5. Diagnostikoaren Paradigma Klasikotik Naïve Bayes Sailkatzailera

Konputazio Zientzia eta Adimen Artifiziala Saila
Universidad del Pas Vasco–Euskal Herriko Unibertsitatea

5.1 Naïve Bayes

Klasea emanik, aldagai iragarleen baldintzapeko independentziaren hipotesia eta Bayesen teorema elkarrekin erabiltzen dituen sailkapenerako paradigma modu desberdinez izendatua izan da: *Bayes inozoa* (Ohmann eta abar, 1988), *naïve Bayes* (Kononenko, 1990), *Bayes sinplea* (Gammerman eta Thatcher, 1991) eta *Bayes independentea* (Todd eta Stamper, 1994).

Formen ezagutzaren arloan tradizio handikoa izan bada ere (Duda eta Hart, 1973), ikasketa automatikoari buruzko literaturan laurogeiko hamarkadaren amaieran agertu zen lehen aldiz naïve Bayes sailkatzailea (Cestnik eta abar, 1987); iragarpenak egiteko ahalmena metodo sofistikatuekin konparatzea izan zen helburua. Poliki-poliki, ikasketa automatikoan diarduten ikerlariak konturatu dira gainbegiratutako sailkapeneko problematarako ahalmen handikoa eta sendoa dela.

Atal honetan *naïve Bayes* paradigmaren errebisioa egingo dugu. Klase-aldagaia emanik, aldagai iragarleen baldintzapeko independentziaren hipotesian oinarrituz eraikitzen da sailkatzailea; hain zuzen ere, hainbesteko sinplifikazioak eragiten dituzten hipotesi horiei zor die bere izena. Diagnostikoaren paradigma klasikotik abiatuko gara, eta parametro-kopuru oso handiaren estimazioa egin beharko dela egiaztatzean, hipotesien sinplifikazioak egingo ditugu naïve Bayes eredura iritsi arte. Hasteko, naïve Bayes sailkatzailearen ezaugarriak hobeto ulertzen lagunduko digun emaitza teoriko bat ikusiko dugu.

Hasiera batean, Bayesen teorema gogoratuko dugu, gertakarien formulazio baten bidez, eta geroago, zorizko aldagaientzako formulazioa emango dugu. Bayesen teorema ikusi ondoren, diagnostikoaren paradigma klasikoa aurkeztuko da, paradigma eraikitzeko erabiltzen diren premisak sinplifikatzeko beharra sortuko delarik, paradigma problema errealen ebazpenerako aplikagarria izan dadin nahi badugu. Atal honetan agertzen den edukia, bibliografiako Díez eta Nell (1998) materialean gai honi buruz dagoenaren egokitzapen bat da.

5.1 Teorema *(Bayes, 1764)* [1] *Izan bitez A eta B bi zorizko gertakari, beren probabilitateak $p(A)$ eta $p(B)$ izanik, hurrenez hurren, eta $p(B) > 0$ betetzen delarik. Demagun A eta B gertakarien a priori probabilitateak, hau da $p(A)$ eta $p(B)$, ezagunak direla, eta baita A gertakaria emanda B gertakariaren probabilitate baldintzatua, hau da, $p(B|A)$. B gertakaria emanda A gertakariaren a posteriori probabilitatea, hau da $p(A|B)$, formula honi jarraituz kalkula daiteke:*

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A)p(B|A)}{p(B)} = \frac{p(A)p(B|A)}{\sum_{A'} p(A')p(B|A')}$$

□

Bayesen teoremaren formulazioa zorizko aldagaietarako ere eman daiteke, bai dimentsio bakarretarako eta baita dimentsio anitzekoetarako ere.

[1] Thomas Bayes (1702–1761) Ingalaterrako apaiza izan zen. Bere aitari laguntzen hasi zen, 1720an Kent-eko artzain izendatua izan zen arte. 1752an artzai izateari uko egin zion. Bere teoria gatazkatsuak Laplacek onartuak izan ziren, eta gerora Boolek zalantzan jarri zituen. *Royal Society of London*-eko kide izendatua izan zen 1742an.

