

Bayesche Netzwerke

- ◆ Thomas Bayes Theorie
- ◆ conditional independence, explain away, d-Separation
- ◆ joint probability, DAG aus Markovian Parents
- ◆ Inference „brute force“
- ◆ Lernen einer binomialen Variable, Naive Bayes classifier
- ◆ Lernen einer multinomialen Variable
- ◆ Beispiel Lernen eines Netzwerks
- ◆ Lernen der einzelnen CPT
- ◆ Optimieren der Struktur
- ◆ Inference mit PPTC probability propagation in trees of clusters

Bayesche Denkweise



- ◆ Thomas Bayes 1702-1761
- ◆ 1764 p.m. veröffentlicht
- ◆ Royal Society of London
- ◆ unbeeinsprucht bis George Bool 1854 „Laws of Thought“

Bayesche Denkweise



- ◆ Thomas Bayes 1702-1761
- ◆ Subjektive Wahrscheinlichkeit
- ◆ Background ξ
- ◆ a priori: $p(e|\xi)$
- ◆ a posteriori: $p(e|D, \xi)$

$$p(e|D, \xi) = \frac{p(e|\xi)p(D|e, \xi)}{p(D|\xi)}$$

research by

◆ David Heckerman

Microsoft Research, Lernen in BN

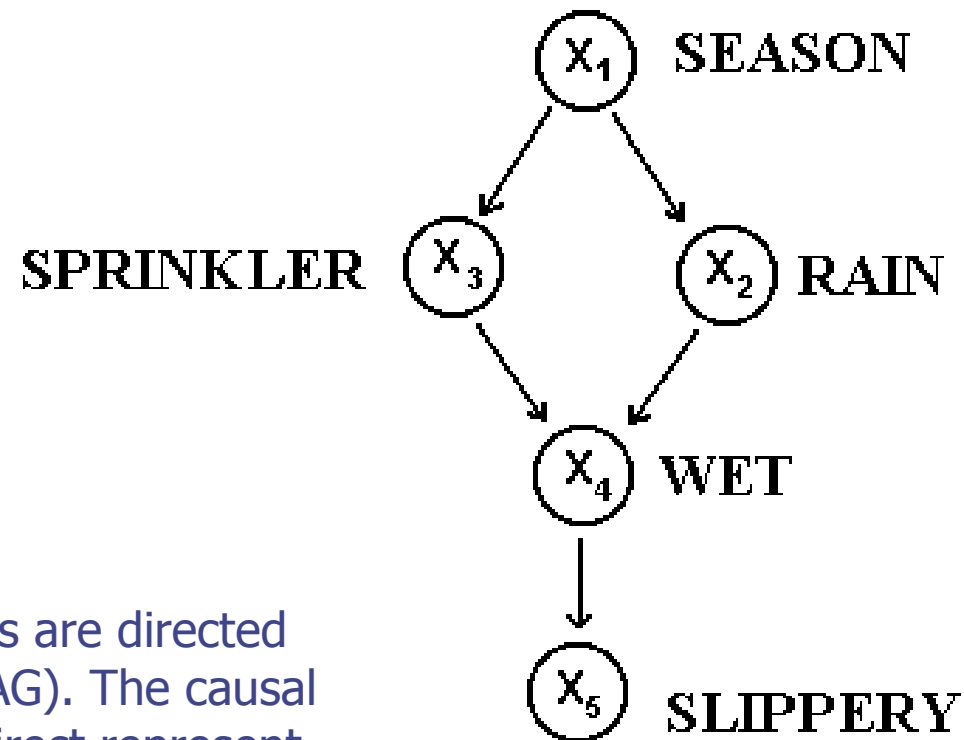
◆ Judea Pearl

University of California, Theorie von Kausalität u.a.

◆ Spiegelhalter

University of Cambridge, Biostatistic, Klinische Studien

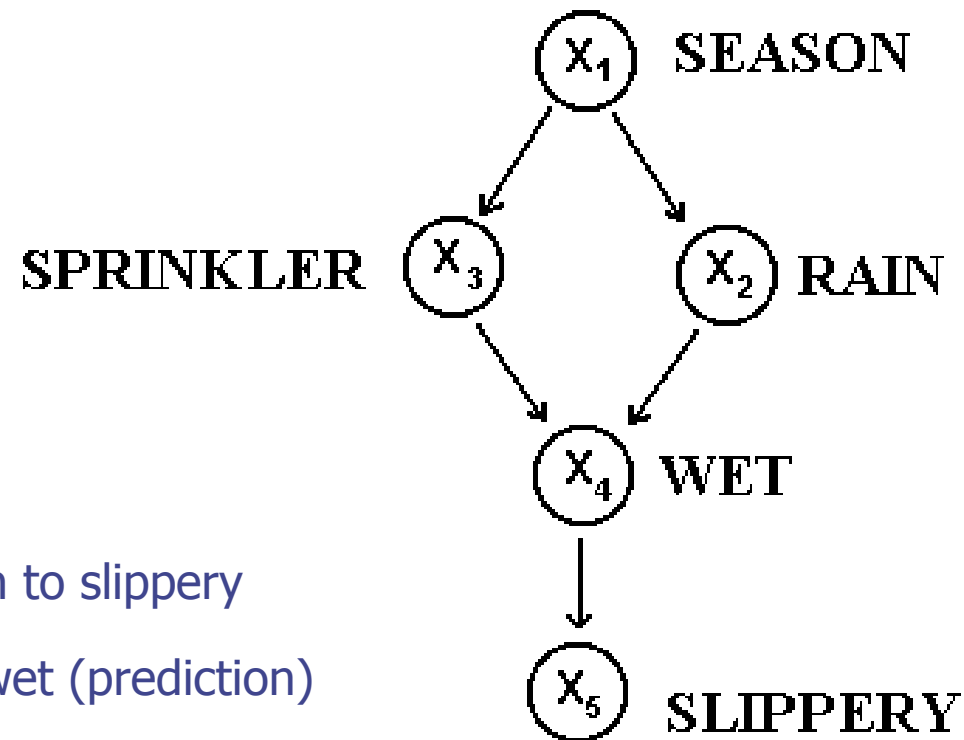
Beispiel seasons (Pearl, Russel 2000)



