

5. Bayesov klasifikator

prof. dr. sc. Bojana Dalbelo Bašić
doc. dr. sc. Jan Šnajder

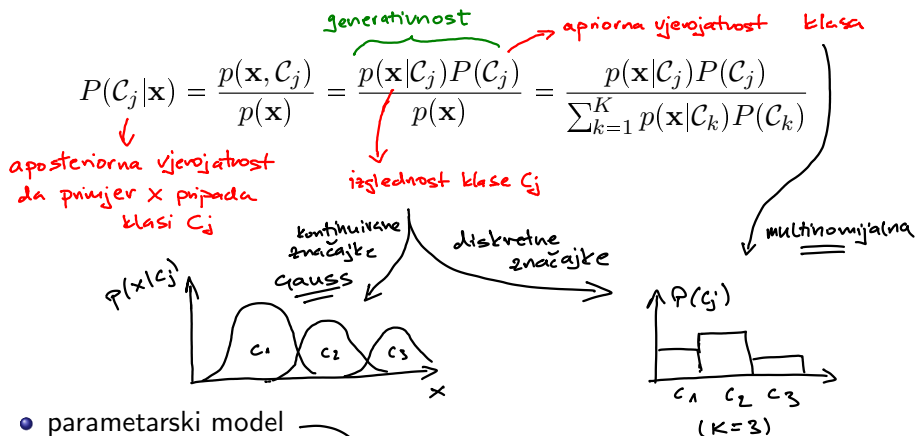
Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

Ak. god. 2012/13.

- 1 Bayesovska klasifikacija
- 2 Naivan Bayesov klasifikator
- 3 Polunaivan Bayesov klasifikator

- 1 Bayesovska klasifikacija
- 2 Naivan Bayesov klasifikator
- 3 Polunaivan Bayesov klasifikator

Bayesovo pravilo



- parametarski model
- generativni model

↳ 1) modeliramo gustoću $P(x, C_j)$
2) opisujemo generativne podatke: $P(x|C_j) \cdot P(C_j)$

→ pretpostavljamo raspodjele za
1) $P(C_j)$ (multinomijalnu) i za $P(x|C_j)$ (multinom./Gaussovu)
2) # parametara je fiksiran!

Klasifikacijska odluka

MAP-hipoteza:

(maximum a posteriori)
$$h(\mathbf{x}) = \operatorname{argmax}_{\mathcal{C}_k} p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) \quad (1)$$

$$h: \mathcal{X} \rightarrow \{c_1, c_2, \dots, c_K\}$$

Vjerojatnost klasifikacije u \mathcal{C}_j :

$$h_j(\mathbf{x}) = P(\mathcal{C}_j|\mathbf{x})$$

$$h_j: \mathcal{X} \rightarrow [0, 1]$$

Pouzdanost klasifikacije u \mathcal{C}_j :

(confidence)

$$h_j(\mathbf{x}) = p(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j)$$

$$h_j: \mathcal{X} \rightarrow \mathbb{R}^+$$

Bayesovska klasifikacija – primjer

$$P(C_1) = P(C_2) = 0.3, P(C_3) = 0.4 \quad (K=3)$$

$$\text{Za neki primjer } x: p(x|C_1) = 0.9, p(x|C_2) = p(x|C_3) = 0.4$$

MAP-hipoteza za svaku od klasa?

$$p(x|C_1) \cdot P(C_1) = 0.9 \times 0.3 = 0.27 \quad \leftarrow \text{MAP}$$

$$p(x|C_2) \cdot P(C_2) = 0.4 \times 0.3 = 0.12$$

$$p(x|C_3) \cdot P(C_3) = 0.4 \times 0.4 = 0.16$$

$$P(x) = \sum_{k=1}^3 p(x|C_k) \cdot P(C_k) = 0.55$$

$$P(C_1|x) = \frac{0.27}{0.55} = 0.49 \quad P(C_2|x) = 0.22 \quad P(C_3|x) = 0.29$$

