MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2017

Prof. E.G. Schukat-Talamazzini

Stand: 25. August 2017

Teil VI

Attributabhängigkeiten: graphische & kausale Modelle

Korrelation Assoziation Dependenz Markovnetze Bayesnetze Inferenz P-Lernen S-Lernen Gaußnetze Σ

Analyse von Attributabhängigkeiten

Dependenzanalyse $\stackrel{?}{=}$ Spaltengruppierung

	$\mathcal{X}_1$	$\mathcal{X}_2$	$\mathcal{X}_3$	$\mathcal{X}_4$
o_1	$+$	*low*	1.2	$+4$
o_2	$-$	*hi*	0.5	-3
o_3	$+$	*hi*	2.3	-3
o_4	$+$	*med*	2.1	-7

Abhängigkeit $\not\equiv$ Ähnlichkeit

- Lineare Abhängigkeiten
  $E = m \cdot c^2$
- Skalenempfindlichkeit
  Temperatur in °C oder °K
- Skalenübergreifend
  $\mathcal{X}_i$ Geschlecht, $\mathcal{X}_j$ Gehalt

Struktur $\not\equiv$ Partition

- keine Äquivalenzrelation
  Zeitreihen, Ortsgitter
- keine binäre Relation
  Alter, Geschlecht, Größe
- Kausalitätsrichtung ?
  $\mathcal{X}_n$ Niederschlag, $\mathcal{X}_m$ Ertrag

Korrelation Assoziation Dependenz Markovnetze Bayesnetze Inferenz P-Lernen S-Lernen Gaußnetze Σ

Analyse von Attributabhängigkeiten

Mit welchem Ziel — zu welchem Zweck ?

Attributwerteprädiktion

Starke Abhängigkeit ➪ $\left\{\begin{array}{l}\text{geringe a posteriori Streuung}\\ \text{kleiner Regressionsfehler}\end{array}\right\}$

- Voraussage · Imputation · Klassifikation

Strukturaufklärung

Lernen des am einfachsten strukturierten Datenmodells (*Occams Razor*)

- Interaktionen · Kausalitäten · Assoziationsregeln

Robuste Datenmodelle

Netzwerk ausgewählter Abhängigkeiten statt **saturierter** W-Modelle

- geringe Kapazität · hohe Effizienz (Zeit/Speicher) · gute Induktivität

Korrelation Assoziation Dependenz Markovnetze Bayesnetze Inferenz P-Lernen S-Lernen Gaußnetze Σ

Statistische Unabhängigkeit von Ereignissen

Paarweise statistische Unabhängigkeit

Faktorisierbarkeit oder (falls $\mathrm{P}(A) \neq 0$) **Neutralität**:

$$A \perp\!\!\!\perp B \quad \Leftrightarrow \quad \mathrm{P}(A,B) = \mathrm{P}(A)\cdot\mathrm{P}(B) \quad \Leftrightarrow \quad \mathrm{P}(B|A) = \mathrm{P}(B)$$

Beispiel: der Wurf zweier fairer Würfel

$A =$ *„gerade Augensumme"* $\quad \mathrm{P}(A,B) = 1/12 = 1/2 \cdot 1/6 = \mathrm{P}(A)\cdot\mathrm{P}(B)$
$B =$ *„erster Wurf ist sechs"* $\quad \mathrm{P}(B,C) = 1/36 = 1/6 \cdot 1/6 = \mathrm{P}(B)\cdot\mathrm{P}(C)$
$C =$ *„Augensumme ist sieben"* $\quad \mathrm{P}(A,C) = 0 \neq 1/2 \cdot 1/6 = \mathrm{P}(A)\cdot\mathrm{P}(C)$

Statistische Unabhängigkeit

der Ereignisse $A_1,\dots,A_I$, falls für alle Indexmengen $\mathcal{I} \subseteq \{1,\dots,I\}$:

$$\mathrm{P}\Big(\bigwedge_{i\in\mathcal{I}} A_i\Big) = \prod_{i\in\mathcal{I}} \mathrm{P}(A_i)$$

Stat. Unabhäng. $\Rightarrow$ paarweise s.U.
Stat. Unabhäng. $\not\Leftarrow$ paarweise s.U.

$A =$ *„erster Wurf hat gerade Augenzahl"*
$B =$ *„zweiter Wurf hat gerade Augenzahl"*
$C =$ *„Augensumme ist ungerade"*

$\mathrm{P}(A,B,C) = 0 \neq 1/8 = \mathrm{P}(A)\cdot\mathrm{P}(B)\cdot\mathrm{P}(C)$

Korrelation Assoziation Dependenz Markovnetze Bayesnetze Inferenz P-Lernen S-Lernen Gaußnetze Σ

Statistische Unabhängigkeit von Zufallsvariablen

Statistische Unabhängigkeit

von Zufallsvariablen $\mathbb{X}_1,\dots,\mathbb{X}_N$, falls für alle $x_1,\dots,x_N \in \mathbb{R}^N$ gilt:

$$\mathrm{P}(\mathbb{X}_1 = x_1,\dots,\mathbb{X}_N = x_N) = \prod_{n=1}^{N} \mathrm{P}(\mathbb{X}_n = x_n)$$

Statistische Unkorreliertheit

von Zufallsvariablen $\mathbb{X}_1,\dots,\mathbb{X}_N$, falls für alle $x_1,\dots,x_N \in \mathbb{R}^N$ gilt:

$$\mathcal{E}\Big[\prod_{n=1}^{N} \mathbb{X}_n\Big] = \prod_{n=1}^{N} \mathcal{E}[\mathbb{X}_n]$$

Bemerkungen

1. Aus der Unabhängigkeit folgt die Unkorreliertheit, aber nicht umgekehrt.
2. Für normalverteilte $\mathbb{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$ gilt: $\mathbb{X}_i$, $\mathbb{X}_j$ korreliert gdw. $\sigma_{ij} \neq 0$.

Beweis.