Bayesian Networks are directed acyclic graphs (DAG). The causal connections are direct representations of the real world.

implementiert in Netica 1.12

Beispiel seasons



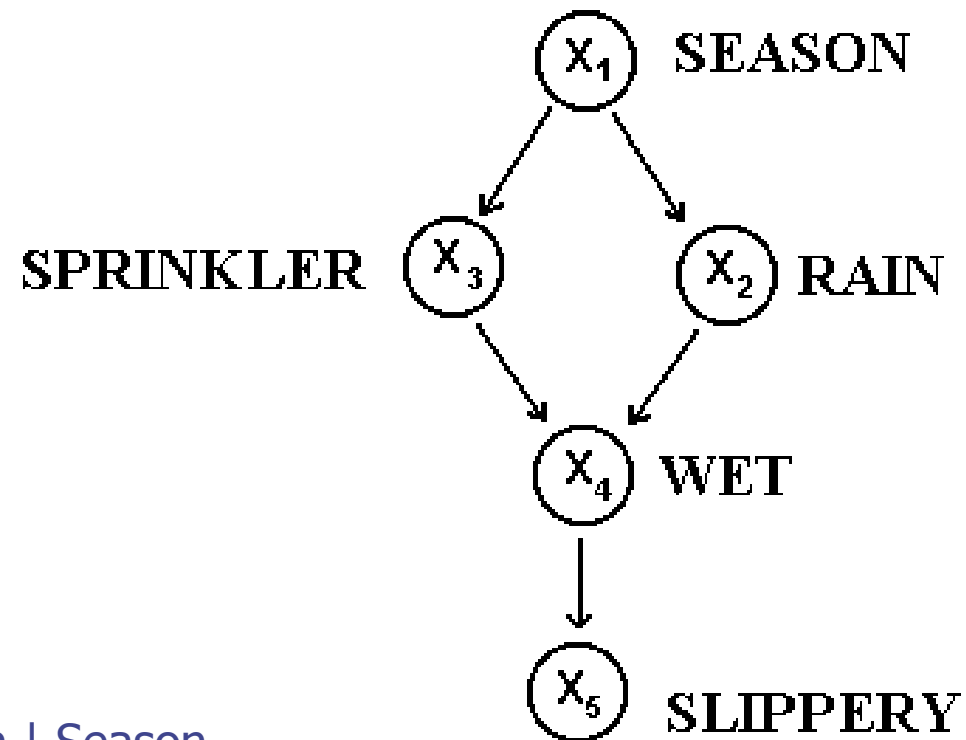
no link from Season to slippery

Sprinkler \Rightarrow prob. wet (prediction)

slippery \Rightarrow prob. wet (abduction)

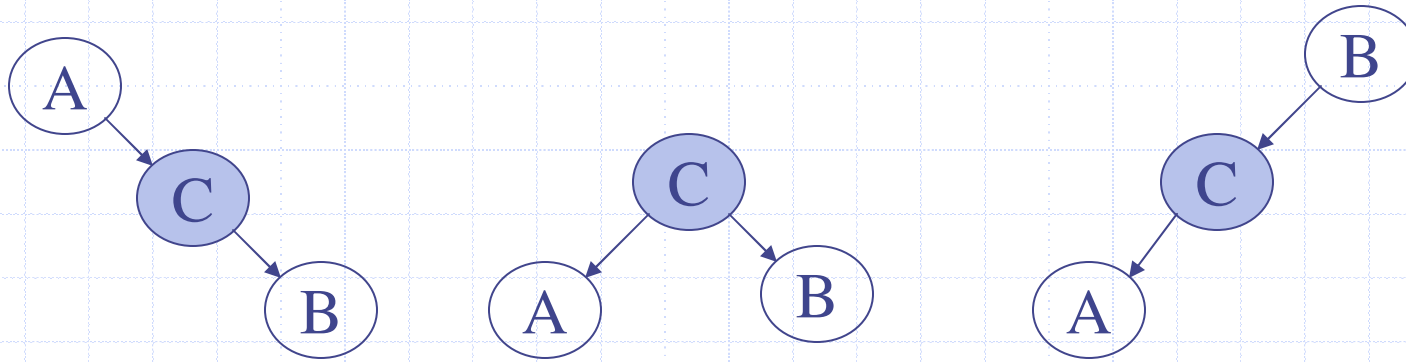
wet \Rightarrow prob. sprinkler or rain (abduction)

Beispiel seasons



Sprinkler \perp Rain | Season

Conditional Independence

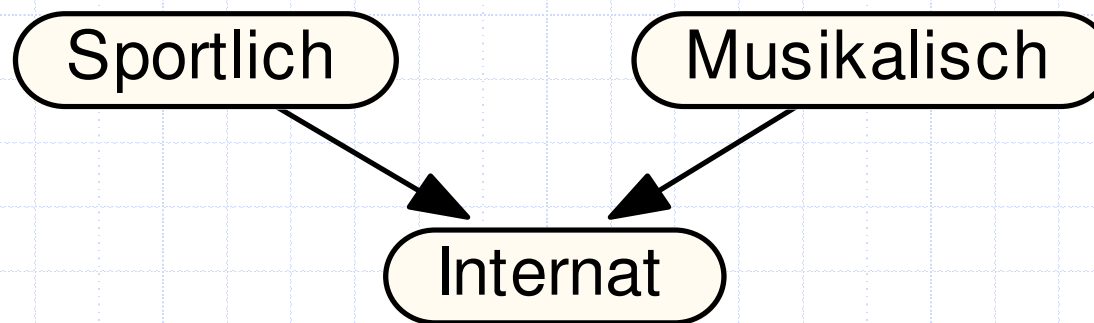


$$p(A|C, \xi) = p(A|B, C, \xi)$$
$$p(B|C, \xi) = p(B|A, C, \xi)$$

Schreibweise: $A \perp B \mid C$

Beispiel explaining away

Aufnahmekriterium des Internats: sportliche Leistung **oder** musikalische Begabung