X eta Y dimentsio bakarreko bi zorizko aldagaietarako formulazioa honela geratzen da:

$$p(Y = y|X = x) = \frac{p(Y = y)p(X = x|Y = y)}{\sum_{y'} p(Y = y')p(X = x|Y = y')}$$

Bayesen teorema adieraz daiteke baita dimentsio anitzeko $\boldsymbol{X}$ eta $\boldsymbol{Y}$ aldagaien osagaien bidez:

$$p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}) = p(Y_1 = y_1, \ldots, Y_m = y_m|X_1 = x_1, \ldots, X_n = x_n) =$$

$$= \frac{p(Y_1 = y_1, \ldots, Y_m = y_m)p(X_1 = x_1, \ldots, X_n = x_n|Y_1 = y_1, \ldots, Y_m = y_m)}{\sum_{y'_1,\ldots,y'_m} p(X_1 = x_1, \ldots, X_n = x_n|Y_1 = y'_1, \ldots, Y_m = y'_m)p(Y_1 = y'_1, \ldots, Y_m = y'_m)}$$

5.1 Taulan erakusten den gainbegiratutako sailkapen-probleman $\boldsymbol{Y} = C$ aldagaia dimentsio bakarrekoa da, $\boldsymbol{X} = (X_1, \ldots, X_n)$ aldiz n-dimentsiokoa.

	X_1	$\ldots$	X_n	Y
$(\boldsymbol{x}^{(1)}, y^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_n^{(1)}$	$y^{(1)}$
$(\boldsymbol{x}^{(2)}, y^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_n^{(2)}$	$y^{(2)}$
$\vdots$		$\ddots$		$\vdots$
$(\boldsymbol{x}^{(N)}, y^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_n^{(N)}$	$y^{(N)}$

5.1 Taula: Gainbegiratutako sailkapen-problema.

Diagnostiko-problema baten formulazio klasikoa emango dugu medikuntzan ohikoa den terminologia erabiliz. Terminologia zientziaren eta teknikaren beste arloetara eramatea posible da, jakina. Aipatutako terminologian, honako terminoak erabiliko ditugu:

- *Aurkikuntza.* X_r aldagai iragarle batek balio jakin bat izatea da. Horrela adibidez, X_r aldagaiaren x_r balioak gaixo bat gonbitoka ari dela adieraz dezake.
- *Kasua.* Izaki jakin batentzako aurkikuntza guztien multzoa adierazten du. Horrela, $\boldsymbol{x} = (x_1, \ldots, x_4)$ notazioaz adieraz daiteke izaki hori gaztea dela, gizonezkoa, gonbitoka ari dela eta ez duela aurrekaririk familian.
- *Diagnostikoa.* $Y_1, \ldots, Y_m$ zorizko m aldagaiek hartutako balioak adierazten ditu, horietako bakoitza gaixotasun bati buruzkoa delarik.
- *Diagnostikoaren a priori probabilitatea,* $p(\boldsymbol{y})$ edo $p(Y_1 = y_1, \ldots, Y_m = y_m)$. Diagnostiko jakin baten probabilitatea adierazten du, aurkikuntzei buruzko inolako informaziorik ez dagoenean, hau da, kasua ezaguna ez denean.
- *Diagnostikoaren a posteriori probabilitatea,* $p(\boldsymbol{y}|\boldsymbol{x})$ edo $p(Y_1 = y_1, \ldots, Y_m = y_m|X_1 = x_1, \ldots, X_n = x_n)$. Diagnostiko jakin baten probabilitatea adierazten du, n aurkikuntza ezagutzen direnean (kasua ezaguna denean).

Diagnostikoaren planteamendu klasikoan (ikus 5.2 Taula) posibleak diren m diagnostikoak *bateragarriak* direla suposatzen da, hau da, batera gerta daitezkeela, bakoitza

dikotomikoa izango delarik. Medikuntzaren arlorako interpretatuz, bateragarriak diren m diagnostiko posible horietako bakoitza gaixotasun batekin erlazionatuta dagoela pentsa genezake, eta bi balio posible har ditzakeela: 0 (gaixotasunik ez dago) eta 1 (gaixotasuna dago). n aurkikuntzei edo sintomei dagokienez, zorizko $X_1, \ldots, X_n$ aldagaien bidez adieraziko dira, eta aldagai iragarle bakoitza ere dikotomikoa dela suposatuko dugu, 0 eta 1 balioak hartuko dituztelarik. X_i aldagaiak 0 balioa hartzen badu i. aurkikuntzarik ez dagoela (sintomarik ez) esan nahi du, 1 balioa hartzen badu, aldiz, aurkikuntza edo sintoma hori badagoela.