Minimizacija pogreške klasifikacije

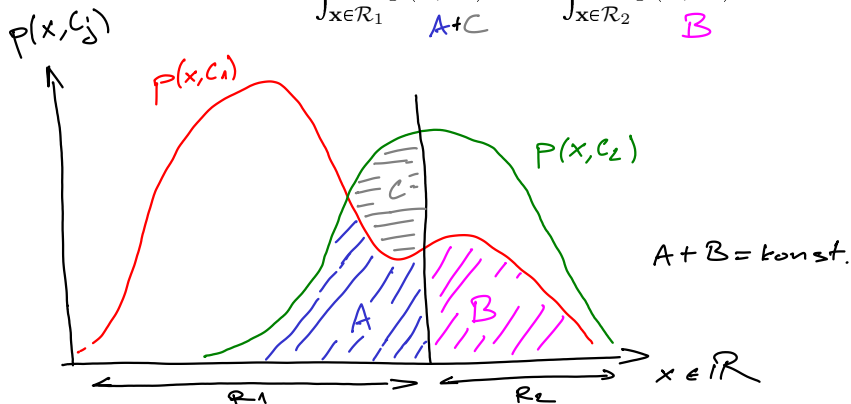
$$\mathcal{R}_1 = \{\mathbf{x} \in \mathcal{X} \mid h_1(\mathbf{x}) = c_1\}, \mathcal{R}_2 = \mathcal{X} \setminus \mathcal{R}_1$$

primjer iz \mathcal{C}_2 označen kao \mathcal{C}_1 ← obrnuto
↓

$$P(\text{pogreška}) = p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1)$$

$$= \int_{\mathbf{x} \in \mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathbf{x} \in \mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x}$$

$A+C$ B



Minimizacija rizika

L_{kj} – gubitak uslijed pogrešne klasifikacije primjera iz klase \mathcal{C}_k u klasu \mathcal{C}_j

↳ matrica gubitka

Očekivani gubitak (funkcija rizika):

$$L = \begin{pmatrix} 0 & 10 \\ 1 & 1 \end{pmatrix}$$

odabrano
↓
Stvarno

$$\mathbb{E}[L] = \sum_{k=1}^K \sum_{j=1}^K \underbrace{\int_{\mathbf{x} \in \mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}}_{\text{element matrice } L} \quad (2)$$

Očekivani rizik pri klasifikaciji \mathbf{x} u \mathcal{C}_j :

element matrice L

$$R(\mathcal{C}_j|\mathbf{x}) = \sum_{k=1}^K L_{kj} P(\mathcal{C}_k|\mathbf{x})$$

Optimalna klasifikacijska odluka:

po retcima matrice L

$$h(\mathbf{x}) = \operatorname{argmin}_{\mathcal{C}_k} R(\mathcal{C}_k|\mathbf{x}) \quad (3)$$

Minimizacija rizika – primjer

$$P(C_1|\mathbf{x}) = 0.25, P(C_2|\mathbf{x}) = 0.6, P(C_3|\mathbf{x}) = 0.15$$

$$K=3$$



$$L = \begin{pmatrix} 0 & 1 & 5 \\ 1 & 0 & 5 \\ 10 & 100 & 0 \end{pmatrix}$$

$$L = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

MAP-hipoteza

$$\begin{aligned} j=1 \quad R(C_1|\mathbf{x}) &= \sum_k L_{k,1} P(C_k|\mathbf{x}) \\ &= 0 \cdot P(C_1|\mathbf{x}) + 1 \cdot P(C_2|\mathbf{x}) + 10 \cdot P(C_3|\mathbf{x}) \\ &= 1 \cdot 0,6 + 10 \cdot 0,15 = \underline{\underline{2,1}} \end{aligned}$$

$$j=2 \quad R(C_2|\mathbf{x}) = 1 \cdot 0,25 + 100 \cdot 0,15 = 15,25$$

$$j=3 \quad R(C_3|\mathbf{x}) = 5 \cdot 0,25 + 5 \cdot 0,6 = 4,25$$

- 1 Bayesovska klasifikacija
- 2 Naivan Bayesov klasifikator
- 3 Polunaivan Bayesov klasifikator

