5. Sailkatzaile Bayestarrak

Irakasgaia: Konputazio Zientzien Metodo Matematikoak

Titulazioa: Informatikan Ingeniaria
Konputazio Zientzia eta Adimen Artifiziala saila
Universidad del País Vasco - Euskal Herriko Unibertsitatea

Sailkatzaile Bayestarrak

Sarrera

- Probabilitate teorian (Bayes-en teorema) oinarritutako sailkatzaileak, gainbegiratutako 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 eraikitzera

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 $c_1, \ldots, c_m$ 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
$(\mathbf{x}_1, c_1)$	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	?

$\mathbf{x}=(x_1, x_2, \ldots, x_n)$ kasu berriaren sailkapenerako:

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

- $p(c_j) \rightarrow$ datubasean c_j klasearen maiztasuna
- $p(x_1, x_2, \ldots, x_n|c_j) \rightarrow$ 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,

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

Naïve Bayes sailkatzailea

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

Naïve Bayes sailkatzailea

Oharrak

- Estimatu beharreko parametro kopurua izugarri jaitsi da
- Eta, C klase-aldagaiaren baldintzapean aldagai iragarleak ez badira independente?

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

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

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

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

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:

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

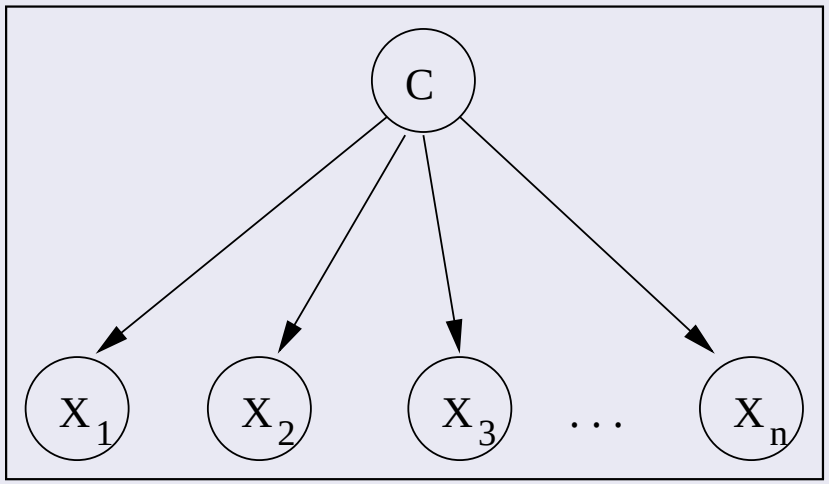
μ_i itxaropen matematikoa; σ_i desbideratze estandarra.

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

Naïve Bayes sailkatzailearen egitura

Adierazpen grafikoa

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



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

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 eredua sortzeko algoritmoa

1. urratsa. $I(X_i, X_j \mid C)$ kalkulatu $i < j$ aldagaietarako $i, j = 1, \ldots, n$.

2. urratsa. Grafo bat definitu. Erpinak $X_1, \ldots, X_n$ aldagai iragarleak izango dira.

3. urratsa. $I(X_i, X_j \mid C)$ elkarrekiko informazioak handienetik txikienera ordenatu

4. urratsa. $I(X_i, X_j \mid C)$ handiena aukeratu eta konektatu X_i eta X_j erpinak ertz ez-zuzendu baten bidez.

5. urratsa. Hurrengo $I(X_i, X_j \mid C)$ handiena aukeratu. X_i eta X_j erpinak konektatzean zikloa sortzen bada, ez konektatu ertzak.

6. urratsa. Errepikatu 5. urratsa $n - 1$ ertzen bidez grafo konektatua lortuko den arte (zuhaitza)