	X_1	$\ldots$	X_n	Y_1	$\ldots$	Y_m
$(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_n^{(1)}$	$y_1^{(1)}$	$\ldots$	$y_m^{(1)}$
$(\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_n^{(2)}$	$y_1^{(2)}$	$\ldots$	$y_m^{(2)}$
$\ldots$		$\ldots$			$\ldots$	
$(\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_n^{(N)}$	$y_1^{(N)}$	$\ldots$	$y_m^{(N)}$

5.2 Taula: Diagnostiko-problema klasikoa.

Kasua ezagututa, a posteriori probabilitate handieneko diagnostikoa aurkitzean datza diagnostikoaren problema. Notazio matematikoa erabiliz, $(y_1^*, \ldots, y_m^*)$ diagnostiko optimoa honakoa betetzen duena izango da:

$$(y_1^*, \ldots, y_m^*) = arg \max_{(y_1,\ldots,y_m)} p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$$

$p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$ kalkulatzeko Bayesen teorema erabiliz, zera lortzen da:

$$p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n) =$$

$$= \frac{p(Y_1 = y_1, \ldots, Y_m = y_m) p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_m = y_m)}{\sum_{y_1', \ldots, y_m'} p(Y_1 = y_1', \ldots, Y_m = y_m') p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1', \ldots, Y_m = y_m')}$$

Kalkula dezagun $(y_1^*, \ldots, y_m^*)$ aldagaien balioa lortu ahal izateko zenbat parametroren estimazioa egin beharko dugun. Kontuan izan behar da parametro horietako bakoitzaren estimazioa 5.2 Taulan agertzen den N kasuko fitxategitik abiatuz egin beharko dela.

$p(Y_1 = y_1, \ldots, Y_m = y_m)$ balioaren estimazioa egin ahal izateko, Y_i aldagai bakoitza dikotomikoa dela kontuan izanik, guztira $2^m - 1$ parametro beharko ditugu. Modu berean, $p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_m = y_m)$ baldintzapeko probabilitate-banaketa bakoitzerako $2^n - 1$ parametroren estimazioa egin beharko da. Guztira 2^m baldintzapeko probabilitate-banaketa daudenez, $(2^n - 1)2^m$ parametroren estimazioa egin behar da. Hortaz, diagnostikoaren paradigma klasikoko eredu jakin bat finkatu ahal izateko $2^m - 1 + 2^m(2^n - 1)$ parametro beharko dira guztira. Estimatu beharreko parametro kopuruaren ideia bat izateko, 5.3 Taulan bildu dira m-ren (gaixotasun kopuruaren) eta n-ren (aurkikuntza kopuruaren) balio desberdinetarako estimatu beharko diren gutxi gora-beherako parametro kopuruak.

Hain parametro kopuru handiaren estimazioa egitea ezinezkoa gertatzen denez, diagnostikoaren paradigma klasikoa eraikitzeko erabilitako premisen sinplifikazioa egin

m	n		parametroak
3	10	$\simeq$	$8 \cdot 10^3$
5	20	$\simeq$	$33 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{17}$

5.3 Taula: Diagnostikoaren paradigma klasikoan estimatu beharreko parametro kopurua, m (gaixotasun kopuruaren) eta n (sintoma kopuruaren) funtzioan.

beharko da.

Hasteko *diagnostikoak beren artean bateraezinak* direla kontsideratuko dugu, hau da, bi diagnostiko batera ezin direla eman. Ondorioz, diagnostikoa m–dimentsioko zorizko aldagai bat izan beharrean, m balio posible dituen eta banaketa polinomial bati jarraitzen dion dimentsio bakarreko zorizko aldagai baten moduan ikus dezakegu.

n aldagai iragarleak $X_1, \ldots, X_n$ notazioaz izendatuko ditugu, eta guztiak bitarrak direla suposatuko dugu. Diagnostiko-aldagaia C notazioaz adieraziko dugu, eta m balio posible har ditzakeela suposatuko dugu (ikus 5.1 Taula). A posteriori probabilitate handieneko c^* diagnostikoa, gaixo jakin baten $\boldsymbol{x} = (x_1, \ldots, x_n)$ sintomak ezagunak izanik, a posteriori probabilitate handieneko C aldagaiaren balioaren bilaketak emango digu, hau da:

$$c^* = arg\max_c p(C = c|X_1 = x_1, \ldots, X_n = x_n)$$

$p(C = c|X_1 = x_1, \ldots, X_n = x_n)$ probabilitatearen kalkulua Bayesen teorema erabiliz egin daiteke, eta helburua C aldagaiaren a posteriori probabilitate handieneko c^* balioa kalkulatzea denez, Bayesen teoremaren zatitzailea kalkulatu beharrik ez da izango, hau da,

$$p(C = c|X_1 = x_1, \ldots, X_n = x_n) \propto p(C = c)p(X_1 = x_1, \ldots, X_n = x_n|C = c)$$

