Strojno učenje

Tehnike učenja bazirane na Bayesovom pristupu

Tomislav Šmuc

Literatura

Bayesian Learning ; Hidden Markov Models

- **The Elements of Statistical Learning**

Hastie, Tibshirani, Friedman (Ch 17 Undirected Graphical Models

Chris Bishop: Pattern Recognition and Machine Learning

- Chapter 8: Graphical Models
- Chapter 12: Sequential Models

L.R. Rabiner: ”A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”

Sadržaj

Dva pristupa

Frekvencijski pristup

- „Fizička” izglednost da će se događaj s ishodom *i* zbiti
- Aproksimira se kao omjer broja događaja s ishodom *i* / ukupan broj događaja
- „Točkaste” procjene n/N

Bayesov pristup

- Vjerojatnost je **stupanj uvjerenja** da će se događaj s ishodom *i* zbiti
- Vjerojatnosti moraju biti uvjetovane podacima, t.j. $p(y|\boldsymbol{x})$
- B. pristup = Optimalan pristup (ako je model dobar, apriorna distribucija dobra, funkcija gubitaka dobra…
- „Točkaste” procjene su loše. Valjane vjerojatnosti su dobivene a posteriori i bazirane su na evidenciji (podacima) uz dane apriorne vjerojatnosti

Vjerojatnost

- Distribucije vjerojatnosti $P(X|\xi)$
  - X slučajna varijabla
    - diskretne
    - kontinuirane
  - ξ neko stanje ili informacija o stanju u „pozadini”

Diskretne slučajne varijable

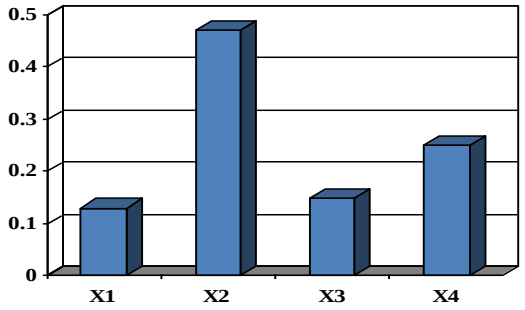
- Konačni broj ishoda

$$X \in \{x_1, x_2, x_3, \ldots, x_n\}$$

$$P(x_i) \geq 0$$

$$\sum_{i=1}^{n} P(x_i) = 1$$

binarna X: $P(x) + P(\bar{x}) = 1$



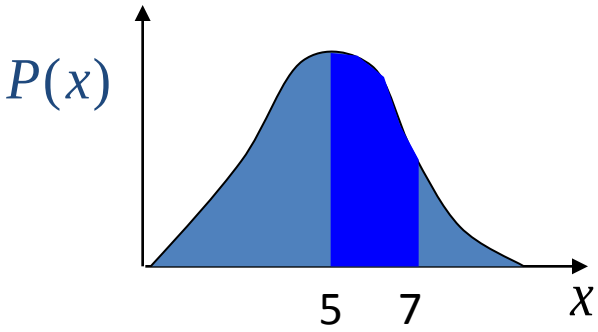
Kontinuirane slučajne varijable

- Distribucije vjerojatnosti (funkcije gustoće vjerojatnosti) preko kontinuiranih vrijedosti

$$X \in [0,10] \qquad P(x) \geq 0$$

$$\int_0^{10} P(x)dx = 1$$

$$P(5 \leq x \leq 7) = \int_5^7 P(x)dx$$



Vjerojatnost (drugi oblici)

- Skupna vjerojatnost

$$P(x, y) \equiv P(X = x \wedge Y = y)$$

  – Vjerojatnost da su (istovremeno) *X*=*x* i Y=y

- Uvjetna vjerojatnost

$$P(x \mid y) \equiv P(X = x \mid Y = y)$$

  – Vjerojatnost da je *X*=*x* u slučajevima kada je *Y*=*y*

Bayes i vjerojatnost, osnovni koncepti

- „Produktno” pravilo

$$P(X,Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

- Marginalizacija

$$P(Y) = \sum_{i=1}^{n} P(Y, x_i)$$ Binarna X: $$P(Y) = P(Y, x) + P(Y, \bar{x})$$

- Bayesovo pravilo

$$P(X,Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

$$\boxed{P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}}$$

Problem

Iz N pokušaja bacanja novčića želimo odrediti vjerojatnost pojave glave u N+1 bacanju.

Rješenje

Bayesovo pravilo: aposteriorna vjerojatnost modela θ uz dane podatke D i znanje (background knowledge) ρ:

$$p(\theta|D,\rho) = \frac{p(\theta|\rho)\cdot p(D|\theta,\rho)}{p(D|\rho)}$$

Gdje je $$p(D|\rho) = \int p(D|\theta,\rho)\cdot p(\theta|\rho)d\theta$$

θ je (u ovom slučaju) varijabla čija vrijednost odgovara mogućoj stvarnoj vrijednosti „fizičke” vjerojatnosti

Funkcija izglednosti(?) („Likelihood”)

Koliko je dobra/ispravna vrijednost θ ?

To ovisi o tome koliko dobro može generirati uočenu sekvencu !

$$L(\theta|D) = P(D|\theta)$$

T.j. izglednost sekvence G,P,G,P,P je npr.

$$L(\theta|D) = \theta \cdot (1-\theta) \cdot \theta \cdot (1-\theta) \cdot (1-\theta)$$

MLE - Maximum Likelihood Estimation

MLE princip:

- Učenje modela => učenje parametara koji maksimiziraju izglednost-vjerojatnost generiranja podataka
- Jedan od najčešćih u statistici – i u strojnom učenju
- Da bi odredili ML – potrebno je imati funkciju (dovoljnu statistiku) koja iz podataka sumira potrebnu informaciju za izračunavanje izglednosti (u slučaju pismo-glava sekvenci => broj G/P

Određivanje P(X|D,κ) ….