1. Die uniforme Verteilungsdichte auf dem Träger
$$\{(x,y) \mid |x| + |y| \leq 1\} \subseteq \mathbb{R}^2$$
ist wegen
$$\mathcal{E}[\mathbb{X}\mathbb{Y}] = 0 = \mathcal{E}[\mathbb{X}]\cdot\mathcal{E}[\mathbb{Y}]$$
zwar unkorreliert, aus ihrer (hypothetischen) Unabhängigkeit folgt aber wegen
$$\mathrm{P}(0,\cdot)\cdot\mathrm{P}(\cdot,0) = \mathrm{P}(0,0) = \mathrm{P}(0,1) = \mathrm{P}(0,\cdot)\cdot\mathrm{P}(\cdot,1)$$
und $\mathrm{P}(0,\cdot) \neq 0$ sofort der Widerspruch $\mathrm{P}(\cdot,0) = \mathrm{P}(\cdot,1)$.
2. Es gilt nach Kovarianzdefinition
$$\sigma_{ij} = \mathrm{Cov}[\mathbb{X}_i,\mathbb{X}_j] = \mathcal{E}[\mathbb{X}_i\mathbb{X}_j] - \mathcal{E}[\mathbb{X}_i]\cdot\mathcal{E}[\mathbb{X}_j]\,;$$
daraus folgt die Behauptung — auch für nicht-normal verteilte Variablen.

□

Korrelation und Kovarianz

Definition

Es sei $\omega \subset \mathbb{R}^N$ ein Datensatz mit der (empirischen) Kovarianzmatrix $\boldsymbol{S} = [\sigma_{ij}]$. Die Zahlen

$$\rho_{ij} \stackrel{\text{def}}{=} \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}} \cdot \sqrt{\sigma_{jj}}}\,, \quad 1 \leq i,j \leq N$$

heißen Pearsonsche **Korrelationskoeffizienten** der Attributpaare (x_i, x_j).

Bemerkungen

1. Betragsmäßig kleine (große) Werte σ_{ij} markieren einen schwachen (starken) Zusammenhang zwischen x_i und x_j.
2. Die **Kovarianzen** sind aber extrem skalierungsempfindlich $(\sigma_{ii}, \sigma_{jj})$.
3. Die Korrelationskoeffizienten ρ_{ij} liegen stets im Intervall $[-1, +1]$.
4. Der Wert $\rho_{ij} = 0$ markiert Unkorreliertheit, die Werte $\rho_{ij} \in \{+1, -1\}$ hingegen **deterministische Abhängigkeit** (mit positiver/negativer Steigung).

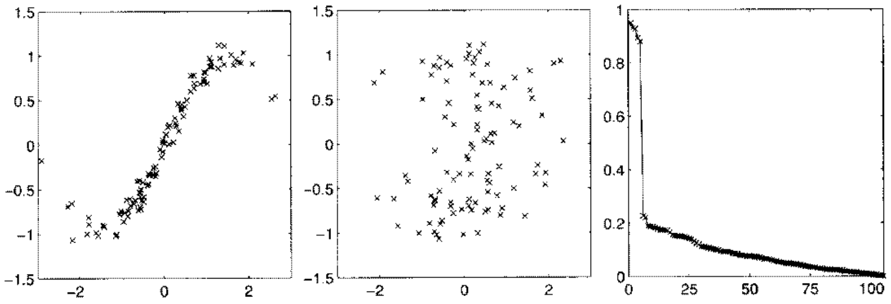
Beispiel — Korrelationsanalyse synthetischer Daten

Zufällig generierte Datenvektoren

$\omega \;=\; \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{100}\} \;\subset\; \mathbb{R}^{15}$ mit Wertetupeln der Zufallsvariablen

$$\mathbb{X}_n \;=\; \begin{cases} \mathcal{N}(0,1) & n = 1, \ldots, 10 \\ \sin(\mathbb{X}_{n-10}) + \mathcal{N}(0, 1/10) & n = 11, \ldots, 15 \end{cases}$$

(10 Kanäle weißes Rauschen & 5 Kanäle verrauschte Sinuskopien)



$\mathbb{X}_{11}$ über $\mathbb{X}_1$ $\quad$ $\mathbb{X}_{11}$ über $\mathbb{X}_2$ $\quad$ alle $|\rho_{ij}|$ sortiert

Korrelationsgruppierung

(Algorithmus)

GEGEBEN:
Daten $\omega \subset \mathbb{R}^N$, Schwelle θ_ρ, *„leere“* Gruppierung $\gamma : i \mapsto \perp$.

1 INITIALISIERUNG
Berechne alle Korrelationskoeffizienten ρ_{ij}.

2 ABSTEIGEND SORTIEREN

$$|\rho_{i_1 j_1}| \;\geq\; |\rho_{i_2 j_2}| \;\geq\; |\rho_{i_3 j_3}| \;\geq\; |\rho_{i_4 j_4}| \;\geq\; \ldots \;\geq\; \ldots \;\geq$$

3 FÜR ALLE $r = 1, 2, \ldots, N(N-1)/2$:

1. Wenn $|\rho_{i_r j_r}| < \theta_\rho$ dann $\rightsquigarrow$ ENDE.
2. Wenn $\gamma(n) \neq \perp$ für alle n dann $\rightsquigarrow$ ENDE.
3. Wenn $\gamma(i_r) = \perp = \gamma(j_r)$ dann erzeuge neue Gruppe $\{i_r, j_r\}$.
4. Wenn $\gamma(i_r) = \perp$ dann setze $\gamma(i_r) \leftarrow \gamma(j_r)$.
5. Wenn $\gamma(j_r) = \perp$ dann setze $\gamma(j_r) \leftarrow \gamma(i_r)$.
6. Wenn $\gamma(i_r) \neq \gamma(j_r)$ dann vereinige die Gruppen: $\gamma(i_r) \cup \gamma(j_r)$.

(Algorithmus)

Korrelationsgruppierung

Was tut dieser Algorithmus ?

Single-linkage Agglomeration — aber:

Terminiert bei Unterschreiten der Korrelationsschwelle.
Terminiert sobald alle Einermengen *„verbraucht“* sind.

Synthesedatenbeispiel

Für die Daten $\omega \subset \mathbb{R}^{15}$ werden in den ersten fünf Schritten die Gruppen

$$\{1, 11\},\; \{2, 12\},\; \{3, 13\},\; \{4, 14\},\; \{5, 15\}$$

gebildet; anschließend gibt es jeweils drei gleichwahrscheinliche Möglichkeiten:

1. Eine der *„alten“* Gruppen wird mit einem neuen Index aufgefüllt.
2. Zwei *„alte“* Gruppen werden vereinigt.
3. Aus zwei *„frischen“* Indizes wird eine neue Gruppe gebildet.