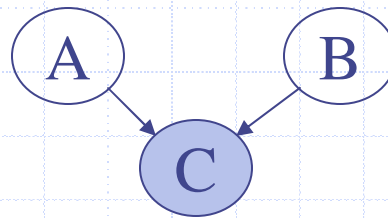


$\text{Sportlich} \perp \text{Musikalisch} \mid \emptyset$

$\neg(\text{Sportlich} \perp \text{Musikalisch} \mid \text{Internat})$

$p(M|I) > p(M|S,I)$

explaining away

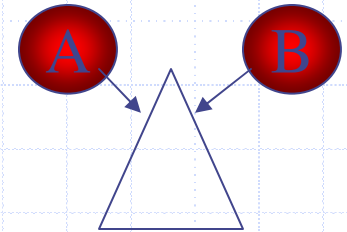
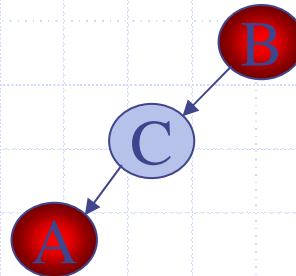
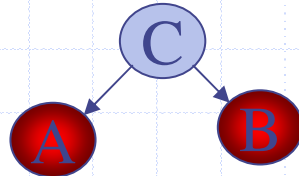
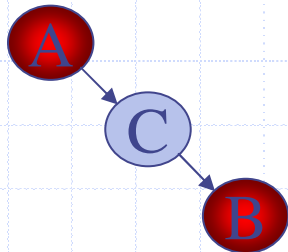


$A \perp B \mid \emptyset$
aber: $\neg(A \perp B \mid C)$

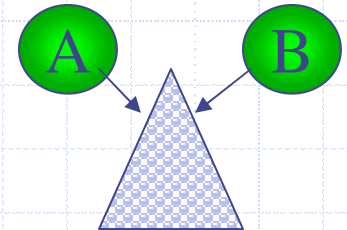
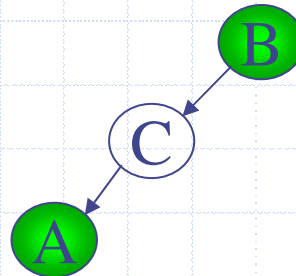
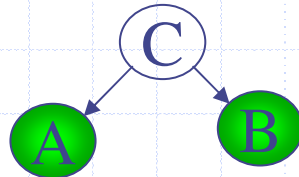
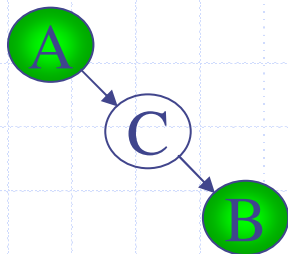
Diese Art des Schließens ist in regelbasierten Systemen oder neuronalen Netzen sehr schwer zu modellieren, in Bayeschen Netzen aber selbstverständlich.

d-separation (Pearl 1988)

$$\{A\} \perp \{B\} \mid \{C\}$$



Pfad blockiert, Information kann nicht passieren ($A \perp B$)



Pfad frei, Information propagiert $\neg(A \perp B)$

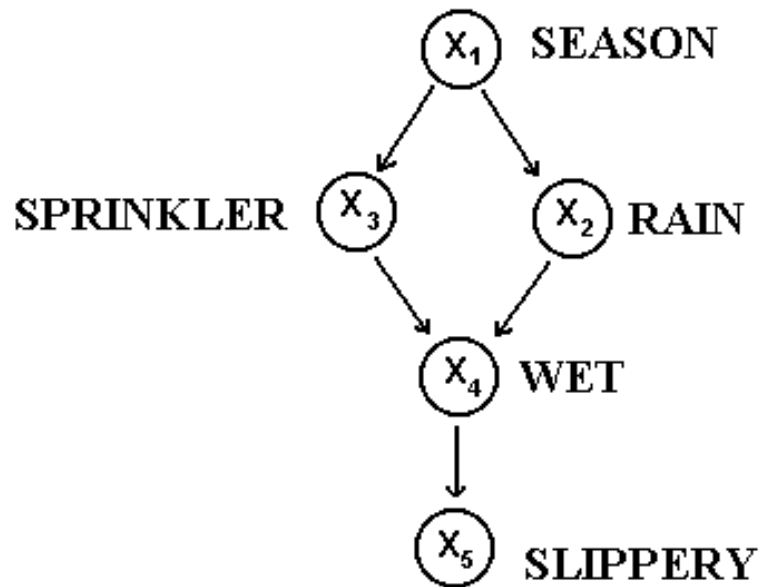
joint probability

- ◆ $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$
- ◆ 2^n verschiedene Realisierungen
- ◆ Typischerweise 50 bis 500 Variablen
- ◆ joint probability ist $O(\exp(n))!!$
- ◆ conditional probability tables in BN sind nur $O(n^k)!!$