Prosjek preko vrijednosti θ da bi se odredila vjerojatnost da će N+1 bacanje biti G:

$$P(X = G \mid D, \kappa) = \int \Theta p(\Theta \mid D, \kappa) d\theta$$

P(X|D,κ) – očekivanje θ s obzirom na distribuciju p(θ/D, κ)

Osnove

- Definicije
- Reprezentacija
- Uvjetna nezavisnost

Bayesova mreža

- grafički model probabilističkih odnosa između grupa varijabli, koji efikasno definira/kodira združenu distribuciju vjerojatnosti vrijednosti svih varijabli u skupu

Zašto je pristup (učenju) preko Bayesovih mreža interesantan ?

- Okvir za učenje o kauzalnim odnosima
- Zaključivanje na nekompletnim podacima
- Olakšano kombiniranje znanja u domeni (background knowledge) i podataka
- Efikasan način izbjegavanja overfitting-a podataka

Definicija:

- Bayesova mreža nad skupom varijabli $\boldsymbol{X} = \{X_1, X_2, \ldots X_n\}$ je:
  - Mrežna struktura **S** koja kodira „tvrdnje” o uvjetnoj nezavisosti među varijablama iz $\boldsymbol{X}$
  - Skup lokalnih distribucija (uvjetnih) vjerojatnosti
- Mrežna struktura **S** je usmjereni aciklički graf (DAG);
- Čvorovi su varijable;
- Veze između čvorova označuju probabilističku zavisnost između dvije varijable;
- Uvjetne vjerojatnosti kodiraju jačinu veze između vrijednosti varijabli

Konstrukcija Bayesove mreže (ekspertski pristup)

Pristup se zasniva na slijedećim praktičnim uvidima:

- Ljudi često dobro poznaju kauzalne odnose između varijabli;
- Kauzalni odnosi tipično korespondiraju sa tvrdnjama uvjetne vjerojatnosti;
- Da bi konstruirali BM jednostavno crtamo veze između varijabli i to od uzročnih prema njihovim neposrednim efektima;
- U konačnom koraku određujemo lokalne distribucije vjerojatnosti.

Određivanje u praksi (korištenje) može dovesti do promjene strukture mreže

Kriteriji za odabir modela

Kriterij mora biti dovoljno dobar da odredi stupanj do kojeg mreža odgovara prethodnom znanju i podacima

Prmjer kriterija

- Relativna posteriorna vjerojatnost

Logaritam relativne posteriorne vjerojatnosti :

$$\log(p(D|S)) = \underbrace{\log\big(p(S)\big)}_{\text{log apriorne vjerojatnosti}} + \underbrace{\log(p(D|S))}_{\text{log marginalne vjerojatnosti}}$$

Apriorne vjerojatnosti

Potrebne su za izračun relativne posteriorne vjerojatnosti

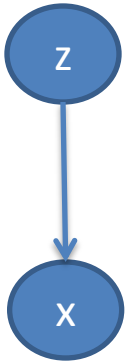
- Apriorne strukturne vjerojatnosti p(S)
  - Pretpostavka da je svaka hipoteza jednako vjerojatna
  - Varijable mogu biti poredane a prisustvo/odsustvo veza su međusobno nezavisni
  - Mogu se koristiti i prije određene mreže (sa sličnom namjenom)
- Parametarske apriorne vjerojatnosti $p(\theta|S)$

Koristi od učenja strukture BM

- Efikasno učenje/modeliranje → točniji modeli sa manje podataka
  - P(A) i P(B) → velika ušteda podataka u odnosu na P(A,B) !
- Otkrivanje strukturnih (kauzalnih) svojstava domene
- Uređivanje događaja koji se događaju sekvencionalno
- Pomaže u analizama osjetljivosti te u zaključivanju
- Predviđanje efekata nekih akcija

Terminologija i semantika BM

Z je roditelj od X



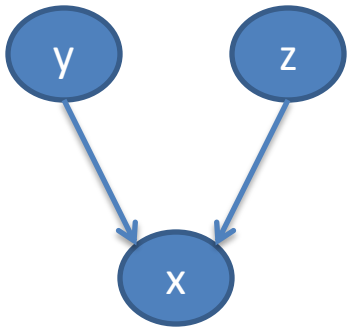
Z utječe na X;
Z uzrokuje X;
X ovisi o Z

X je dijete od X

BM => DAG => izjave o zavisnosti/nezavisnosti među varijablama
Nezavisnost => familija distribucija vjerojatnosti

Primjer: bacanje dva novčića

P(y):
G:0.5
P:0.5

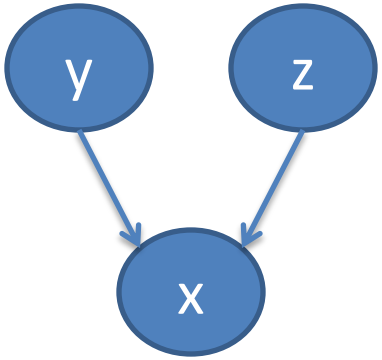


P(z):
G:0.5
P:0.5

x	GG	GP	PG	PP
P(x|y,z):	0.25	0.25	0.25	0.25

Definiranje BM

1. Struktura (zavisnost ,nezavisnost)
2. Distribucija vjerojatnosti



BM

Kompaktna reprezentacija odnosa varijabli

Faktorizacija distribucije vjerojatnosti

Svaka distribucija vjerojatnosti konzistentna sa nekim DAG – morase faktorizirati prema:

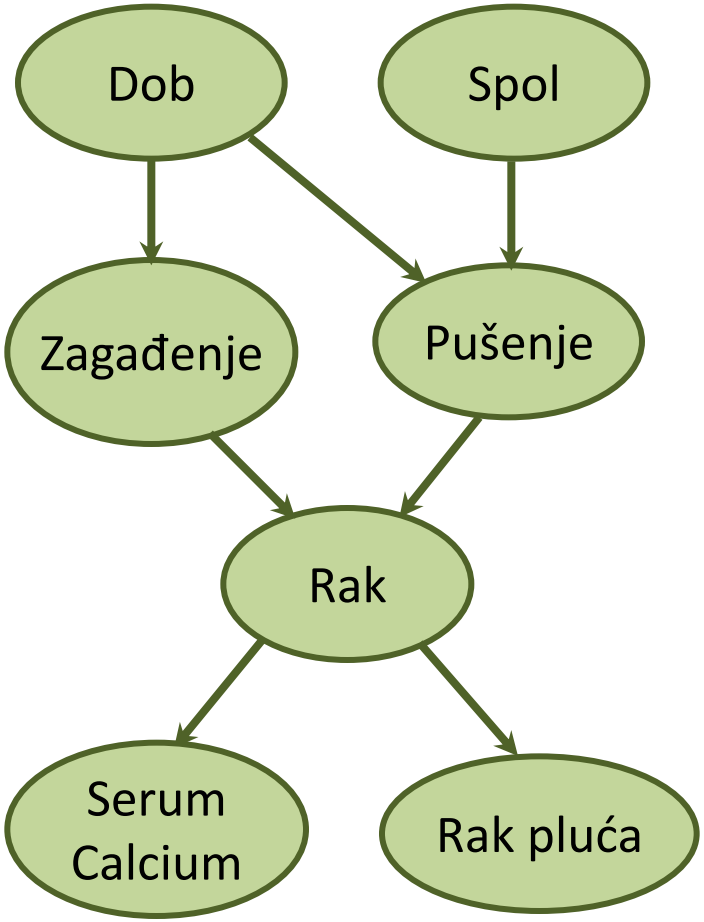
$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \boldsymbol{R_i})$$

gdje je

$$\boldsymbol{R_i} - \text{oznaka za roditelje } X_i$$

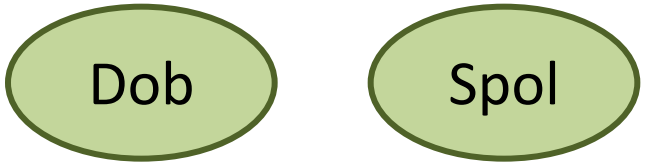
(pravilo produkta za BM!)

Združena vjerojatnost - faktorizacija



$$P(D,S,Z,Pu,R,SC,Rp)$$
$$= P(D)\cdot P(S)\cdot P(Z\mid D)$$
$$\cdot P(Pu\mid D,S)\cdot P(R\mid Pu,Z)$$
$$\cdot P(SC\mid R)\cdot P(Rp\mid R)$$

