Syntaktische und Statistische Mustererkennung

VO 1.0 840.040
(UE 1.0 840.041)

Bernhard Jung

bernhard@jung.name
http://bernhard.jung.name/VUSSME/

Rückblick

- Nicht lineare Entscheidungsfunktionen
- SVM, Kernel Trick
- Sequentielle Klassifikation
- Boosting
- Clustering

K-means

- Vorgabe:
  - Anzahl an Klassen k
  - Funktion zur Bestimmung eines Clustermittelpunkts
  - Algorithmus:
    1. Wahl von k (beliebigen) Zentren
    2. Jedes Objekt wird dem ihm am nächsten liegenden Zentrum zugeordnet
    3. Clusterzentren werden neu berechnet
    4. Wiederholung der Schritte 2 und 3 bis sich die Zuordnung nicht mehr ändert

K-means

- Probleme:
  - Konvergiert nicht notwendigerweise
  - Ein Cluster kann leer bleiben und somit kann sein Zentrum nicht berechnet werden
  - Lösungsansatz: Abbruch und Neustart mit anderen Clusterzentren
- Vorteil: Praktisch einfach zu berechnen und gute Resultate

Expectation-Maximization (EM)

- Basierend auf k-means mit parametrischen Wahrscheinlichkeitsverteilungen
  - Initialisierung der Parameter der Verteilungen
  - Wiederhole bis Verfahren konvergiert:
    - Bestimme die Klassenzugehörigkeit aufgrund der Likelihood
    - Schätze die Verteilungsparameter aus der Stichprobe (Maximum Likelihood Estimation)

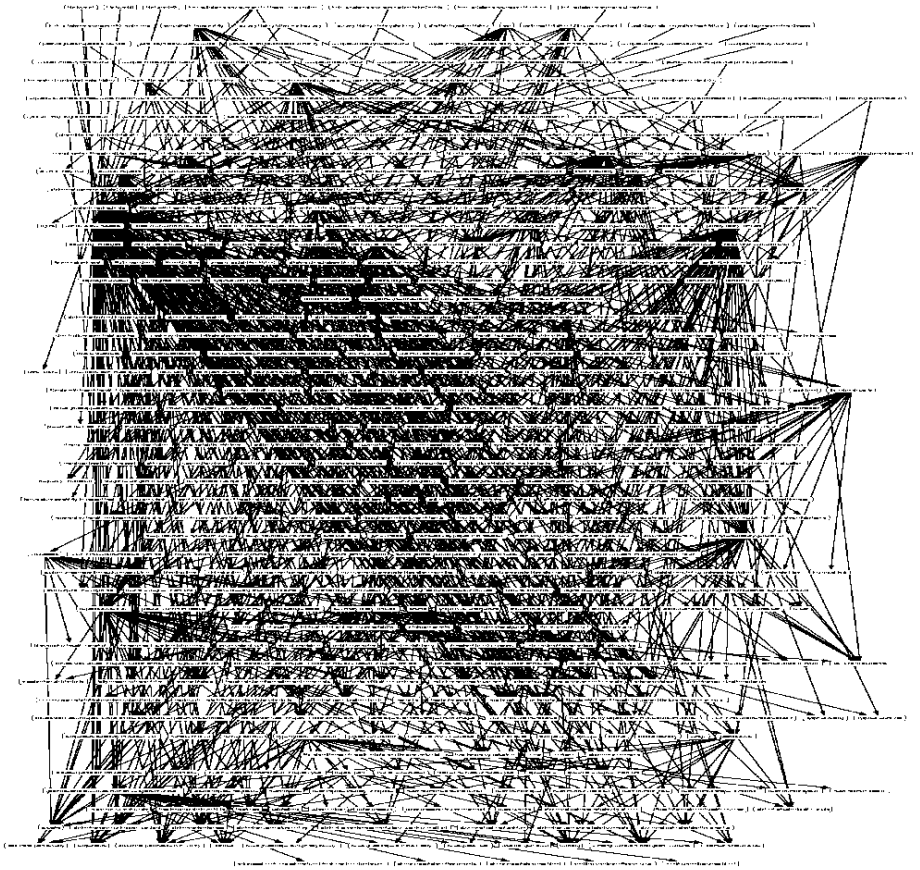
Nicht-hierarchisches Clustering

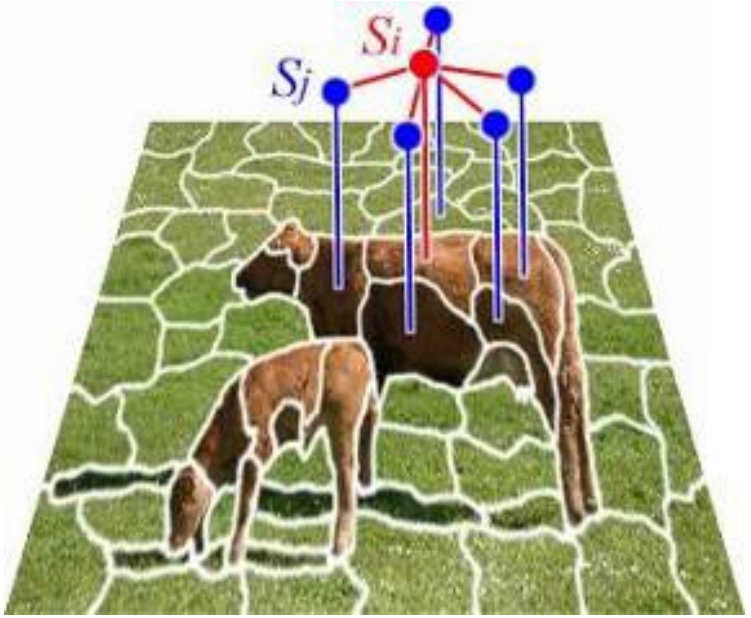
- Parameter $\epsilon >= 0$
- Algorithmus:
  - Start: wähle beliebiges Objekt B_i und fasse alle Objekte im der ϵ-Umgebung von B_i zur Klasse K_j hinzu
  - Wiederholung des Hinzufügens von Objekten in der ϵ-Umgebung von allen bisherigen Objekten der Klasse K_j
  - Wenn kein neues Element hinzufügbar, Wiederholen des gesamten Verfahrens für ein bisher noch nicht klassifiziertes Objekt B_k und neue Klasse K_{j+1}
- Verfahren abhängig von geeigneter Wahl von ϵ

