

5. Sailkatzaile Bayestarrak

Irakasgaia: Datu Meatzaritza

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

Sailkatzaile Bayestarrak

Sarrera

- Probabilitate teorian (**Bayes-en teorema**) oinarritutako sailkatzaileak, gainbegiraturako sailkapenerako.
- **Abantailak**: Kasu berri bat izanik, klase-aldagaiaren iragarpenari probabilitate bat egokituko zaio. Beste sailkatzaileekin lortutako emaitzen parekoak lortzen dira
- **Eragozpenak**: Konputazionalki garestiak izan daitezke, probabilitate asko estimatu beharko direlako
- Gai honetan aztertuko ditugunak:
 - **Naïve Bayes**
 - Zuhaitzera zabaldutako Naïve Bayes (**Tree Augmented Naïve Bayes, TAN**), (Friedman eta abar, 1997)
 - k -mendekotasuneko sailkatzaile Bayestarra (**k -Dependence Bayesian classifier, k -DB**) (Sahami, 1996)

Bayes-en teorema

C iragarri beharreko hipotesia eta X data ezaguna izanik,

$$p(C|X) = \frac{p(C)p(X|C)}{p(X)} = \frac{p(C)p(X|C)}{\sum_{j=1}^m p(c_j)p(X|c_j)}$$

- $p(C)$: C klase-aldagaiaren **a priori probabilitatea**
- $p(X)$: X dataren **a priori probabilitatea**
- $p(X|C)$: X-ren probabilitate baldintzatua C emanda ; X-ren **a posteriori probabilitatea**
- $p(C|X)$: C-ren probabilitate baldintzatua, X emanda; C-ren **a posteriori probabilitatea**

Eredu probabilistikotik sailkatzailea eraikitzea

Bayes-en teorema sailkapen gainbegiraturako: C klase-aldagaiaren a posteriori probabilitate maximoko balioa

A posteriori probabilitate maximoa: MAP

$$c^* = \arg \max_{c_j \in C} p(c_j|X) = \arg \max_{c_j \in C} \frac{p(c_j)p(X|c_j)}{\sum_{j=1}^m p(c_j)p(X|c_j)}$$

- C klase-aldagaiaren (hipotesiaren) balio posibleak $c_1, \dots, c_m$ izanik, c_j bakoitzerako $p(c_j|X)$ kalkulatu
- $p(c_j|X)$ handieneko c^* klasea iragarriko da
- $P(X) = \sum_{j=1}^m p(c_j)p(X|c_j) \rightarrow$ hipotesiekiko independentea

Parametroen estimazioa

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

$x=(x_1, x_2, \dots, x_n)$ kasu berriaren sailkapenerako:

$$c^* = \arg \max_{c_j=c_1, \dots, c_m} p(c_j | x_1, x_2, \dots, x_n) = \arg \max_{c_j=c_1, \dots, c_m} p(c_j) p(x_1, x_2, \dots, x_n | c_j)$$

- ▶ $p(c_j) \rightarrow$ datubasean c_j klasearen maiztasuna
- ▶ $p(x_1, x_2, \dots, x_n | c_j) \rightarrow$ datubasea oso oso handia den kasuan bakarrik estima daiteke

Estimazio kopuru handiegia...

Naïve Bayes sailkatzailea

Eredu probabilistikoaren sinplifikazioa

C klase-aldagaiaren baldintzapean **aldagai iragarleak independente** direla suposatzen badugu,

$$p(x_1, x_2, \dots, x_n | c_j) = \prod_{i=1}^n p(x_i | c_j)$$

Naïve Bayes sailkatzailea

$$c^* = \arg \max_{c_j=c_1, \dots, c_m} p(c_j | x_1, x_2, \dots, x_n) = \arg \max_{c_j=c_1, \dots, c_m} p(c_j) \prod_{i=1}^n p(x_i | c_j)$$

Naïve Bayes sailkatzailea

Oharrak

- ▶ Estimatu beharreko **parametro kopurua izugarri jait**si da
- ▶ Eta, C klase-aldagaiaren baldintzapean aldagai iragarleak **ez badira independente?**

$$p(x_1, x_2, \dots, x_n | c_j) \neq \prod_{i=1}^n p(x_i | c_j)$$

Hala ere, Naïve Bayes-ek emaitza onak ematen ditu...

