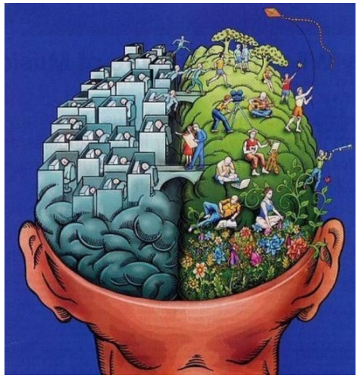
Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs rendszerek Tanszék

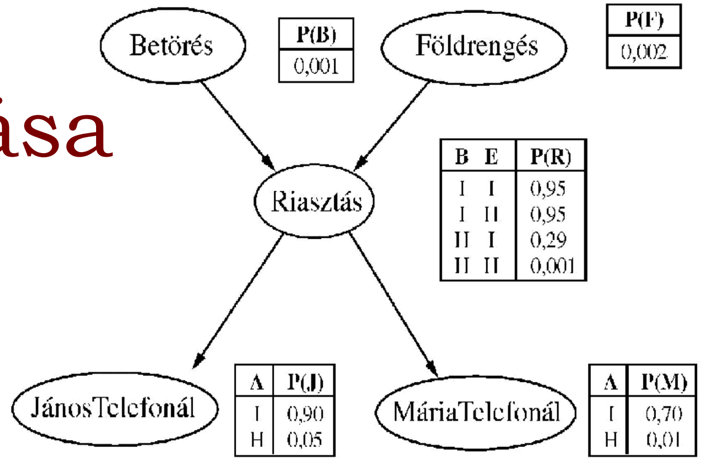


Bayes-hálók tanulása

Előadó:	Hullám Gábor
Előadás anyaga:	Dobrowiecki Tadeusz Antal Péter

MŰEGYETEM 1782

Valószínűségi hálók tanítása

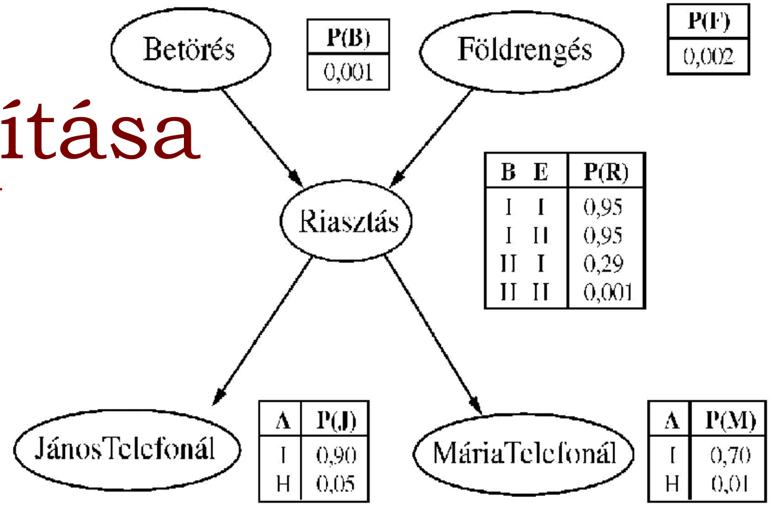


Ismert struktúra, teljesen megfigyelhető változók:

- tanulandó: a feltételes valószínűségek táblázata, közvetlenül becsülhető a példahalmaz alapján

paramétertanulás

Valószínűségi hálók tanítása



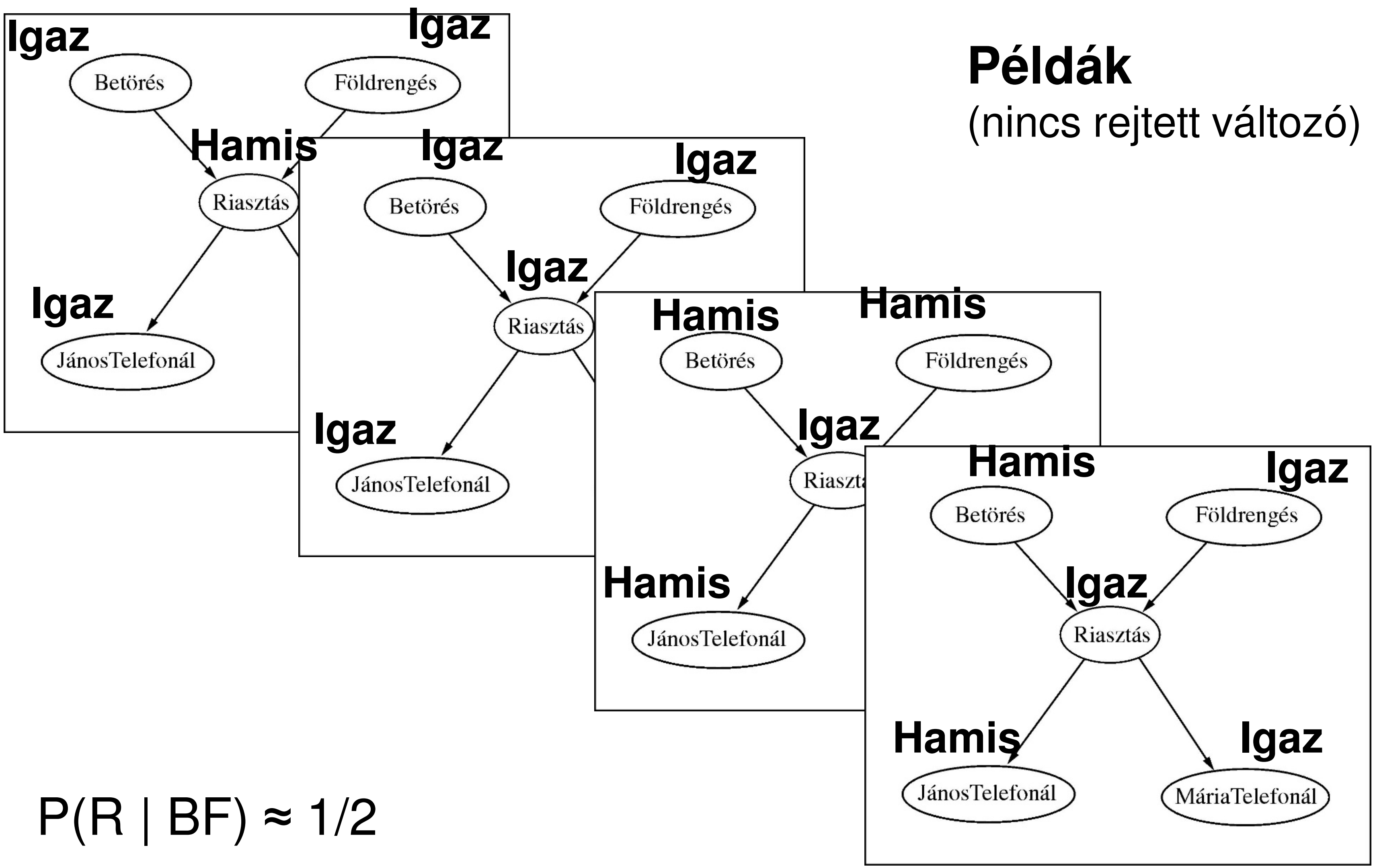
Ismeretlen struktúra, teljesen megfigyelhető változók:

tanulandó: a háló topológiája

- struktúrák terében való keresés, a keresést az egyes struktúráknak az adatok modellezési képessége irányítja

struktúratanulás

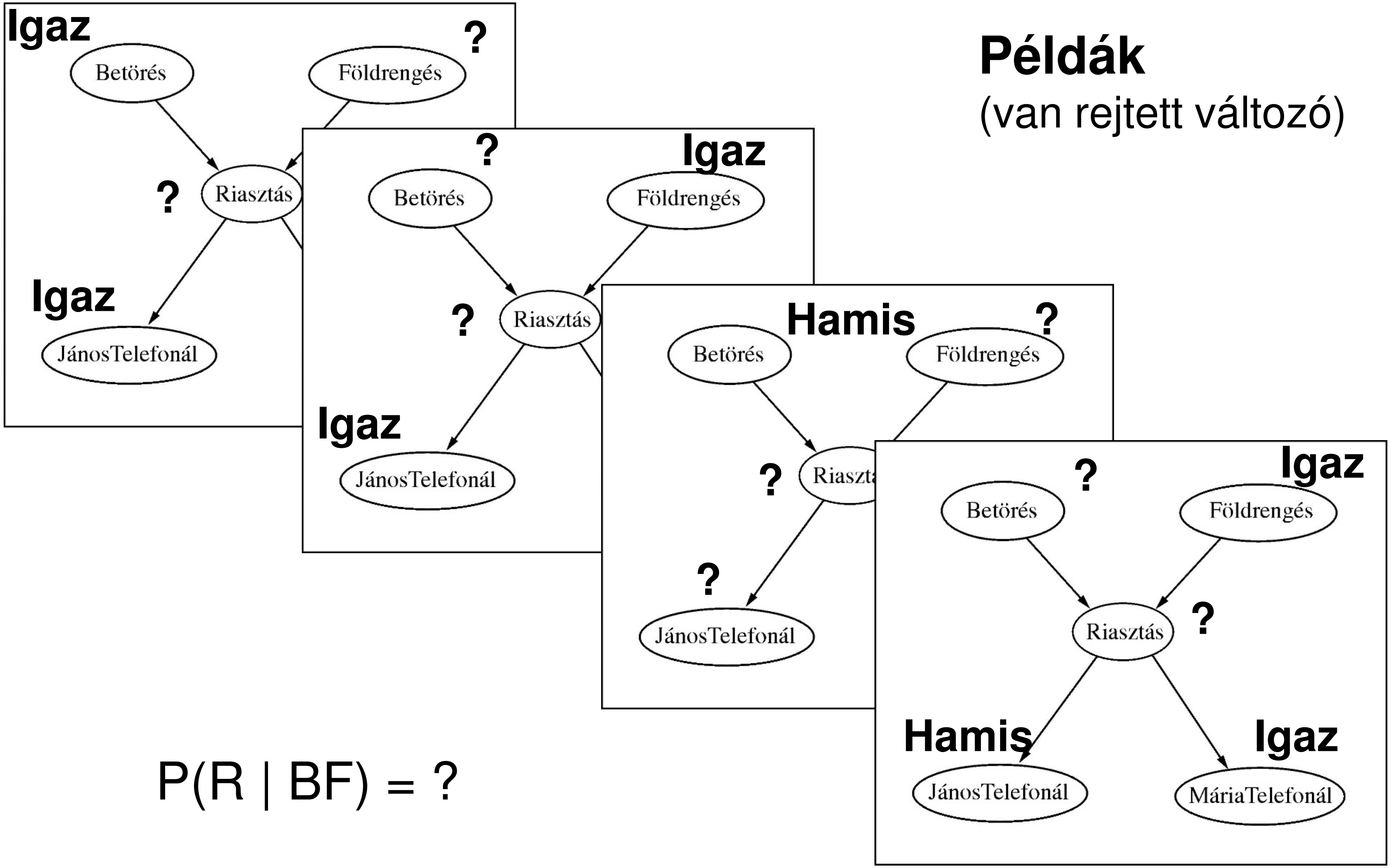
Példák
(nincs rejtett változó)



$P(R \mid BF) \approx 1/2$

…

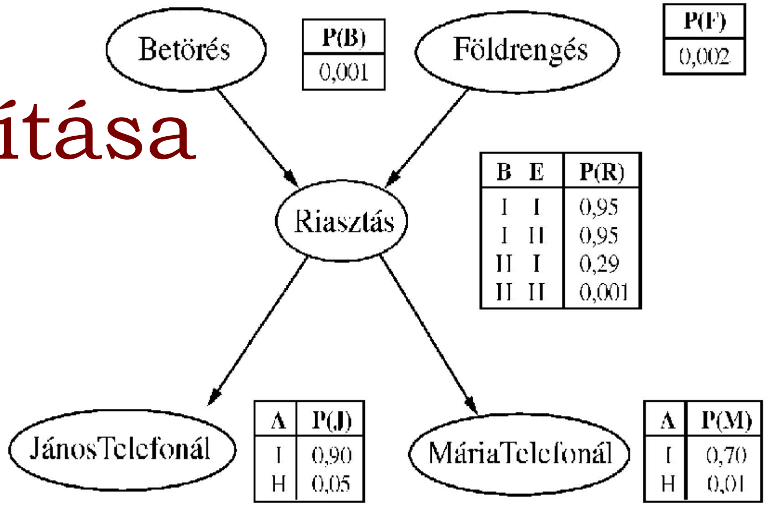
Példák
(van rejtett változó)



$P(R \mid BF) = ?$

...

Valószínűségi hálók tanítása



Ismert struktúra, rejtett változók:

- EM (expectation maximization) algoritmus

Ismeretlen struktúra, rejtett változók:

- Strukturális EM
- Egyéb közelítő megoldások

Gépi tanulás (MSc)

Valószínűségi következtető és döntéstámogató rendszerek (MSc)

Bayes tanulás

Általános predikció - hipotézis konstruálása adatok alapján:

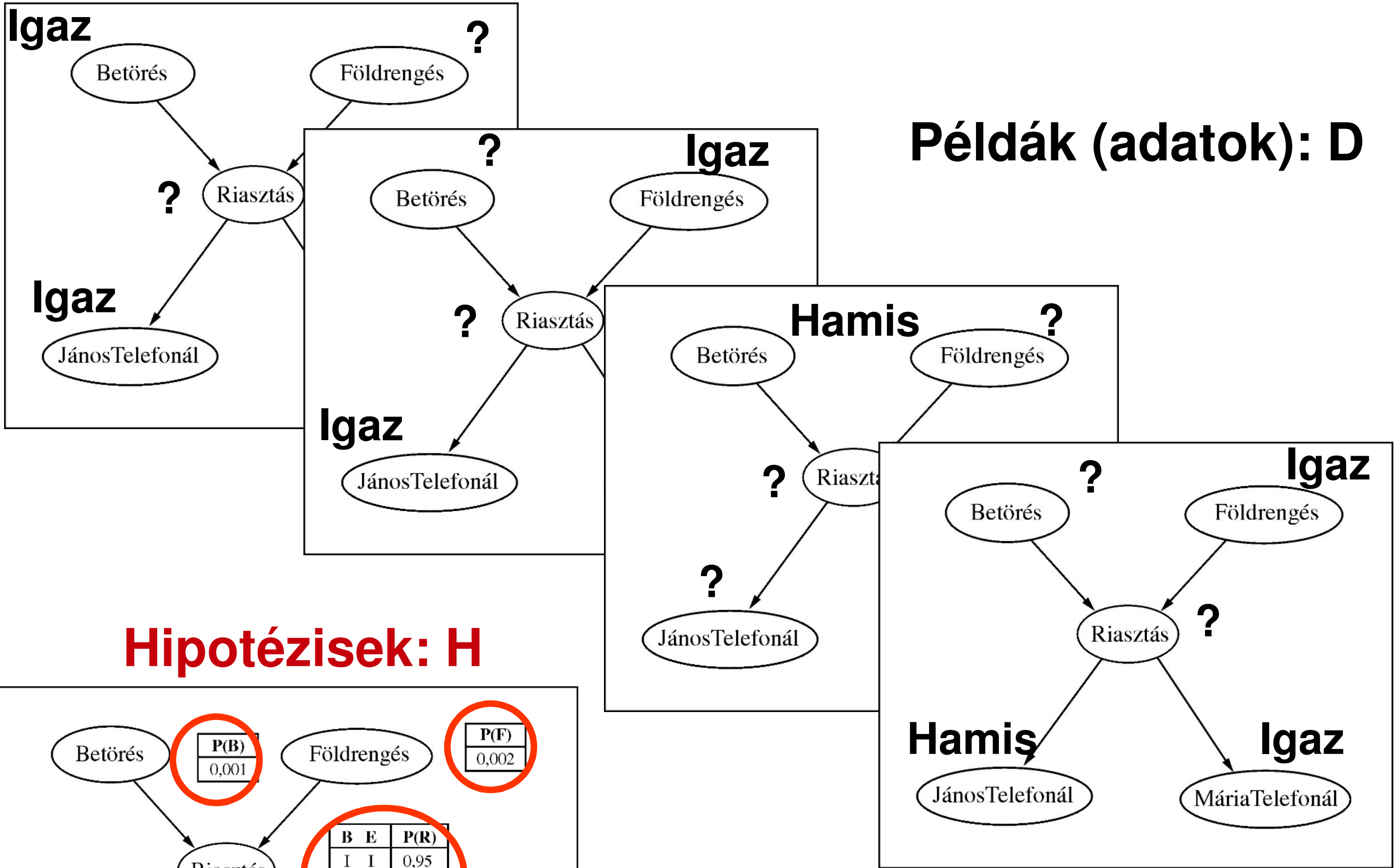
1. az összes hipotézis valószínűségét megbecsüljük az adatokból.
2. a hipotézisek alapján predikciókat készítünk.
3. a predikciókat a hipotézisek a posteriori valószínűségével súlyozzuk.

D adatok, $\boldsymbol{H_1}$, $\boldsymbol{H_2}$,... hipotézisek, ismeretlen ***X*** predikciója
- minden egyes $\boldsymbol{H_i}$ az ***X*** egy teljes eloszlását specifikálja

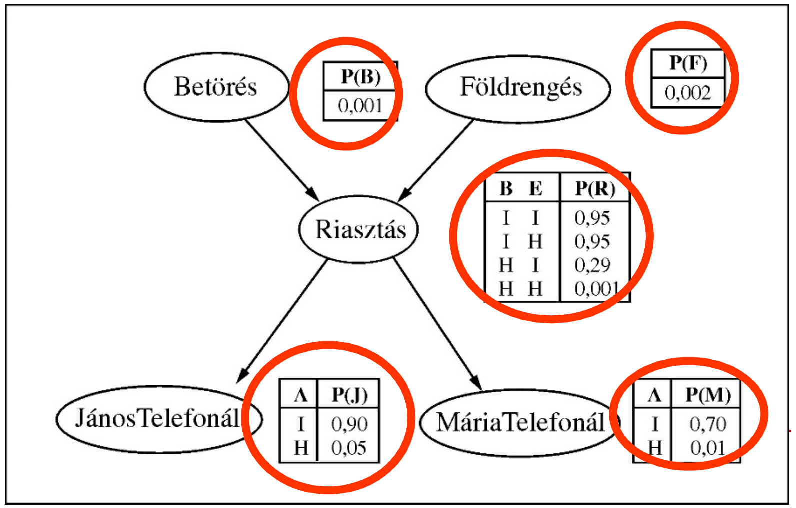
$$P(X) = \sum_i P(X \mid H_i) P(H_i)$$

$$P(X \mid D) = \sum_i P(X \mid D, H_i) P(H_i \mid D) = \sum_i P(X \mid H_i) P(H_i \mid D)$$

Példák (adatok): D



Hipotézisek: H



Bayes-tanulás

D adatok, ***H*₁**, ***H*₂**,... hipotézisek, ismeretlen ***X*** predikciója
- minden egyes ***H*ᵢ** az ***X*** egy teljes eloszlását specifikálja

$$P(X) = \sum_i P(X \mid H_i) P(H_i)$$

$$P(X \mid D) = \sum_i P(X \mid D, H_i) P(H_i \mid D) = \sum_i P(X \mid H_i) P(H_i \mid D)$$

- $P(H_i|D) = \alpha \cdot P(D|H_i) \cdot P(H_i)$

posterior likelihood prior

Teljes Bayes tanulás: P(H_i | D) kiszámítása az összes H_i-re.
A legtöbb esetben kezelhetetlen, de nincs jobb módja az optimalis predikció készítésnek.

Teljes bayesi tanulás közelítése

$$P(X) = \sum_i P(X \mid H_i) P(H_i)$$

$$P(X \mid D) = \sum_i P(X \mid D, H_i) P(H_i \mid D) = \sum_i P(X \mid H_i) P(H_i \mid D)$$

A legelterjedtebb közelítés - a **legvalószínűbb hipotézis:** max P(H_i | D)

P(H_{MAP} | D) ≈ 1, P($H_{más}$ | D) ≈ 0

Maximum a posteriori (**MAP**) hipotézis melletti predikció:

$$P(X \mid D) \approx P(X \mid H_{MAP})$$

A probléma a H_{MAP} megtalálása.