Mit der Ausnahme von 1. sind all diese Optionen höchst **unerwünscht**.

Gestörte (lineare) Abhängigkeit

$\mathbb{Y} = f(\mathbb{X}) + \mathbb{E}$ mit Funktionsprototyp $f : \mathbb{R} \to \mathbb{R}$ und Residuum $\mathbb{E}$

Lemma

Für zwei normalverteilte Zufallsvariablen $\mathbb{X}$, $\mathbb{Y}$ mit

$$\mathbb{Y} = a\mathbb{X} + b + \mathbb{E} , \qquad \mathbb{X} \sim \mathcal{N}(\mu_x, \sigma_x^2) , \qquad \mathbb{E} \sim \mathcal{N}(0, \sigma_e^2)$$

gehorcht $\mathbb{Y}$ der Verteilungsaussage

$$\mathbb{Y} \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2 + \sigma_e^2) .$$

*Die **Kovarianz** und die **Korrelation** zwischen $\mathbb{X}$ und $\mathbb{Y}$ betragen*

$$\sigma_{xy} = a \cdot \sigma_x^2 \quad \textit{bzw.} \quad \rho_{xy} = \operatorname{sign}(a) \cdot \left(1 + \frac{\sigma_e^2}{a^2 \cdot \sigma_x^2}\right)^{-\frac{1}{2}} .$$

Bemerkung

Die Korrelation ρ_{xy} erbt das Vorzeichen von a.
Der Betrag wächst und fällt mit σ_e^{-2} im Einheitsintervall.

Beweis.

Berechnung der Kovarianz (o.B.d.A. ist $\mu_x = 0$):

$$\begin{aligned}
\sigma_{xy} &= \mathrm{Cov}[\mathbb{X}, \mathbb{Y}] = \mathcal{E}[\mathbb{X}\mathbb{Y}] - \mu_x\mu_y \\
&= \mathcal{E}[a\mathbb{X}^2 + b\mathbb{X} + \mathbb{E}\mathbb{X}] - \mu_x\mu_y \\
&= a \cdot (\sigma_x^2 + \mu_x^2) + b\mu_x + \mu_e\mu_x - \mu_x\mu_y \\
&= a\sigma_x^2 + a\mu_x^2 + b\mu_x + \mu_e\mu_x - \mu_x\mu_y \\
&= a\sigma_x^2
\end{aligned}$$

Berechnung der Korrelation:

$$\begin{aligned}
\rho_{xy} &= \frac{a\sigma_x^2}{\sqrt{\sigma_x^2 \cdot (a^2\sigma_x^2 + \sigma_e^2)}} \\
&= \operatorname{sign}(a) \cdot \frac{a\sigma_x^2}{a\sigma_x^2 \cdot \sqrt{1 + \sigma_e^2 \,/\, (a^2\sigma_x^2)}} \\
&= \operatorname{sign}(a) \cdot \left(1 + \frac{\sigma_e^2}{a^2\sigma_x^2}\right)^{-\frac{1}{2}}
\end{aligned}$$

□

Kausalität und Scheinzusammenhang

Verursacht Diät-Cola wirklich Übergewicht ?

Ursache und Wirkung

Korrelation und Abhängigkeit haben keine Vorzugsrichtung:

$$\left\{ \begin{array}{c} \mathbb{X}_i = \text{„}\textit{Körpergewicht [kg]}\text{“} \\ \mathbb{X}_j = \text{„}\textit{Konsum kalorienreduzierter Getränke } [\ell]\text{“} \end{array} \right\}$$

Hohe (positive) Korrelation(en) $\rho_{ij} = \rho_{ji}$ ohne Hinweis auf Kausalrichtung.

Versteckte gemeinsame Ursache oder Lederallergie ?

Das Korrelationsmaß hat keine Vorzugsrichtung:

$$\left\{ \begin{array}{c} \mathbb{X}_i = \text{„}\textit{Kaffee gerührt ?}\text{“} \\ \mathbb{X}_j = \text{„}\textit{Kaffee schmeckt süß ?}\text{“} \\ \dots\ \dots\ \dots \\ \mathbb{X}_k = \text{„}\textit{Zuckerwürfel drin ?}\text{“} \end{array} \right\} \qquad \left\{ \begin{array}{c} \mathbb{X}_i = \text{„}\textit{Lederschuhe an ?}\text{“} \\ \mathbb{X}_j = \text{„}\textit{Kopfschmerzen ?}\text{“} \\ \dots\ \dots\ \dots \\ \mathbb{X}_k = \text{„}\textit{Wagnergasse bis 3h ?}\text{“} \end{array} \right\}$$

Hohe (positive) Korrelation ρ_{ij} ohne jeden (direkten) kausalen Zusammenhang.

(Bi-) Partielle Korrelation

Vergleich nach Subtraktion der Ausgleichsgeraden

Definition

Es seien Zufallsvariablen $\mathbb{X}_1, \dots, \mathbb{X}_N$ gegeben; ferner bezeichne

$$\mathbb{X}_{i|k} = a_{i|k} \cdot \mathbb{X}_k + b_{i|k} , \qquad (i, k \in \{1, \dots, N\},\ i \neq k)$$

den linearen **Quadratmittelprädiktor** für $\mathbb{X}_i$ aus $\mathbb{X}_k$
(„*Ausgleichsgerade*“).

Dann heißt

$$\rho_{ij|k} \stackrel{\text{def}}{=} \mathrm{Corr}[\mathbb{X}_i - \mathbb{X}_{i|k}, \mathbb{X}_j - \mathbb{X}_{j|k}]$$

die **partielle Korrelation** zwischen $\mathbb{X}_i$ und $\mathbb{X}_j$ hinsichtlich $\mathbb{X}_k$ und es heißt

$$\rho_{i|k,j|\ell} \stackrel{\text{def}}{=} \mathrm{Corr}[\mathbb{X}_i - \mathbb{X}_{i|k}, \mathbb{X}_j - \mathbb{X}_{j|\ell}]$$

die **bipartielle Korrelation** zwischen $\mathbb{X}_i$ und $\mathbb{X}_j$ hinsichtlich $\mathbb{X}_k$ und $\mathbb{X}_\ell$.