Naivan Bayesov klasifikator

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N, y^{(i)} \in \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$$

$$P(\mathcal{C}_j | x_1, \dots, x_n) \propto P(x_1, \dots, x_n | \mathcal{C}_j) P(\mathcal{C}_j)$$

$$h(\mathbf{x} = x_1, \dots, x_n) = \underset{j}{\operatorname{argmax}} P(\mathbf{x} = x_1, \dots, x_n | y = \mathcal{C}_j) P(y = \mathcal{C}_j)$$

MAP-hipoteza

multinomijalna
varijabla

ML-procjena za y (multinomijalna varijabla):

$$\hat{P}(\mathcal{C}_j) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{y^{(i)} = \mathcal{C}_j\} = \frac{N_j}{N}$$

relativna
frekvencija

Broj parametara za $\hat{P}(\mathcal{C}_j)$ $j = 1, \dots, K$?

→ $K-1$ parametar

Naivan Bayesov klasifikator

Procjena za $P(x_1, \dots, x_n | C_j)$?

$\mathbf{x} = (x_1, \dots, x_n)$ kao multinomijalna varijabla?

Broj parametara?

Generalizacija?

Nemoguća!

$$P(x_1, x_2, \dots, x_n | C_j)$$

Sve kombinacije:

0	0	—	0	1	0
0	0	—	1	1	0
⋮					
1	1	—	1	1	0
0	0	—	0	1	1
					⋮

Općenito:

$$\left(\prod_{k=1}^n K_k - 1 \right) \cdot K$$

parametara

$$\Rightarrow O(K^n)$$

Naivan Bayesov klasifikator

Pravilo lanca (uz uvjetnu varijablu C_j):

$$P(x_1, \dots, x_n | C_j) = \prod_{k=1}^n P(x_k | x_1, \dots, x_{k-1}, C_j).$$

Pretpostavka: $x_i \perp x_k | C_j$ ($i \neq k$) $\Leftrightarrow P(x_i | x_k, C_j) = P(x_i | C_j)$

$$P(x_1, \dots, x_n | C_j) = \prod_{k=1}^n P(x_k | C_j)$$

Naivan Bayesov klasifikator:

$$h(x_1, \dots, x_n) = \underset{j}{\operatorname{argmax}} P(C_j) \prod_{k=1}^n P(x_k | C_j)$$

$K-1$ $n \cdot K \cdot (K-1)$ $\Rightarrow O(n \cdot K)$

Naivan Bayesov klasifikator

ML-procjena:

$$\hat{P}(x_k | \mathcal{C}_j) = \frac{\sum_{i=1}^N \mathbf{1}\{x_k^{(i)} = x_k \wedge y^{(i)} = \mathcal{C}_j\}}{\sum_{i=1}^N \mathbf{1}\{y^{(i)} = \mathcal{C}_j\}} = \frac{N_{kj}}{N_j}$$

Laplaceov procjenitelj:

$$\hat{P}(x_k | \mathcal{C}_j) = \frac{\sum_{i=1}^N \mathbf{1}\{x_k^{(i)} = x_k \wedge y^{(i)} = \mathcal{C}_j\} + \lambda}{\sum_{i=1}^N \mathbf{1}\{y^{(i)} = \mathcal{C}_j\} + K_k \lambda} = \frac{N_{kj} + \lambda}{N_j + K_k \lambda}$$

Broj parametara: $\sum_{k=1}^n (K_k - 1)K$ ~~$+ K - 1$~~

Binarne značajke: nK

 $K_k = 2$

NB – primjer

i	x_1 Mjesto radnje	x_2 Glavni lik	x_3 Vrijeme radnje	x_4 Vanzemaljci	y Dobar film
1	svemir	znanstvenica	sadašnjost	da	ne
2	Zemlja	kriminalac	budućnost	ne	ne
3	drugdje	dijete	prošlost	da	ne
4	svemir	znanstvenica	sadašnjost	ne	da
5	svemir	kriminalac	prošlost	ne	ne
6	Zemlja	dijete	prošlost	da	da
7	Zemlja	policajac	budućnost	da	ne
8	svemir	policajac	budućnost	ne	da
	$k_1=3$	$k_2=4$	$k_3=3$	$k_4=2$	$k=2$