MAP hipotézis:

$$\max\left(P(H_i \mid D) = \frac{P(D \mid H_i) P(H_i)}{P(D)}\right)$$

MAP és ML tanulás

Maximálizálás: ‚probléma' az ***a priori*** valószínűség

$$P(H_i|D) = \frac{P(D|H_i) \cdot P(H_i)}{P(D)}$$

- *a priori* valószínűségeket rendeljünk az egyes hipotézisekhez, összegük 1
- bizonyos esetekben: **egyenletes** *a priori* eloszlás
  - tudás hiánya, de néha jobb nem felhasználni korábbi eredményeket

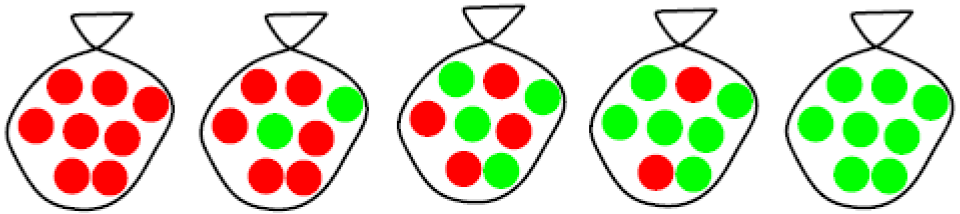
$$\max_i P(H_i \mid D) = \max_i P(D \mid H_i)$$

maximum likelihood (ML) hipotézis: $\boldsymbol{H_{ML}}$

Bayesian Model Averaging example

Suppose there are five kinds of bags of candies:

10% are h_1: 100% cherry candies
20% are h_2: 75% cherry candies + 25% lime candies
40% are h_3: 50% cherry candies + 50% lime candies
20% are h_4: 25% cherry candies + 75% lime candies
10% are h_5: 100% lime candies

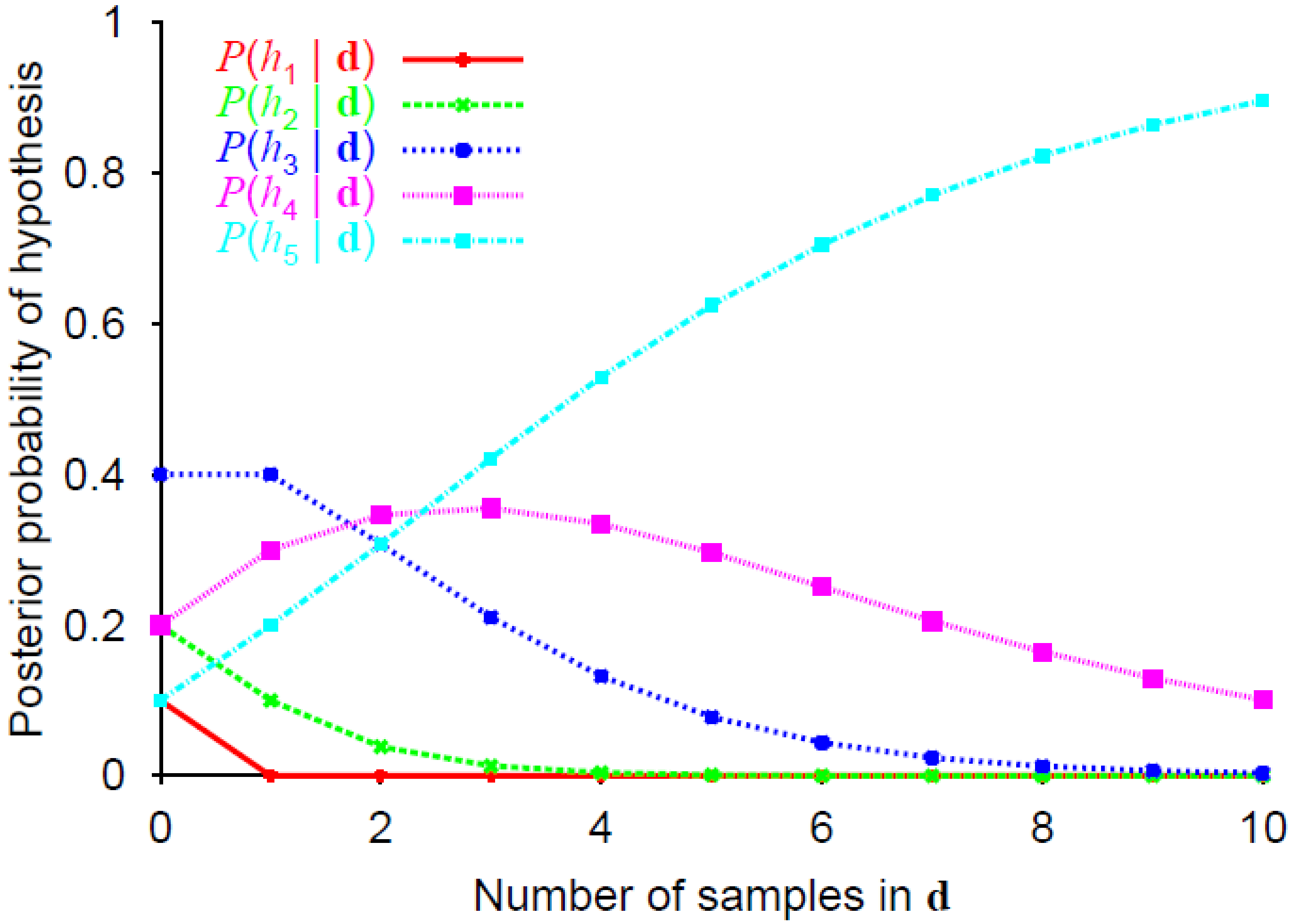


Then we observe candies drawn from some bag:

What kind of bag is it? What flavour will the next candy be?