- ▶ Eta, $p(x_i | c_j)$ probabilitatearen estimazioa kalkulatzeko, datubasean **c_j klaseko kasuen artean $X_i = x_i$ kasurik ez badago?**

$$\hat{p}(x_i | c_j) = 0 \rightarrow \hat{p}(c_j) \prod_{i=1}^n \hat{p}(x_i | c_j) = 0$$

Naïve Bayes sailkatzailea

Oharrak

$\hat{p}(x_i | c_j) = 0$ arazoa ekiditeko, estimazioa horrela kalkulatu:

$$\hat{p}(x_i | c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k}$$

non

- ▶ $N(C = c_j)$: c_j klaseko kasu kopurua
- ▶ $N(X_i = x_i, C = c_j)$: c_j klaseko kasuen artean zenbatek duten $X_i = x_i$
- ▶ k : X_i aldagai iragarlearen balio kopurua

Naïve Bayes sailkatzailea

Oharrak

- Eta, **aldagaiak jarraiak** badira? Honakoa bilatzen da:

$$c^* = \arg \max_c p(C = c) \prod_{i=1}^n f_{X_i|C=c}(X_i|c)$$

non $f_{X_i|C=c}(X_i|c)$ probabilitate-banaketa normala izan ohi den:

$$f_{X_i|C=c}(X_i|c) \rightsquigarrow \mathcal{N}(X_i; \mu_i, \sigma_i)$$

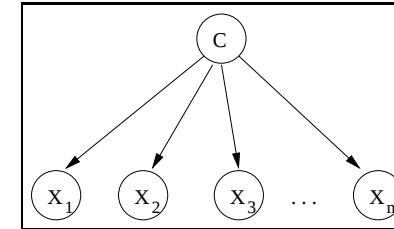
μ_i itxaropen matematikoa; σ_i desbideratze estandarra.

$$c^* = \arg \max_c p(C = c) \prod_{i=1}^n \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2}$$

Naïve Bayes sailkatzailearen egitura

Adierazpen grafikoa

C klase-aldagaiaren baldintzapean **aldagai iragarleak independente** direla suposatzen denez,



$$p(c_j | X_1, \dots, X_n) \propto p(c_j) \prod_{i=1}^n p(X_i | c_j)$$

Zuhaitzera zabaldutako Naïve Bayes (TAN)

Tree Augmented Naïve Bayes (TAN)

- Naïve Bayes sailkatzailearen hedapen bat da.
- C klase-aldagaiaren baldintzapean **aldagai iragarleen arteko mendekotasuna** egon daitekeela aintzakotzat hartzen da; elkarrekiko informazioa erabiltzen da

$$I(X, Y|C) = \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^r p(X_i, Y_j, C_k) \log_2 \frac{p(X_i, Y_j, C_k)}{p(X_i, C_k) \cdot p(Y_j, C_k)}$$

- Mendekotasunak **zuhaitz egitura** batera mugatzen dira; sinplea izaten jarraitzen du

Zuhaitzera zabaldutako Naïve Bayes (TAN)

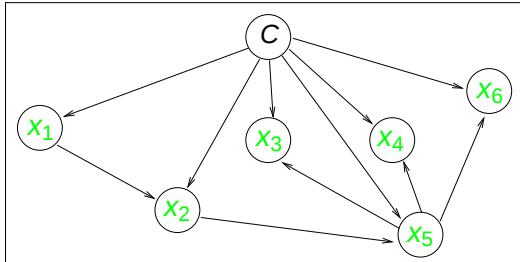
TAN eredia sortzeko algoritmoa

- urratsa.** $I(X_i, X_j | C)$ kalkulatu $i < j$ aldagaietarako $i, j = 1, \dots, n$.
- urratsa.** Grafo bat definitu. Erpinak $X_1, \dots, X_n$ aldagai iragarleak izango dira.
- urratsa.** $I(X_i, X_j | C)$ elkarrekiko informazioak handienetik txikienera ordenatu
- urratsa.** $I(X_i, X_j | C)$ handiena aukeratu eta konektatu X_i eta X_j erpinak ertz ez-zuzendu baten bidez.
- urratsa.** Hurrengo $I(X_i, X_j | C)$ handiena aukeratu. X_i eta X_j erpinak konektatzean zikloa sortzen bada, ez konektatu ertzak.
- urratsa.** Errepikatu 5. urratsa $n - 1$ ertzen bidez grafoa konektatua lortuko den arte (zuhaitza)
- urratsa.** Lortu berri den zuhaitz ez-zuzendua zuhaitz zuzendu bihurtu, erpin aukeratuz zuhaitzaren erro izateko eta horren arabera ertzei noranzkoa jarri.
- urratsa.** TAN eredia eraiki, C etiketadun erpina gehituz eta C-tik X_i erpin bakoitzera ertz bat gehituz.