$$K-1 + \sum_{k=1}^n (k_k-1) \cdot K = 1 + 2 \cdot (2+3+2+1) = 17$$

Bez naivne pretpostavke: $1 + 2 \cdot (3 \cdot 4 \cdot 3 \cdot 2 - 1) = 143$

Broj parametara - pojašnjenje

za faktor $P(X_1, X_2 | c)$ broj parametara je $(2 \cdot 3 - 1) \cdot 2 = 10$
 $2 \quad 3 \quad 2$ (\leftarrow broj vrijednosti)

zašto? zato što u CPT (engl. conditional prob. table) moramo paziti sve ove kombinacije:

$X_1 \quad X_2 \quad C \quad P(x_1, x_2 c)$				$x_1 \quad x_2 \quad c \quad P(x_1, x_2 c)$				
$(K_1 \cdot K_2 - 1)$ kombinacija	0	0	0	0,2	0	0	1	0
	0	1	0	0,3	0	1	1	0,1
	0	2	0	0,1	0	2	1	0,2
	1	0	0	0,1	1	0	1	0,1
	1	1	0	0,2	1	1	1	0,2
	1	2	0	0,1	1	2	1	0,4

ne moramo paziti jer $\sum P(x) = 1$

Općenito: za $P(X_1, \dots, X_n | c) \Rightarrow \left(\prod_{k=1}^n K_k - 1 \right) \cdot K$ parametara

NB – primjer

$$X = (\underset{x_1}{\text{svemir}}, \underset{x_2}{\text{dijete}}, \underset{x_3}{\text{sadašnjost}}, \underset{x_4}{\text{da}})$$

$$P(\text{da} | x) = ? \quad P(\text{ne} | x) = ?$$

$$P(\text{da} | x) \propto P(\text{da}) \prod_{k=1}^n P(x_k | \text{da}) = 0,003$$

$$\hat{P}_{ML}(\text{da}) = \frac{N_{\text{da}}}{N} = 3/8$$

$$\hat{P}_{ML}(x_1 = \text{svemir} | \text{da}) = \frac{N_{1=\text{svemir}}}{N_{\text{da}}} = 2/3 = 0,66$$

$$\hat{P}_{ML}(x_2 = \text{dijete} | \text{da}) = 1/3$$

$$\hat{P}_{ML}(x_3 = \text{sadašnjost} | \text{da}) = 1/3$$

$$\hat{P}_{ML}(x_4 = \text{da} | \text{da}) = 1/3$$

$$h(x)_{MAP} = \text{da}$$

$$P(\text{da} | x) = \frac{0,003}{0,003 + 0,006}$$

$$P(\text{ne} | x) = 0,006$$

- 1 Bayesovska klasifikacija
- 2 Naivan Bayesov klasifikator
- 3 Polunaivan Bayesov klasifikator

Ako $x_2 \not\perp x_3 | \mathcal{C}_j$, umjesto

$$P(\mathcal{C}_j | x_1, x_2, x_3) \propto P(x_1 | \mathcal{C}_j) P(x_2 | \mathcal{C}_j) P(x_3 | \mathcal{C}_j) P(\mathcal{C}_j) \quad \text{NB}$$

faktorizirati kao:



$$P(\mathcal{C}_j | x_1, x_2, x_3) \propto P(x_1 | \mathcal{C}_j) P(x_2, x_3 | \mathcal{C}_j) P(\mathcal{C}_j) \quad \text{Semi-NB}$$

ili:

$$P(\mathcal{C}_j | x_1, x_2, x_3) \propto P(x_1 | \mathcal{C}_j) P(x_2 | \mathcal{C}_j) P(x_3 | x_2, \mathcal{C}_j) P(\mathcal{C}_j)$$

- prednosti?
- broj parametara?
- koje varijable združiti?

Koje varijable združiti?