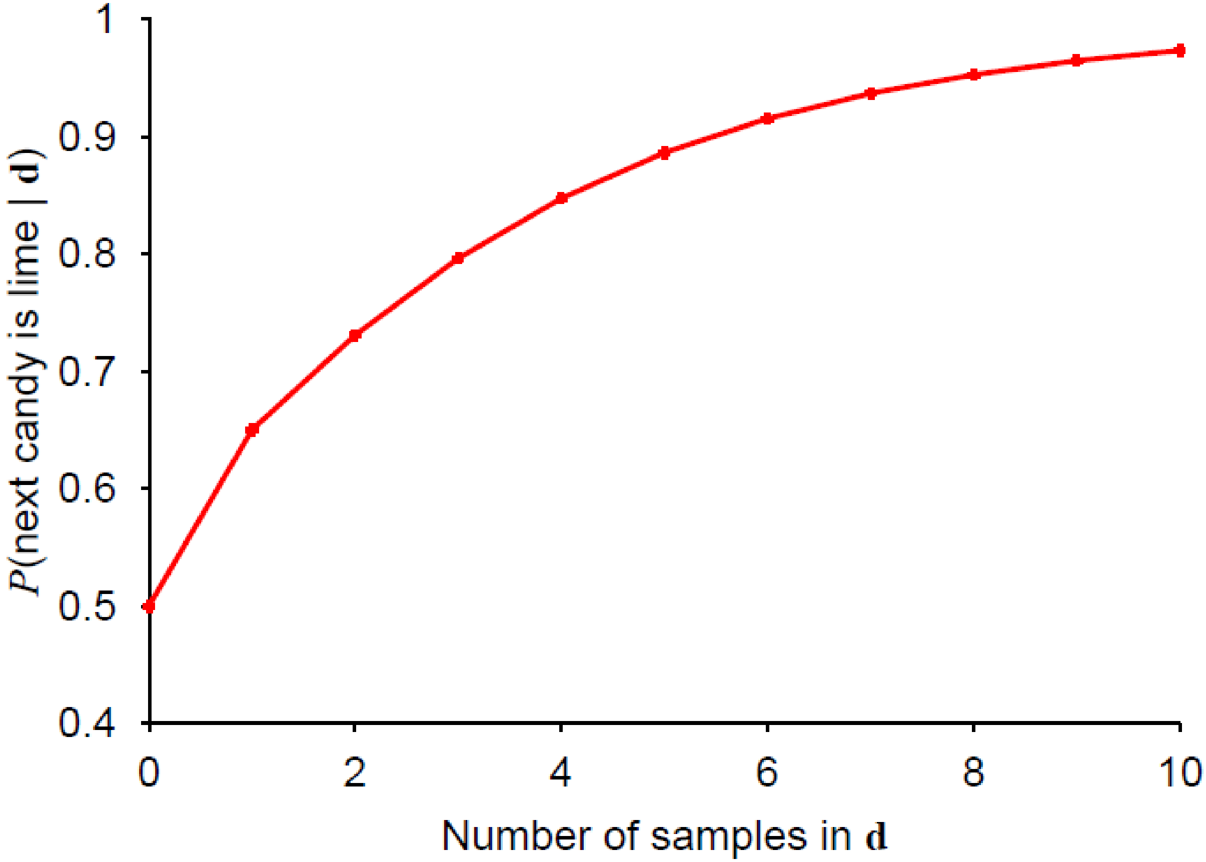
Russel&Norvig: Artificial intelligence

Az egyes modellek predikciójának a posteriori valószínűsége



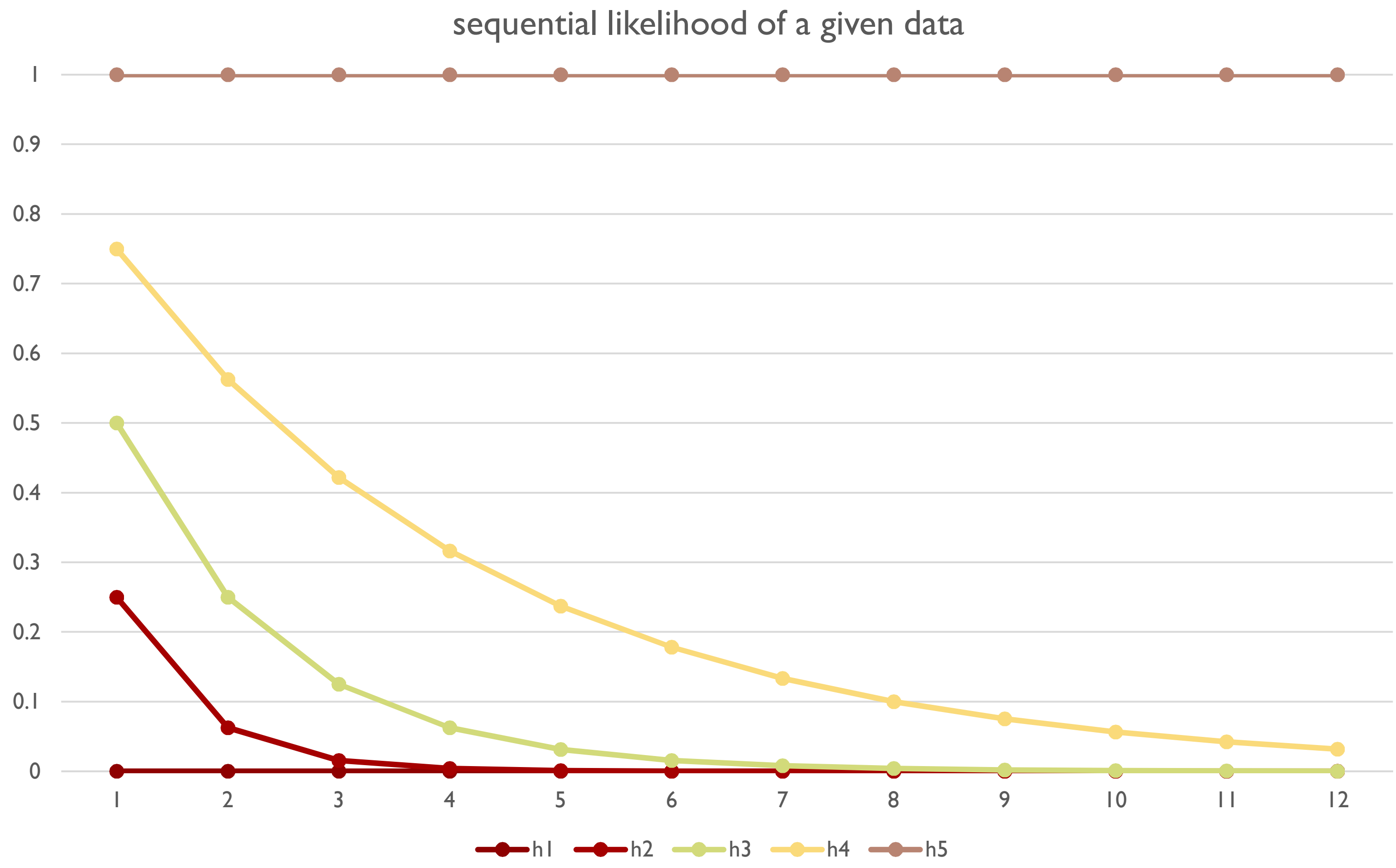
Russel&Norvig: Artificial intelligence

A bayesi módon átlagolt modell predikciójának *a posteriori* valószínűsége



Russel&Norvig: Artificial intelligence

Maximum likelood model selection



ML paramétertanulás

A cukorkás példa esetében

$$P(D|H_\theta) = \prod_{j=1}^{N} P(d_j|H_\theta) = \theta^c \cdot (1-\theta)^l$$

- Sokszor előnyös a log-likelihooddal számolni

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

$$= \sum_{j=1}^{N} P(d_j|H_\theta) = c \cdot \log\theta + l \cdot \log(1-\theta)$$

- Maximum számítás: deriválással

$$\frac{dL(D|H_\theta)}{d\theta} = \frac{c}{\theta} - \frac{l}{1-\theta} = 0 \to \theta = \frac{c}{c+l} = \frac{c}{N}$$

Struktúratanulás

- Irányított aciklikus gráfok n csomópontra: DAG(n)
- Csomópontok/változók sorrendezése: $\prec$
- Szülői halmazok száma: Π
- Kardinalitás (számosság):

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} 2^{i(n-1)} f(n-i) \text{ with } f(0) = 1$$

n	$\lvert DAG(n) \rvert$	$\lvert G_{\prec} \rvert$	$\lvert G_{\prec}^{\lvert\pi\rvert \le 4} \rvert$	$\lvert G_{\prec}^{\lvert\pi\rvert \le 2} \rvert$	$\lvert \prec \rvert$	$\lvert \pi^{\prec} \rvert$	$\lvert \lvert \pi^{\prec} \rvert \le 4 \rvert$	$\lvert \lvert \pi^{\prec} \rvert \le 2 \rvert$
5	2.9e+004	1e+003	1e+003	6.2e+002	1.2e+002	30	30	24
6	3.8e+006	3.3e+004	3.2e+004	9.9e+003	7.2e+002	62	61	40
7	1.1e+009	2.1e+006	1.8e+006	2.2e+005	5e+003	1.3e+002	1.2e+002	62
8	7.8e+011	2.7e+008	1.8e+008	6.3e+006	4e+004	2.5e+002	2.2e+002	91
9	1.2e+015	6.9e+010	2.9e+010	2.3e+008	3.6e+005	5.1e+002	3.8e+002	1.3e+002
10	4.2e+018	3.5e+013	7.5e+012	1.1e+010	3.6e+006	1e+003	6.4e+002	1.7e+002
15	2.4e+041	4.1e+031	2.1e+027	3.1e+019	1.3e+012	3.3e+004	4.9e+003	5.7e+002
35	2.1e+213	1.3e+179	1.8e+109	8.5e+068	1e+040	3.4e+010	3.8e+005	7.2e+003

