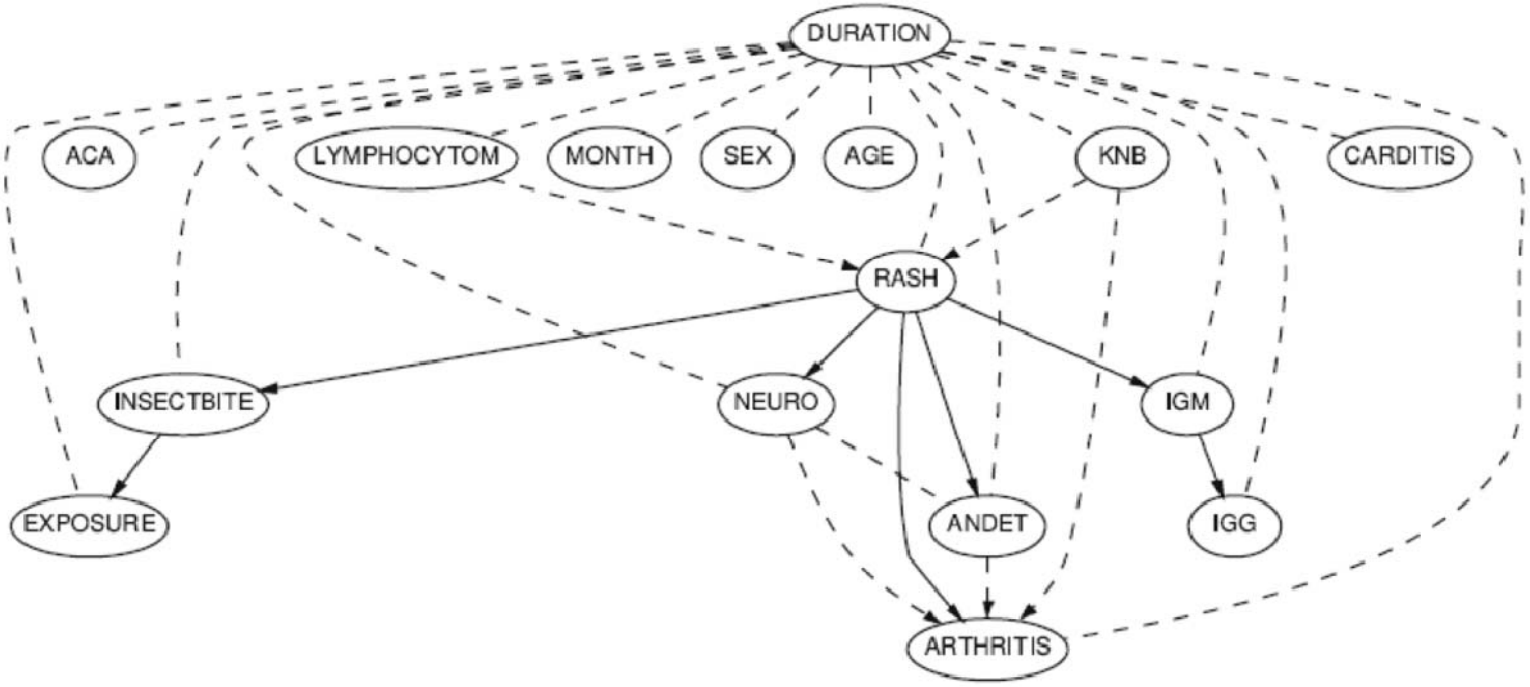
Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Graphische Modelle

Christoph Sawade/Niels Landwehr/Tobias Scheffer

Graphische Modelle

- Modellierung einer Domäne mit verschiedenen Zufallsgrössen
- Gemeinsame Verteilung, insb. Abhängigkeiten



Überblick

- Gerichtete Graphische Modelle: Bayessche Netze
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen
- Ungerichtete Graphische Modelle: Markov Netze

Überblick

- Gerichtete Graphische Modelle: Bayessche Netze
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen
- Ungerichtete Graphische Modelle: Markov Netze

Zufallsvariablen, Verteilungen

- Zufallsvariable: Abbildung eines Ereignisses auf reelle Zahl

$$X : \Omega \to \mathbb{R} \qquad \Omega \text{ Raum der Elementarereignisse}$$

- Verteilung über ZV: wie wahrscheinlich, bestimmte Werte zu beobachten?

- Diskrete Zufallsvariablen: diskreter Wertebereich (z.B. Münzwurf)

$$\text{Diskrete Wahrscheinlichkeit } p(x) \in [0,1], \quad \sum_x p(x) = 1 \qquad x \text{ Wert der ZV}$$

- Kontinuierliche Zufallsvariablen: kontinuierlicher Wertebereich (z.B. Körpergröße)

$$\text{Dichtefunktion } p(x) \in \mathbb{R}_{\geq 0}, \quad \int_{-\infty}^{\infty} p(x)dx = 1 \qquad x \text{ Wert der ZV}$$

Zufallsvariablen, Verteilungen

- Beispiele für diskrete/kontinuierliche Verteilungen
  - Bernoulli-Verteilung: binäre ZV $X \in \{0,1\}$

    Parameter $\mu = p(X = 1 \mid \mu)$

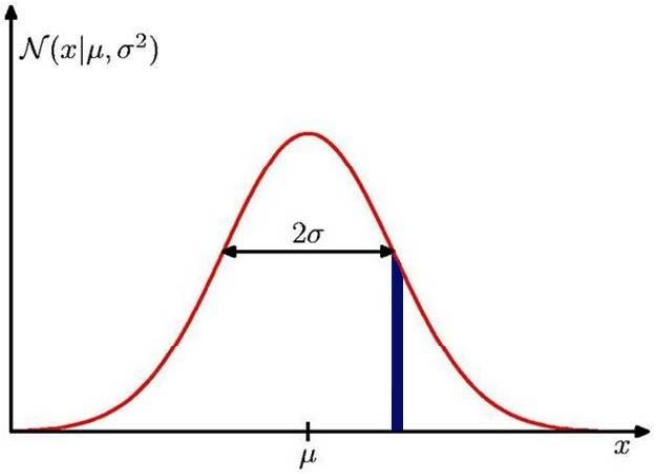
    $X \sim \text{Bern}(X \mid \mu) = \mu^X (1-\mu)^{1-X}$

  - Normalverteilung: kontinuierliche ZV $X \in \mathbb{R}$

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

μ Mittelwert

σ Standardabweichung

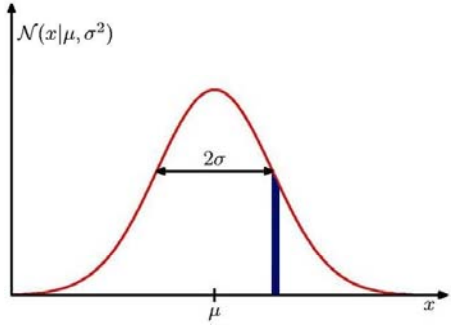


Zufallsvariablen, Verteilungen

- Zusammenhang Dichte/Wahrscheinlichkeit

$$p(X \in [x-\varepsilon, x+\varepsilon]) = \int_{x-\varepsilon}^{x+\varepsilon} p(z)dz$$