Nezavisnost varijabli



Dob i spol su nezavisni ⇒ $P(D,S) = P(D)P(S)$

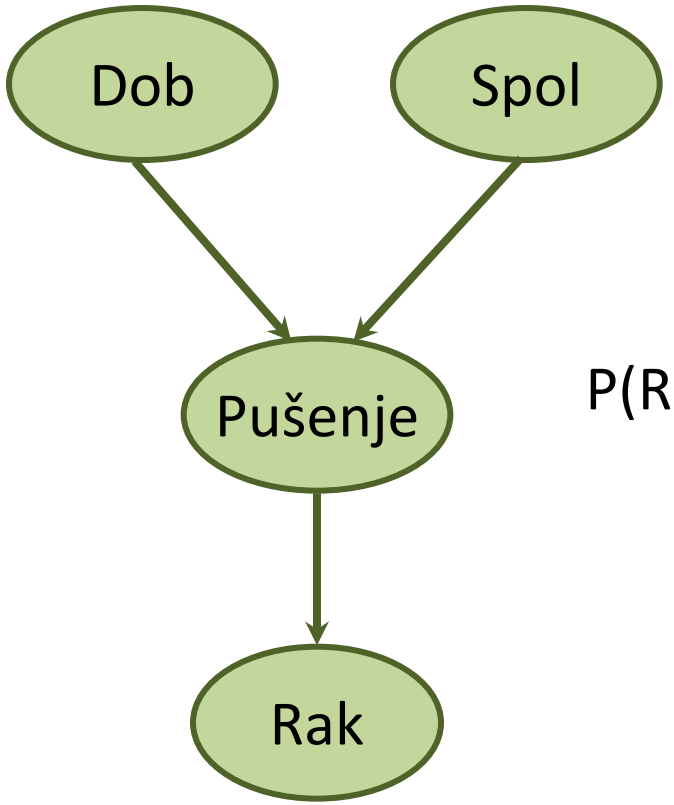
$P(D|S) = P(D)$ $\quad D \perp S$

$P(S|D) = P(S)$ $\quad S \perp D$

$P(D,S) = P(S|D)\,P(D) = P(S)P(D)$

$P(D,S) = P(D|S)\,P(S) = P(D)P(S)$

Uvjetna nezavisnost u BM

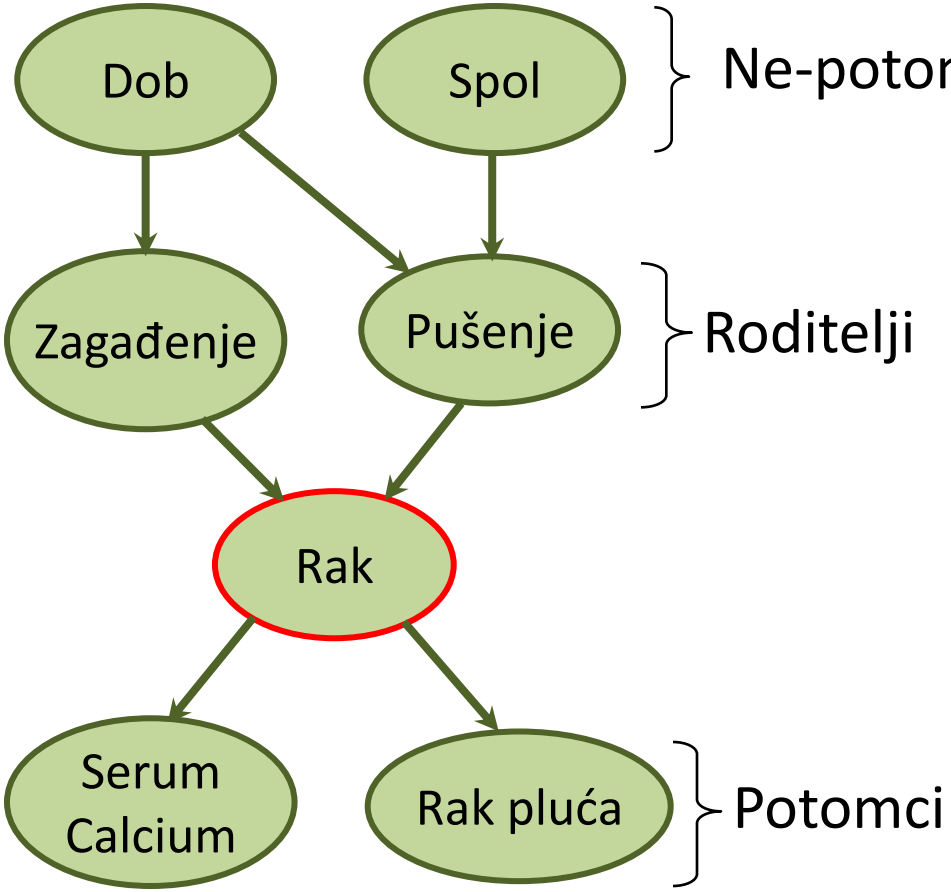


Ukoliko znamo status varijable *Pušenje, Rak* postaje nezavisna o *Dobi i Spolu.*

$P(R|D,S,Pu) = P(R|Pu) \quad R \perp D,S \mid Pu$

Uvjetna nezavisnost u BM

Varijabla (čvor) je uvjetno nezavisan od svojih ne-potomaka, uz zadane roditelje

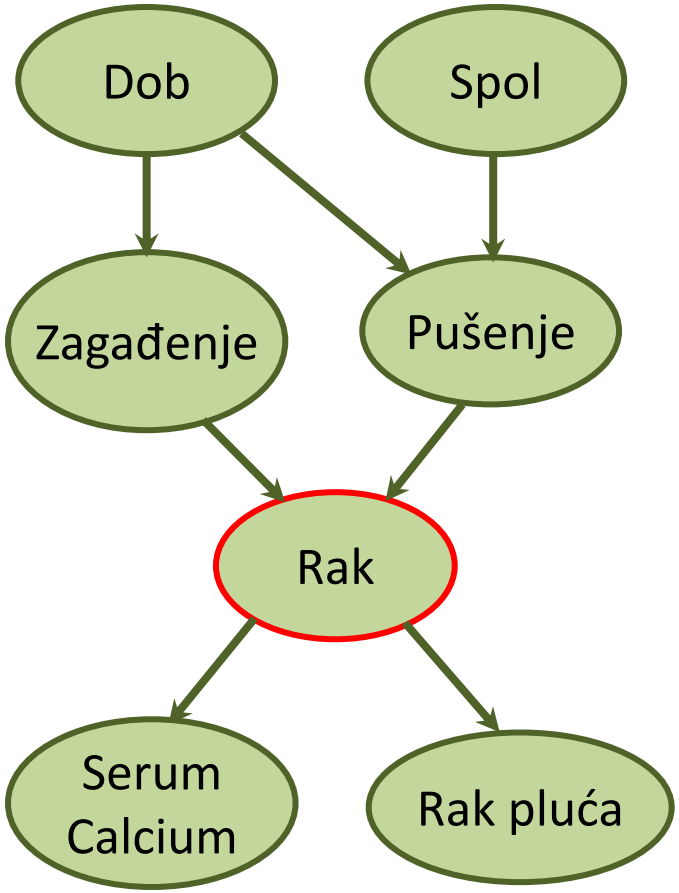


Rak je nezavisan od *Dobi* i *Spola* uz zadano *Zagađenje* i *Pušenje*.

Nezavisnost i odvojenost u DAG

- Uz dane (neke) opservacije, da li je neki skup varijabli neovisan o drugom skupu ?
- Opservacije induciraju zavisnosti !
- **d-separacija** (Pearl 1988) – kako grafički provjeriti uvjetnu nezavisnost u grafu

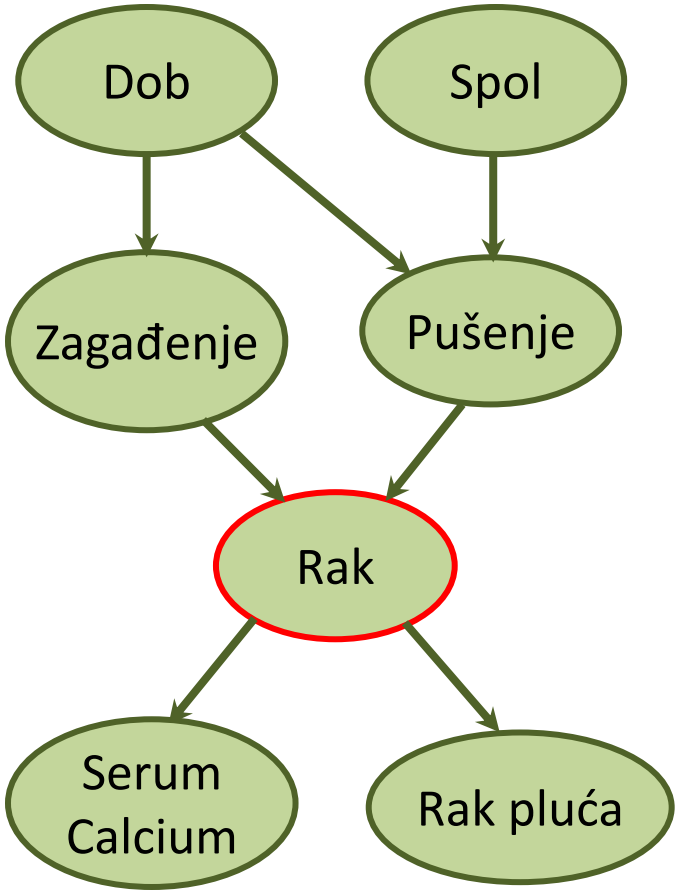
Prediktivno zaključivanje



Koliko je vjerjatno da
Stariji muškarci
Obole od *Raka*?