Probabilistic Graphical Models (PGM)

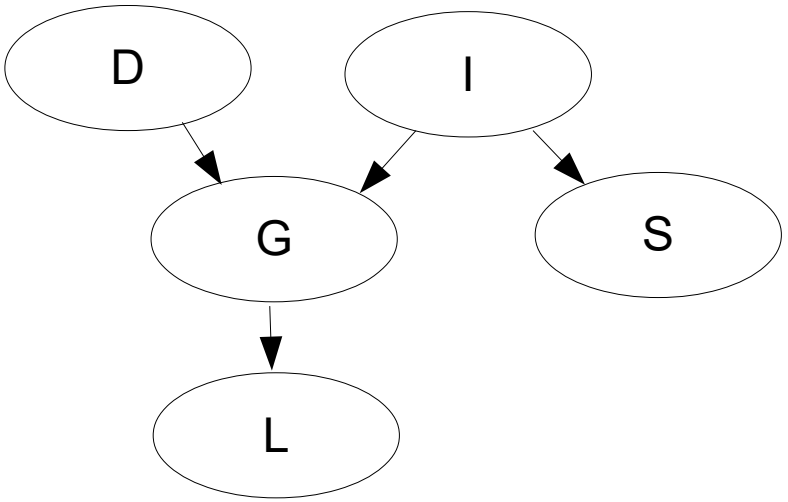
- Kombination aus
  - Wahrscheinlichkeitstheorie
  - Graphentheorie
- Stellt allgemeine Frameworks zur Verfügung, subsumiert u.a.:
  - Kalman filters
  - Hidden Markov models
  - Ising models
  - ...

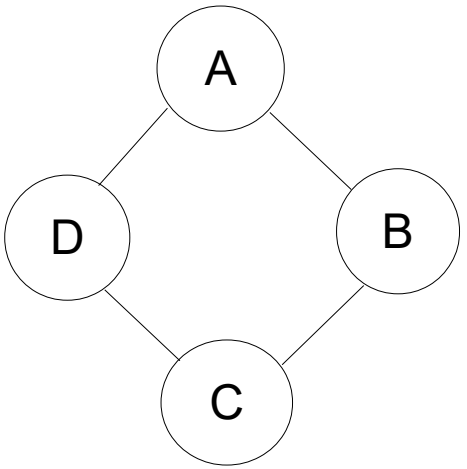
Probabilistic Graphical Models (PGM)





Probabilistic Graphical Models (PGM)

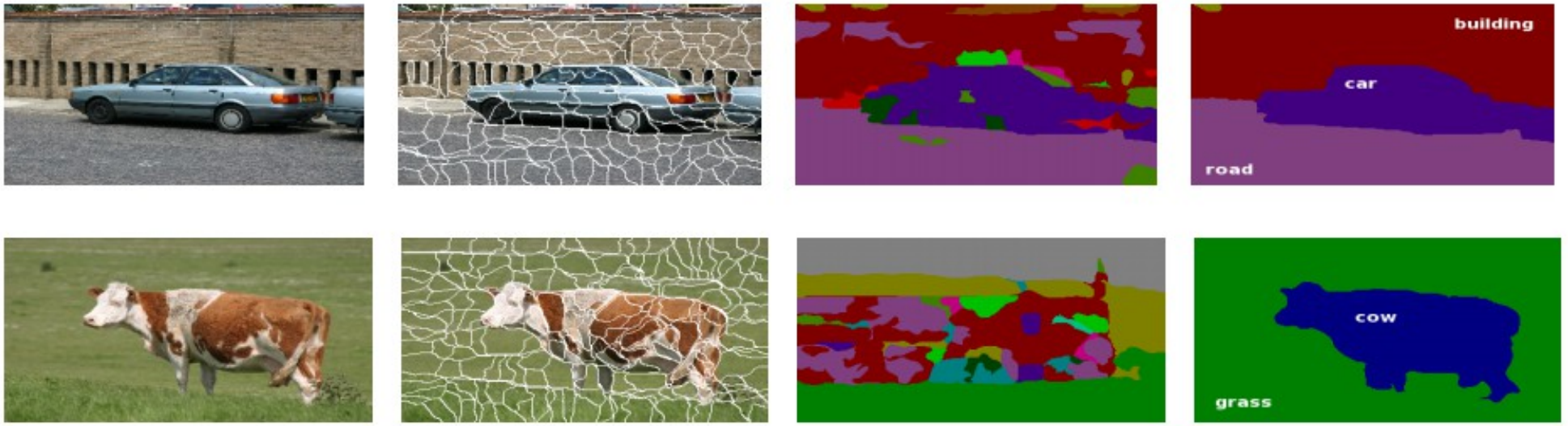




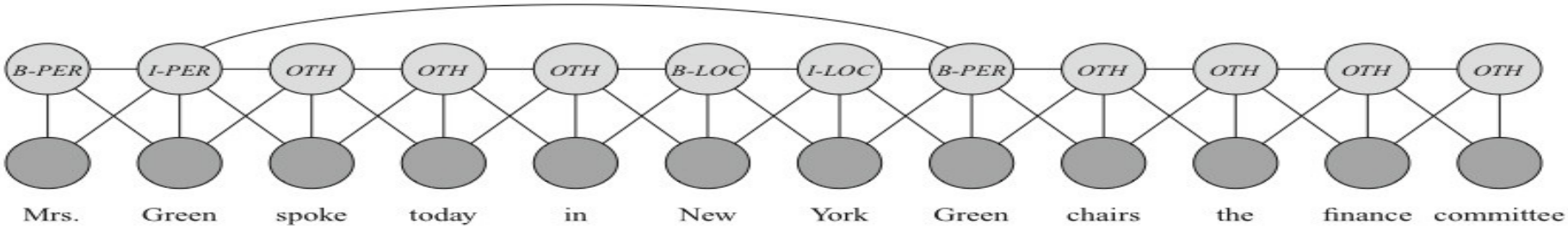
Anwendungsbereiche

- Medizinische Diagnose
- Fehleranalyse
- Sprachverarbeitung
- Verkehrsanalyse
- Soziale Netzwerkanalyse
- Nachrichtendekodierung
- Computer Vision
- Bildsegmentation
- 3D Rekonstruktion
- Szenenanalyse
- Spracherkennung
- Roboter Lokalisierung
- ...