$p(X \in [x-\varepsilon, x+\varepsilon])$ ist
(diskrete) Wahrscheinlichkeit



- Oft abstrahieren wir vom Typ der Zufallsvariable
  - Rechenregeln sind im Wesentlichen gleich für diskrete/kontinuierliche Variablen („ersetze Summe durch Integral“)
  - Grundlegende Begriffe wie Abhängigkeit, Erwartungswert, Varianz sind auf beide Typen von Variablen anwendbar

Gemeinsame Verteilung, bedingte Verteilung

- Gemeinsame Verteilung $p(X,Y)$ über ZV X,Y

$$p(x,y) \in [0,1], \quad \sum_{x,y} p(x,y) = 1 \qquad p(x,y) \in \mathbb{R}_{\geq 0}, \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)dxdy = 1$$

- Bedingte Verteilung

$$p(X \mid Y) = \frac{p(X,Y)}{p(Y)}$$ diskret oder kontinuierlich

Für $p(y) > 0$ ist $p(X \mid y)$ wieder eine Verteilung über X:

$$\sum_{x} p(x \mid y) = 1 \text{ für } p(y) > 0 \text{ (diskret)}$$

$$\int_{-\infty}^{\infty} p(x \mid y)dx = 1 \textit{ für } p(y) > 0 \quad \text{(kontinuierlich)}$$

Unabhängigkeit von Zufallsvariablen

- Unabhängigkeit (diskret oder kontinuierlich)

  X,Y unabhängig genau dann wenn $p(X,Y) = p(X)p(Y)$

  X,Y unabhängig genau dann wenn $p(X \mid Y) = p(X)$

  X,Y unabhängig genau dann wenn $p(Y \mid X) = p(Y)$

- Bedingte Unabhängigkeit (diskret oder kontinuierlich)

  X,Y unabhängig gegeben Z genau dann wenn $p(X,Y \mid Z) = p(X \mid Z)p(Y \mid Z)$

  X,Y unabhängig gegeben Z genau dann wenn $p(Y \mid X,Z) = p(Y \mid Z)$

  X,Y unabhängig gegeben Z genau dann wenn $p(X \mid Y,Z) = p(X \mid Z)$

  ... einfach Anwendung des Unabhängigkeitsbegriffs auf die bedingte gemeinsame Verteilung p(X,Y|Z)

Rechenregeln Verteilungen

- Rechenregeln Wahrscheinlichkeiten/Verteilungen
  - Summenregel

$$p(x) = \sum_{y} p(x, y) \quad \text{diskrete Verteilungen}$$

$$p(x) = \int_{-\infty}^{\infty} p(x, y)dy \quad \text{kontinuierliche Verteilungen}$$

$p(X)$ heisst auch Randverteilung, Marginalverteilung

  - Produktregel

$$p(X, Y) = p(X \mid Y)p(Y) \quad \text{diskret oder kontinuierlich}$$

- Wahrscheinlichkeitsrechnungen beruhen auf Anwendungen dieser beiden Regeln

Graphische Modelle: Idee/Ziel

- Ziel: Modellierung der gemeinsame Verteilung $p(X_1,\ldots,X_N)$ einer Menge von ZV $X_1,\ldots,X_N$
- Aus $p(X_1,\ldots,X_N)$ lassen sich berechnen…
  - Alle Randverteilungen (Summenregel)

$$p(X_{i_1},\ldots,X_{i_m}),\quad \{i_1,\ldots,i_m\}\subseteq\{1,\ldots,N\}$$

  - Alle bedingten Verteilungen (aus Randverteilungen)

$$p(X_{i_1},\ldots,X_{i_m}\mid X_{i_{m+1}},\ldots,X_{i_{m+k}}),\quad \{i_1,.,i_{m+k}\}\subseteq\{1,\ldots,N\}$$

- Damit lassen sich alle probabilistischen Fragestellungen (*Inferenzprobleme*) über $X_1,\ldots,X_N$ beantworten

Graphische Modelle: Idee/Ziel

- Graphische Modelle: Kombination von Wahrscheinlichkeitstheorie und Graphentheorie
- Kompakte, intuitive Modellierung von $p(X_1,...,X_N)$
  - Graphstruktur repräsentiert Abhängigkeiten zwischen Variablen $X_1,...,X_N$
  - Einsicht in Struktur des Modells; einfach, Vorwissen einzubringen
  - Effiziente Algorithmen für Inferenz, die Graphstruktur ausnutzen
- Viele Methoden des maschinellen Lernens lassen sich in Form von GM darstellen
- Fragestellungen wie MAP Lernen, Bayessche Vorhersage lassen sich als Inferenzprobleme in GM formulieren

Graphische Modelle: Beispiel

- Beispiel: „Alarm“ Szenario
  - Unser Haus in LA hat eine Alarmanlage.
  - Wir sind im Urlaub. Unser Nachbar ruft an, falls er den Alarm hört. Wenn eingebrochen wurde, wollen wir zurück kommen.
  - Leider ist der Nachbar nicht immer zu Hause
  - Leider geht die Alarmanlage auch bei kleinen Erdbeben los
- 5 binäre Zufallsvariablen

B Burglary – Einbruch hat stattgefunden

E Earthquake – Erdbeben hat stattgefunden

A Alarm – Alarmanlage geht los

N NeighborCalls – Nachbar ruft an

R RadioReport – Bericht über Erdbeben im Radio

Graphische Modelle: Beispiel

- Zufallsvariablen haben eine gemeinsame Verteilung $p(B,E,A,N,R)$. Wie angeben? Welche Abhängigkeiten gelten?
- Beispiel für Inferenzproblem: Nachbar hat angerufen ($N=1$), wie wahrscheinlich ist Einbruch ($B=1$)?
  - Hängt von verschiedenen Faktoren ab
    - Wie wahrscheinlich Einbruch a priori?
    - Wie wahrscheinlich Erdbeben a priori?
    - Wie wahrscheinlich, dass Alarmanlage auslöst?
    - …

(Naive) Inferenz: $p(B=1 \mid N=1) = \dfrac{p(B=1,N=1)}{p(N=1)}$

$$= \frac{\sum_E \sum_A \sum_R p(B=1,E,A,N=1,R)}{\sum_B \sum_E \sum_A \sum_R p(B,E,A,N=1,R)}$$

Graphische Modelle: Beispiel

- Wie modellieren wir $p(B,E,A,N,R)$?
  - 1. Versuch: vollständige Tabelle

2^N

B	E	A	N	R	$P(B,E,A,N,R)$
0	0	0	0	0	0.6
1	0	0	0	0	0.005
0	1	0	0	0	0.01
...	...	...	...	...	...