$$P(R=Da \mid Dob > 60, Spol= muški)$$

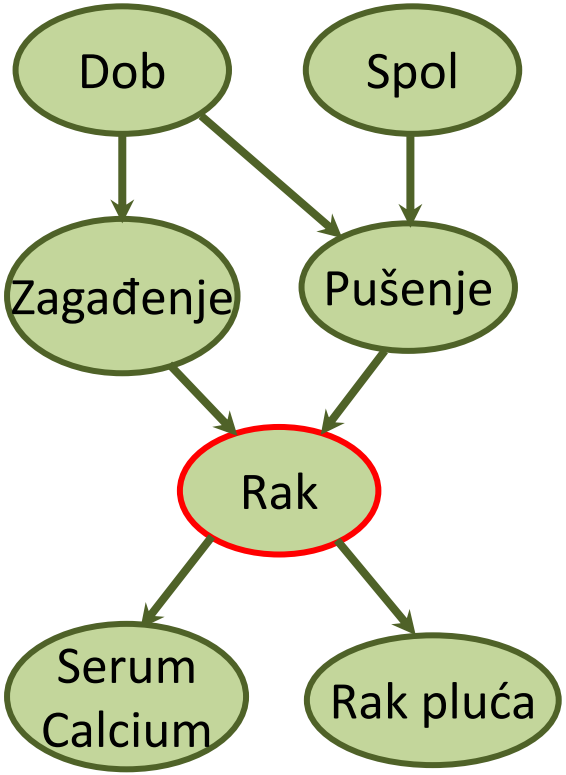
Prediktivno zaključivanje



Koliko je vjerojatno da *stariji muškarci* Sa povišenim *SC* dobiju *Rak*?

$$P(R=Da \mid Dob > 60, Spol= muški, SC=visok)$$

Efekt “objašnjavanja” (explaining away)



- Ako smo otkrili *Rak pluća*, vjerojatnost *povišenog pušenja* i *zagađenja* rastu !
- Ako smo dodatno ustanovili prisutnost *zagađenja* vjerojatnost povišenog pušenja pada !

Osnovno zaključivanje: Pravilo produkta

$P(R_j, Pu_i)$

$Pu\Downarrow \quad R\Rightarrow$	ne	da
ne	0.70	0.05
malo	0.13	0.07
puno	0.03	0.02

$Pu \in \{ne, malo, puno\}$ $\qquad$ $Rak \in \{ne, da\}$

$P(R_j \mid Pu_i)$

Pušenje=	ne	malo	puno
P(R=ne)	0.933	0.65	0.6
P(R=Da)	0.067	0.35	0.4

Pravilo produkta

$$P(R, Pu) = P(R \mid Pu)P(Pu)$$

$P(R, Pu)$

Pu⇓ R⇒	ne	da
ne	0.70	0.05
malo	0.13	0.07
puno	0.03	0.02

Marginalizacija

Pu⇓ R⇒	ne	da	ukupno
ne	0.70	0.05	0.75
malo	0.13	0.07	0.20
puno	0.03	0.02	0.05
ukupno	0.86	0.14	

$P(Pu)$

$P(R)$

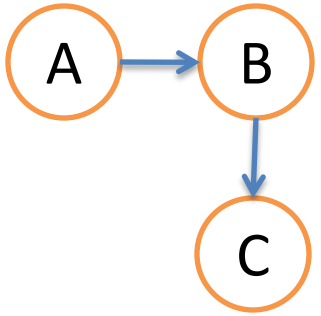
$$P(Pu \mid R) = \frac{P(R \mid Pu)P(Pu)}{P(R)} = \frac{P(R, Pu)}{P(R)}$$

Pu⇓ R⇒	ne	da
ne	0.70/0.86	0.05/0.14
malo	0.13/0.86	0.07/0.14
puno	0.03/0.86	0.02/0.14

$$P(R \mid Pu) = \frac{P(Pu \mid R)P(R)}{P(Pu)} = \frac{P(R, Pu)}{P(Pu)}$$

Pu⇓ R⇒	ne	da
ne	0.70/0.75	0.05/0.75
malo	0.13/0.20	0.07/0.20
puno	0.03/0.05	0.02/0.05