Problem pretraživanja prostora stanja

$\{\{a\}, \{b\}, \{c\}\}$

$\{\{a\}, \{b, c\}\}$

$\{\{b\}, \{a, c\}\}$

$\{\{c\}, \{a, b\}\}$

$\{\{a, b, c\}\}$

$\# \text{particija} = \text{broj stanja } |S|$

Bellov broj: $B_3 = 5$, $B_4 = 15$, $B_5 = 52$, ..., $B_{10} = 115975$, ...

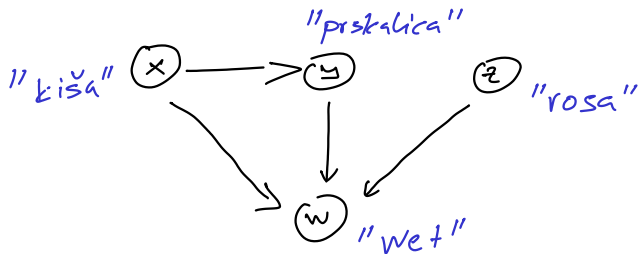
Trebamo heuristiku!

→ Kriterij

- mijenimo zavisnost varijabli
- isprobavamo točnost modela na skupu za provjeru

Bayesova mreža

$$P(X, Y, Z, W) = P(X)P(Y|X)P(Z)P(W|X, Y, Z)$$



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

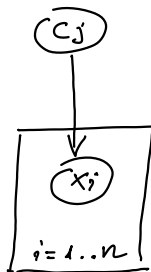
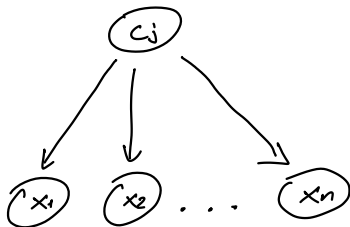
↳ roditelji čvora x_i

NB kao Bayesova mreža

$$h(x_1, \dots, x_n) = \underset{j}{\operatorname{argmax}} P(C_j) \prod_{k=1}^n P(x_k | C_j)$$

$$P(C_j | x_1, \dots, x_n) \propto \underbrace{P(C_j) \prod_{k=1}^n P(x_k | C_j)}_{\text{„plate diagram“}}$$

„plate diagram“



Algoritam FSSJ

- 1 Inicijaliziraj $X = \emptyset$. Početna faktORIZACIJA:

$$P(x_1, \dots, x_n, \mathcal{C}_j) = P(x_1) \cdots P(x_n) P(\mathcal{C}_j)$$

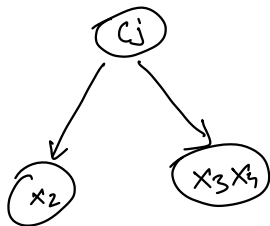
$$P(\mathcal{C}_j | x_1, \dots, x_n) = P(\mathcal{C}_j)$$

Klasificiraj primjere iz skupa za provjeru: $\mathcal{C}^* = \operatorname{argmax}_j P(\mathcal{C}_j)$

- 2 Za svaku varijablu $x_i \notin X$ koja još nije uključena u model:
 - (a) Uključi x_i kao uvjetno nezavisnu u odnosu na ostale varijable za danu klasu \mathcal{C}_j (dodavanje lûka (\mathcal{C}_j, x_i))
 - (b) Uključi x_i tako da se ona doda u zajednički čvor ("superčvor") s nekom već uključenom varijablom
- 3 Izaberi x_i i opciju koja minimizira pogrešku generalizacije
- 4 Ponavljaj od koraka (2) do konvergencije pogreške

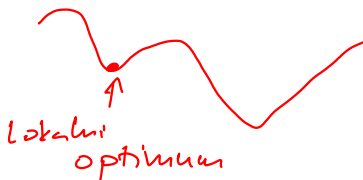
Algoritam FSSJ – primjer

x_1, x_2, x_3, x_4



$$P(c_j | x_1, x_2, x_3, x_4) \propto P(c_j) \cdot P(x_2 | c_j) \cdot P(x_3 x_4 | c_j)$$

Pohlepno pretraživanje!
(engl. greedy)



Klasifikator TAN


Ideja: združiti zavisne varijable. Kako mjeriti zavisnost?