+Alle Verteilungen $p(B,E,A,N,R)$ können repräsentiert werden

- Anzahl Parameter exponentiell
- Schwierig anzugeben

  - 2. Versuch: alles unabhängig

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A)\,p(N)\,p(R)$$

+Anzahl Parameter linear

- Zu restriktiv, Unabhängigkeitsannahme erlaubt keine sinnvolle Inferenz

Graphische Modelle: Beispiel

- Graphisches Modell: Selektive Unabhängigkeitsannahmen, durch Vorwissen motiviert
- Wähle Variablenordnung: z.B. $B<E<A<N<R$
- Produktregel:

$$
\begin{aligned}
p(B,E,A,N,R) &= p(B,E,A,N)\,p(R \mid B,E,A,N) \\
&= p(B,E,A)\,p(N \mid B,E,A)\,p(R \mid B,E,A,N) \\
&= p(B,E)\,p(A \mid B,E)\,p(N \mid B,E,A)\,p(R \mid B,E,A,N) \\
&= p(B)\,p(E \mid B)\,p(A \mid B,E)\,p(N \mid B,E,A)\,p(R \mid B,E,A,N)
\end{aligned}
$$

Faktoren beschreiben die Verteilung einer Zufallsvariablen in Abhängigkeit anderer Zufallsvariablen.

Können wir diese Faktoren vereinfachen?
Welche dieser Abhängigkeiten bestehen wirklich?

Graphische Modelle: Beispiel

- Zerlegung in Faktoren nach Produktregel:

$$p(B,E,A,N,R) = p(B)\,p(E \mid B)\,p(A \mid B,E)\,p(N \mid B,E,A)\,p(R \mid B,E,A,N)$$

- Annahme bedingter Unabhängigkeiten (Entfernen von Variablen aus Bedingungsteil)

$p(E \mid B) = p(E)$	*Erdbeben hängt nicht von Einbruch ab*
$p(A \mid B,E) = p(A \mid B,E)$	*Alarm hängt von Einbruch und Erdbeben ab*
$p(N \mid B,E,A) = p(N \mid A)$	*Anruf von Nachbar hängt nur von Alarm ab*
$p(R \mid B,E,A,N) = p(R \mid E)$	*Nachricht im Radio hängt nur Erdbeben ab*

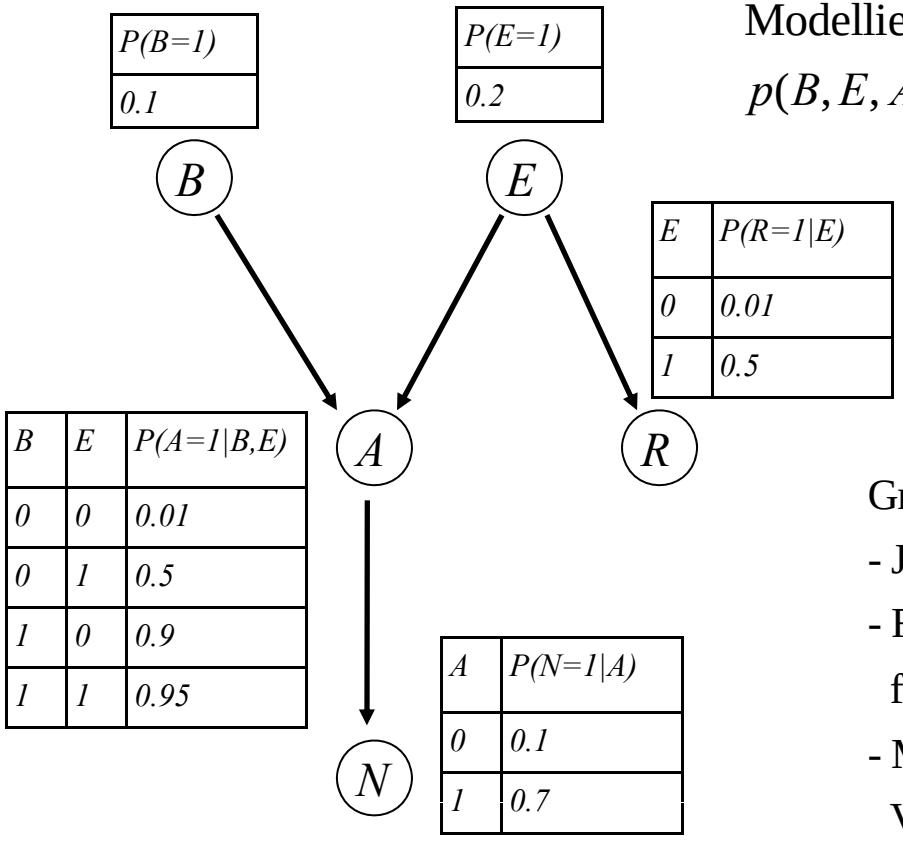
- Vereinfachte Darstellung der gemeinsamen Verteilung:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

Vereinfachte Faktoren

Graphische Modelle: Beispiel

- Graphisches Modell für „Alarm“ Szenario



Modellierte Verteilung:

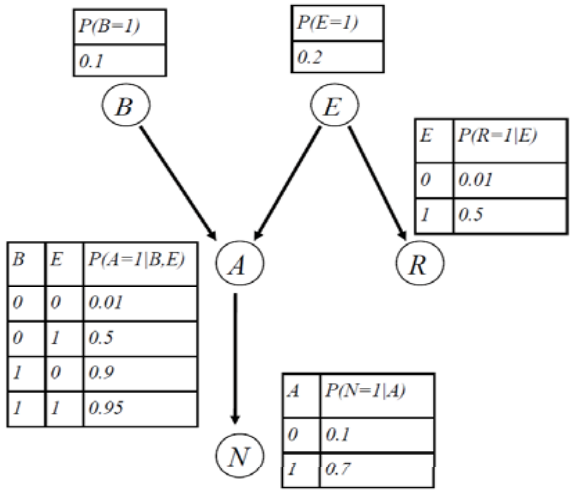
$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

Graphisches Modell:

- Jede ZV ist ein Knoten
- Für jeden Faktor der Form $p(X \mid X_1,...,X_k)$ fügen wir gerichtete Kanten von den X_i zu X ein
- Modell ist parametrisiert mit den bedingten Verteilungen $p(X \mid X_1,...,X_k)$

Graphische Modelle: Beispiel

- Graphisches Modell für „Alarm“ Szenario



  - Anzahl Parameter: $O(N*2^K)$, K= max. Anzahl von Elternknoten
  - Hier 1+1+2+2+4 statt 2^5-1 Parameter
- Gerichtete graphische Modelle heißen auch **Bayessche Netze**

Bayessche Netze: Definition

- Gegeben eine Menge von ZV $\{X_1,...,X_N\}$
- Ein Bayessches Netz über den ZV $\{X_1,...,X_N\}$ ist ein gerichteter Graph mit
  - Knotenmenge $X_1,...,X_N$
  - Es gibt keine gerichteten Zyklen $X_{i_1} \to X_{i_2} \to ... \to X_{i_k} \to X_{i_1}$
  - Knoten sind mit parametrisierten bedingten Verteilungen $p(X_i \mid pa(X_i))$ assoziiert, wobei $pa(X_i) = \{X_j \mid X_j \to X_i\}$ die Menge der Elternknoten eines Knoten ist

- Das Bayessche Netz modelliert eine gemeinsame Verteilung über $X_1,...,X_N$ durch

$$p(X_1,...,X_N) = \prod_{i=1}^{N} p(X_i \mid pa(X_i))$$

Bayessche Netze: Definition

- Warum muss der Graph azyklisch sein?
  - Satz aus der Graphentheorie:

    G ist azyklisch $\Leftrightarrow$ es gibt Ordnung $\leq_G$ der Knoten, so dass gerichtete Kanten die Ordnung respektieren ($N \to N' \Rightarrow N \leq_G N'$)

  - Damit ergibt sich

$$p(X_1,...,X_N) = \prod_{i=1}^{N} p(X_i \mid pa(X_i))$$

    aus Produktregel + bedingten Unabhängigkeitsannahmen (Variablen entsprechend $\leq_G$ umsortieren)

- Gegenbeispiel (kein Bayessches Netz):

$$p(X,Y) \neq p(X \mid Y)p(Y \mid X)$$

Bayessche Netze: Unabhängigkeit

- Die Graphstruktur eines Bayesschen Netzes impliziert (bedingte) Unabhängigkeiten zwischen ZV
- Notation: Für Variablen X, Y, Z schreiben wir

$$X \perp Y \mid Z \Leftrightarrow p(X \mid Y, Z) = p(X \mid Z)$$

"X unabhängig von Y gegeben Z"

- Erweiterung auf disjunkte Mengen A, B, C von ZV

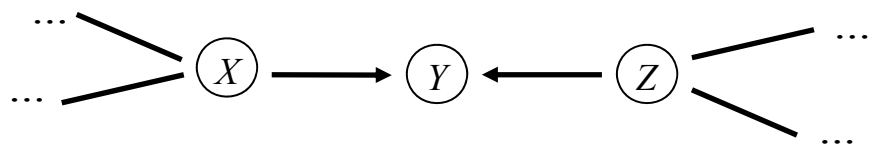
$$\mathrm{A} \perp \mathrm{B} \mid C \Leftrightarrow p(A \mid B, C) = p(A \mid C)$$

Bayessche Netze: Unabhängigkeit

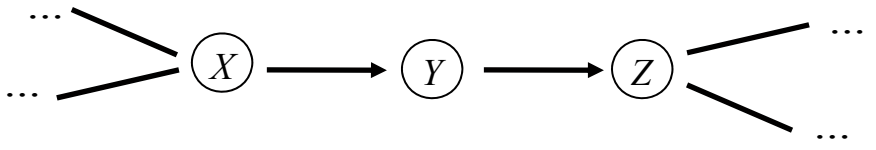
- Welche Unabhängigkeiten der Form $\mathrm{A} \perp \mathrm{B} \mid C$ werden durch die Graphstruktur modelliert?
  - Im Prinzip auszurechnen durch Summen/Produktregel aus der gemeinsamen Verteilung
  - Bei graphischen Modellen aber direkt aus der Graphstruktur ableitbar→ viel einfacher
  - „D-separation“: Menge einfacher Regeln, nach denen sich alle Unabhängigkeiten ableiten lassen

Bayessche Netze: Unabhängigkeit

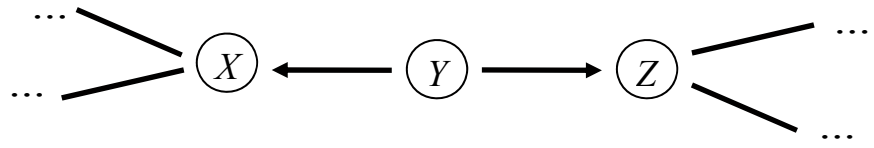
- D-separation: Welche Unabhängigkeiten der Form $A \perp B \mid C$ werden durch die Graphstruktur modelliert?
- Wichtige Rolle spielen Pfade im Graphen, die ZV verbinden
- Notation:



Pfad zwischen X und Z hat eine „**konvergierende**“ Verbindung bei Y („head to head“)

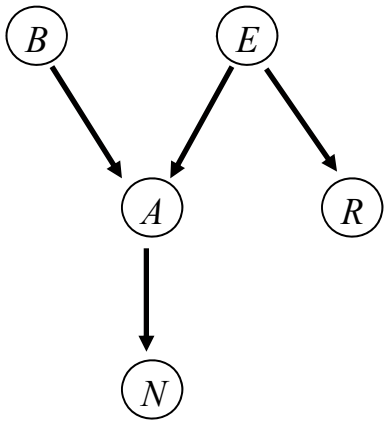


Pfad zwischen X und Z hat eine „**serielle**“ Verbindung bei Y („head to tail“)



Pfad zwischen X und Z hat eine „**divergierende**“ Verbindung bei Y („tail-to-tail“)

Divergierende Verbindungen



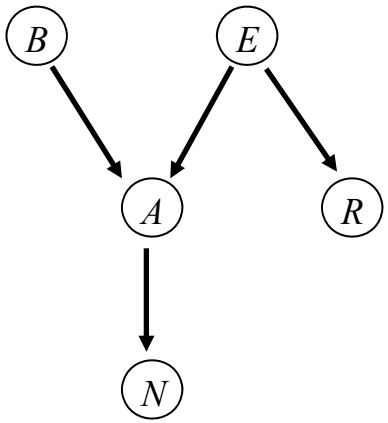
Gemeinsame Verteilung:

$$p(B,E,A,N,R) = p(B)p(E)p(A \mid E,B)p(N \mid A)p(R \mid E)$$

B=„Einbruch“ N=„Nachbar ruft an“
E=„Erdbeben“ R=„Radio Bericht“
A=„Alarm“

- Betrachte Pfad $A \leftarrow E \rightarrow R$. Gilt $\mathrm{A} \perp R \mid \varnothing$?
  -

Divergierende Verbindungen



Gemeinsame Verteilung:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

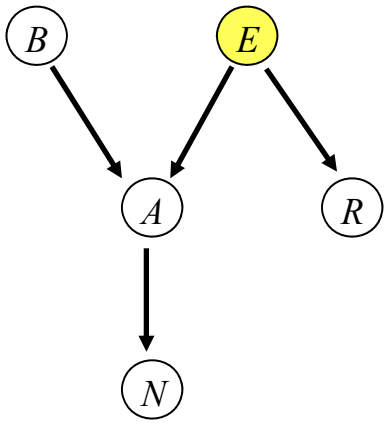
B=„Einbruch“ *N=„Nachbar ruft an“*
E=„Erdbeben“ *R=„Radio Bericht“*
A=„Alarm“

- Betrachte Pfad $A \leftarrow E \rightarrow R$. Gilt $A \perp R \mid \varnothing$?
  - Nein, $p(A \mid R) \neq p(A)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:

    RadioReport $\Rightarrow$ wahrscheinlich Erdbeben $\Rightarrow$ wahrscheinlich Alarm

    $p(A=1 \mid R=1) > p(A=1 \mid R=0)$
  - ZV R beeinflusst ZV A über die divergierende Verbindung $R \leftarrow E \rightarrow A$

Divergierende Verbindungen



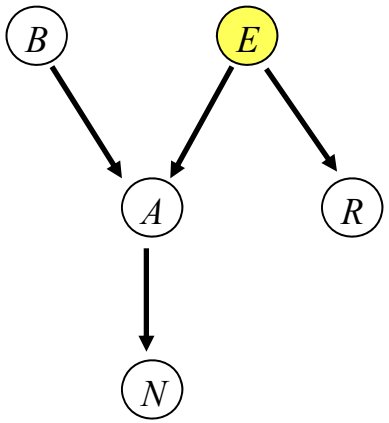
Gemeinsame Verteilung:

$$p(B,E,A,N,R) =$$
$$p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

beobachteter Knoten

- Betrachte Pfad $A \leftarrow E \rightarrow R$. Gilt $A \perp R \mid E$?