Osnovno zaključivanje



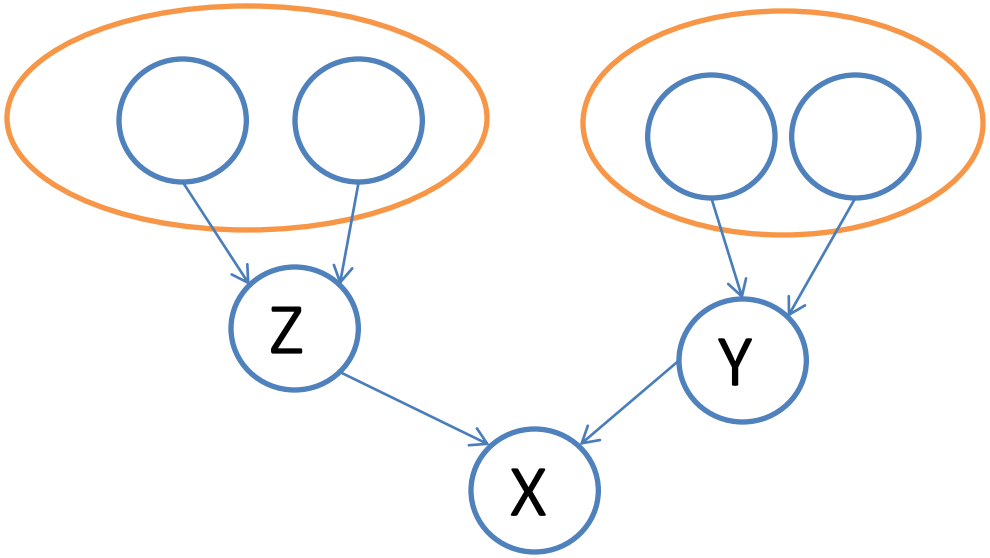
$$\underbrace{P(b)} = \sum_a P(a, b) = \sum_a P(b \mid a)\, P(a)$$