Hortaz, diagnostikoak beren artean bateraezinak direnean, m diagnostiko posible egonik, eta X_i aldagai iragarle bakoitza dikotomikoa dela kontsideratuz, $(m-1) + m(2^n - 1)$ parametroren estimazioa egin beharko da, non:

- $m - 1$: C aldagaiaren a priori probabilitate kopurua da.
- $m(2^n - 1)$: aldagai iragarleen konbinaketa posible bakoitzaren probabilitate baldintzatu kopurua da, C aldagaiaren balio posible bakoitzerako.

5.4 Taulan ikus daiteke m eta n balio desberdinetarako estimatu beharko den parametro kopuruaren gutxi gora beherakoa.

Estimatu beharko den parametro kopurua oraindik ere handia dela ikus daiteke. Hori dela eta, murrizketa zorrotzagoak ezartzen dituzten suposaketak egin beharko ditugu, lortutako paradigmak eredu inplementagarri bihur daitezen.

Amaitzeko, *Naïve Bayes: diagnostiko bateraezinak eta diagnostikoaren baldintzapean independente diren aurkikuntzak* paradigma aurkeztuko dugu. Naïve Bayes paradigma aldagai iragarleen (aurkikuntzen, sintomen) gain eta iragarri beharreko aldagaiaren (diagnostikoaren) gain ezarritako honako bi premisetan oinarritzen da:

m	n		parametroak
3	10	$\simeq$	$3 \cdot 10^3$
5	20	$\simeq$	$5 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{15}$

5.4 Taula: Diagnostikoaren paradigma klasikoan diagnostikoak bateraezinak izanik estimatu beharreko parametro kopurua, m (gaixotasun kopuruaren) eta n (sintoma kopuruaren) funtzioan.

- Diagnostikoak bateraezinak dira, hau da, iragarri beharreko C aldagaiak har ditzakeen $c_1, \ldots, c_m$ balio posibleen artetik bakar bat hartuko du.
- Diagnostikoaren baldintzapean independente dira aurkikuntzak, hau da, diagnostiko-aldagaiaren balioa ezaguna bada, orduan aurkikuntza baten balioa ezagutzeak ez du garrantziarik beste aurkikuntzetarako. Baldintza hori matematikoki formula honen bidez adierazten da:

$$p(X_1 = x_1, \ldots, X_n = x_n | C = c) = \prod_{i=1}^{n} p(X_i = x_i | C = c) \tag{1}$$

katearen erregelaren bidez zera lortzen baita:

$$\begin{aligned} p(X_1 = x_1, \ldots, X_n = x_n | C = c) \quad = \quad & p(X_1 = x_1 \mid X_2 = x_2, \ldots, X_n = x_n, C = c) \\ & p(X_2 = x_2 \mid X_3 = x_3, \ldots, X_n = x_n, C = c) \\ & \ldots p(X_n = x_n \mid C = c) \end{aligned}$$

Bestalde, klase-aldagaia emanda aldagai iragarleen artean dagoen baldintzapeko independentzia kontuan izanda, zera daukagu:

$$p(X_i = x_i \mid X_{i+1} = x_{i+1}, \ldots, X_n = x_n, C = c) = p(X_i = x_i \mid C = c)$$

$i = 1, \ldots, n$ guztietarako. Hori da (1) ekuazioa betetzearen arrazoia.

Hortaz, naïve Bayes paradigman gaixo jakin baten $(x_1, \ldots, x_n)$ sintomak ezagututa probabilitate handieneko c^* diagnostikoaren bilaketa honela murriztuta geratu da:

$$c^* = arg \max_c p(C = c | X_1 = x_1, \ldots, X_n = x_n)$$

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

Aldagai iragarle guztiak dikotomikoak direla suposatuz, naïve Bayes eredua zehazteko $(m-1) + mn$ parametro beharko dira, zera betetzen baita:

- C aldagaiaren a priori probabilitatea zehazteko $(m-1)$ parametro behar dira.
- X_i aldagai iragarle bakoitzerako m parametro behar dira baldintzapeko probabilitate-banaketak zehazteko.