TAN ereduaren egitura

Adibidea. Aldagai iragarleak: $X_1, X_2, X_3, X_4, X_5, X_6$

$I(X_1, X_2 | C) > I(X_2, X_5 | C) > I(X_1, X_5 | C) > I(X_3, X_5 | C) > I(X_2, X_3 | C) > I(X_4, X_5 | C) > I(X_1, X_4 | C) > I(X_5, X_6 | C) > \dots$



$$p(c_j | X_1, \dots, X_6) \propto p(c_j) p(X_1 | c_j) p(X_2 | X_1, c_j) p(X_3 | X_2, c_j) p(X_4 | X_3, c_j) p(X_5 | X_4, c_j) p(X_6 | X_5, c_j)$$

k -mendekotasuneko sailkatzaile Bayestarra (k -DB)

k -Dependence Bayesian classifiers, k -DB

- ▶ Naïve Bayes sailkatzailearen beste hedapen bat da, non aldagai iragarle bakoitzak gehienez k aldagai guraso izango dituen, C klase-aldagaia kontatu gabe.
- ▶ k parametroarekin finkatuko dugu ereduari adierazi nahi dugun aldagai iragarleen arteko **mendekotasun maila**
 - ▶ Naïve Bayes: 0-mendekotasuneko sailk. Bayestarra
 - ▶ TAN eredu: 1-mendekotasuneko sailkatzaile Bayestarra
- ▶ Konputazionalki nahiko eraginkorra da
- ▶ Aldagai iragarleen arteko mendekotasunak maila altuagoan adierazi ahal izateak **emaitza hobek** ematen ditu zenbait problema konplexutan.
- ▶ k optimoa?

k -DB ereduaren lortzeko algoritmoa

- urratsa.** $I(X_i, C)$ elkarrekiko informazioak kalkulatu, $i = 1, \dots, n$.
- urratsa.** X_i eta X_j pare guztietarako $i \neq j, i, j = 1, \dots, n$, kalkulatu $I(X_i, X_j | C)$.
- urratsa.** Erabili aldagaien zerrenda, $\mathbb{X}$, multzo hutsarekin hasieratu.
- urratsa.** Hasieratu BN egitura Bayestarra erpin bakarrekin; C .
- urratsa.** $\mathbb{X}$ multzoan aldagai guztiak egongo diren arte errepikatu:
 - X_{max} aukeratu, $I(X_{max}, C) = \max_{X \in \mathbb{X}} I(X, C)$.
 - BN sareari erpin bat gehitu X_{max} aldagaientzat.
 - BN sarean C erpinetik X_{max} erpinera ertza gehitu.
 - $\mathbb{X}$ multzoko X_j aldagaientatik $I(X_{max}, X_j | C)$ balio handieneko $m = \min(|\mathbb{X}|, k)$ aldagai aukeratu eta ertza gehitu X_{max} erpinera.
 - Gehitu X_{max} aldagaia $\mathbb{X}$ multzoari.
- urratsa.** Lortutako BN egiturari jarraituz, sailkapen-eredua adierazi baldintzapeko probabilitateen bidez.

Oinarrizko bibliografia

- ▶ **Introducción a la Minería de Datos**, 10: Métodos Bayesianos, J. Hernández, M^aJ. Ramírez, C. Ferri, Pearson Prentice Hall, 2004
- ▶ "Datu Meatzaritza" irakasgaiaren web orriko: "5. Diagnostikoaren Paradigma Klasikotik Naïve Bayes Sailkatzaileara"
- ▶ "An Analysis of Bayesian Classifiers". Langley et al. 1992
- ▶ "Building Classifiers using Bayesian Networks". Friedman and Goldszmidt. Machine Learning 29:131-163, (1997) → TAN eredu aurkezten dute
- ▶ "Learning Limited Dependence Bayesian Classifiers". Sahami. Knowledge Discovery and Data Mining, KDD-96, 335-338, (1996) → k -DB eredu
- ▶ http://en.wikipedia.org/wiki/Bayes'_theorem
- ▶ http://en.wikipedia.org/wiki/Naive_Bayes_classifier
- ▶ http://en.wikipedia.org/wiki/Rule_of_succession
- ▶ http://eu.wikipedia.org/wiki/Banakuntza_normal