$$p(x) = p(x_1, x_2, x_3, x_4, x_5)$$

63 verschiedene Werte

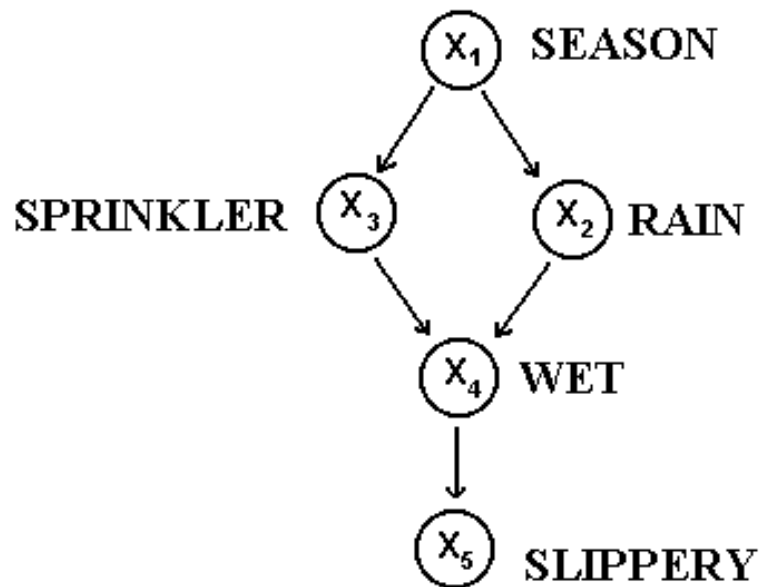
Markovian Parents



Use product rule to decompose joint probability into product of conditional probabilities. Order is important!!

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2)p(x_4 | x_1, x_2, x_3)p(x_5 | x_1, x_2, x_3, x_4)$$

Markovian Parents



Order : x_1, x_2, x_3, x_4, x_5

$$p(x_1)$$

$$p(x_2 \mid x_1)$$

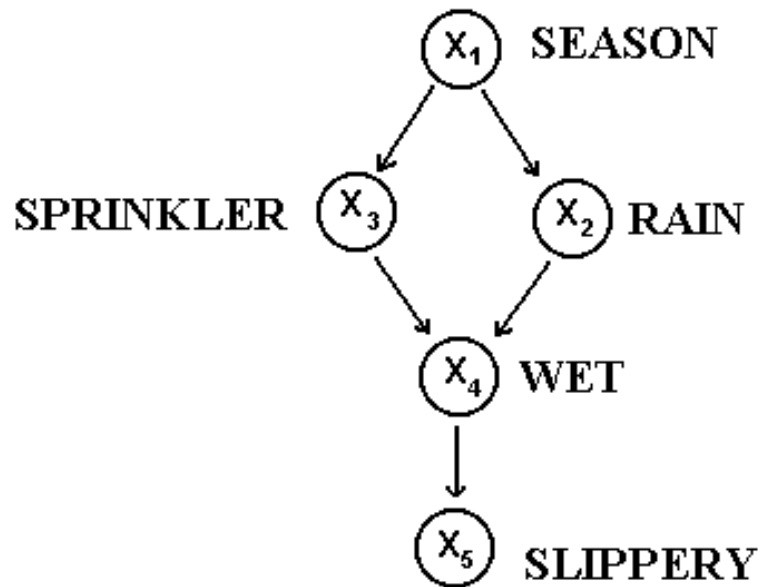
$$p(x_3 \mid x_1, x_2) = p(x_3 \mid x_1)$$

$$p(x_4 \mid x_1, x_2, x_3) = p(x_4 \mid x_2, x_3)$$

$$p(x_5 \mid x_1, x_2, x_3, x_4) = p(x_5 \mid x_4)$$

$$p(x) = p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1) p(x_4 \mid x_2, x_3) p(x_5 \mid x_4)$$

Markovian Parents



Order : x_1, x_2, x_3, x_4, x_5

$$p(x_1)$$

$$p(x_2 \mid x_1)$$

$$p(x_3 \mid x_1, x_2) = p(x_3 \mid x_1)$$

$$p(x_4 \mid x_1, x_2, x_3) = p(x_4 \mid x_2, x_3)$$

$$p(x_5 \mid x_1, x_2, x_3, x_4) = p(x_5 \mid x_4)$$

$$p(x) = p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1) p(x_4 \mid x_2, x_3) p(x_5 \mid x_4)$$

$3 + 4 + 4 + 4 + 2 = 17$ verschiedene Werte

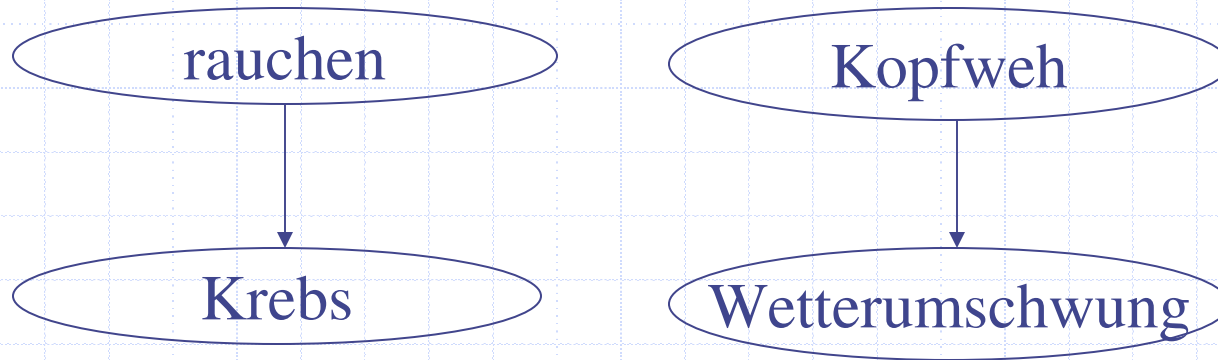
Markovian Parents

◆ PA_i sind markovian parents von X_i ,
wenn gilt:

$$X_i \perp \left(\bigvee_{j=1}^{i-1} \{X_j\} \setminus PA_i \right) \mid PA_i \quad \text{und } PA_i \text{ sind minimal}$$

daraus folgt:
$$p(x_1, \dots, x_n) = \prod_i p(x_i \mid pa_i)$$

Kausalität



Judea Pearl: „correlation does not imply causation“
statistical knowlege is not causal knowlege

a priori Hypothese

- ◆ Expertenwissen fließt ein
- ◆ Kausale Zusammenhänge sind „logisch“
- ◆ Assessment liefert daher gute Ergebnisse
- ◆ Struktur von Experten
- ◆ CPT durch Lernen aus Datenbank