Divergierende Verbindungen



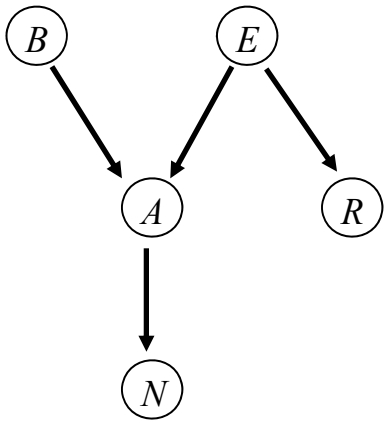
Gemeinsame Verteilung:

$$p(B,E,A,N,R) =$$
$$p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

beobachteter Knoten

- Betrachte Pfad $A \leftarrow E \rightarrow R$. Gilt $A \perp R \mid E$?
  - Ja, $p(A \mid R,E) = p(A \mid E)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    Wenn wir schon wissen, dass ein Erdbeben eingetreten ist, wird die Wahrscheinlichkeit für Alarm nicht höher/niedriger durch RadioReport
  - Der divergierende Pfad $R \leftarrow E \rightarrow A$ wird durch Beobachtung von E blockiert

Serielle Verbindungen

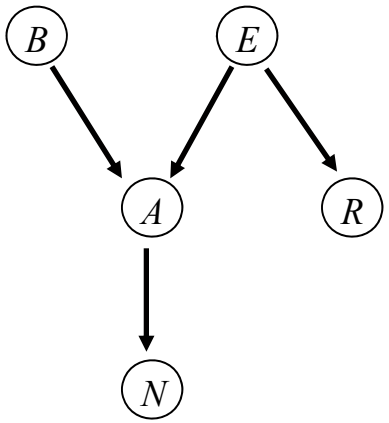


Gemeinsame Verteilung:

$$p(B,E,A,N,R) =$$
$$p(B)p(E)p(A \mid E,B)p(N \mid A)p(R \mid E)$$

- Betrachte Pfad $N \leftarrow A \leftarrow B$. Gilt $\mathrm{B} \perp \mathrm{N} \mid \varnothing$?
  -

Serielle Verbindungen



Gemeinsame Verteilung:

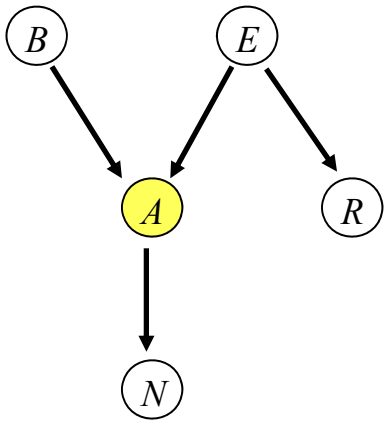
$$p(B,E,A,N,R) =$$
$$p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

- Betrachte Pfad $N \leftarrow A \leftarrow B$. Gilt $\mathrm{B} \perp \mathrm{N} \mid \varnothing$?
  - Nein, $p(B \mid N) \neq p(B)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:

    NeighborCalls $\Rightarrow$ wahrscheinlich Alarm $\Rightarrow$ *w*ahrscheinlich Burglary

    $p(B=1 \mid N=1) > p(B=1 \mid N=0)$
  - ZV N beeinflusst ZV B über den seriellen Pfad $\mathrm{N} \leftarrow A \leftarrow B$

Serielle Verbindungen



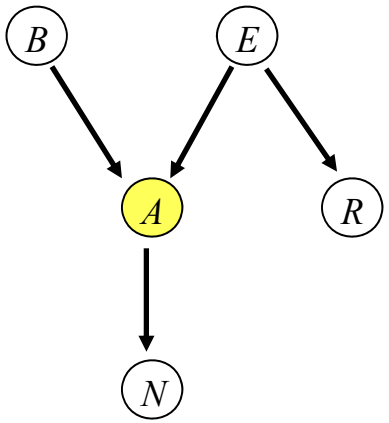
Gemeinsame Verteilung:

$$p(B,E,A,N,R) =$$
$$p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

beobachteter Knoten

- Betrachte Pfad $N \leftarrow A \leftarrow B$. Gilt $\mathrm{B} \perp \mathrm{N} \mid A$?
  -

Serielle Verbindungen



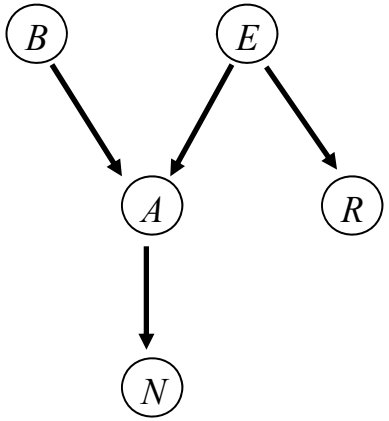
Gemeinsame Verteilung:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

beobachteter Knoten

- Betrachte Pfad $N \leftarrow A \leftarrow B$. Gilt $\mathrm{B} \perp \mathrm{N} \mid A$?
  - Ja, $p(B \mid N, A) = p(B \mid A)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    Wenn wir schon wissen, dass der Alarm ausgelöst wurde, sinkt/steigt die Wahrscheinlichkeit für Einbruch nicht dadurch, dass Nachbar anruft
  - Der serielle Pfad $N \leftarrow A \leftarrow B$ wird durch Beobachtung von A blockiert.

Konvergierende Verbindung

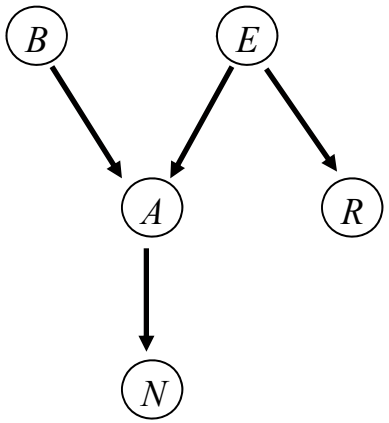


Gemeinsame Verteilung:

$$p(B,E,A,N,R) =$$
$$p(B)p(E)p(A \mid E,B)p(N \mid A)p(R \mid E)$$

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E} \mid \varnothing$?
  -

Konvergierende Verbindung

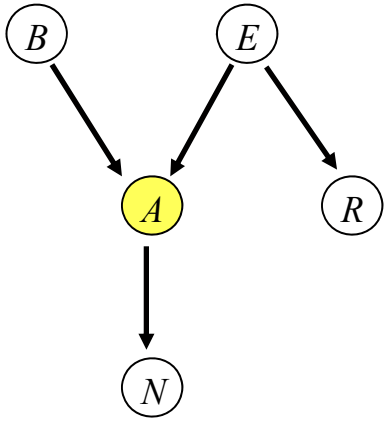


Gemeinsame Verteilung:

$$p(B,E,A,N,R) =$$
$$p(B)p(E)p(A \mid E,B)p(N \mid A)p(R \mid E)$$

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E} \mid \varnothing$?
  - Ja, $p(B \mid E) = p(B)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    Einbrüche treten nicht häufiger/seltener auf an Tagen mit Erdbeben
  - Der konvergierende Pfad $\mathrm{B} \rightarrow A \leftarrow E$ ist blockiert wenn A **nicht** beobachtet ist