$$P(c) = \sum_b P(c \mid b)\, \overbrace{P(b)}$$

$$P(c) = \sum_{b,a} P(a, b, c) = \sum_{b,a} P(c \mid b)\, P(b \mid a)\, P(a)$$

$$= \sum_b P(c \mid b) \underbrace{\sum_a P(b \mid a)\, P(a)}_{P(b)}$$

Osnovno zaključivanje



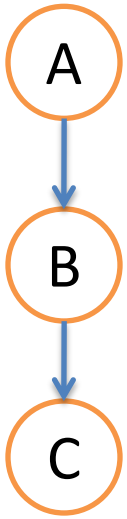
Stabla

$$P(x) = \sum_{z,y} P(x \mid z, y)\, P(z, y)$$

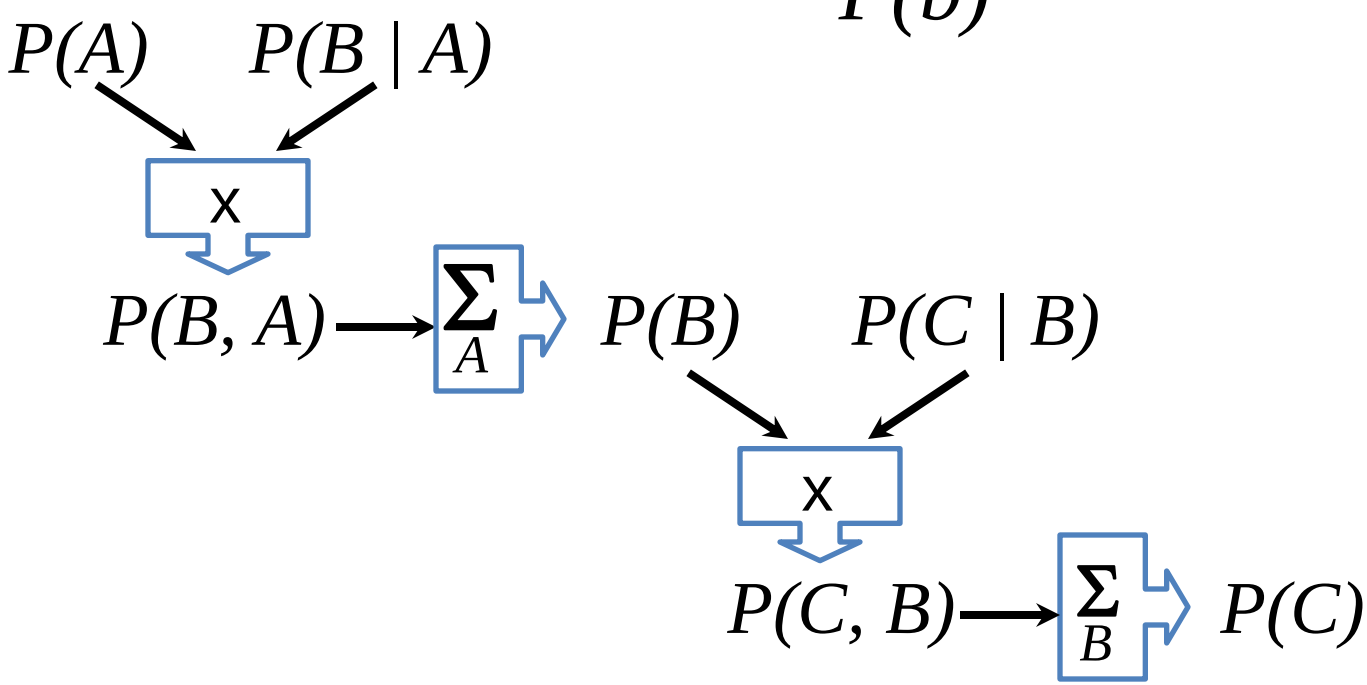
Zato što su z, y nezavisne

$$= \sum_{z,y} P(x \mid z, y)\, P(z)\, P(y)$$

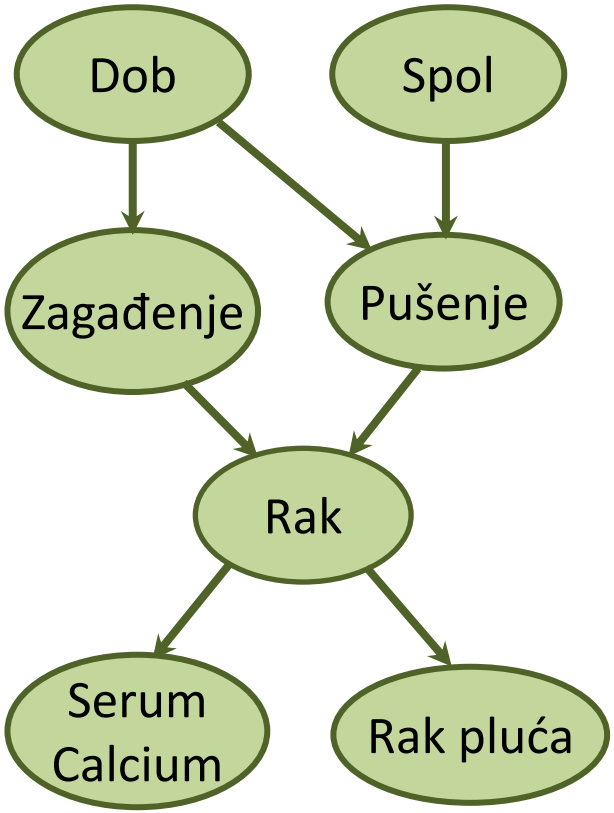
Eliminacija varijabli



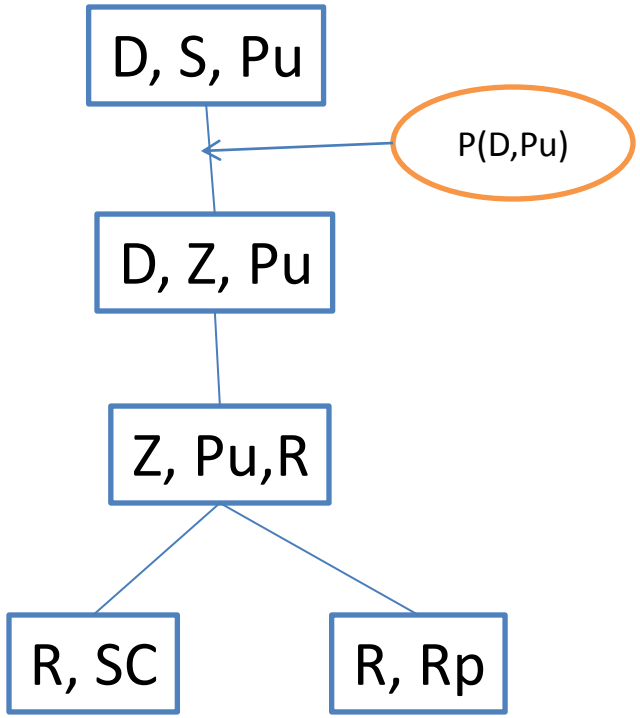
$$P(c) = \sum_b P(c \mid b) \underbrace{\sum_a P(b \mid a)\, P(a)}_{P(b)}$$



Združena stabla (Join trees)



Parcijalna faktorizacija



BM zaključivanje kao eliminacija varijabli

- Faktorizacija preko **X** je funkcija koja preslikava vrijednosti **X** na [0,1]:
  - Tablica uvjetnih vjerojatnosti
  - Združena distribucija vjerojatnosti
- BM zaključivanje:
  - Faktore možemo množiti da bi dobili nove
  - Možemo sumirati po varijablama (marginalizacija) i time ih "izbaciti" iz daljnjeg razmatranja
  - Varijabla se može izbaciti (sumirati) kada su svi faktori koji je koriste izmnoženi

Učenje Bayesovih mreža iz podataka

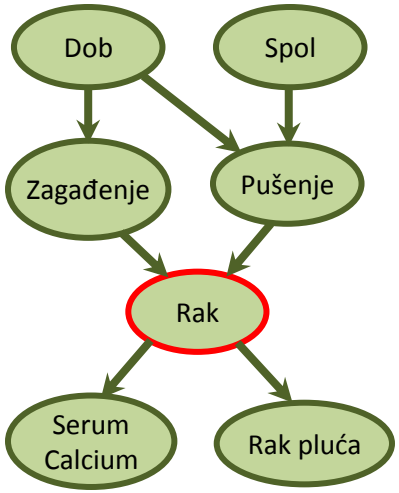
Zadaci

- Učenje parametara
  - Kompletni podaci
  - Nekompletni podaci (missing values)
- Učenje strukture

Učenje BM

D	*S*	*Z*	*Pu*	*R*
<30	z	v	p	ne
<40	m	?	m	da
⋮				

skup za učenje



- Ulaz: kompletni ili nekompletni podaci?
- Model: Parametri uz zadanu mrežu ili parametri +struktura?

MLE (podsjetnik)

- Ulaz: rezultat bacanja

$$X_1, \ldots, X_n = \{\underbrace{G,G,G,P,P,P,\ldots\ldots\ldots\ldots,G,P}\}$$

  - Procjena θ
  - Izglednost $P(X_1, \ldots, X_n \mid \theta) = \theta^{G} (1-\theta)^{P}$
  - ML rezultat: $\theta^* = \frac{G}{G+P}$

Učenje BM parametara

- Kompletni podaci:

$$\theta_{x|z} = \frac{\#\ primjera\ x,\ z}{\#primjera\ z} = \frac{\sum_j I\ (x,z \mid d_j)}{\sum_j I(z \mid d_j)}$$

$$I(z|d_j) = \begin{cases} 1 & Ako\ Z{=}z\ za\ podatak\ d_j \\ 0 & inače \end{cases}$$

- Nekompletni podaci $I(z|d_j)$ nije poznato?!

Procijeniti I : $\hat{I}(x,z|d_j) = P_\theta\ (x,z|d_j)$

No – ne znamo θ !