Inference, evaluation

- ◆ gegeben sind einige Fakten
- ◆ gesucht ist die Wahrscheinlichkeit von einer gewissen Konfiguration

$$p(\text{sprinkler} = \text{on} \mid \text{slippery} = \text{yes}) = ?$$

$$p(X_3 = \text{on} \mid X_5 = \text{yes}) = ?$$

Inference, evaluation

◆ Berechnung durch cond.prob.

$$\begin{aligned} p(X_3 = on \mid X_5 = yes) &= \frac{p(X_3 = on, X_5 = yes)}{p(X_5 = yes)} \\ &= \frac{\sum_{x_1, x_2, x_4} p(x_1, x_2, X_3 = on, X_5 = true)}{\sum_{x_1, x_2, x_3, x_4} p(x_1, x_2, x_3, X_5 = true)} \\ &= \frac{\sum_{x_1, x_2, x_4} p(x_1) p(x_2 \mid x_1) p(X_3 = on \mid x_1) p(x_4 \mid x_2, X_3 = on) p(X_5 = true \mid x_4)}{\sum_{x_1, x_2, x_3, x_4} p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1) p(x_4 \mid x_2, x_3) p(X_5 = true \mid x_4)} \end{aligned}$$

Inference, evaluation

- ◆ im allgemeinen NP-hart (Cooper, 1987)
- ◆ in realistischen Netzen ist exakte Lösung ebenfalls NP-hart
- ◆ lineare Lösung nur für polytrees
„singly connected networks“

Applications

◆ speech recognition

Thiesson, Meek, Chickering, Heckerman 98

◆ causal discovery

Lernen von Strukturen, s. Heckerman 95

◆ expert systems

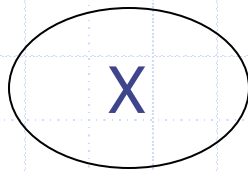
Microsoft Printer-Troubleshooter, Office Wizards

◆ preference prediction

Microsoft Commerce Server 4.0

Lernen, einfachstes Netz

- ◆ nur ein Knoten in Graph
- ◆ z.B. Werfen eines Reißnagels

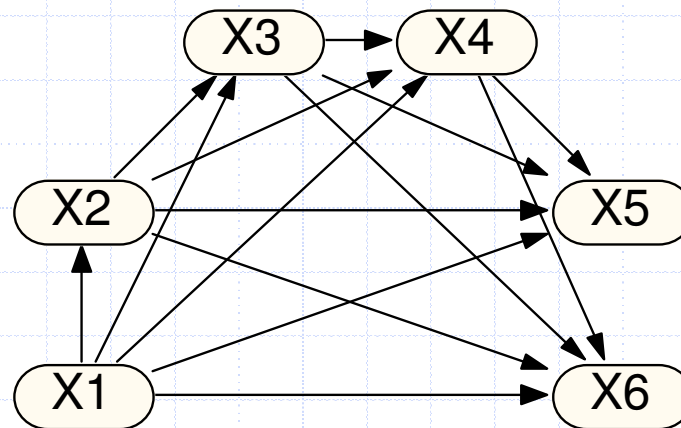


X kann die Werte
head oder tail annehmen

- ◆ nach n Versuchen soll der
n+1ste vorhergesagt werden

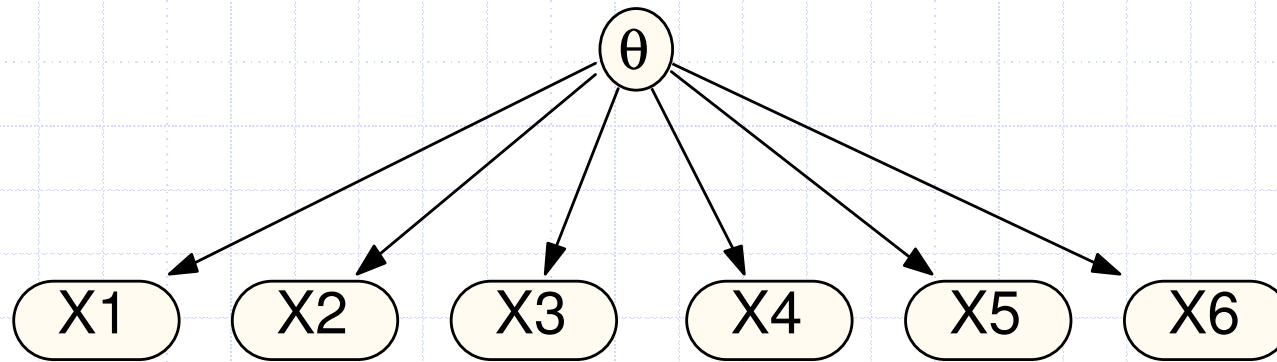
Lernen als Bayesches Netz

- ◆ mit jedem Wurf steigt mein Wissen
- ◆ leider steigen auch die Abhängigkeiten



- ◆ hidden variable einführen

Naive Bayes



Bayesisches Netz mit einer hidden variable

Lernen, Hypothese

$$D = \{X_1, X_2, \dots, X_n\}$$

$$\theta := p(X_{n+1} = \text{head} \mid X_1, X_2, X_3, \dots, X_n, \xi)$$

$$X_{n+1} \perp X_1, X_2, X_3, \dots, X_n \mid \theta$$

$$p(X_{n+1} = \text{head} \mid \theta, \xi) = \theta \quad p(X_{n+1} = \text{tail} \mid \theta, \xi) = 1 - \theta$$