(Bi-) Partielle Korrelation

Berechnung aus den gewöhnlichen Korrelationskoeffizienten

Lemma

Es seien die Zufallsvariablen $\mathbb{X}_1, \ldots, \mathbb{X}_N$ und ihre Korrelationen ρ_{ij}, $i, j \in \{1, \ldots, N\}$ gegeben.

1. *Die partielle Korrelation zwischen $\mathbb{X}_i$ und $\mathbb{X}_j$ ohne den Einfluß von $\mathbb{X}_k$ hat den Wert*

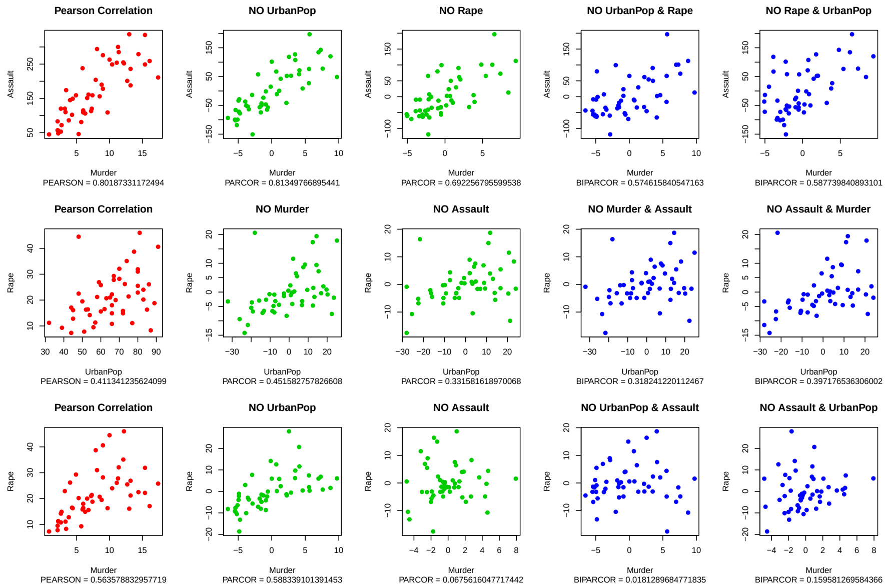
$$\rho_{ij|k} \;=\; \frac{\rho_{ij} - \rho_{ik} \cdot \rho_{jk}}{\sqrt{(1-\rho_{ik}^2) \cdot (1-\rho_{jk}^2)}} \,.$$

2. *Die bipartielle Korrelation zwischen $\mathbb{X}_i$ und $\mathbb{X}_j$ ohne den Einfluß von $\mathbb{X}_k$ bzw. $\mathbb{X}_\ell$ hat den Wert*

$$\rho_{i|k,j|\ell} \;=\; \frac{\rho_{ij} - \rho_{ik}\rho_{jk} - \rho_{i\ell}\rho_{j\ell} + \rho_{i\ell}\rho_{k\ell}\rho_{j\ell}}{\sqrt{(1-\rho_{ik}^2) \cdot (1-\rho_{j\ell}^2)}} \,.$$

Beispiel — U.S. Arrests

Mord/Überfall · Metropol/Vergewaltigung · Mord/Vergewaltigung



Regressionsanalyse

Definition

Eine Familie

$$\Big[f(\cdot|\boldsymbol{a}) : \mathbb{R}^N \to \mathbb{R} \Big]_{\boldsymbol{a} \in \mathcal{M}}$$

von Abbildungen heißt **Funktionsprototyp** der Dimension N; ein Element $f(\cdot|\boldsymbol{a})$ der Familie heißt **Funktionsinstanz** zu $\boldsymbol{a}$.

Für einen Datensatz $\omega \subset \mathbb{R}^N \times \mathbb{R}$ definieren wir den **Regressionsfehler**

$$\varepsilon(f, \boldsymbol{a}, \omega) \;\stackrel{\text{def}}{=}\; \sum_{(\boldsymbol{x},y) \in \omega} (y - f(\boldsymbol{x}, \boldsymbol{a}))^2$$

von $f(\cdot|\boldsymbol{a})$ über ω. Eine Funktionsinstanz $f(\cdot|\boldsymbol{a}^*)$ mit minimalem Regressionsfehler heißt **Regressionfunktion** von ω, ihre Parameter $\boldsymbol{a}^*$ heißen **Regressionsparameter**.

Beispiel — lineare Regression

Die spezielle Familie der $f(\cdot|\boldsymbol{a}) : (x_1, \ldots, x_N) \;\mapsto\; a_0 + \sum_{n=1}^N a_n x_n$ mit $\boldsymbol{a} \in \mathbb{R}^{N+1}$ heißt **affiner** oder — im Fall $a_0 \equiv 0$ — **linearer** Funktionsprototyp.

Beispiel — Ausgleichsgerade

Funktionsprototyp der Dimension $N = 1$ ⇨ Geradengleichungen $y = a + bx$

Regressionsparameter

für einen gegebenen Datensatz $\omega \subset \mathbb{R} \times \mathbb{R}$

$$b \;=\; \frac{\sigma_{xy}}{\sigma_{xx}} \qquad \text{und} \qquad a \;=\; \mu_y - b\mu_x \;=\; \mu_y - \frac{\sigma_{xy}}{\sigma_{xx}} \cdot \mu_x$$

Regressionsfehler

einer Geraden $y = a + bx$ (Verschiebung $\rightsquigarrow$ o.B.d.A. $\mu_x = 0$)

$$\begin{aligned}
\frac{1}{T} \cdot \varepsilon(a, b, \omega) \;&=\; \frac{1}{T} \sum_t (y_t - a - bx_t)^2 \;=\; \ldots \\
&=\; \sigma_{yy} + \mu_y^2 + a^2 + b^2\sigma_{xx} - 2a\mu_y - 2b\sigma_{xy} \\
&=\; \sigma_{yy} \cdot (1 - \rho_{xy}^2) \qquad \text{(Einsetzen } a = \mu_y \text{ und } b = \sigma_{xy}/\sigma_{xx})
\end{aligned}$$

Aufgeklärte Varianz

Die quadrierte Korrelation $\rho_{xy}^2 \in [0,1]$ ist der proportionale Anteil der Varianz σ_{yy} von $\mathbb{Y}$, der durch die ZV $\hat{\mathbb{Y}} = a + b \cdot \mathbb{X}$ aufgeklärt werden konnte.