Beispiel: Bildsegmentierung



Beispiel: Sprachverarbeitung

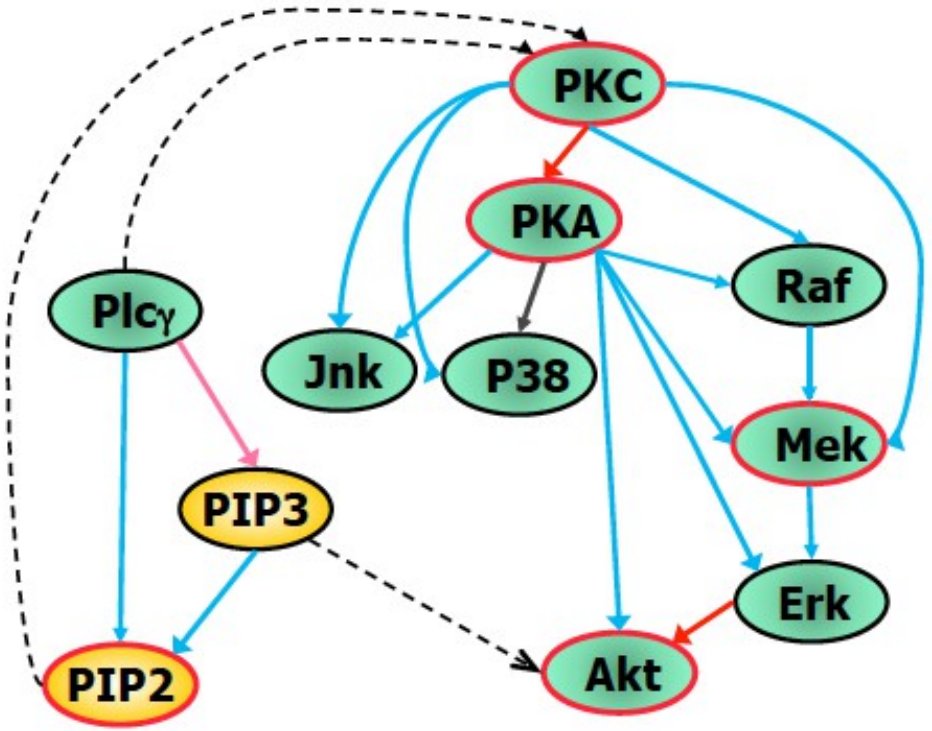


KEY

B-PER	Begin person name	*I-LOC*	Within location name
I-PER	Within person name	*OTH*	Not an entitiy
B-LOC	Begin location name		

(a)

Beispiel: Protein-Signaling Network



Themen bei PGMs

- Repräsentation
  - Gerichtete, ungerichtete Modelle
  - Temporale Modelle
  - „Plate“ Modelle
- Inferenz, Schlussfolgern
  - Exakt vs. Approximativ
  - Entscheidungsfindung
- Lernen
  - Parameter vs. Struktur
  - Vollständige vs. Unvollständige Daten

2 Arten von PGMs

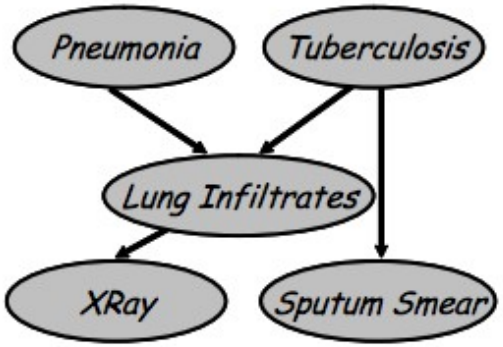
- Bayesian networks
  (auch Belief networks oder kausale Netzwerke genannt)
    - Gerichtete graphische Modelle
- Markov networks
  (auch Markov random fields (MRFs) genannt).
    - Ungerichtete graphische Modelle

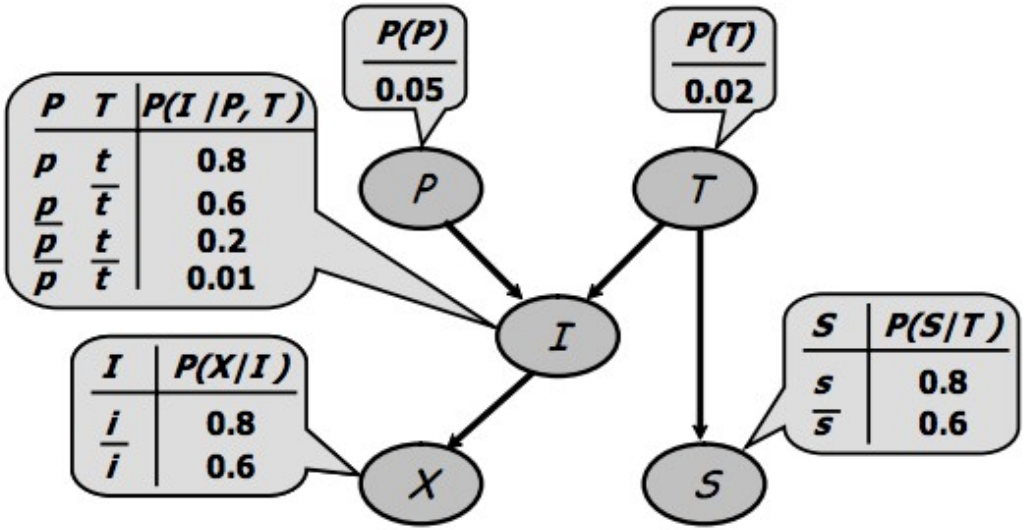
Bayes Network

Gerichteter azyklischer Graph (directed acyclic graph – DAG)

- Knoten von G sind Zufallsvariablen X_i
- Kanten: direkter Einfluss von einem Knoten auf den anderen
- Jeder Zufallsvariable X_i ist eine bedingte Wahrscheinlichkeitsverteilung (conditional probability distribution – CPD).
- Die CPD für X_i gegeben die Elternknoten (Pa_{Xi}) ist $P(X_i \mid Pa_{xi})$.