5.5 Taulako datuak ikusita kontura gaitezke zein den naïve Bayes paradigman beharrezko gertatzen den parametro kopurua, diagnostiko posibleen kopuruaren eta sintoma kopuruaren funtzioan.

m	n	parametroak
3	10	32
5	20	104
10	50	509

5.5 Taula: Naïve Bayes paradigman estimatu beharko den parametro kopurua, diagnostiko posibleen kopuruaren (m) eta sintoma kopuruaren (n) funtzioan.

$X_1, \ldots, X_n$ aldagai iragarleak jarraiak diren kasuan, naïve Bayes paradigma horrela geratuko da: C aldagaiaren a posteriori probabilitatea maximizatuko duen c^* balioa bilatu behar da, $X_1, \ldots, X_n$ aldagaien $\boldsymbol{x} = (x_1, \ldots, x_n)$ instantziazioa emanik (kasua ezaguna izanik).

Hau da, *aldagai jarraietarako naïve Bayes* paradigmak honakoa betetzen duen c^* bilatzen du:

$$c^* = arg\max_c p(C = c|X_1 = x_1, \ldots, X_n = x_n) =$$

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

non $f_{X_i|C=c}(x_i|c)$ funtzioa $i = 1, \ldots, n$ guztietarako X_i aldagaiaren dentsitate-funtzioa izango den, diagnostikoaren balioa c izatera baldintzatuta dagoelarik.

X_i aldagaiaren portaera eredutzeko ohikoa izaten da zorizko aldagai normal bat erabiltzea C-ren balio bakoitzerako. Hau da, c guztietarako eta $i \in \{1, \ldots, n\}$ guztietarako, zera onartuko dugu:

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

Kasu horretan, naïve Bayes paradigmak c^* balioa horrela lortzen du:

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

Estimatu beharreko parametro-kopurua orain $(m-1) + 2nm$ izango da:

- $m - 1$: $p(C = c)$ a priori probabilitateak
- $2nm$: baldintzapeko dentsitate-funtzioak.

Amaitzeko, gerta daiteke aurkikuntza batzuk aldagai diskretuetan jasotzea beste batzuk jarraiak diren bitatean. Horrelakoetan *aldagai iragarle jarraiak eta diskretuak dituen naïve Bayes* paradigma izango dugu.

Demagun n aldagai iragarleen artetik n_1 diskretuak direla, $X_1, \ldots, X_{n_1}$, eta gainontzeko $n - n_1 = n_2$ aldagaiak $Y_1, \ldots, Y_{n_2}$ jarraiak direla. Besterik gabe, naïve Bayes paradigmaren formula egoera honetan aplikatuz gero zera lortzen da:

$$p(c|x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i|c) \prod_{j=1}^{n_2} f(y_j|c)$$

Adierazpen horretan aldagai jarraiei diskretuei baino garrantzia handiagoa ematea posiblea da, $p(x_i|c)$ probabilitateek $0 \leq p(x_i|c) \leq 1$ betetzen den bitartean $f(y_j|c) > 1$

izan daitezkeelako. Egoera hori ekiditeko asmoz, aldagai jarraien normalizazioa egitea proposatzen da, balio bakoitza $\max_{y_j} f(y_j|c)$ balioaz zatituz. Zera lortzen da:

$$p(c|x_1,\ldots,x_{n_1},y_1,\ldots,y_{n_2}) \propto p(c)\prod_{i=1}^{n_1} p(x_i|c)\prod_{j=1}^{n_2}\frac{f(y_j|c)}{\max_{y_j} f(y_j|c)} \tag{2}$$