Mjera **uzajamne informacije** (engl. *mutual information*):

$$I(X, Y) = \sum_{j=1}^{K_Y} \sum_{i=1}^{K_X} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$$

Uvjetna uzajamna informacija:

$$\hookrightarrow KL(P(x, y) \parallel P(x) \cdot P(y))$$


$$\begin{aligned} I(X, Y|Z) &= \sum_{k=1}^{K_Z} P(z_k) I(X, Y|z_k) \\ &= \sum_{k=1}^{K_Z} \sum_{j=1}^{K_Y} \sum_{i=1}^{K_X} P(x_i, y_j, z_k) \log \frac{P(x_i, y_j|z_k)}{P(x_i|z_k)P(y_j|z_k)} \end{aligned}$$

$$X \perp Y | Z \Leftrightarrow I(X, Y|Z) = 0$$

Algoritam TAN

Algoritam TAN

- 1 Izračunaj $I(x_i, x_j | \mathcal{C})$ za $i < j$, $i = 1, \dots, n$ i sortiraj silazno
- 2 Izgradi nepovezanu Bayesovu mrežu s čvorovima x_1, \dots, x_n
- 3 Razmotri par (x_i, x_j) s najvećom vrijednošću $I(x_i, x_j | \mathcal{C})$ i dodaj brid (x_i, x_j) ako time ne nastaje ciklus; inače razmatraj idući par u listi
- 4 Ponavljaj korak 3 dok ne izgradiš $n - 1$ bridova
- 5 Pretvori neusmjeren graf u usmjeren tako da nasumično odabereš jedan čvor kao korijen
- 6 Dodaj čvor \mathcal{C} i poveži ga lukovima sa svim ulaznim varijablama

↙ smjer je nebitan!

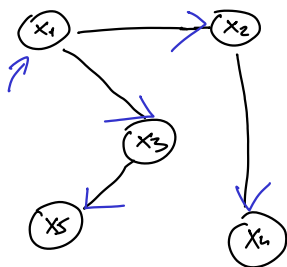


$$p(x_1) \cdot p(x_2 | x_1) = p(x_1, x_2) = p(x_2) \cdot p(x_1 | x_2)$$

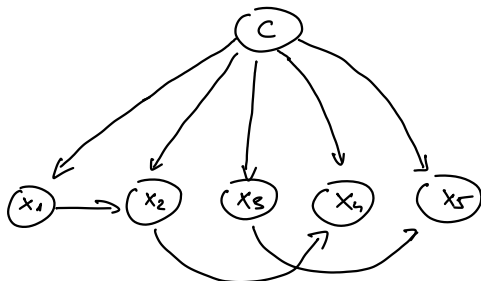
Algoritam TAN – primjer

Procjena na \mathcal{D} : X_1, X_2, X_3, X_4, X_5

$$I(x_1, x_3 | \mathcal{C}) > I(x_2, x_4 | \mathcal{C}) > I(x_1, x_2 | \mathcal{C}) > \overset{\text{ciklus}}{\cancel{I(x_3, x_4 | \mathcal{C})}} > \overset{\text{ciklus}}{\cancel{I(x_1, x_4 | \mathcal{C})}} > \\ I(x_3, x_5 | \mathcal{C}) > I(x_1, x_5 | \mathcal{C}) > I(x_2, x_3 | \mathcal{C}) > I(x_2, x_5 | \mathcal{C}) > \underline{I(x_4, x_5 | \mathcal{C})}$$



\Rightarrow



$$P(\mathcal{C} | x_1 \text{---} x_5) \propto P(\mathcal{C}) \cdot P(x_1 | \mathcal{C}) \cdot P(x_2 | x_1, \mathcal{C}) \cdot P(x_3 | x_1, \mathcal{C}) \cdot \\ P(x_4 | x_2, \mathcal{C}) \cdot P(x_5 | x_3, \mathcal{C})$$