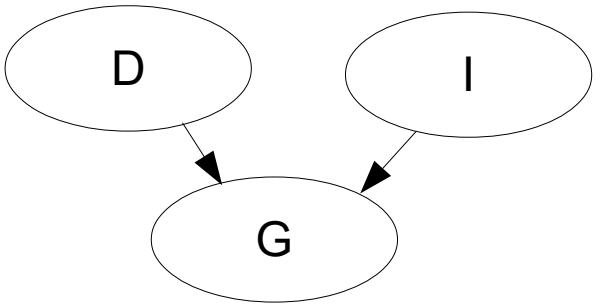
Bayes Network mit CPDs





Wahrscheinlichkeitsverteilung

I	D	G	P(I,D,G)
i0	d0	g1	0.126
i0	d0	g2	0.168
i0	d0	g3	0.126
i0	d1	g1	0.009
i0	d1	g2	0.045
i0	d1	g3	0.126
i1	d0	g1	0.252
i1	d0	g2	0.0224
i1	d0	g3	0.0056
i1	d1	g1	0.06
i1	d1	g2	0.036
i1	d1	g3	0.024



Konditionierung auf g1

P(I,D,g1)

I	D	G	P(I,D,G)
i0	d0	g1	0.126
i0	d1	g1	0.009
i1	d0	g1	0.252
i1	d1	g1	0.06

Reduktion und Normalisierung

P(I,D,g1)
→
P(I,D|g1)

I	D	G	P(I,D,g1)
i0	d0	g1	0.126
i0	d1	g1	0.009
i1	d0	g1	0.252
i1	d1	g1	0.06
			0.447

I	D	P(I,D|g1)
i0	d0	0.282
i0	d1	0.02
i1	d0	0.564
i1	d1	0.134
		1.000

Marginalisierung

P(I,D)
→
P(I)

I	D	P(I,D)
i0	d0	0.282
i0	d1	0.02
i1	d0	0.564
i1	d1	0.134
		1.000

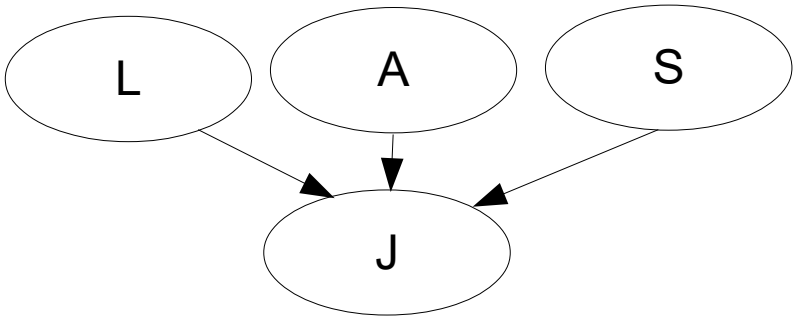
I	P(I)
i0	0.282+0.02 = 0.302
i1	0.564+0.134 = 0.698
	1.000

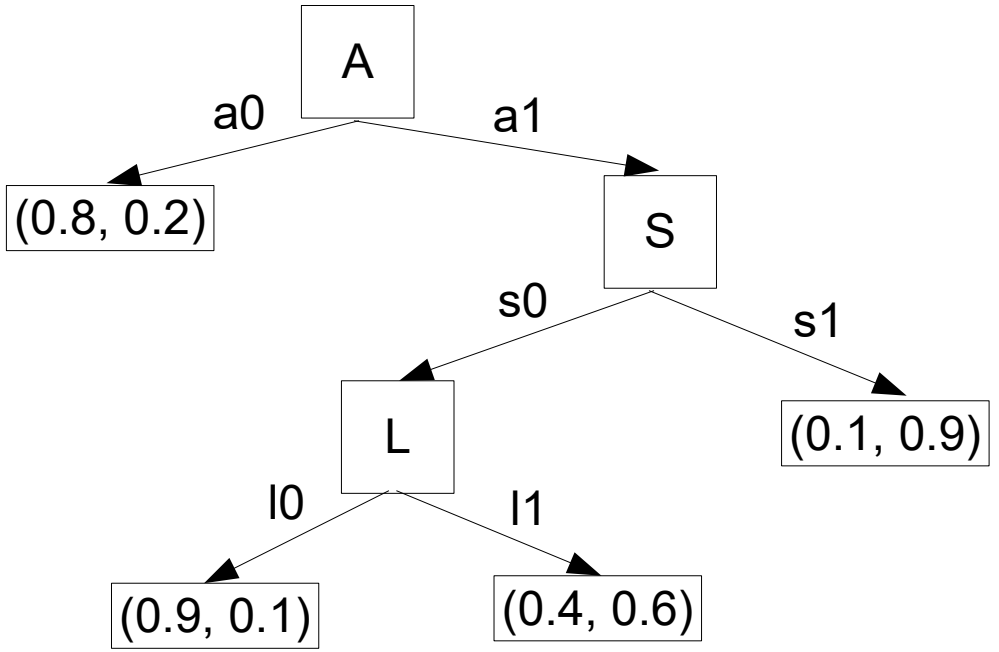
Darstellung von CPDs

- Tabellarische CPDs
- Baum-strukturierte CPDs
- Lokale Wahrscheinlichkeitsmodelle
- Logistische CPDs & Generalisierungen
- Noisy OR / AND
- Lineare Gauss Verteilungen & Generalisierungen
- ...