Lineare und nichtlineare Regression

Kein Fall für Ausgleichsgeraden

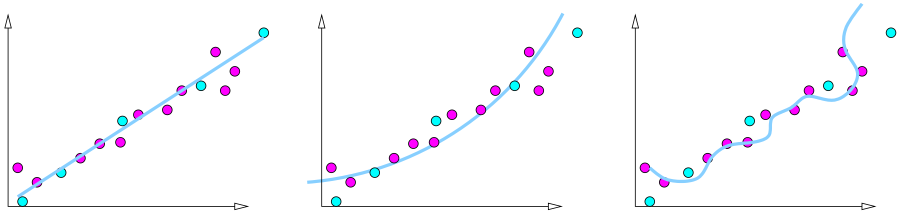
Betrachte die Taylorreihenentwicklung der **sinusoidalen** Abhängigkeit

$$y \;=\; \sin(x+\alpha) \;=\; \underbrace{\sin\alpha + x\cos\alpha}_{\text{linear}} \;-\; x^2\frac{\sin\alpha}{2} \;-\; x^3\frac{\cos\alpha}{6} \;\pm\; \ldots$$

Ausgleichspolynome

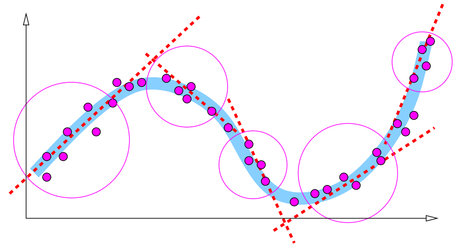
Affiner Regressionsansatz mit Termexpansion, z.B. polynomial für $N=3$:

$$(x_1,x_2,x_3) \;\mapsto\; (1,x_1,x_2,x_3,x_1^2,x_2^2,x_3^2,x_1x_2,x_1x_3,x_2x_3,\ldots)$$



robuste Ausgleichsgerade — leichte Überanpassung — starke Überanpassung

Lokale Regression



Verzögertes Lernen

- lokales Modell *„just in time"*
- kein globales Modell

$$f(\boldsymbol{x}_t|\boldsymbol{a}^*) \approx y_t \;(\forall t)$$

(Algorithmus)

GEGEBEN:
Lerndatenprobe $\omega = [(\boldsymbol{x}_t, y_t)]_1^T \subset \mathbb{R}^N \times \mathbb{R}$ und Eingabevektor $\boldsymbol{z} \in \mathbb{R}^N$

1. NACHBARSCHAFT FIXIEREN
   Berechne Nachbarschaftsmenge $\omega^{(z)} \subset \omega$, eventuell mit Gewichten $\{w_t\}_1^T$.
2. LOKALE AUSGLEICHSRECHNUNG
   Schätze lokale Regressionsfunktion $f(\cdot|\boldsymbol{a}^{(z)})$ für den $\omega^{(z)}$-Datensatz.
3. VORHERSAGE TREFFEN
   Setze $\hat{y}(\boldsymbol{z}) := f(\boldsymbol{z}|\boldsymbol{a}^{(z)})$.

Lokale Regression

Eine Frage der guten Nachbarschaft

Nächster-Nachbar-Regel

Belegmenge $\omega^{(x)} = \{\boldsymbol{x}_s\}$ ist einelementig.

$$\varepsilon(f,\boldsymbol{a},\omega \mid \boldsymbol{x}) \overset{\text{def}}{=} (y_s - f(\boldsymbol{x}_s|\boldsymbol{a}))^2 \;, \quad s = \underset{t=1..T}{\operatorname{argmin}}\, d(\boldsymbol{x},\boldsymbol{x}_t)$$

k-Nächste-Nachbarn-Regel

Scharfe Belegmenge $\omega^{(x)}$ mit genau k Elementen.

$$\varepsilon(f,\boldsymbol{a},\omega \mid \boldsymbol{x}) \overset{\text{def}}{=} \sum_{i=1}^{k} (y_{s_i} - f(\boldsymbol{x}_{s_i}|\boldsymbol{a}))^2$$

Gewichtete Mittelung

Unscharfe Belegmenge $\omega^{(x)}$ mit T Elementen.

$$\varepsilon(f,\boldsymbol{a},\omega \mid \boldsymbol{x}) \overset{\text{def}}{=} \sum_{t=1}^{T} w_t\cdot(y_t - f(\boldsymbol{x}_t|\boldsymbol{a}))^2 \;, \quad w_t \propto \exp\left\{-\frac{1}{2\sigma^2}\cdot\|\boldsymbol{x}-\boldsymbol{x}_t\|^2\right\}$$

Lokale Regression

Konstante Funktionsprototypen · Disjunkte Nachbarschaften

Konstanter Funktionsprototyp

$$f(\cdot|a):\begin{cases}\mathbb{R}^N & \to \;\mathbb{R}\\ \boldsymbol{x} & \mapsto \;a\end{cases}\;, \qquad a \in \mathbb{R}$$

NN-Regel	k-NN-Regel	Distanzgewichte
$f_n(\boldsymbol{x}) = y_{t(x)}$	$f_k(\boldsymbol{x}) = \frac{1}{k}\sum_{i=1}^{k} y_{t_i(x)}$	$f_g(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{y} \,/\, \|\boldsymbol{w}\|_1$
„Kopie"	*„Ortsmittel"*	*„Schwerpunkt"*

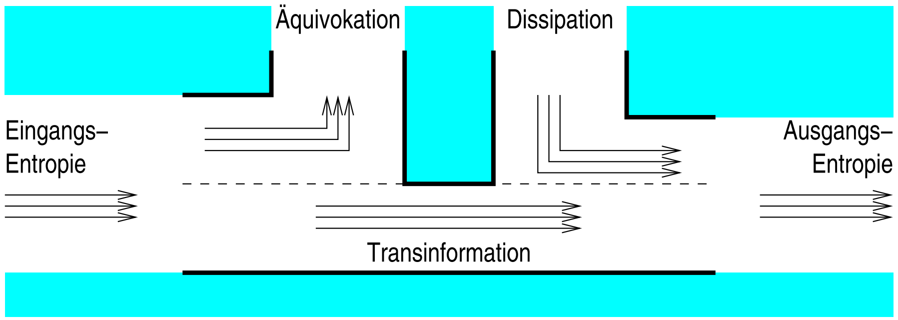
Stückweise lineare Regression

(Algorithmus)