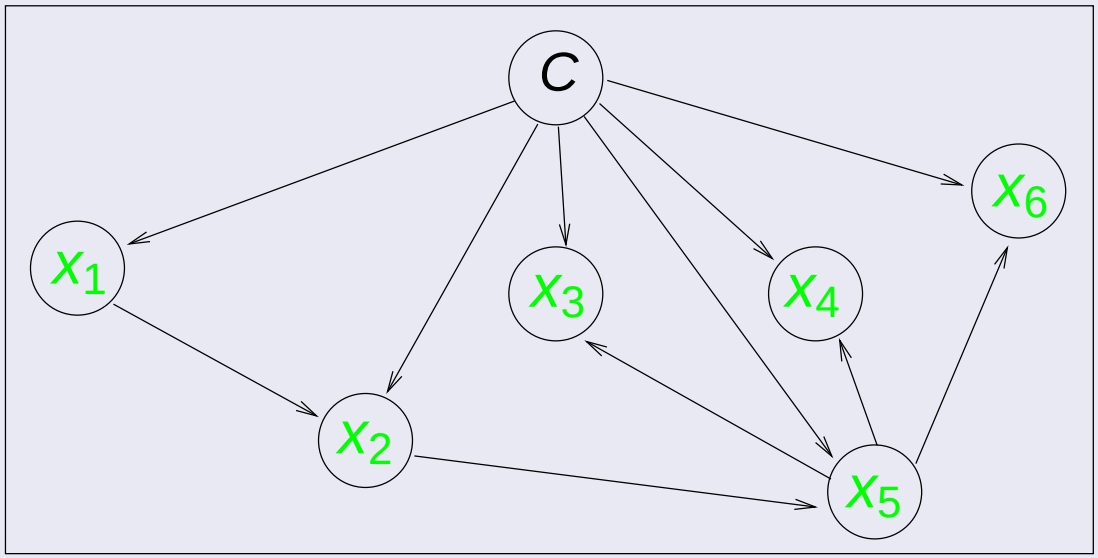
7. urratsa. Lortu berri den zuhaitz ez-zuzendua zuhaitz zuzendu bihurtu, erpin aukeratuz zuhaitzaren erro izateko eta horren arabera ertzei noranzkoa jarriz.

8. urratsa. *TAN* eredua eraiki, C etiketadun erpina gehituz eta C-tik X_i erpin bakoitzera ertz bat gehituz.

TAN ereduaren egitura

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

$I(X_1, X_2 \mid C) > I(X_2, X_5 \mid C) > I(X_1, X_5 \mid C) > I(X_3, X_5 \mid C) > I(X_2, X_3 \mid C) >$
$I(X_4, X_5 \mid C) > I(X_1, X_4 \mid C) > I(X_5, X_6 \mid C) > \ldots$



$$p(c_j|x_1, \ldots, x_6) \propto p(c_j)p(x_1|c_j)p(x_2|x_1, c_j)p(x_3|x_5, c_j)p(x_4|x_5, c_j)p(x_5|x_2, c_j)p(x_6|x_5, c_j)$$

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 ereduan adierazi nahi dugun aldagai iragarleen arteko mendekotasun maila
  - Naïve Bayes: 0-mendekotasuneko sailk. Bayestarra
  - TAN eredua: 1-mendekotasuneko sailkatzaile Bayestarra
- Konputazionalki nahiko eraginkorra da
- Aldagai iragarleen arteko mendekotasunak maila altuagoan adierazi ahal izateak emaitza hobeak ematen ditu zenbait problema konplexutan.
- k optimoa?

k-DB eredua lortzeko algoritmoa

1. urratsa. $I(X_i, C)$ elkarrekiko informazioak kalkulatu, $i = 1, \ldots, n$.

2. urratsa. X_i eta X_j pare guztietarako $i \neq j$, $i, j = 1, \ldots, n$, kalkulatu $I(X_i, X_j | C)$.

3. urratsa. Erabilitako aldagaien zerrenda, $\aleph$, multzo hutsarekin hasieratu.

4. urratsa. Hasieratu BN egitura Bayestarra erpin bakarrarekin; C.

5. urratsa. $\aleph$ multzoan aldagai guztiak egongo diren arte errepikatu:

5.1. X_{max} aukeratu, $I(X_{max}, C) = \max_{X \notin \aleph} I(X, C)$.

5.2. BN sareari erpin bat gehitu X_{max} aldagaiarentzat.

5.3. BN sarean C erpinetik X_{max} erpinera ertza gehitu.

5.4. $\aleph$ multzoko X_j aldagaietatik $I(X_{max}, X_j | C)$ balio handieneko $m = \min(|\aleph|, k)$ aldagai aukeratu eta ertza gehitu X_{max} erpinera.

5.5. Gehitu X_{max} aldagaia $\aleph$ multzoari.

6. urratsa. Lortutako BN egiturari jarraituz, sailkapen-eredua adierazi baldintzapeko probabilitateen bidez.

Oinarrizko bibliografia

- Introducción a la Minería de Datos, 10: Métodos Bayesianos, J. Hernández, Mª J. Ramírez, C. Ferri, Pearson Prentice Hall, 2004
- "Datu Meatzaritza" irakasgaiaren web orriko: "5. Diagnostikoaren Paradigma Klasikotik Naïve Bayes Sailkatzailera"
- "An Analysis of Bayesian Classifiers". Langley et al. 1992
- "Building Classifiers using Bayesian Networks". Friedman and Goldszmidt. Machine Learning 29:131-163, (1997) $\rightarrow$ TAN eredua aurkezten dute
- "Learning Limited Dependence Bayesian Classifiers". Sahami. Knowledge Discovery and Data Mining, KDD-96, 335-338, (1996)$\rightarrow$ k-DB eredua
- 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