Konvergierende Verbindung



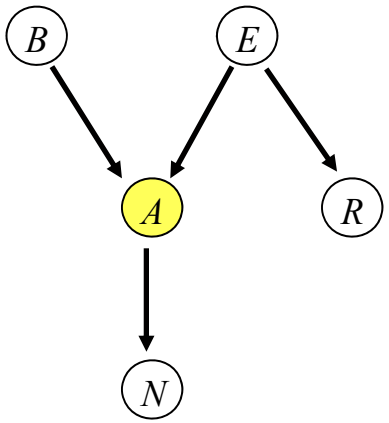
Gemeinsame Verteilung:

$$p(B,E,A,N,R) =$$
$$p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

beobachteter Knoten

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E} \mid A$?

Konvergierende Verbindung



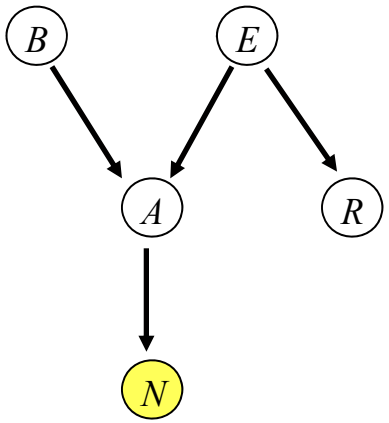
Gemeinsame Verteilung:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

beobachteter Knoten

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E} \mid A$?
  - Nein, $p(B \mid E,A) \neq p(B \mid A)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    Alarm wurde ausgelöst. Falls wir ein Erdbeben beobachten, erklärt das den Alarm, Wahrscheinlichkeit für Einbruch sinkt ("explaining away").
  - Der konvergierende Pfad $\mathrm{B} \rightarrow A \leftarrow E$ wird **freigegeben** durch Beobachtung von A

Konvergierende Verbindung



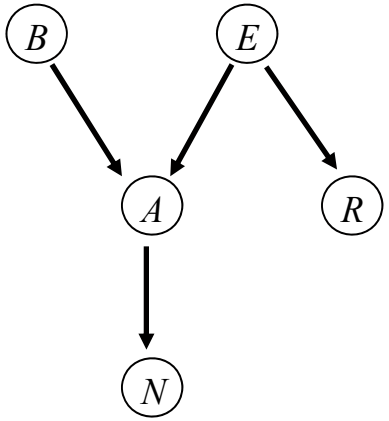
Gemeinsame Verteilung:

$$p(B,E,A,N,R) =$$
$$p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

beobachteter Knoten

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E} \mid N$?
  - Nein, $p(B \mid N,A) \neq p(B \mid A)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    NeighborCalls indirekte Beobachtung von Alarm. Erdbebenbeobachtung erklärt den Alarm, Wahrscheinlichkeit für Einbruch sinkt ("explaining away").
  - Der konvergierende Pfad $\mathrm{B} \rightarrow A \leftarrow E$ wird **freigegeben** durch Beobachtung von N

Zusammenfassung Pfade



Gemeinsame Verteilung:

$$p(B,E,A,N,R) = p(B)p(E)p(A \mid E,B)p(N \mid A)p(R \mid E)$$

- Zusammenfassung: ein Pfad …-*X*-*Y*-*Z*-... ist
  - Blockiert bei Y, wenn
    - Divergierende Verbindung, und Y beobachtet, oder
    - Serielle Verbindung, und Y beobachtet, oder
    - Konvergierende Verbindung, und weder Y noch einer seiner Nachfahren $Y' \in Descendants(Y)$ beobachtet
    - $Descendants(Y) = \{Y' \mid \text{es gibt gerichteten Pfad von } Y \text{ zu } Y'\}$
  - Sonst ist der Pfad frei bei Y

D-Separation: Blockierte Pfade

- Seien X,X' ZV, C eine beobachtete Menge von ZV, $X,X' \notin C$
- Ein ungerichteter Pfad $X - X_1 - \ldots - X_n - X'$ zwischen X und X' ist blockiert gegeben C gdw es einen Knoten X_i gibt so dass Pfad bei X_i blockiert ist gegeben C

- D-Separation basiert auf blockierten Pfaden:
  - Seien A,B,C disjunkte Mengen von ZV.
  - Definition: A und B sind d-separiert durch C gdw jeder Pfad von einem Knoten $X \in A$ zu einem Knoten $Y \in B$ blockiert ist gegeben C.

D-Separation: Korrektheit

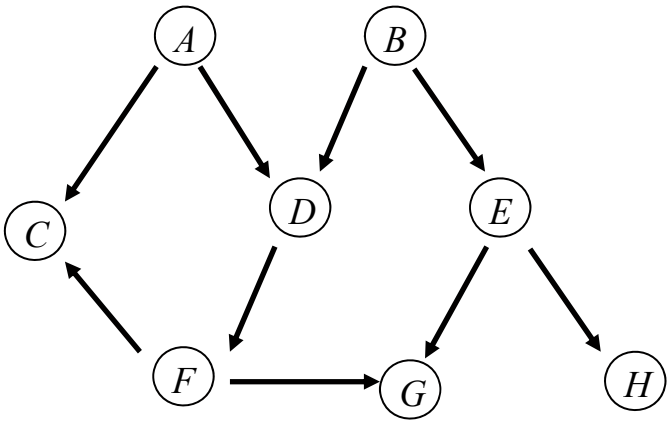
- Gegeben ein Bayessches Netz über $\{X_1,...,X_N\}$ mit Graphstruktur G.
- Das Bayessche Netz modelliert eine Verteilung durch

$$p(X_1,...,X_N) = \prod_{i=1}^{N} p(X_i \mid pa(X_i))$$

abhängig von den bedingten Verteilungen $p(X_n \mid pa(X_n))$.

- Theorem (Korrektheit, Vollständigkeit d-separation)
  - Falls A,B d-separiert gegeben C in G, dann $p(A \mid B,C) = p(A \mid C)$
  - Es gibt keine anderen Unabhängigkeiten, die für jede Wahl der bedingten Verteilungen $p(X_i \mid pa(X_i))$ gelten.