Bayes-háló struktúratanulása NP-nehéz

- **Tétel:** Legyen ***V*** a változók egy halmaza, melynek együttes valószínűségeloszlása p(***V***). Tegyük fel, hogy létezik egy orákulum, amely *O(1)* időben megválaszolja, hogy egy függetlenségi állítás igaz p-ben.
- Legyen $0 < k \leq |V|$ és $s = \frac{1}{2}n(n-1) - \frac{1}{2}k(k-1)$
- Ekkor annak eldöntése az orákulum segítségével, hogy létezik-e vagy sem egy olyan nem minimális Bayes-háló, ami reprezentálja *p*-t kisebb vagy egyenlő, mint *s* éllel, NP-nehéz

= *Megfelelő Bayes-háló megtalálása a megfigyelési adatokhoz NP-nehéz*

Bayes-háló struktúratanulása NP-nehéz

- **Tétel:** Legyen ***V*** a változók egy halmaza D_N a teljes adathalmaz $S(G,D_N)$ egy pontszámfüggvény (scoring function) és *c* egy valós szám.
- Ekkor annak eldöntése, hogy létezik-e egy ***V*** változók felett definiált G_0 Bayes-háló, amelynél minden csomópont legfeljebb $1<k$ szülővel rendelkezik, úgy hogy $c \leq S(G_0, D_N)$, az NP-nehéz

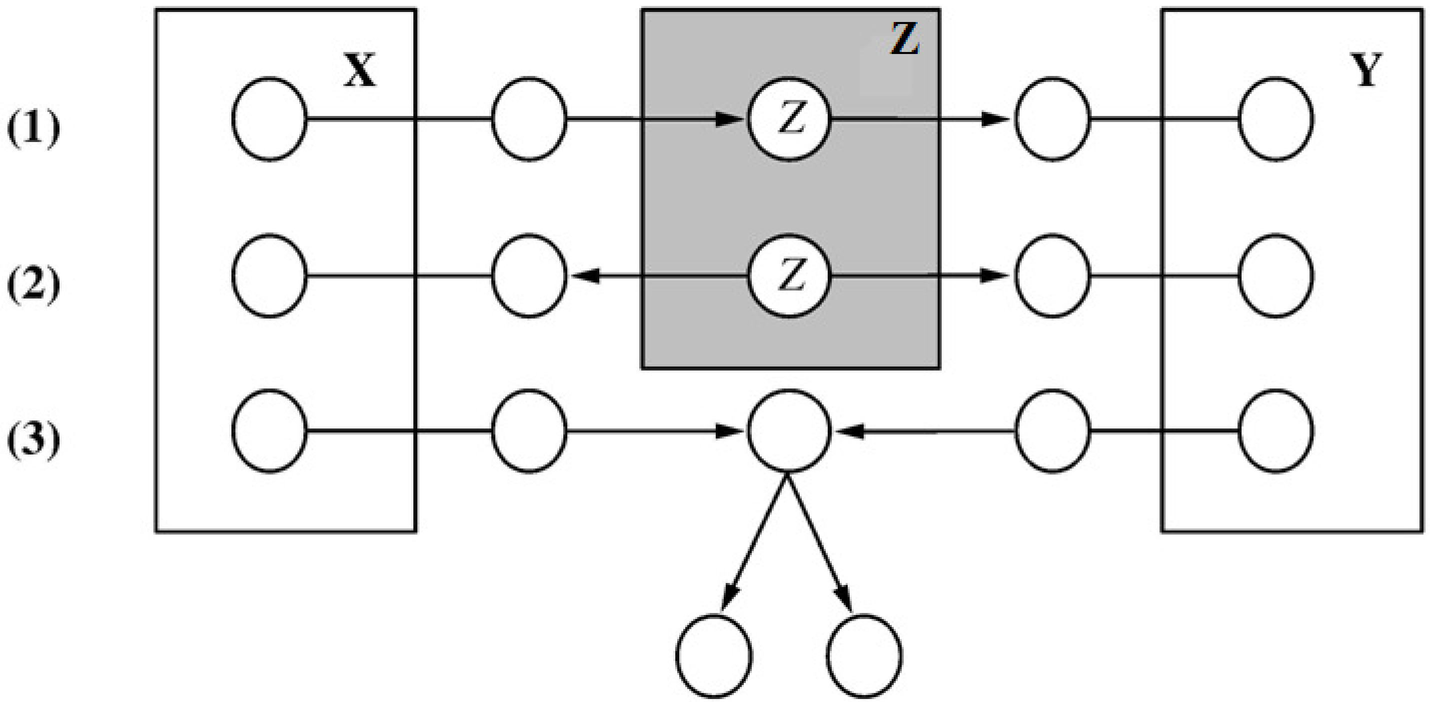
= A legjobb illeszkedési pontszámú Bayes-háló megtalálása NP-nehéz

Bayes-háló - függőségek

- **Bayes-háló =** feltételes függetlenségek (függőségek) térképe
- A kérdés: egy adott struktúrából milyen feltételes függetlenségeket/függőségeket lehet kiolvasni?
- A válasz: **d-szeparációs szabályok** szerint
  - Directed separation ~ d-elválasztás

Bayes-háló: D-szeparáció

- Jelölje $I_G(X,Y|\mathbf{Z})$ hogy *X* d-szeparált *Y*-tól feltéve ***Z***-t a *G* irányított gráfban.



D-szeparáció – globális Markov-feltétel

Definition 7 *A distribution $P(X_1, \ldots, X_n)$ obeys the global Markov condition w.r.t. DAG G, if*

$$\forall\, X, Y, Z \subseteq U \; (X \perp\!\!\!\perp Y | Z)_G \Rightarrow (X \perp\!\!\!\perp Y | Z)_P, \tag{9}$$

where $(X \perp\!\!\!\perp Y|Z)_G$ denotes that X and Y are d-separated by Z, that is if every path p between a node in X and a node in Y is blocked by Z as follows

1. *either path p contains a node n in Z with non-converging arrows (i.e. $\rightarrow n \rightarrow$ or $\leftarrow n \rightarrow$),*
2. *or path p contains a node n not in Z with converging arrows (i.e. $\rightarrow n \leftarrow$) and none of its descendants of n is in Z.*

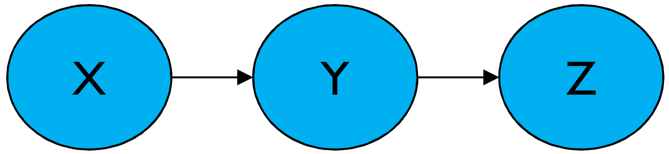
Feltételes függetlenségi modell

- A P eloszlás M függetlenségi térképe (modellje) az érvényes feltételes függetlenségeket jelölő tripletek halmaza:

$$M_P=\{I_{P,1}(X_1,Y_1|\mathbf{Z}_1),\ldots,\ I_{P,k}(X_k,Y_k|\mathbf{Z}_k)\}$$

- Ha $P(X,Y,Z)$ egy Markov-lánc, akkor

$$\mathrm{M_P}=\{\mathrm{D(X,Y),\ D(Y,Z),\ I(X,Z|Y),\ és\ D(X,Z)^*}\},$$



- ahol D(…) = ┐ I(..)