Klase-aldagaiaren balio posible bakoitzaren baldintzapean dauden aldagai jarraien dentsitate-funtzioek banaketa normalei jarraitzen dieten kasuan, hau da, baldin

$$Y_j|C=c \quad \rightsquigarrow \quad \mathcal{N}(y_j;\mu_j^c,(\sigma_j^c)^2)$$

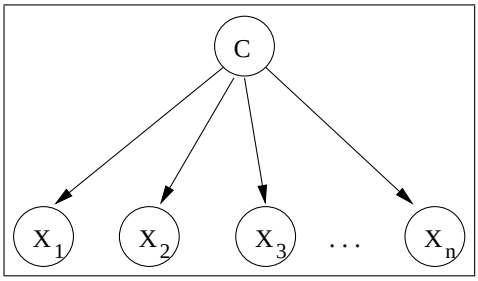
orduan zera lortzen da,

$$\frac{f(y_j|c)}{\max_{y_j} f(y_j|c)} = \frac{\frac{1}{\sqrt{2\pi}\sigma_j}e^{-\frac{1}{2}\left(\frac{y_j-\mu_j^c}{\sigma_j^c}\right)^2}}{\frac{1}{\sqrt{2\pi}\sigma_j}e^{-\frac{1}{2}\left(\frac{\mu_j^c-\mu_j^c}{\sigma_j^c}\right)^2}} = e^{-\frac{1}{2}\left(\frac{y_j-\mu_j^c}{\sigma_j^c}\right)^2}$$

eta (2) formula honela adierazita geratuko da:

$$p(c|x_1,\ldots,x_{n_1},y_1,\ldots,y_{n_2}) \propto p(c)\prod_{i=1}^{n_1} p(x_i|c)\prod_{j=1}^{n_2} e^{-\frac{1}{2}\left(\frac{y_j-\mu_j^c}{\sigma_j^c}\right)^2}$$

5.1 Irudian naïve Bayes ereduaren egitura grafikoa azaltzen da.



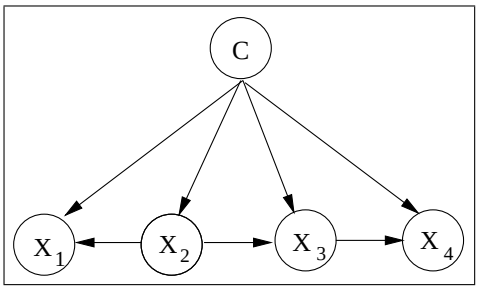
5.1 Irudia: Naïve Bayes ereduaren egitura grafikoa.

5.2 Zuhaitzera zabaldutako Naïve Bayes

Atal honetan zuhaitzera zabaldutako naïve Bayes egitura duten sailkatzaileak eraikitzen dituzten lan batzuk aurkeztuko ditugu (*Tree Augmented Network (TAN)*). Egitura mota hori lortzeko, aldagai iragarleen zuhaitz-egitura batetik hasten da, eta ondoren klase-aldagaia aldagai iragarle bakoitzarekin konektatzen da. 5.2 Irudian zuhaitzera zabaldutako naïve Bayes egitura baten adibidea ikus daiteke.

Friedman eta abar. (1997) lanean *Tree Augmented Network (TAN)* izeneko algoritmoa aurkezten da, funtsean Chow–Liu (1968) lanean agertzen den algoritmoaren egokitzapen bat dena. Bertan klase-aldagaiaren baldintzapean dagoen elkarrekiko informazio kantitatea erabiltzen da, Chow–Liu-ren algoritmoa elkarrekiko informazio kantitatean oinarritzen den bitartean. C klase-aldagaiaren baldintzapean X eta Y bi aldagai diskretuen artean dagoen elkarrekiko informazio-kantitatea honela definitzen da:

$$I(X,Y|C) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{r=1}^{w} p(x_i,y_j,c_r)\log_2\frac{p(x_i,y_j|c_r)}{p(x_i|c_r)p(y_j|c_r)}$$



5.2 Irudia: Zuhaitzera zabaldutako naïve Bayes egitura baten adibidea (*Tree Augmented Network (TAN)*).

---

1. urratsa. $I(X_i, X_j \mid C)$ kalkulatu $i < j$ aldagaietarako $i, j = 1, \ldots, n$.
2. urratsa. Grafo ez-zuzendu osotua eraiki, grafoaren erpinak $X_1, \ldots, X_n$ aldagai iragarleak izanik. X_i eta X_j aldagaiak lotzen dituen ertzari $I(X_i, X_j \mid C)$ pisua esleitu.
3. urratsa. Kruskall-en algoritmoa erabiliz, eraiki berri den grafo osotuari dagokion pisu maximoko zuhaitz hedatua eraiki.
4. urratsa. Lortu berri den zuhaitz ez-zuzendua zuhaitz zuzendu bihurtu, aldagai bat aukeratuz zuhaitzaren erro izateko eta horren arabera ertzei noranzkoa jarriz.
5. urratsa. *TAN* eredua eraiki, C etiketadun erpina gehituz eta C-tik X_i aldagai iragarle bakoitzera ertz bat gehituz.