BM učenje parametara u slučaju nekompletnih podataka – EM algoritam

Iteriraj do konvergencije ($|\ \Delta\theta\ | \sim 0$)

- Izračunaj očekivanje (Expectation – E step)
  - Koristi trenutne parametre θ_k da bi popunio MV u podacima

$$\hat{I}(x, z|d_j) = P_{\theta_k}(x, z|d_j)$$

- Izračunaj novi θ_{k+1} (Maximization -M step)
  - Koristi komplet(ira)ne podatke da bi dobio novi MLE θ

$$\theta_{x|z} = \frac{\sum_j \hat{I}(x, z|d_j)}{\sum_j \hat{I}(x|d_j)}$$

Učenje strukture BM

Cilj:
"Dobra" BM (u odnosu na podatke)

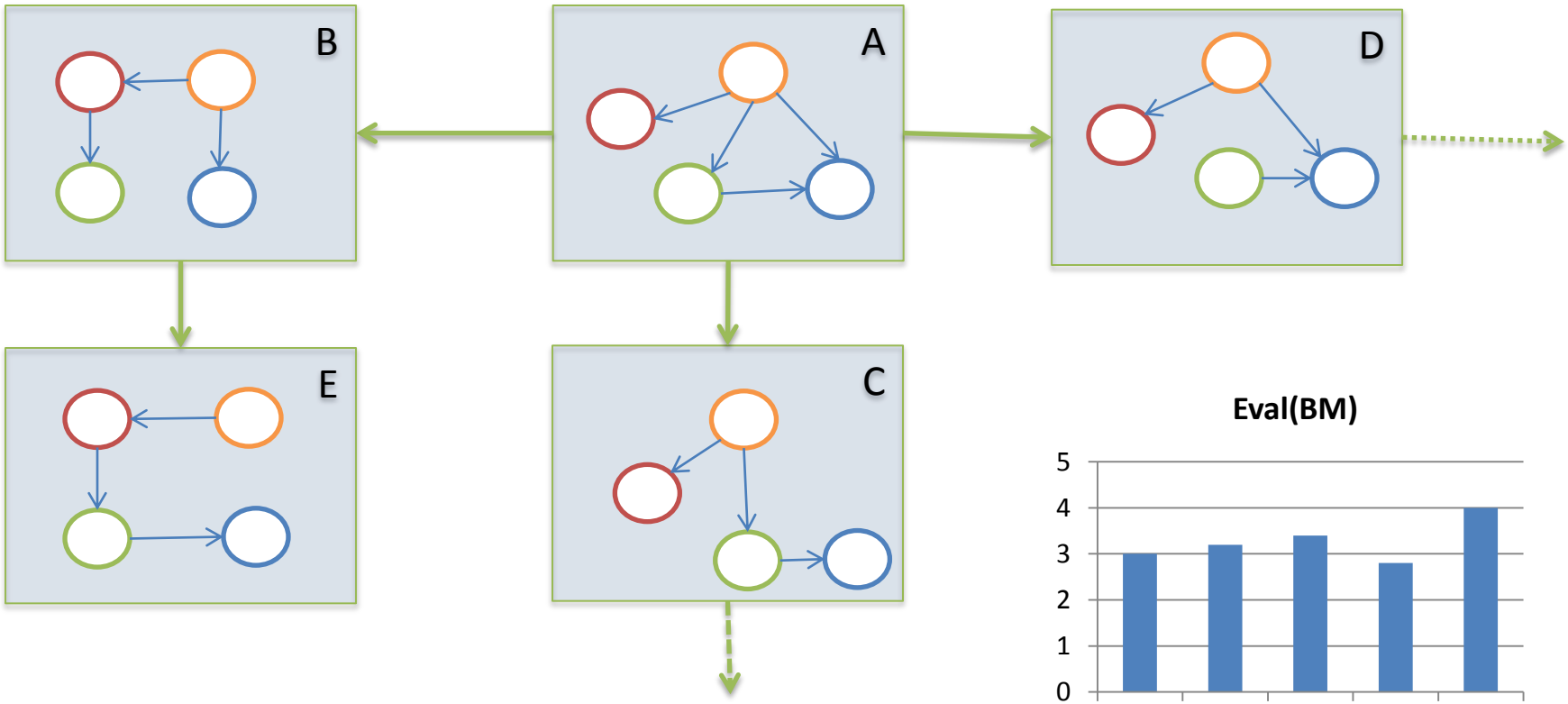
Tipično rješenje:
Heurističko pretraživanje prostora struktura

Učenje strukture BM = pretraživanje

Prostor pretraživanja = prostor mrežnih struktura

Operatori = dodaju/brišu veze; mijenjaju usmjerenje veza

Algoritmi pretraživanja= greedy, hill-climbing, GA…



Evaluacija BM strukture

- Uz popunjene parametre (CPT) korištenjem tehnika (faktorizacija, eliminacija varijabli) moguća je evaluacija kompletne mreže
- Evaluacija => ML funkcija:

$$ML\ (BM) = P(podaci \mid BM)$$

Tipično => potpuni graf !? (overfitting)

MDL mjere

- MDL - Minimum Description Length:

  (regularizacija!) uzima u obzir kompleksnost modela (broj parametara ~ veza u BM)

$$Eval(BM) = P(data \mid BM) - kompleksnost\ (BM)$$

Online materijali i tutoriali

K. Murphy (2001) http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html

W. Buntine. Operations for learning with graphical models. Journal of Artificial Intelligence Research, 2, 159-225 (1994).

D. Heckerman (1999). A tutorial on learning with Bayesian networks. In Learning in Graphical Models (Ed. M. Jordan). MIT Press.

Knjige:

Daphne Koller and Nir Friedman, "Probabilistic graphical models: principles and techniques", MIT Press 2009

J. Pearl (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press.

M. I. Jordan (ed, 1988). Learning in Graphical Models. MIT Press.

Algoritmi:

C. Huang: Inference in Belief Networks: A Procedural Guide PPTC (Probability Propagation in Trees of Clusters)

Software

- B-Course (Online course + demo; CoSco – University of Helsinki)
- BAYDA: http://www.cs.Helsinki.FI/research/cosco

  Classification
- GeNIe + SMILE (http://genie.sis.pitt.edu/)
- BN Power Constructor: BN PowerConstructor
- Microsoft Research: WinMine

  http://research.microsoft.com/~dmax/WinMine/Tooldoc.htm
- BUGS: http://www.mrc-bsu.cam.ac.uk/bugs
- Hugin: http://www.hugin.dk