*normális esetben D(X,Z), kivételesen lehet I(X,Z)

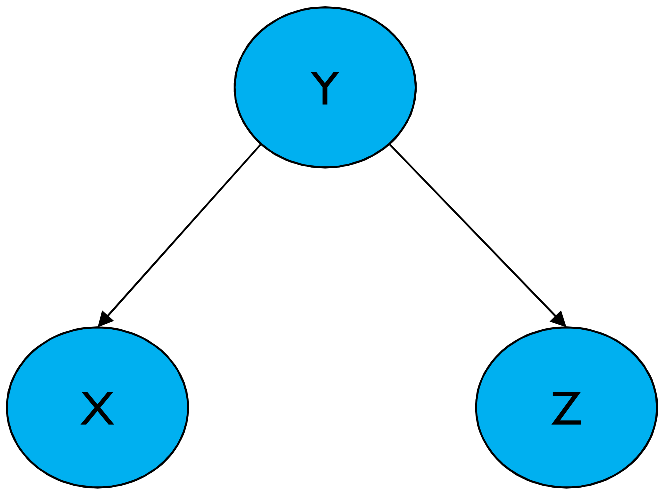
Feltételes függetlenségi modell - NBN

- Ha $P(X,Y,Z)$ egy naiv Bayes-háló (NBN), akkor

$$M_P=\{D(X,Y),\ D(Y,Z),\ I(X,Z|Y),\ \text{és}\ D(X,Z)^*\},$$

- ahol $D(\ldots) = \neg\, I(..)$

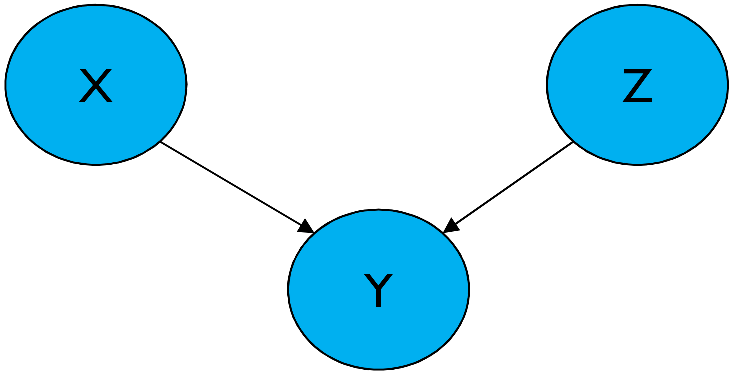
*normális esetben D(X,Z), kivételesen lehet I(X,Z)



Feltételes függetlenségi modell – V-struktúra

- Ha *P(X,Y,Z)* egy V-struktúra, akkor

$$M_P=\{D(X,Y),\ D(Y,Z),\ I(X,Z),\ \text{és}\ D(X,Z|Y)\},$$



Feltételes függőségeket/függetlenségeket kódoló tripletek

$p(X), p(Y|X), p(Z|Y)$

$X \rightarrow Y \rightarrow Z$

$p(X|Y), p(Y|Z), p(Z)$

$X \leftarrow Y \leftarrow Z$

$p(X|Y), p(Y), p(Z|Y)$

$X \leftarrow Y \rightarrow Z$

"tranzitív" M ≠ „intranzitív" M

$p(X), p(Z|X,Y), p(Y)$

$X \rightarrow Z \leftarrow Y$

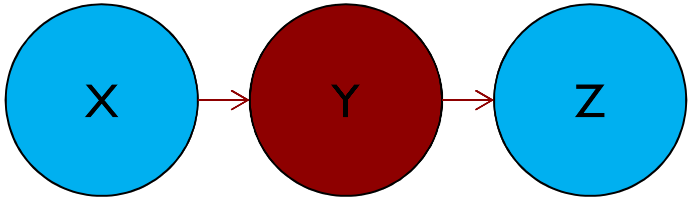
„V-struktúra"

$M_P = \{D(X,Z), D(Z,Y), D(X,Y), I(X,Y|Z)\}$

Gyakran: a jelenlegi állapot (tudás) ismerete a jövőbeli állapotokat feltételesen függetlenné teszi a korábbiaktól

Különleges esetek

Elsőrendű Markov-lánc esetén az X-Y és Y-Z függősége ellenére lehetnek X és Z függetlenek (feltéve, hogy Y nem bináris).



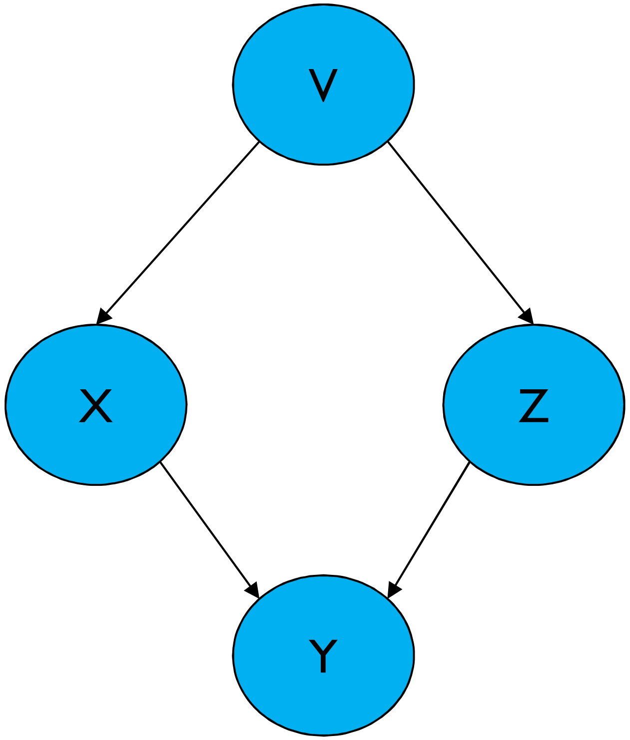
Bizonyos eloszlásokat nem lehet pontosan reprezentálni Bayes-hálókkal

1. Intranzitív Markov-lánc: X➔Y➔Z
2. Tisztán többváltozós okozati kapcsolat: {X,Z}➔Y
3. Gyémánt struktúra:

Különleges esetek

Gyémánt struktúra:

$P(X,Y,Z,V)$ esetén $M_P=\{D(X,Z), D(X,Y), D(V,X), D(V,Z), I(V,Y|\{X,Z\}), I(X,Z|\{V,Y\}).. \}$.



