia gertatzen, ordezkaria ez delako klasearen adierazgarri egokia izango

4. Aldaera: K-NN auzokideen pisaketarekin

- Kasu guztiek ez dute garrantzia bera. **Auzokide hurbilenei pisu bat egokitu** (distantziaren mendekoa, adibidez)
- Pisuen baturan oinarrituz emango da klasearen iragarpena

Adibidea



K-NN auzokideen pisaketarekin. Adibidea

- Pisua distantziaren alderantzizko proportzionala

Adibidea

Auzokide hurbilenak	Klasea C	Distantzia $d(\mathbf{x}_i, \mathbf{x})$	Pisua w_i
$\mathbf{x}_1$	O	2	$\frac{1}{2} = 0,5$
$\mathbf{x}_2$	O	2	$\frac{1}{2} = 0,5$
$\mathbf{x}_3$	O	2	$\frac{1}{2} = 0,5$
$\mathbf{x}_4$	O	2	$\frac{1}{2} = 0,5$
$\mathbf{x}_5$	+	0,7	$1/0,7 = 1,42$
$\mathbf{x}_6$	+	0,8	$1/0,8 = 1,25$

$1,42 + 1,25 > 0,5 + 0,5 + 0,5 + 0,5 \rightarrow +$ klasea iragarriko da

5. Aldaera: K-NN aldagaien pisaketarekin

Aldagai iragarleen pisaketa

- Aldagai guztiek C klase-aldagaiaren iragarpenerako garrantzia bera ez dutenean erabiltzea komeni da
- Distantzia eta distantzia ponderatua:

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

- X_j aldagaiari w_j pisua egokitu, X_j aldagaiaren eta C klase-aldagaiaren $I(X_j, C)$ elkarrekiko informazioaren neurria erabiliz, adibidez. Informazio-teoria

Kasuen aukeraketa

- Databaseetatik kasuen aukeraketa egin daiteke, sailkatzaile **azkarragoak** eta **hobeak** lortzeko.
 - Distantzien kalkuluan oinarritutako K-NN teknikak **kostu konputazional handia** izaten dute database handietarako.
 - Databasean **gaizki etiketatutako kasuak** badaude, K-NN sailkatzailearen **asmatze-tasa jaitsi** egingo da.
- Bi teknika: edizio-teknikak eta kondentsazio-teknikak.
 - **Edizio-teknikak**: databasean zarata sortzen duten kasuak kenduz, erabaki-muga leunak sortzen saiatzen dira
 - **Kondentsazio-teknikak**: jatorrizko databaseko kasuetatik induzitutako erabaki-mugak mantentzen saiatzen dira
- Database murriztua erabiliko da sailkatzailea sortzeko

Ediziorako-teknikak: Wilson-en edizioa

- Klaseen arteko mugak garbitu, gainjarritako kasuak kenduz
- K-NN sailkatzaileak gaizki sailkatuko lituzkeen kasuak datubasetik kendu

Wilson-en ediziorako sasikodea

- Errepikatu i guztietarako
 - Datubaseko $\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_N$ kasuen artean K auzokide hurbilenak aurkitu
 - $\mathbf{x}_i$ kasuari bere auzokide hurbilenen artean sarrien agertzen den klasea esleitu
- Esleitutako klasea eta klase erreala $\text{bat ez datozen kasu guztiak datubasetik ezabatu}$

Kondentsazio-teknikak: Hart-en kondentsazioa

Hart-en kondentsaziorako sasikodea

- 1 Datubaseko lehenengo kasua **STORE** zerrendan gorde
- 2 Datubaseko bigarren kasuaren iragarpena egin **STORE**ko kasuetan oinarrituz. Sailkapena zuzena bada, **GRABBAG** zerrendan sartu, bestela **STORE**n sartu
- 3 Modu berean jarraitu datubaseko kasu guztiekin
- 4 Datubaseko kasu guztiak aztertu ondoren, **GRABBAG** zerrendakoekin prozedura errepikatu behin eta berriz. Prozesuaren amaiera bi modutara lor daiteke:
 - **GRABBAG** zerrenda hutsa dago (kasu guztiak **STORE** zerrendan daude)
 - **GRABBAG** zerrendako kasu guztiak aztertu ondoren bat bera ere **STORE** zerrendara ez da pasa
- 5 **GRABBAG**-eko kasuak datubasetik ezabatu

Oinarrizko bibliografia

Wikipedia

- http://en.wikipedia.org/wiki/K-nearest-neighbor_algorithm

Liburua

- **Aprendizaje Automático: conceptos básicos y avanzados**
 - **Capítulo 3: Algoritmos de clasificación por vecindad**
- Koordinatzailea: Basilio Sierra Araujo
- Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad del País Vasco
- Argitaratzailea: Pearson Prentice Hall, 2004
- ISBN 10: 84-8322-318-X, ISBN 13: 978-84-8322-318-5