(beliebige) Dichtefunktion $p(\theta \mid \xi) = f(\theta)$

Lernen mit conjugate prior

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

$$\begin{aligned} p(\theta|D_{h,t}, \xi) &= c p(\theta|\xi) p(D_{h,t}|\theta, \xi) \\ &= c p(\theta|\xi) \theta^h (1-\theta)^t \\ &= c f(\theta) \theta^h (1-\theta)^t \end{aligned}$$

$$p(D_{h,t}|\theta, \xi) = \theta^h (1-\theta)^t$$

a priori: $p(\theta|\xi) = f(\theta) = \text{Beta}_{\alpha_h, \alpha_t}(\theta) = \frac{\Gamma(\alpha_h + \alpha_t)}{\Gamma(\alpha_h)\Gamma(\alpha_t)} \theta^{\alpha_h-1} (1-\theta)^{\alpha_t-1}$

$$p(\theta|D_{h,t}, \xi) = \frac{\Gamma(\alpha_h + \alpha_t + h + t)}{\Gamma(\alpha_h + h)\Gamma(\alpha_t + t)} \theta^{\alpha_h+h-1} (1-\theta)^{\alpha_t+t-1} = \text{Beta}_{\alpha_h+h, \alpha_t+t}(\theta)$$

Lernen, Ergebnis

$$\begin{aligned} p(X = head | D_{h,t}, \xi) &= \int_{\theta} p(X = head | \theta, D_{h,t}, \xi) p(\theta | D_{h,t}, \xi) d\theta = \\ &= \int_{\theta} \theta p(\theta | D_{h,t}, \xi) d\theta = \\ &= E[\theta] = E[\text{Beta}_{\alpha_h+h, \alpha_t+t}] \\ &= \frac{\alpha_h + h}{\alpha_h + \alpha_t + h + t} \end{aligned}$$

Marginalisierung über die Wahrscheinlichkeitsverteilung
entspricht hier dem Erwartungswert

Lernen einer multinomialen Variable

$$D = \{X_1, X_2, \dots, X_n\}, \quad X_i = \{x^1, x^2, \dots, x^r\}$$

gesucht: $p(X_{n+1} = x^k \mid D, \xi)$

$$\vec{\theta} = [\theta_1, \theta_2, \dots, \theta_r], \quad \sum_{k=1}^r \theta_k = 1$$

$$p(X_{n+1} = x^k \mid \vec{\theta}, \xi) = \theta_k$$

$$p(\vec{\theta} \mid \xi) = \text{Dir}_{\alpha_1, \dots, \alpha_r}(\vec{\theta}) = \frac{\Gamma(\alpha_1 + \dots + \alpha_r)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \dots \Gamma(\alpha_r)} \prod_{k=1}^r \theta_k^{\alpha_k - 1}$$

Lernen einer multinomialen Variable



$$p(\vec{\theta} | \xi) = \text{Dir}_{\alpha_1, \dots, \alpha_r}(\vec{\theta}) = \frac{\Gamma(\alpha_1 + \dots + \alpha_r)}{\Gamma(\alpha_1) \Gamma(\alpha_2) \dots \Gamma(\alpha_r)} \prod_{k=1}^r \theta_k^{\alpha_k - 1}$$

$$\begin{aligned} p(\vec{\theta} | D_{N_1, \dots, N_r}, \xi) &= p(\vec{\theta} | \xi) \cdot p(D_{N_1, \dots, N_r} | \vec{\theta}, \xi) \cdot \frac{1}{p(D_{N_1, \dots, N_r} | \xi)} \\ &= \text{Dir}_{\alpha_1, \dots, \alpha_r}(\vec{\theta}) \cdot \prod_{k=1}^r \theta_k^{N_k} \cdot c \\ &= \text{Dir}_{\alpha_1 + N_1, \alpha_2 + N_2, \dots, \alpha_r + N_r} \end{aligned}$$

Korrekturfaktor c
durch Marginalisierung:

$$p(D | \xi) = \int_{\Theta} p(D | \vec{\theta}, \xi) d\vec{\theta}$$

$$p(D | \xi) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^r \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)}$$

Lernen einer multinomialen Variable

$$\begin{aligned} p(X_{n+1} = x^k \mid D_{N_1, \dots, N_r}, \xi) &= \int_{\Theta} p(X_{n+1} = x^k \mid \vec{\theta}, D, \xi) p(\vec{\theta} \mid D, \xi) \\ &= \int_{\Theta} \theta_k \text{Dir}_{\alpha_1 + N_1, \dots, \alpha_r + N_r}(\vec{\theta}) d\vec{\theta} \\ &= \frac{\alpha_k + N_k}{\alpha + N} \end{aligned}$$

Bestimmung der a priori Parameter

- ◆ leere Hypothese $\text{Beta}_{1,1}$ bzw. $\text{Dir}_{1, 1, \dots, 1}$
- ◆ equivalent samples
- ◆ imagined future data
- ◆ parameter assessment

Lernen der CPT

$$\mathbf{X} = \{X_1, X_2, \dots, X_n\}$$

$$X_i \in \{x_i^1, x_i^2, \dots, x_i^r\}$$

Datenbank D mit N Fallbeispielen

Arbeitshypothese DAG S_h

Lernen der CPT

$\vec{\theta}_S$ ist der Wert des Vektor-Vektors $(\vec{\theta}_{ij})$ aller Wahrscheinlichkeiten aus allen CPT der Struktur S_h .