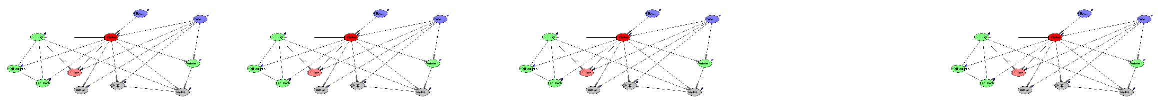
Függetlenségek reprezentálása

D-separation provides a sound and complete, computationally efficient algorithm to read off an (in)dependency model consisting the independencies that are valid in all distributions Markov relative to G, that is $\forall\ X, Y, Z \subseteq V$

$$(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow ((X \perp\!\!\!\perp Y|Z)_P \text{ in all P Markov relative to G}). \tag{10}$$

Oksági modellek megfigyelési ekvivalenciája

Oksági modellek:



J.Pearl:
~„3D objects”

Passzív megfigyelések:

$P(X_1,\dots, X_n)$

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\dots, I_{P,K}(X_K;Y_K|Z_K)\}$

„2D projection”

- Különböző oksági modelleknek lehet ugyanaz a függetlenségi térképe
- Jellemzően az oksági modelleket nem lehet tökéletesen megtanulni/meghatározni kizárólag passzív megfigyelési adatokból
- Ezek egy része **megfigyelés-ekvivalens**

Markov-lánc, mint oksági modell

Oksági modell:

$P(X_1,\ldots)$

$M_P=\{I(X_{i+1},X_{i-1}|X_i)\}$

„elsőrendű Markov-tulajdonság”

Időbeliség meghatározható?

Megfigyelési ekvivalencia – V-struktúrák

Definition 11 *Two DAGs G_1, G_2 are* observationally equivalent *, if they imply the same set of independence relations (i.e. $(X \perp\!\!\!\perp Y|Z)_{G_1}) \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$).*

The implied equivalence classes may contain $n!$ number of DAGs (e.g. all the full networks representing no independencies) or just 1.

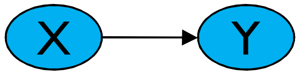
Theorem 2 *Two DAGs G_1, G_2 are* observationally equivalent *, iff they have the same skeleton (i.e. the same edges without directions) and the same set of* v-structures *(i.e. two converging arrows without an arrow between their tails).*

Definition 12 *The* essential graph *representing observationally equivalent DAGs is a partially oriented DAG (PDAG), that represents the identically oriented edges called* compelled edges *of the observationally equivalent DAGs (i.e. in the equivalence class), such a way that in the common skeleton only the compelled edges are directed (the others are undirected representing inconclusiveness).*

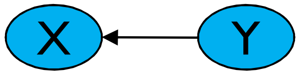
Reichenbach közös ok elve

Reichenbach's Common Cause Principle:

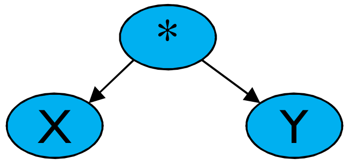
A korreláció *X* és *Y* esemény között azt jelenti, hogy vagy *X* okozza *Y-t*, vagy *Y* okozza *X-et*, vagy *X* és *Y* –nak közös oka van.



X okozza Y-t

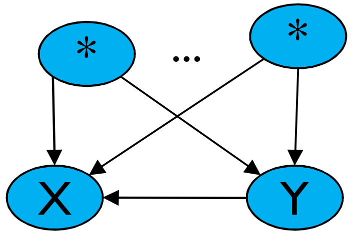


Y okozza X-et



Van egy közös ok
(pure confounding)

...

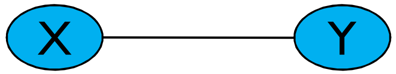


Egy ok-okozati kapcsolatot
„zavarnak" más hatások

Passzív megfigyelésekből:

P(X,Y)

$M_P=\{D(X;Y)\}$



„X és Y asszociált"

Kényszeralapú Bayes-háló tanulás

The Inductive Causation algorithm (assuming a stable distribution P):

1. *Skeleton:* Construct an undirected graph (skeleton), such that variables $X, Y \in \mathbf{V}$ are connected with an edge iff $\forall S (X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \mathbf{V} \setminus \{X, Y\}$.
2. *v-structures:* Orient $X \rightarrow Z \leftarrow Y$ iff X, Y are nonadjacent, Z is a common neighbour and $\neg\exists S$ that $(X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \mathbf{V} \setminus \{X, Y\}$ and $Z \in S$.
3. *propagation:* Orient undirected edges without creating new v-structures and directed cycle.

Theorem

The following four rules are necessary and sufficient.

R_1 *if* $(a \not- c) \wedge (a \rightarrow b) \wedge (b - c)$, *then* $b \rightarrow c$

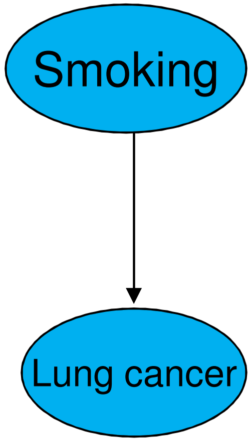
R_2 *if* $(a \rightarrow c \rightarrow b) \wedge (a - b)$, *then* $a \rightarrow b$

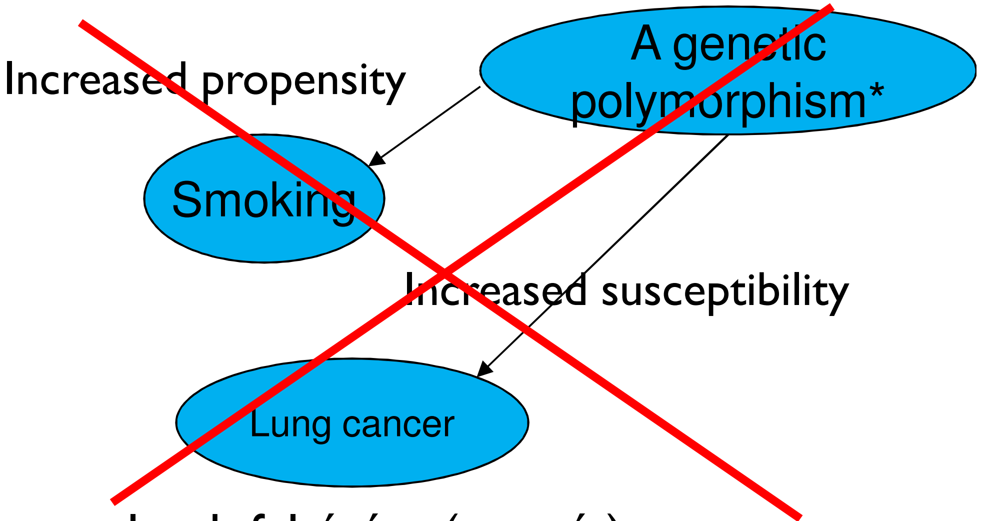
R_3 *if* $(a - b) \wedge (a - c \rightarrow b) \wedge (a - d \rightarrow b) \wedge (c \not- d)$, *then* $a \rightarrow b$

R_4 *if* $(a - b) \wedge (a - c \rightarrow d) \wedge (c \rightarrow d \rightarrow b) \wedge (c \not- b) \wedge (a - d)$, *then* $a \rightarrow b$

Lokális oksági felfedezés (LCD)

- Lehet-e oksági kapcsolatokat tanulni megfigyelési adatokból zavaró változók jelenlétében?





- Automatizált, „tabula rasa" oksági kapcsolatok feltárása (passzív) megfigyelésekből lehetséges, vagyis zavaró változók kizárhatók