1. GRUPPIERUNG
   Lerne extensionale Partition $\omega_1,\ldots,\omega_K$ von $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T$ (K-means).
2. STÜCKWEISE REGRESSION
   Lerne lokale Regressionsfunktionen $f(\cdot|\boldsymbol{a}_1), f(\cdot|\boldsymbol{a}_2),\ldots,f(\cdot|\boldsymbol{a}_K)$.
3. VORHERSAGEPHASE
   · Bestimme zu $\boldsymbol{x} \in \mathbb{R}^N$ den Gruppenindex λ, also mit $\boldsymbol{x} \in \boldsymbol{\Omega}_\lambda \supset \omega_\lambda$.
   · Berechne den Vorhersagewert $\hat{y}(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{a}_\lambda)$.

Informationstheorie

Der gedächtnislose Informationskanal — Claude Shannon, 1949



Der Informationskanal

ist durch die gemeinsame Verteilung $f_{xy}(\cdot,\cdot)$ seiner **Eingangsvariablen** $\mathbb{X}$ und seiner **Ausgangsvariablen** $\mathbb{Y}$ charakterisiert.

Kanalentropien

Eingangsentropie $\mathcal{H}(\mathbb{X}) = \mathcal{E}[-\log f_x(\mathbb{X})]$

Ausgangsentropie $\mathcal{H}(\mathbb{Y}) = \mathcal{E}[-\log f_y(\mathbb{Y})]$

Gesamtentropie $\mathcal{H}(\mathbb{XY}) = \mathcal{E}[-\log f_{xy}(\mathbb{X},\mathbb{Y})]$

Transinformation normalverteilter Attribute

Lemma

Für die (differentiellen) Entropien und die Transinformation normalverteilter Zufallsvariablen gelten die nachfolgenden Aussagen:

1. *Wenn* $\mathbb{X} \sim \mathcal{N}(\mu,\sigma^2)$, *so gilt:*

$$\mathcal{H}(\mathbb{X}) = \frac{1}{2}\cdot(\log\sigma^2 + 1 + \log(2\pi))$$

2. *Wenn* $(\mathbb{X}_1,\ldots,\mathbb{X}_N) \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{S})$, *so gilt:*

$$\begin{aligned}\mathcal{H}(\mathbb{X}_1\ldots\mathbb{X}_N) &= \frac{1}{2}\cdot(\log\det(\boldsymbol{S}) + N + N\log(2\pi))\\ \mathcal{H}(\mathbb{X}_i\mathbb{X}_j) &= \frac{1}{2}\cdot\log(\sigma_{ii}\cdot\sigma_{jj} - \sigma_{ij}^2) + 1 + \log(2\pi)\end{aligned}$$

3. *Für jedes bivariat normale Variablenpaar* $(\mathbb{X}_i,\mathbb{X}_j)$ *gilt:*

$$\Im(\mathbb{X}_i;\mathbb{X}_j) = -\frac{1}{2}\cdot\log\left(1-\rho_{ij}^2\right)$$

Bedingte Kanalentropien

Was Sie schon immer über Entropien wissen wollten, aber noch nie zu fragen wagten

Definition

Der Informationskanal sei durch f_{xy} charakterisiert.

- $\mathcal{H}(\mathbb{X}|\mathbb{Y}) = \mathcal{E}[-\log f_{x|y}(\mathbb{X}|\mathbb{Y})]$ heißt **Äquivokation** des Kanals.
- $\mathcal{H}(\mathbb{Y}|\mathbb{X}) = \mathcal{E}[-\log f_{y|x}(\mathbb{Y}|\mathbb{X})]$ heißt **Dissipation** des Kanals.
- $\Im(\mathbb{X};\mathbb{Y}) = \mathcal{E}[-\log \frac{f_x(\mathbb{X})\cdot f_y(\mathbb{Y})}{f_{xy}(\mathbb{X},\mathbb{Y})}]$ heißt **Transinformation** des Kanals.

Lemma

In einem gedächtnislosen Informationskanal gelten die Aussagen:

1. $\mathcal{H}(\mathbb{X}|\mathbb{Y}) = \mathcal{H}(\mathbb{XY}) - \mathcal{H}(\mathbb{Y})$
2. $\mathcal{H}(\mathbb{Y}|\mathbb{X}) = \mathcal{H}(\mathbb{XY}) - \mathcal{H}(\mathbb{X})$
3. $\Im(\mathbb{X};\mathbb{Y}) = \mathcal{H}(\mathbb{X}) + \mathcal{H}(\mathbb{Y}) - \mathcal{H}(\mathbb{XY})$
4. $\Im(\mathbb{X};\mathbb{Y}) = \mathcal{D}(f_{xy}\|f_x\cdot f_y)$

Divergenz

(Kullback-Leibler)

$$\mathcal{D}(f\|g) = \mathcal{E}_f[\log {}^f\!/_g]$$

Beweis.

1. **Univariater Fall**:

$$\begin{aligned}\mathcal{H}(\mathbb{X}) &= \mathcal{E}[-\log\mathcal{N}(\mathbb{X}\mid\mu,\sigma^2)] = \mathcal{E}[\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\left(\frac{\mathbb{X}-\mu}{\sigma}\right)^2]\\ &= \mathcal{E}[\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\tilde{\mathbb{X}}^2] = \frac{1}{2}\cdot(\log\sigma^2 + 1 + \log(2\pi))\end{aligned}$$

Beachte, daß $\tilde{\mathbb{X}} = {}^{(\mathbb{X}-\mu)}\!/_{\sigma}$ standardnormalverteilt ist, d.h. $\tilde{\mathbb{X}} \sim \mathcal{N}(0,1)$.