D-Separation: Beispiel



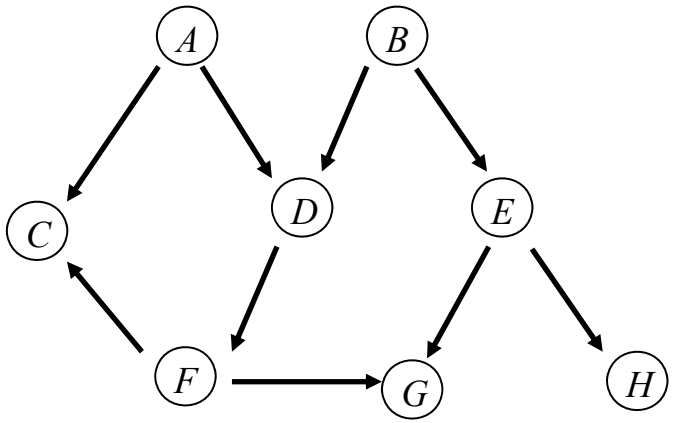
Gilt $B \perp H \mid E$?

Gilt $A \perp H \mid F$?

Gilt $C \perp E \mid \varnothing$?

- Ein Pfad …-*X*-*Y*-*Z*-... ist
  - Blockiert bei Y, wenn
    - Divergierende Verbindung, und Y beobachtet, oder
    - Serielle Verbindung, und Y beobachtet, oder
    - Konvergierende Verbindung, und weder Y noch einer seiner Nachfahren $Y' \in Descendants(Y)$ beobachtet
  - Sonst ist der Pfad frei bei Y

D-Separation: Beispiel



Gilt $B \perp H \mid E$? Ja

Gilt $A \perp H \mid F$? Nein: $A - D - B - E - H$

Gilt $C \perp E \mid \varnothing$? Nein: $C - F - D - B - E$

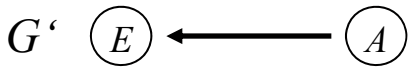
- Ein Pfad …-*X*-*Y*-*Z*-... ist
  - Blockiert bei Y, wenn
    - Divergierende Verbindung, und Y beobachtet, oder
    - Serielle Verbindung, und Y beobachtet, oder
    - Konvergierende Verbindung, und weder Y noch einer seiner Nachfahren $Y' \in Descendants(Y)$ beobachtet
  - Sonst ist der Pfad frei bei Y

Bayessche Netze: Kausalität

- Oft werden Bayessche Netze so konstruiert, dass gerichtete Kanten kausalen Einflüssen entsprechen

„Erdbeben lösen die Alarmanlage aus“

- Äquivalentes Modell



„die Alarmanlage löst Erdbeben aus“

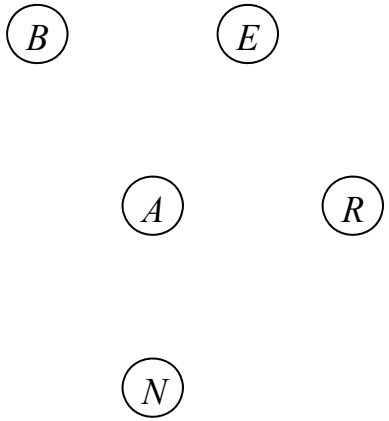
- Definition: $\mathrm{I(G)} = \{ (A \perp \mathrm{B} \mid C) : A$ und B sind d-separiert gegeben C in $G\}$

„Alle Unabhängigkeitsannahmen, die durch G kodiert werden“

- $I(G) = I(G') = \varnothing$:
  - Keine statistischen Gründe, eines der Modelle vorzuziehen
  - Kann nicht aufgrund von Daten zwischen Modellen unterscheiden
  - Aber „kausale“ Modelle oft besser verständlich

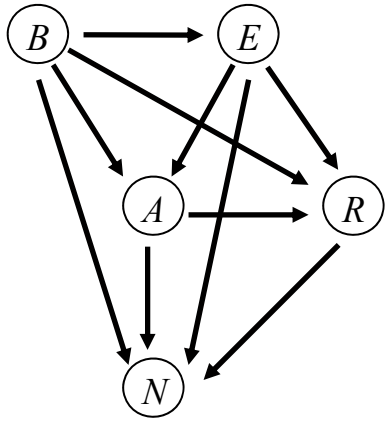
Netze Unterschiedlicher Komplexität

- Komplexität: bestimmt durch Menge der Kanten im Graph
  - Beeinflusst die Anzahl der Parameter im Modell. Für binäre ZV gilt: Anzahl Parameter in $O(N*2^K)$, $K = \max_i |pa(X_i)|$
  - Hinzufügen von Kanten: Familie der darstellbare Verteilungen wird grösser, $I(G)$ wird kleiner
- Extremfälle: Graph ohne Kanten, (ungerichtet) vollständig verbundener Graph



N Parameter

$I(G) = \{(A \perp B \mid C) : A, B, C$ disj. Mengen von ZV$\}$



2^N -1 Parameter

$I(G) = \varnothing$

Überblick

- Gerichtete Graphische Modelle: Bayessche Netze
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen
- Ungerichtete Graphische Modelle: Markov Netze

Erinnerung: Lernproblem

- Erinnerung: Lernproblem
  - Trainingsdaten

$$L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) \rangle$$

$\mathbf{x}_i \in \mathbb{R}^m$ Merkmalsvektoren

$y_i \in \{0,1\}$ binäre Klassenlabel

$y_i \in \mathbb{R}$ reelles Label

  - Matrixschreibweise

Merkmalsvektoren

$$X = (\mathbf{x}_1 \quad ... \quad \mathbf{x}_N) = \begin{pmatrix} x_{11} & x_{N1} \\ \vdots & \ddots & \vdots \\ x_{1m} & x_{Nm} \end{pmatrix}$$

Zugehörige Klassenlabels

$$\mathbf{y} = \begin{pmatrix} y_1 \\ ... \\ y_N \end{pmatrix}$$

  - Ziel: Vorhersage des Klassenlabels für Testinstanz $\mathbf{x}$

$$\mathbf{x} \mapsto y$$

Erinnerung: Lernen

- Annahme: gemeinsame Verteilung $p(\mathbf{x},y)$ (unbekannt)
  - Trainingsdaten $(\mathbf{x}_i, y_i) \sim p(\mathbf{x}, y)$ i.i.d.
  - Testinstanzen $(\mathbf{x}, y) \sim p(\mathbf{x}, y)$
- Wir betrachten probabilistische Modelle

  $p(y|\mathbf{x},\theta)$ [diskriminativ] $p(\mathbf{x}, y|\theta)$ [generativ]
- A-priori Verteilung über Modelle $p(\theta)$ (bekannt)
- Vorhersageproblem: MAP Lösung

  $\theta_* = \arg\max_\theta p(\theta|L)$ $y_* = \arg\max_y p(y|\mathbf{x},\theta_*)$
- Vorhersageproblem: Bayes Lösung

  $y_* = \arg\max_y p(y|\mathbf{x},L) = \arg\max_y \int p(y|\mathbf{x},\theta)p(\theta|L)d\theta$