$\vec{\theta}_{ij}$ ist die j -te Zeile aus der CPT der Variablen X_i

θ_{ijk} ist die Wahrscheinlichkeit $p(X_i = x_i^k \mid PA_i = pa_i^j)$, wobei die j -te mögliche Realisierung der Variablen in der Menge PA_i bedeuten.

i zählt die Variablen von 1 bis n .

j zählt die möglichen Realisierungen pa_i^j der Variablen in der Menge PA_i von $j = 1$ bis $j = q_i = \prod_{X_l \in PA_i} r_l$.

k zählt die mögliche Realisierung der Variable X_i von 1 bis r_k .

Lernen der CPT

$$p(\mathbf{X}_{N+1} = (x_1, \dots, x_n) \mid \vec{\theta}_s, S^h) = \prod_{i=1}^n p(x_i \mid pa_i, \vec{\theta}_i, S^h)$$

$$p(\vec{\theta}_S \mid D, S^h) = \frac{p(\vec{\theta}_S \mid S^h) p(D \mid \vec{\theta}, S^h)}{p(D \mid S^h)}$$

$$p(\vec{\theta}_S \mid D, S^h) = \prod_{i=1}^n \prod_{j=1}^{q_i} p(\vec{\theta}_{ij} \mid D, S^h)$$

parameter independence

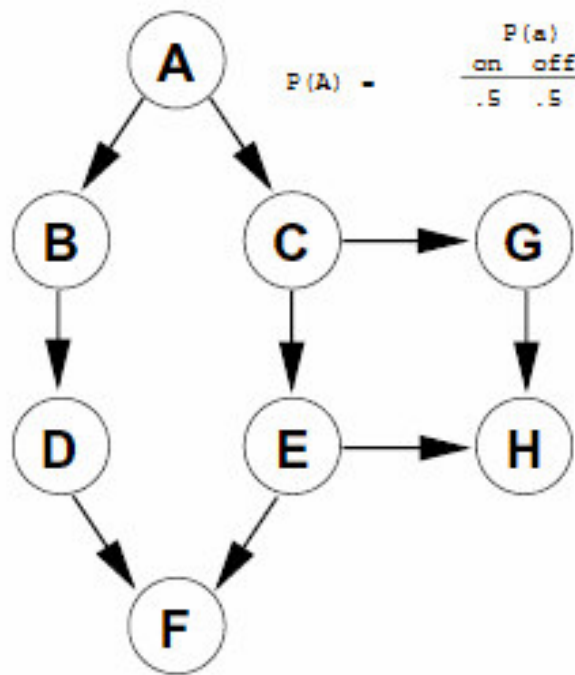
Lernen des Struktur S^h

- ◆ restriktive Bedingungen liefern geschlossene Formeln
- ◆ jede Abweichung nur numerisch lösbar
- ◆ S^h lernen ist auf jeden Fall NP-hart
- ◆ heuristische Suche (greedy search)
- ◆ Bewertungsfunktion pro Variable zerlegbar
- ◆ lokale Maxima?

efficient Inference

- ◆ PPTC Probability Propagation in Trees of Clusters
- ◆ Lauritzen and Spiegelhalter, refined by Jensen et al.
- ◆ compiliern des Netzes in joint trees (clusters, sepsets)
optimal ist NP-hart, heuristic search polynomial-time
- ◆ det. message passing, $2(n-1)$ passes
- ◆ sehr effiziente Observation-Evidence-Berechnung
- ◆ dynamische Observations