2. **Multivariater Fall**:

$$\begin{aligned}\mathcal{H}(\mathbb{X}) &= \mathcal{E}[-\log\mathcal{N}(\mathbb{X}\mid\boldsymbol{\mu},\boldsymbol{S})] = \mathcal{E}[\frac{1}{2}\log\det(2\pi\boldsymbol{S}) + \frac{1}{2}\cdot(\mathbb{X}-\boldsymbol{\mu})^\top\boldsymbol{S}^{-1}(\mathbb{X}-\boldsymbol{\mu})]\\ &= \mathcal{E}[\frac{1}{2}\log\det(2\pi\boldsymbol{S}) + \frac{1}{2}\cdot\tilde{\mathbb{X}}^\top\tilde{\mathbb{X}}] = \frac{1}{2}\cdot(\log\det(\boldsymbol{S}) + N + N\log(2\pi))\end{aligned}$$

Bivariater Fall: gilt wegen $\det\begin{pmatrix}\sigma_{ii} & \sigma_{ij}\\ \sigma_{ji} & \sigma_{jj}\end{pmatrix} = \sigma_{ii}\sigma_{jj} - \sigma_{ij}^2$.

3. **Transinformationen**:

$$\begin{aligned}\Im(\mathbb{X}_i;\mathbb{X}_j) &= \mathcal{H}(\mathbb{X}_i) + \mathcal{H}(\mathbb{X}_j) - \mathcal{H}(\mathbb{X}_i\mathbb{X}_j)\\ &= +\frac{1}{2}\cdot\log\left(\frac{\sigma_{ii}\sigma_{jj}}{\sigma_{ii}\sigma_{jj}-\sigma_{ij}^2}\right) = -\frac{1}{2}\cdot\log\left(1-\rho_{ij}^2\right)\end{aligned}$$

$(1-\rho_{ij}^2)$ ist der Anteil **unaufgeklärter** Varianz.

Transinformation diskreter Attribute

Wertebereiche und Verteilung

Es sei $\mathbb{X} \in \{\xi_1, \dots, \xi_K\}$ und $\mathbb{Y} \in \{\eta_1, \dots, \eta_L\}$ verteilt gemäß

$$p_{k\ell} \;=\; \mathrm{P}(\mathbb{X} = \xi_k, \mathbb{Y} = \eta_\ell)$$

Marginale und gemeinsame Entropien

$$\left.\begin{aligned}
\mathcal{H}(\mathbb{X}\mathbb{Y}) &= -\sum_k \sum_\ell p_{k\ell} \cdot \log p_{k\ell} \\
\mathcal{H}(\mathbb{X}) &= -\sum_k \left(\sum_\ell p_{k\ell}\right) \cdot \log \left(\sum_\ell p_{k\ell}\right) \\
\mathcal{H}(\mathbb{Y}) &= -\sum_\ell \left(\sum_k p_{k\ell}\right) \cdot \log \left(\sum_k p_{k\ell}\right)
\end{aligned}\right\} \Rightarrow \underbrace{-\sum_{k,\ell} p_{k\ell} \cdot \log \frac{p_{k\cdot} \cdot p_{\cdot\ell}}{p_{k\ell}}}_{\textbf{Transinformation } \Im(\mathbb{X};\mathbb{Y})}$$

Transinformation gemischtskaliger Attribute

$\mathbb{X} \in \mathbb{R}$ und $\mathbb{Y} \in \{\eta_1, \dots, \eta_L\}$

Punktweise Transinformation

$$\Im(\mathbb{X};\mathbb{Y}) = \int_x \sum_y f(x,y) \cdot \Im(x;y)$$

Die „*mutual information*" zwischen korrespondierenden Werten x und y:

$$\log \frac{f(x|y)}{f(x)} \;=\; \underbrace{\log \frac{f(x,y)}{f(x) \cdot f(y)}}_{\Im(x;y)} \;=\; \log \frac{f(y|x)}{f(y)}$$

Faktor diskret

Gaußsche Mischverteilung

$$f(x,\eta_\ell) \;=\; \pi_\ell \cdot \mathcal{N}(x \mid \mu_\ell, \sigma_\ell^2)$$

Schätzformel

$$\sum_{t=1}^{T} \frac{1}{T} \cdot \log \frac{\mathcal{N}(x_t \mid \mu_{\ell(t)}, \sigma_{\ell(t)}^2)}{\sum_\ell \pi_\ell \cdot \mathcal{N}(x_t \mid \mu_\ell, \sigma_\ell^2)}$$

Faktor stetig

Diskriminantverteilung

$$f(x,\eta_\ell) \;=\; f_{\mathbb{X}}(x) \cdot p(\eta_\ell|x)$$

Schätzformel

$$\sum_{t=1}^{T} \frac{1}{T} \cdot \log \frac{p(\eta_{\ell(t)}|x_t)}{\pi_{\ell(t)}}$$

Assoziationsanalyse

Agrawal (SIGMOD Conference 1993) — mehr als 6.000× zitiert!

Warenkorbdaten

Objekte = qualitative **Stücklisten** $\rightsquigarrow\ \mathbf{\Omega} = \mathfrak{P}(\mathfrak{A})$

$$\omega \;\subset\; \mathbf{\Omega} \;=\; \{0,1\}^N$$

über einem globalem **Artikelinventar** $\mathfrak{A} = \{\mathfrak{a}_1, \dots, \mathfrak{a}_N\}$

Assoziationsregeln

„*Wer alle Produkte aus A kauft, der kauft auch alle Produkte aus B.*"

$$\texttt{IF } A \texttt{ THEN } B \quad , \quad A, B \in \mathbf{\Omega} \;, \quad A \cap B = \varnothing$$

Beispielregeln

IF {Windeln} THEN {Bier}

IF {Brot, Butter} THEN {Milch}

IF {Rosen, Wein, Goldbären} THEN {Kondome}

Bemerkungen

1. Warenkorbdaten haben **binäre Attribute**.
2. Assoziationsregeln formulieren **multiple Abhängigkeiten**.

Gute und schlechte Regeln

Abdeckungs- und Geltungsgrad einer Regel · Signifikanz ihrer Prämisse

Definition

Es sei $\omega \subset \mathbf{\Omega}$ ein Datensatz, $A, B \in \mathbf{\Omega}$ zwei Stücklisten und `IF` A `THEN` B (kürzer: $A \to B$) eine Assoziationsregel. Die Größe

$$\text{supp}(A \to B) \stackrel{\text{def}}{=} \text{supp}(A \cup B)\ , \qquad \text{supp}(A) \stackrel{\text{def}}{=} \frac{|\{\boldsymbol{x} \in \omega \mid \boldsymbol{x} \supseteq A\}|}{|\omega|}$$

heißt **Support**, $\leadsto \hat{\mathrm{P}}(A \cup B)$

$$\text{conf}(A \to B) \stackrel{\text{def}}{=} \frac{\text{supp}(A \cup B)}{\text{supp}(A)}$$

heißt **Konfidenz** und $\leadsto \hat{\mathrm{P}}(B|A)$

$$\text{lift}(A \to B) \stackrel{\text{def}}{=} \frac{\text{supp}(A \cup B)}{\text{supp}(A) \cdot \text{supp}(B)}$$

heißt **Relevanz** der Assoziation $A \to B$. $\leadsto \dfrac{\hat{\mathrm{P}}(B|A)}{\hat{\mathrm{P}}(B)}$