Baum CPD

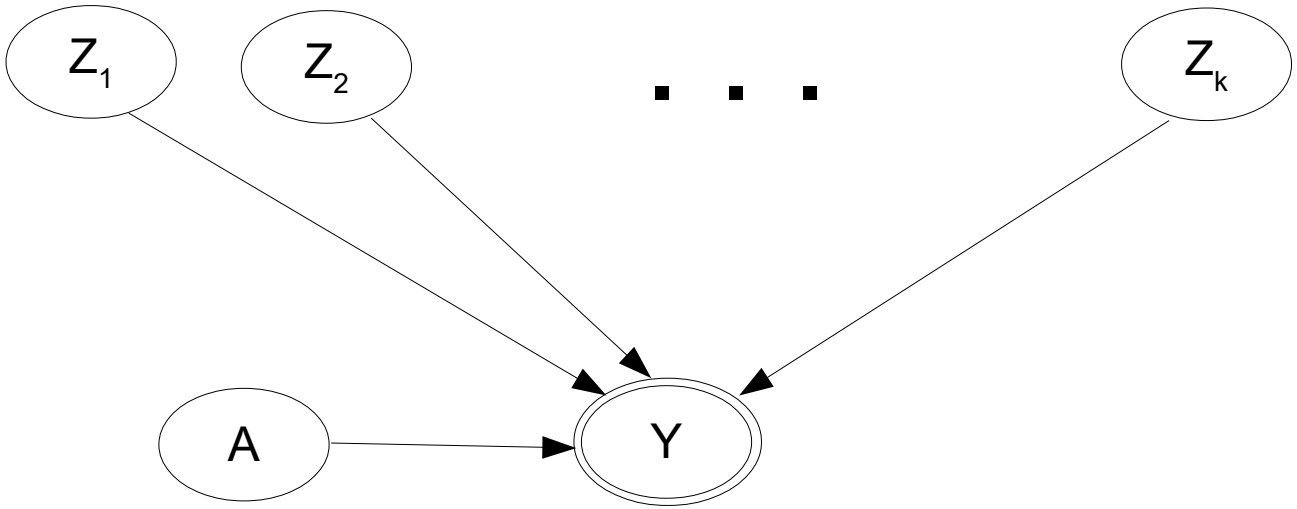
P(J | L, A, S)





Multiplexer CPD

$A = \{1,k\}$ $A = a \Rightarrow Y = Z_a$



$P(Y \mid A, Z_1, \ldots, Z_k) =$ 1 ..... $Y = Z$

0 ..... ansonsten

Faktoren

Faktor $\Phi(X_1,...X_n)$
Funktion $\Phi : Val(X_1,...X_n) \rightarrow R$

Scope = $\{X_1,...X_n\}$

- Jede JPD und CPD ist ein Faktor
- Unnormalisierte Funktion P(I,D,g1) ist Faktor über {I,D}

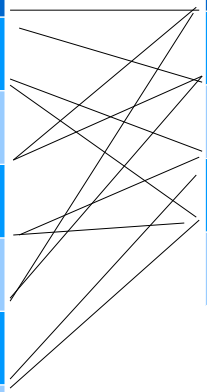
I	D		P(I,D,g1)
i0	d0	g1	0.126
i0	d1	g1	0.009
i1	d0	g1	0.252
i1	d1	g1	0.06
			0.447

Allgemeine Faktoren

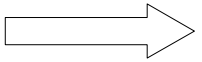
A	B	Φ
a0	b0	30
a0	b1	5
a1	b0	1
a1	b1	30

Faktor Produkt

A	B	Φ(A,B)
a0	b0	0.5
a0	b1	0.8
a1	b0	0.1
a1	b1	0
a2	b0	0.3
a2	b1	0.9



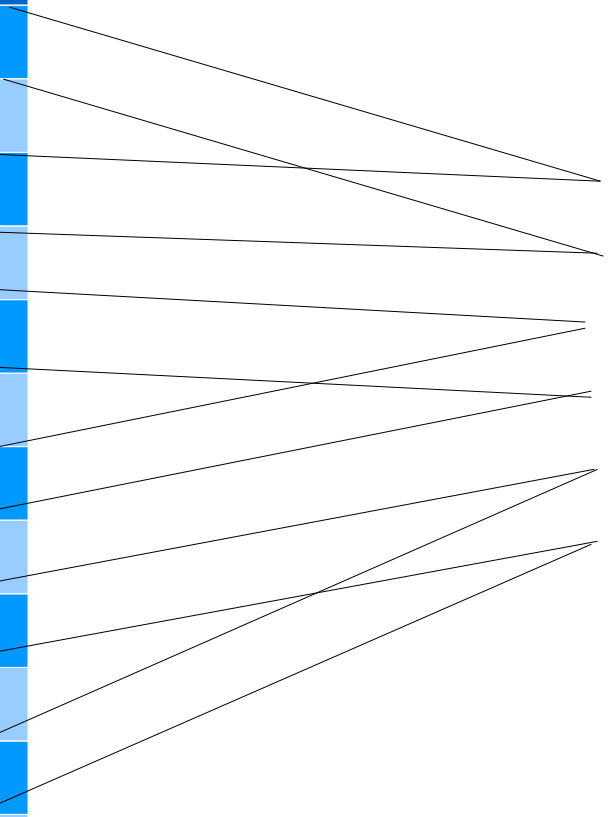
B	C	Φ(B,C)
b0	c0	0.5
b0	c1	0.7
b1	c0	0.1
b1	c1	0.2



A	B	C	Φ(A,B,C)
a0	b0	c0	0.5*0.5=0.25
a0	b0	c1	0.5*0.7=0.35
a0	b1	c0	0.8*0.1=0.08
a0	b1	c1	0.8*0.2=0.16
a1	b0	c0	0.1*0.5=0.05
a1	b0	c1	0.1*0.7=0.07
a1	b1	c0	0*0.1=0
a1	b1	c1	0*0.2=0
a2	b0	c0	0.3*0.5=0.15
a2	b0	c1	0.3*0.7=0.21
a2	b1	c0	0.9*0.1=0.09
a2	b1	c1	0.9*0.2=0.18

Faktor Marginalisierung

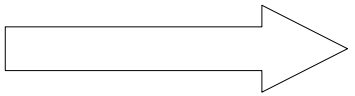
A	B	C	Φ(A,B,C)
a0	b0	c0	0.25
a0	b0	c1	0.35
a0	b1	c0	0.08
a0	b1	c1	0.16
a1	b0	c0	0.05
a1	b0	c1	0.07
a1	b1	c0	0
a1	b1	c1	0
a2	b0	c0	0.15
a2	b0	c1	0.21
a2	b1	c0	0.09
a2	b1	c1	0.18