5.3 Irudia: *TAN* algoritmoaren sasikodea (Friedman eta abar. 1997).

5.3 Irudiko sasikodean ikus daitekeen bezala, *TAN* algoritmoak bost urrats ditu.

Algoritmoaren lehen urratsean X_i eta X_j aldagai pare bakoitzerako C klase-aldagaiaren baldintzapean dagoen elkarrekiko informazio-kopurua kalkulatu behar da. Ondoren aldagai iragarle bakoitzeko erpin bat izango duen n erpineko grafo ez-zuzendu osotua eraikiko da. Grafoaren ertz bakoitzari egokituko zaion pisua ertzak lotzen dituen bi aldagaietarako kalkulatutako C klase-aldagaiaren baldintzapean dagoen elkarrekiko informazio-kopurua izango da.

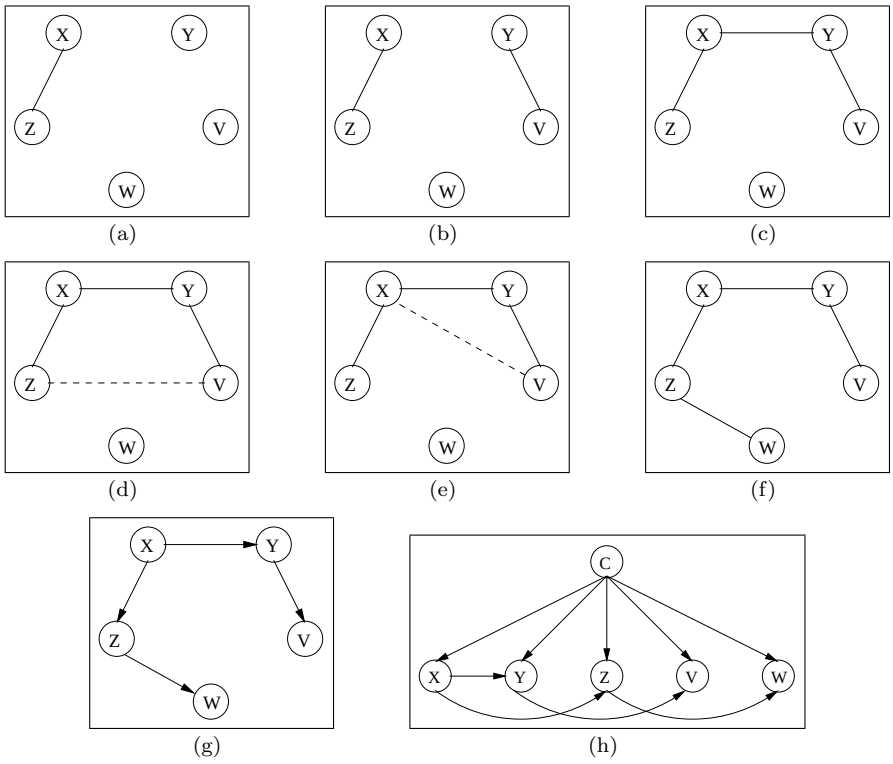
Kruskall-en algoritmoa $n(n-1)/2$ pisu horietatik abiatzen da, eta graforako pisu maximoko zuhaitz hedatua eraikitzen da modu honetan:

1. urratsa. Pisu handieneko bi ertzak eraikitze bidean dagoen zuhaitzerako aukeratu.

2. urratsa. Pisu handieneko hurrengo ertza aztertu. Zuhaitzak dagoeneko dituen ertzekin ziklorik osatzen ez badu, orduan ertza aukeratu. Bestela, ertza baztertu, eta pisu handieneko hurrengo ertzarekin azterketa berberari ekin.

3. urratsa. 2. urratsa errepikatu, $n-1$ ertz aukeratuko diren arte.