Erinnerung: Parameterschätzung Münzwurf

- Erinnerung: Münzwurf
  - Einzelner Münzwurf Bernouilli-verteilt mit Parameter μ

$$X \sim \text{Bern}(X \mid \mu) = \mu^X (1-\mu)^{1-X}$$

$$\mu = p(X = 1 \mid \mu) \quad \text{unbekannter Parameter}$$

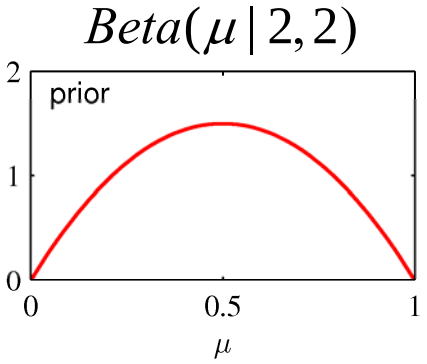
- Parameterschätzproblem:
  - Wir haben N unabhängige Münzwürfe gesehen, als Ausprägung $L = \{x_1, \ldots, x_N\}$ der ZV $X_1, \ldots, X_N$
  - Der echte Parameter μ ist unbekannt, wir wollen eine Schätzung $\hat{\mu}$ bzw. eine posterior-Verteilung $p(\mu \mid L)$
  - Bayesscher Ansatz: Posterior $\propto$ Prior x Likelihood

$$\underbrace{p(\mu \mid L)}_{posterior} \propto \underbrace{p(L \mid \mu)}_{likelihood} \underbrace{p(\mu)}_{prior}$$

Erinnerung: Parameterschätzung Münzwurf

- Prior: Beta-Verteilung über Münzparameter μ

$$p(\mu \mid \alpha_k, \alpha_z) = Beta(\mu \mid \alpha_k, \alpha_z)$$
$$= \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \mu^{\alpha_k - 1}(1-\mu)^{\alpha_z - 1}$$



- Likelihood N unabhängige Münzwürfe:

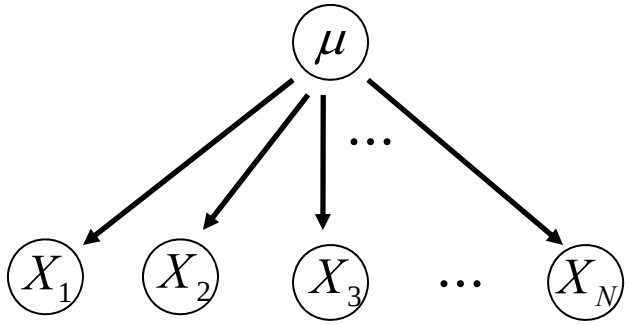
$$P(X_1, ..., X_N \mid \mu) = \prod_{n=1}^{N} p(X_n \mid \mu) \qquad i.i.d.$$
$$= \prod_{n=1}^{N} \text{Bern}(X_n \mid \mu)$$
$$= \prod_{n=1}^{N} \mu^{X_n}(1-\mu)^{1-X_n}$$

Erinnerung: Parameterschätzung Münzwurf

- Zufallsvariablen in Münzwurfszenario sind $X_1, ..., X_N, \mu$
- Gemeinsame Verteilung von Daten und Parameter (gegeben „Hyperparameter“ α_k, α_z): Prior x Likelihood

$$p_{\alpha_k,\alpha_z}(X_1, ..., X_N, \mu) = p_{\alpha_k,\alpha_z}(\mu) \prod_{i=1}^{N} \underbrace{p(X_i \mid \mu)}_{\text{Bernoulli}}$$

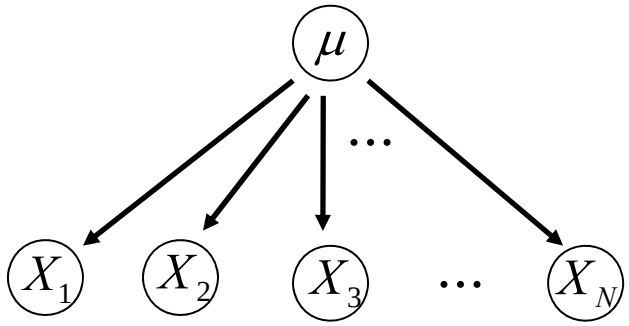
- Darstellung als graphisches Modell:



$$pa(\mu) = \varnothing$$
$$pa(X_i) = \{\mu\}$$

Schätzung eines Münzparameters als Graphisches Modell

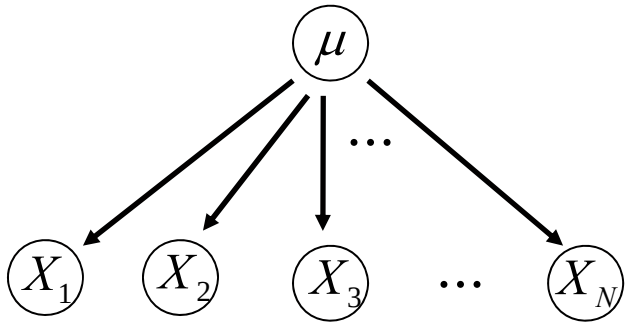
- Unabhängige Münzwürfe: Darstellung als graphisches Modell



- D-separation
  - Gilt $X_N \perp X_1,...,X_{N-1} | \varnothing$?

Schätzung eines Münzparameters als Graphisches Modell

- Unabhängige Münzwürfe: Darstellung als graphisches Modell



- D-separation
  - Gilt $X_N \perp X_1,...,X_{N-1} \mid \varnothing$?
  - Nein, Pfad durch μ ist nicht blockiert.
  - Intuitiv:

$$X_1 = X_2 = ... = X_{N-1} = 1 \Rightarrow \text{Wahrscheinlich } \mu > 0.5 \Rightarrow \text{Wahrscheinlich } X_N = 1$$

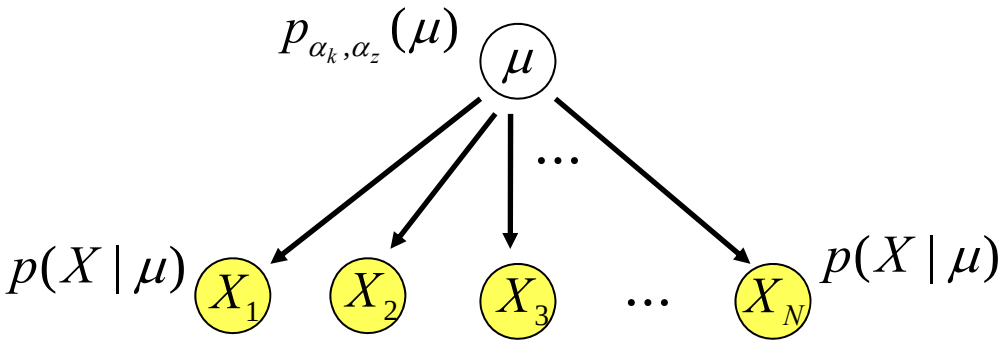
  - Der versteckte Parameter μ koppelt ZV $X_1, ..., X_N$.
  - Aber es gilt $X_N \perp X_1,...,X_{N-1} \mid \mu$

Parameterschätzung als Inferenzproblem

- MAP-Parameterschätzung Münzwurf

$$\hat{\mu} = \arg\max_{\mu} p_{\alpha_k,\alpha_z}(\mu \mid x_1, ..., x_N)$$

- Inferenzproblem:



  - Evidenz auf den Knoten $X_1, ..., X_N$
  - Wahrscheinlichster Zustand des Knotens μ gegeben $X_1, ..., X_N$