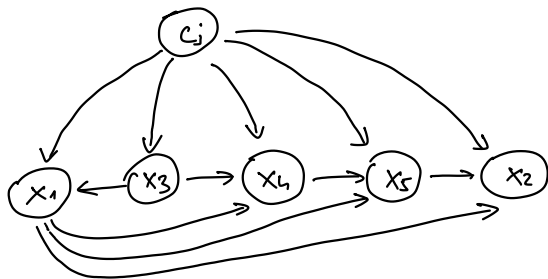
Algoritam k -DB

- 1 Izračunaj $I(x_i, C_j)$ i $I(x_i, x_j | C_j)$ za svaki par varijabli. Sortiraj varijable silazno po $I(x_i, C_j)$
 - 2 Za varijablu x_i koja je prva u listi:
 - 1 Dodaj varijablu x_i u model i izbaci je iz liste
 - 2 Postavi čvor C_j za roditelja čvora x_i
 - 3 Od varijabli x_j koje su već uključene u model, njih k (ili manje, ako ih nema toliko) koje imaju najveću vrijednost $I(x_i, x_j | C_j)$ postavi kao čvorove roditelje od x_i
 - 3 Ponavljaj prethodni korak dok lista nije prazna
- "važne" varijable prve*

Algoritam k -DB – primjer

$$I(x_3, \mathcal{C}) > I(x_1, \mathcal{C}) > I(x_4, \mathcal{C}) > I(x_5, \mathcal{C}) > I(x_2, \mathcal{C})$$

$$I(x_3, x_4 | \mathcal{C}) > I(x_2, x_5 | \mathcal{C}) > I(x_1, x_3 | \mathcal{C}) > I(x_1, x_2 | \mathcal{C}) > I(x_2, x_4 | \mathcal{C}) > \\ I(x_2, x_3 | \mathcal{C}) > I(x_1, x_4 | \mathcal{C}) > I(x_4, x_5 | \mathcal{C}) > I(x_1, x_5 | \mathcal{C}) > I(x_3, x_5 | \mathcal{C})$$



2-DB

$$P(c_j | x_1 \dots x_5) \propto P(c_j) \cdot P(x_1 | c_j) \cdot P(x_2 | x_1, x_5, c_j) \cdot P(x_3 | c_j) \cdot \\ P(x_4 | x_1, x_3, c_j) \cdot P(x_5 | x_1, x_4, c_j)$$

Broj parametara Bayesovog klasifikatora

$\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$, $K = 2$, $K_k = 3$, $x_1 \perp x_2 | \mathcal{C}_j$, $x_2 \perp x_3 | \mathcal{C}_j$

NB:

$$P(\mathbf{x}, \mathcal{C}_j) \propto P(\mathcal{C}_j) \prod_{i=1}^5 P(x_i | \mathcal{C}_j)$$

FSSJ:

$$P(\mathbf{x}, \mathcal{C}_j) \propto P(x_1, x_2, x_3 | \mathcal{C}_j) P(x_4 | \mathcal{C}_j) P(x_5 | \mathcal{C}_j) P(\mathcal{C}_j)$$

TAN:

$$P(\mathbf{x}, \mathcal{C}_j) \propto P(x_1 | x_2, \mathcal{C}_j) P(x_2 | x_3, \mathcal{C}_j) P(x_4 | \mathcal{C}_j) P(x_5 | \mathcal{C}_j) P(\mathcal{C}_j)$$

$$P(X_1, \dots, X_i | X_{i+1}, \dots, X_n) \Rightarrow \mathcal{O}(K^n)$$

↖ D2

- Bayesov klasifikator je **generativni parametarski** model koji primjere klasificira na temelju **MAP-hipoteze**
- Učenje diskretnog Bayesovog klasifikatora svodi se na **procjenu parametara** (MLE, MAP, bayesovski) multinomijalne razdiobe
- Naivan Bayesov klasifikator koristi pretpostavku o **nezavisnosti značajki za zadanu klasu**, koja omogućuje **generalizaciju** i **pojednostavljuje** model (broj parametara je linearan s n)
- **Polunaivni klasifikatori** (FSSJ, TAN, k -DB) modeliraju zavisnost između odabranih varijabli, čime dobivamo **složeniji model**
- **Bayesova mreža** sažeto prikazuje zajedničku vjerojatnost kao graf kauzalnih veza između varijabli



Sljedeća tema: Bayesov klasifikator (nastavak)