Algoritmoaren aplikaziorako adibide bat 5.4 Irudian ikus daiteke. Bertan C klase-aldagaiaren baldintzapean dagoen elkarrekiko informazio-kopuruen artean erlazio hau betetzen da: $I(X,Z|C) \geq I(Y,V|C) \geq I(X,Y|C) \geq I(Z,V|C) \geq I(X,V|C) \geq I(Z,W|C) \geq I(X,W|C) \geq I(Y,Z|C) \geq I(Y,W|C) \geq I(V,W|C)$. (a) iruditik (f) irudira bitartean Kruskall-en algoritmoaren aplikazioa ikus daiteke. (g) irudian *TAN* algoritmoaren 4. urratsa ikus daiteke, eta (h) irudian 5. urratsa. Lortutako sailkapen-



5.4 Irudia: *TAN* algoritmoaren aplikaziorako adibide bat X, Y, Z, V eta W aldagai iragarleekin.

eredua honakoa da:

$$p(c|x, y, z, v, w) \propto p(c) \cdot p(x|c) \cdot p(y|x, c) \cdot p(z|x, c) \cdot p(v|y, c) \cdot p(w|z, c)$$

5.3 k-Mendekotasuneko Sailkatzaile Bayestarra

Sahami-k 1996. urtean *k Dependence Bayesian classifier (kDB)*, izeneko algoritmoa aurkeztu zuen, Sahami (1996). Algoritmoaren oinarrian k-mendekotasuneko sailkatzaile Bayestarraren kontzeptua dago, non naïve Bayes sailkatzailearen egitura izanik aldagai iragarle bakoitzak gehienez k aldagai guraso izango dituen, C klase-aldagaia kontatu gabe. Horrela, naïve Bayes eredua 0-mendekotasuneko sailkatzaile Bayestarra dela esango dugu, eta *TAN* eredua 1-mendekotasuneko sailkatzaile Bayestarra. Sailkatzaile Bayestar osotua $(n-1)$-mendekotasuneko sailkatzaile Bayestarra da; bere egituran ez da inolako independentziarik erakusten. kDB algoritmoaren sasikodea 5.5 Irudian ikus daiteke.

Erreferentziak

1. T. Bayes (1764). Essay towards solving a problem in the doctrine of chances. *The Phylosophical Transactions of the Royal Society of London.*
2. C. Chow, C. Liu (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14, 462–467.
3. F. J. Díez, E. Nell (1998). *Introducción al Razonamiento Aproximado.* Departamento de Inteligencia Artificial. UNED.

1. urratsa. X_i aldagai iragarlearen eta C klase-aldagaiaren $I(X_i, C)$ elkarrekiko informazio-kopurua kalkulatu, $i = 1, \ldots, n$.
2. urratsa. X_i eta X_j aldagai iragarle pare guztietarako $i \neq j$, $i, j = 1, \ldots, n$, kalkulatu C klase-aldagaiaren baldintzapean dagoen elkarrekiko informazio-kopurua, $I(X_i, X_j|C)$.
3. urratsa. Erabilitako aldagaien zerrenda, $\aleph$, multzo hutsarekin hasieratu.
4. urratsa. Eraikiko den sare Bayestarra, BN, erpin bakarraz hasieratu; C klase-aldagaiarena.
5. urratsa. $\aleph$ multzoan aldagai guztiak egongo diren arte errepikatu:
   5.1. $\aleph$ multzoan ez dauden aldagaien artean C aldagaiarekiko informazio handieneko 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.

5.5 Irudia: kDB algoritmoaren sasikodea, (Sahami, 1996).

4. R. Duda, P. Hart (1973). *Pattern Classification and Scene Analysis*. John Wiley and Sons.

5. N. Friedman, D. Geiger, M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning*, 29, 131–163.

6. A. Gammerman, A. R. Thatcher (1991). Bayesian diagnostic probabilities without assuming independence of symptoms. *Methods of Information in Medicine*, 30, 15–22.

7. F. V. Jensen (2001). *Bayesian Networks and Decision Graphs*. Springer Verlag.

8. E. J. Keogh. M. Pazzani (1999). Learning augmented Bayesian classifiers: a comparison of distribution-based and non distribution-based approaches. *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*, 225–230.

9. I. Kononenko (1990).Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition. *Current Trends in Knowledge Acquisition*.

10. C. Ohmann, Q. Yang, M. Kunneke, H. Stolzing, K. Thon, W. Lorenz (1988). Bayes theorem and conditional dependence of symptoms: different models applied to data of upper gastrointestinal bleeding. *Methods of Information in Medicine*, 27, 73–83.

11. M. Sahami (1996). Learning limited dependence Bayesian classifiers. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 335–338.

12. B. S. Todd, R. Stamper (1994). The relative accuracy of a variety of medical diagnostic programs. *Methods of Information in Medicine*, 33, 402–416.