A	C	Φ(A,C)
a0	c0	0.33
a0	c1	0.51
a1	c0	0.05
a1	c1	0.07
a2	c0	0.3
a2	c1	0.9

Faktor Reduktion

A	B	C	Φ(A,B,C)
a0	b0	c0	0.25
a0	b0	c1	0.35
a0	b1	c0	0.08
a0	b1	c1	0.16
a1	b0	c0	0.05
a1	b0	c1	0.07
a1	b1	c0	0
a1	b1	c1	0
a2	b0	c0	0.15
a2	b0	c1	0.21
a2	b1	c0	0.09
a2	b1	c1	0.18



A	B	C	Φ(A,B)
a0	b0	c0	0.25
a0	b1	c0	0.08
a1	b0	c0	0.05
a1	b1	c0	0
a2	b0	c0	0.15
a2	b1	c0	0.09

Warum Faktoren?

- Grundlegende Bestandteile zur Definition von Wahrscheinlichkeitsverteilungen in hoch-dimensionalen Räumen
- Menge von grundlegenden Operationen zur Manipulation der Wahrscheinlichkeitsverteilungen

Unabhängigkeit

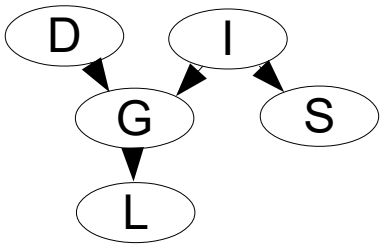
- Unabhängkeit

  $P(X = x, Y = y) = P(X = x)P(Y = y)$

- Bedingte Unabhängigkeit

  $P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$

Faktorisierung



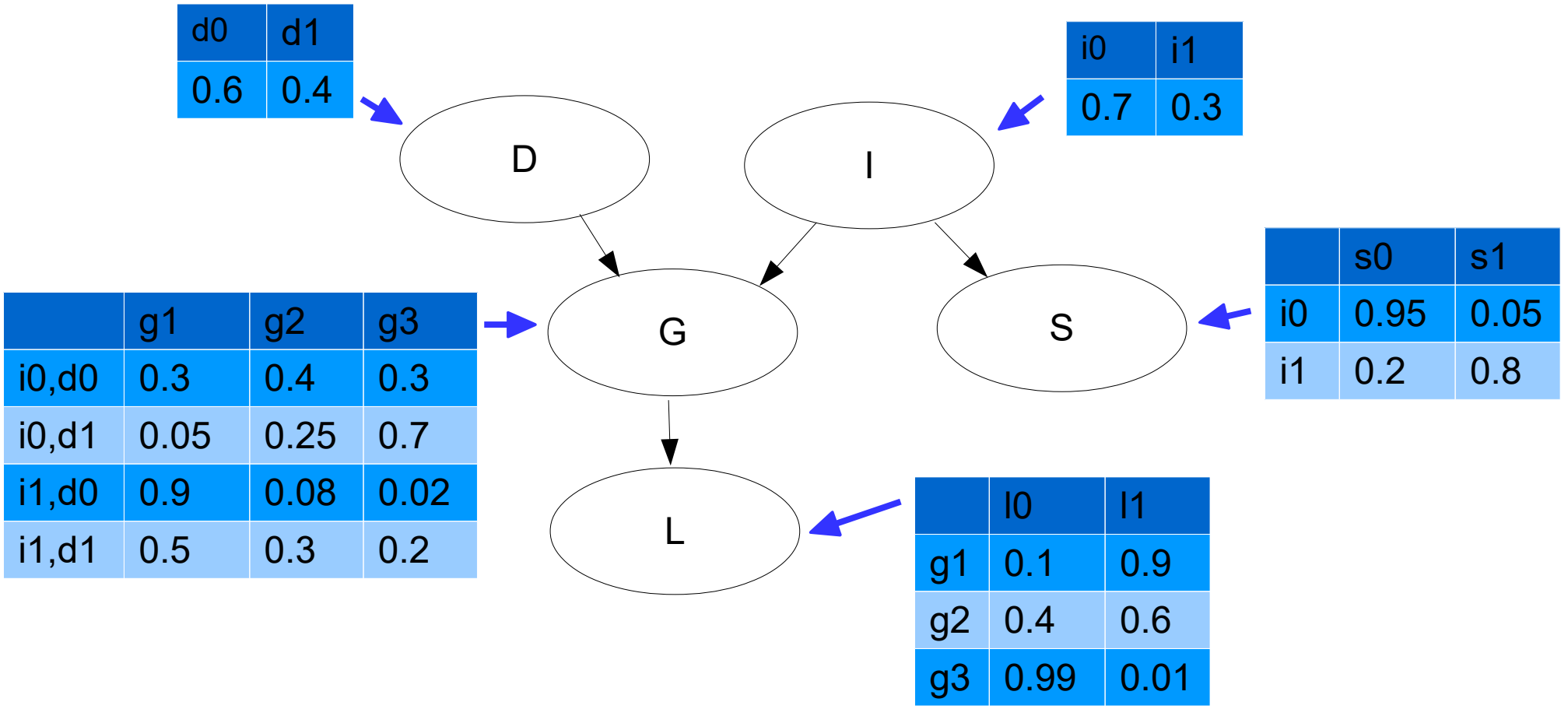
- Kettenregel
  $P(L,G,S,D,I) = P(L|G,S,D,I)*P(G|S,D,I)*P(S|D,I)*P(D|I)*P(I)$
- Kettenregel für Bayes'sche Netze

  $P(L,G,S,D,I) = P(L|G)*P(G|D,I)*P(S|I)*P(D)*P(I)$

  $P(X_1,\ldots,X_n) = \Pi_i P(X_i \mid Par_G(X_i))$

  → Distribution = Product of Factors
- Bedingte Unabhängigkeitsannahme,
  lokale Markov Annahme: $(X_i \perp NonDescendantsX_i \mid Par_G(X_i))$
- P faktorisierung über G, wenn
  $P(X_1,\ldots,X_n) = \Pi_i P(X_i \mid Par_G(X_i))$