PPTC, Huang, Darwiche 1994



$$P(A) =$$

	a	off
	.5	.5

$$P(B|A) =$$

a	b	a	off
on	.5	.5	
off	.4	.6	

$$P(C|A) =$$

a	c	a	off
on	.7	.3	
off	.2	.8	

$$P(D|B) =$$

a	d	b	off
on	.9	.1	
off	.5	.5	

$$P(E|C) =$$

a	e	c	off
on	.3	.7	
off	.6	.4	

$$P(F|DE) =$$

d	e	f	de	off
on	on	.01	.99	
on	off	.01	.99	
off	on	.01	.99	
off	off	.99	.01	

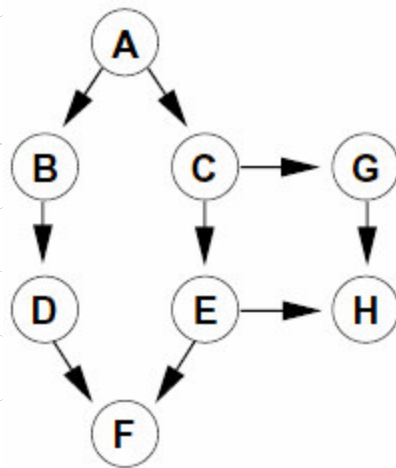
$$P(G|C) =$$

c	g	c	off
on	.8	.2	
off	.1	.9	

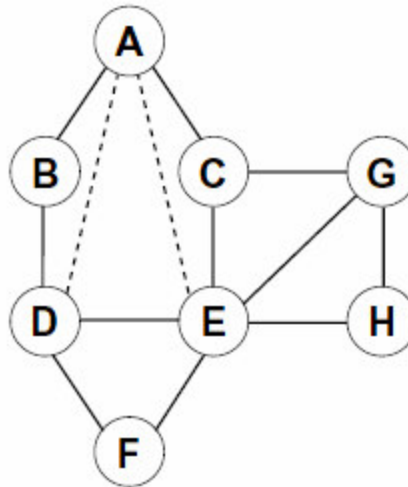
$$P(H|EG) =$$

g	h	eg	off
on	on	.05	.95
on	off	.95	.05
off	on	.95	.05
off	off	.95	.05

Cluster bilden



Belief-Network Structure



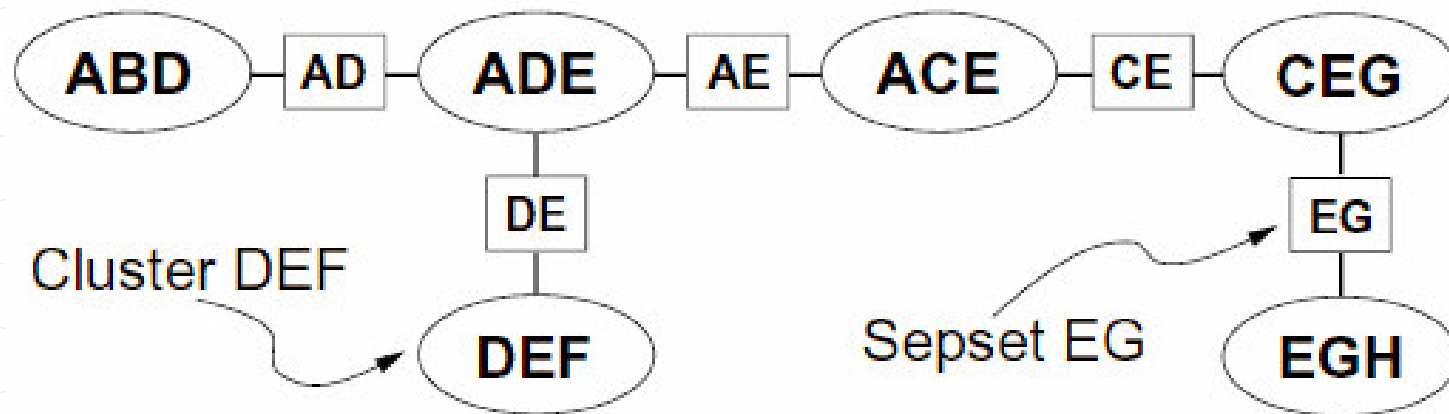
Triangulated Graph

Eliminated Vertex	Induced Cluster	Edges Added
H	EGH	none
G	CEG	none
F	DEF	none
C	ACE	(A, E)
B	ABD	(A, D)
D	ADE	none
E	AE	none
A	A	none

Elimination Ordering

undirected Graph
 connect parent-pairs
 node elimination forms clusters
 add necessary cluster connections

PPTC, joint tree



$$\theta_{ABD} =$$

a	b	d	$\theta_{ABD}(abd)$
on	on	on	.225
on	on	off	.025
on	off	on	.125
on	off	off	.125
off	on	on	.180
off	on	off	.020
off	off	on	.150
off	off	off	.150

$$\theta_{AD} =$$

a	d	$\theta_{AD}(ad)$
on	on	.35
on	off	.15
off	on	.33
off	off	.17

etc.

belief potential, constraints

- ◆ one function (table) for each cluster X , each sepset S : (local consistency)

$$\sum_{X \setminus S} \phi_X = \phi_S \quad X \text{ und benachbarte } S$$

- ◆ joint distribution $P(U)$, global condition

$$P(U) = \frac{\prod_i \phi_{X_i}}{\prod_j \phi_{S_j}}$$

belief potential

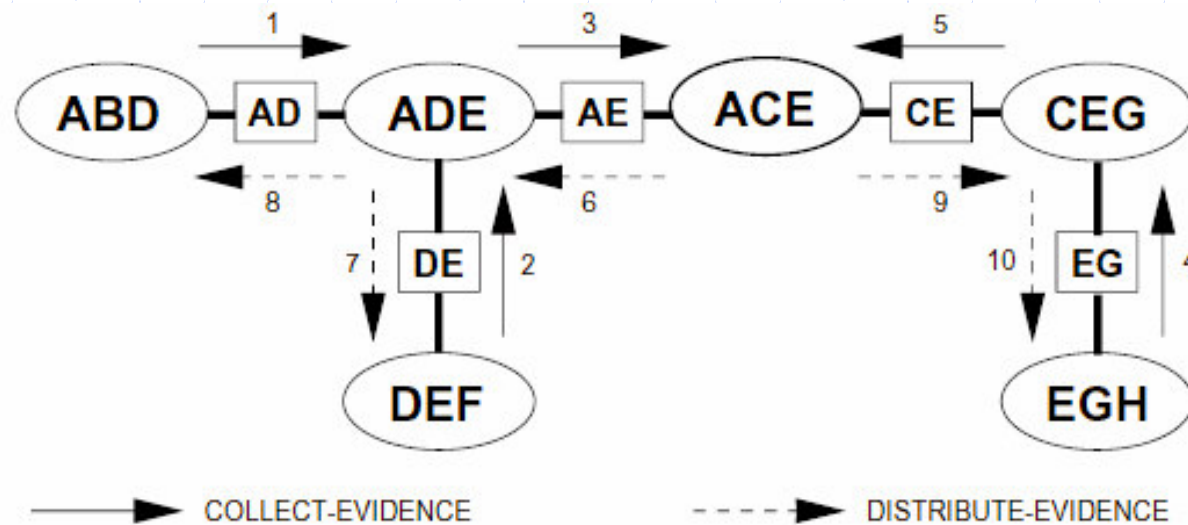
◆ compute one probability distribution

$$P(V) = \sum_{X \setminus \{V\}} \phi_X$$

aus jedem Cluster oder Sepset, das V enthält

message passing

from cluster to cluster through sepset



$$\phi_R^{old} \leftarrow \phi_R \quad \phi_R \leftarrow \sum_{X \setminus R} \phi_X \quad \phi_Y \leftarrow \phi_Y \frac{\phi_R}{\phi_R^{old}}$$