Schlussfolgern in BN

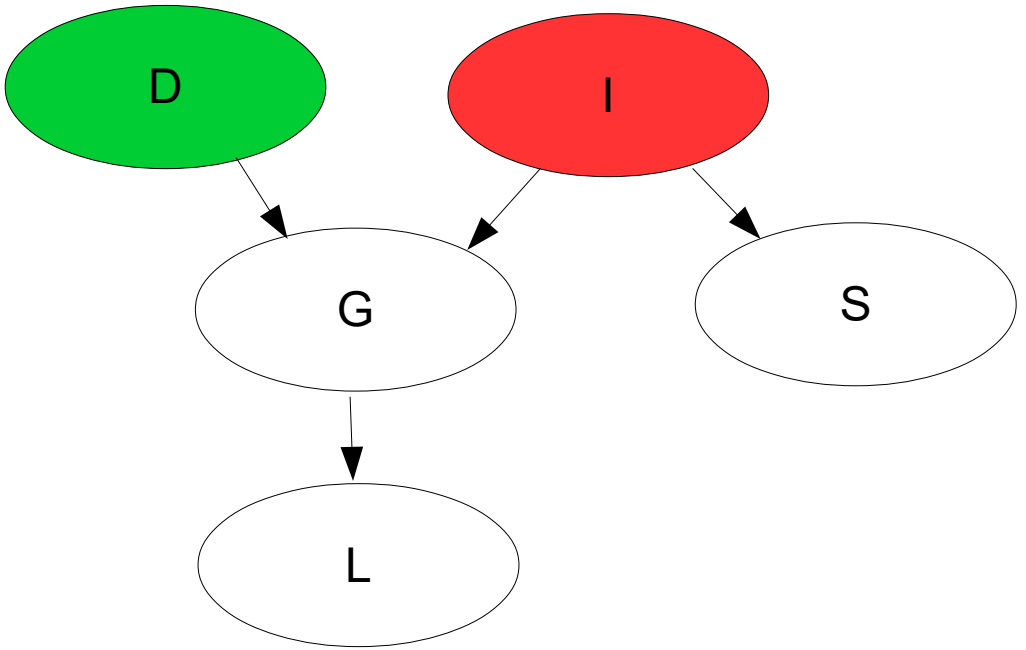


Kausalität

P(I1) ≈0.5

P(I1 | i0) ≈0.39

P(I1 | i0, d0) ≈0.51



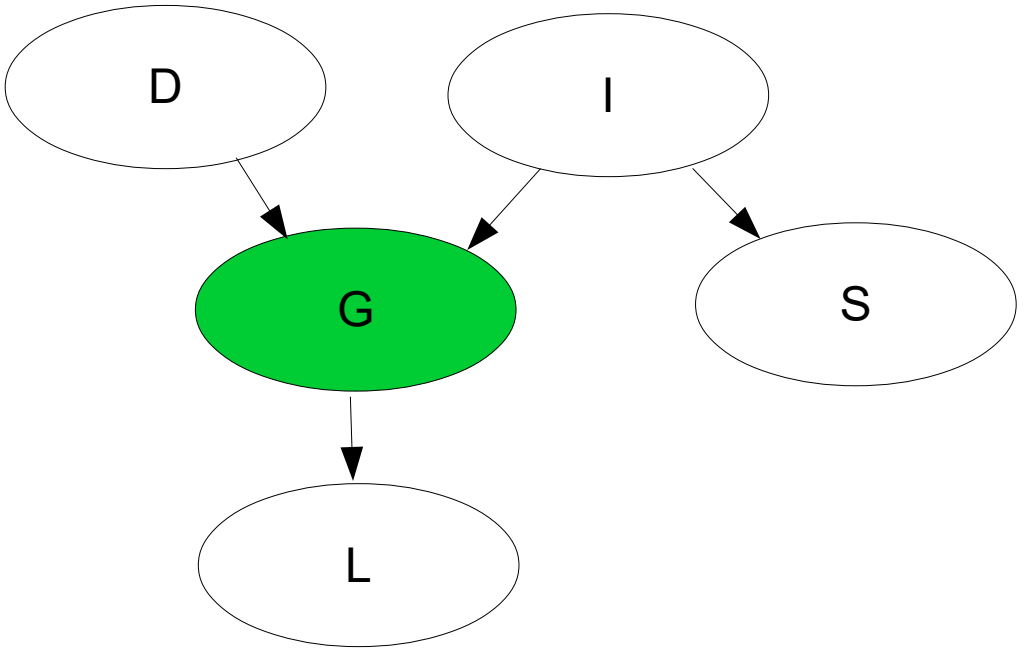
Evidenz

P(d1) = 0.4

P(d1 | g3) ≈0.63

P(i1) = 0.3

P(i1 | g3) ≈0.08



Interkausalität

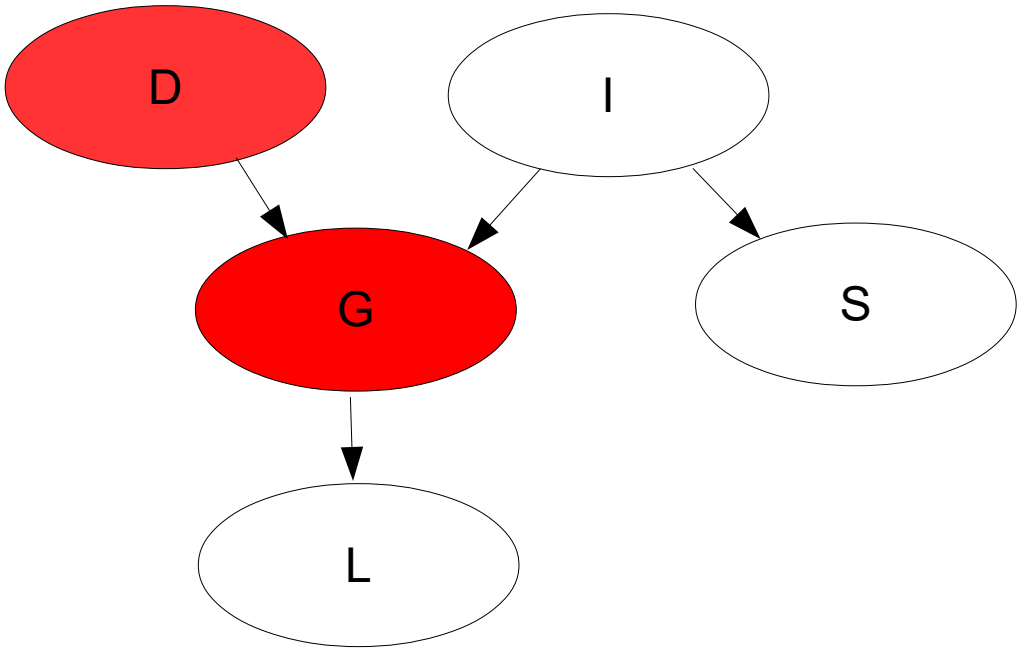
P(d1)=0.4

P(d1|g3)≈0.63

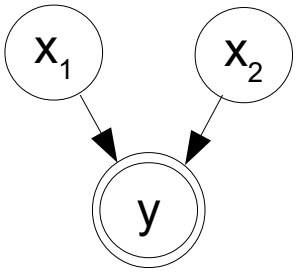
P(i1)=0.3

P(i1|g3)≈0.08

P(i1|g3,d1)≈0.11



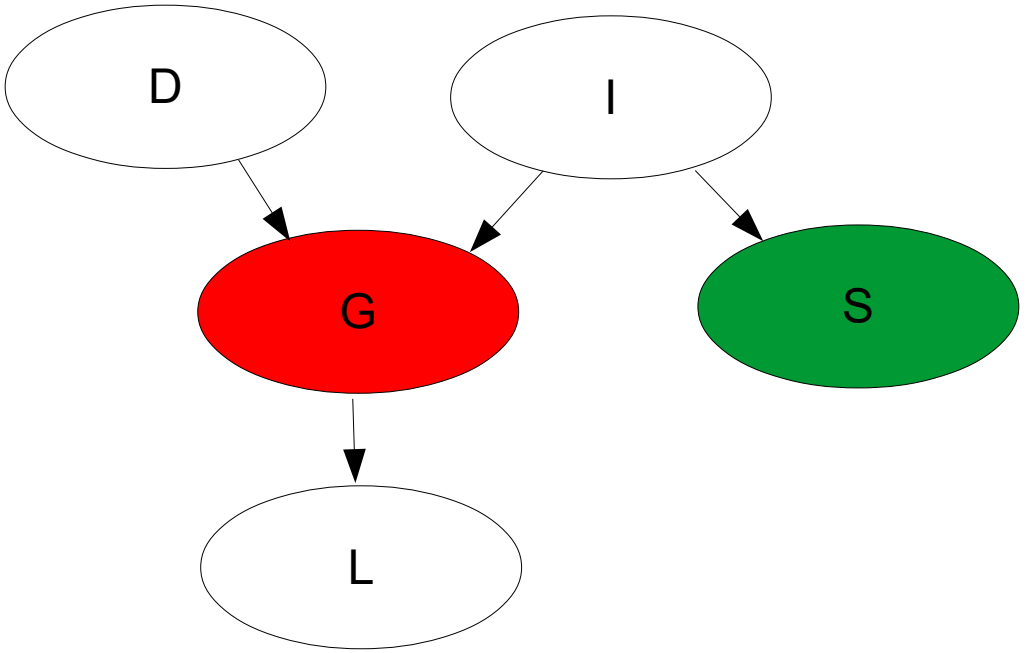
Explaining away



P(x2 | y=1, x1=1)

x1	x2	y	P
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	1	0.25

Interkausalität (2)



Flow of Influence

Verändert die Konditionierung auf X
die Wahrscheinlichkeit von Y?

- X → Y
- X ← Y
- X → W → Y ja
- X ← W ← Y ja
- X ← W → Y ja
- X → W ← Y nein
  (v-structure)

Flow of Influence

Verändert die Konditionierung auf X
die Wahrscheinlichkeit von Y, gegeben Evidenz Z?

Evidence Z	W nicht in Z	W in Z
• X → W → Y	ja	nein
• X ← W ← Y	ja	nein
• X ← W → Y	ja	nein
• X → W ← Y (v-structure)	?	?

  - Ja, wenn W oder ein Nachkomme von W in Z
  - Sonst nein

d-Separierbarkeit

d (=directed)-Separierbarkeit. Definition von aktiven Pfaden:

- Kausaler Pfad X → Z → Y: aktiv ⇔ Z nicht beobachtet
- Evidenz Pfad X ← Z ← Y : aktiv ⇔ Z nicht beobachtet
- Gemeinsamer Ursache X ← Z → Y : aktiv ⇔ Z nicht beobachtet
- Gemeinsamer Effekt X → Z ← Y : aktiv ⇔ Z oder einer seiner Nachkommen beobachtet

d-Separierbarkeit

Ein Pfad X_1— ... —X_n ist aktiv gegeben Evidenz E

- Wenn für jede v-Struktur $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ entweder X_i oder ein Nachkomme von X_i in E ist;
- Kein anderer Knoten entlang des Pfades in E ist.

Für drei Mengen X, Y, Z von Knoten aus G gilt. X und Y sind d-separiert gegeben Z d-sepG(X; Y | Z), wenn es keinen aktiven Pfaden zwischen zwei Knoten $x \in X$ und $y \in Y$ gegeben Z gibt.

Zusammenfassung: Bayesian Network

Zwei equivalente Betrachtungsweisen der Graphenstruktur

- Faktorisierung: G ermöglicht die Representation von P
- I-map: In G kodierte Unabhängigkeiten halten in P

Wenn P über G faktorisiert, dann können aus dem Graphen G die Unabhängigkeiten extrahiert werden, die in P halten müssen (Independence Map).

Ausblick

- Nächster Termin:

Donnerstag, 15.12.2016 13.15-13.45 (s.t.)

Strukturelle, syntaktische Mustererkennung