Extraktion nützlicher Assoziationsregeln

Eine Frage des Aufwandes

Aufgabenstellung

Gesucht ist — bei gegebenen Warenkorbdaten — die Teilmenge solcher Regeln $A \to B$

- mit **signifikantem Abdeckungsgrad** $\text{supp}(A \to B) \geq \theta_s$
- und hohem **Geltungsgrad** $\text{conf}(A \to B) \geq \theta_c$
- und erheblicher **Aussagekraft**. $\text{lift}(A \to B) \geq \theta_r$

Problem

Es gibt 2^N kombinatorisch mögliche Stücklisten und es gibt 3^N mögliche Assoziationsregeln ($N = |\mathfrak{A}|$).

Lösungsansatz

Es gilt die **Antitonie**

$$A \supseteq B \ \Rightarrow\ \text{supp}(A) \leq \text{supp}(B)$$

und es gibt nur N **einelementige** Stücklisten.

Apriori-Basisalgorithmus

Schichtenweise Stücklisten- und Regelgenerierung

⟨Algorithmus⟩

GEGEBEN
Warenkorbdaten ω, Stückzahlgrenze N^*, Schwellen θ_s, θ_c, θ_r.

1 INITIALISIERUNG

$$\mathcal{M}_1 \leftarrow \{\{i\} \mid \text{supp}(\{i\}) \geq \theta_s\}\ , \qquad \mathcal{R} = \varnothing$$

2 SCHICHTEXPANSION $(n = 2, \ldots, N^*)$
Erzeuge alle

$$A = B \cup \{i\} \quad \text{mit } B \in \mathcal{M}_{n-1},\ \{i\} \in \mathcal{M}_1 \text{ und } i \notin B.$$

Bringe A nach $\mathcal{M}_n$ falls $\text{supp}(A) \geq \theta_s$.

3 REGELERZEUGUNG
Für alle $C \in \mathcal{M} = \bigcup_{n=1}^{N^*} \mathcal{M}_n$:

Für alle Artikel $j \in C$ teste

$$\text{conf}(C \setminus \{j\} \to \{j\}) \geq \theta_c \quad \wedge \quad \text{lift}(C \setminus \{j\} \to \{j\}) \geq \theta_r$$

und verbringe die Regel im Erfolgsfall nach $\mathcal{R}$.

⟨Algorithmus⟩

Regelformat und Artikelbeschreibung

Assoziationsregeln mit multipler Conclusio

Statt einfacher Regeln $C \setminus \{j\} \to \{j\}$ produziere

$$A \to B \qquad \text{mit } A \cap B = \varnothing \text{ und } A \cup B = C.$$

Stufenweise Erzeugung („*bottom-up*") unter Verwendung der Monotonie:

$$B_1 \subseteq B_2 \ \Rightarrow\ \begin{cases} \text{supp}(A_1 \to B_1) = \text{supp}(A_2 \to B_2) \\ \text{conf}(A_1 \to B_1) \geq \text{conf}(A_2 \to B_2) \\ \text{lift}(A_1 \to B_1) \stackrel{?}{\geq} \text{lift}(A_2 \to B_2) \end{cases}$$

Aufwandsreduktion durch onthologische Gliederung

Artikeleinträge werden durch ihre Verallgemeinerungen aufgestockt.

Assoziationsanalyse für (mehrwertige) Nominalskalen

Verallgemeinerte Stücklisten

Listen von kontradiktionsfreien Attribut-Wert-Paaren:

$$(windy = false, play = no, outlook = sunny, humidity = high)$$

Es gibt $\prod_n (L_n + 1)$ Stücklisten und $\prod_n (2 \cdot L_n + 1)$ Assoziationsregeln.

Beispielregeln

Klassisch:

- IF $\{Spaghetti\}$ THEN $\{Rotwein, Tomaten, Basilikum\}$
- IF $\{Waits, Dylan, Bush\}$ THEN $\{Spektor\}$

Mehrwertig:

- IF $\{humidity = high, windy = false\}$ THEN $\{outlook = sunny\}$

Zweiwertig:

- IF $\{Pommes, \neg\, Ketchup\}$ THEN $\{Mayonnaise\}$
- IF $\{E.Jelinek, Ch.Roche\}$ THEN $\{\neg\, U.Danella\}$

Beispiel — Tennisdaten mit WEKA

5 Attribute · 14 Objekte · Apriori mit $\theta_s = 15\%$, $\theta_c = 90\%$

Stücklistenaufstellung („*itemsets*“)

12 Einermengen · 47 Paare · 39 Tripel · 6 Quadrupel

Beste Assoziationsregeln ($\theta_c \equiv 100\%$)

4 IF $\{humidity = normal, windy = false\}$ THEN $\{play = yes\}$
IF $\{temperature = cool\}$ THEN $\{humidity = normal\}$
IF $\{outlook = overcast\}$ THEN $\{play = yes\}$

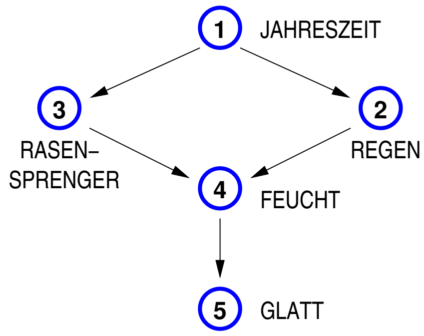
3 IF $\{temperature = cool, play = yes\}$ THEN $\{humidity = normal\}$
IF $\{outlook = rainy, windy = false\}$ THEN $\{play = yes\}$
IF $\{outlook = rainy, play = yes\}$ THEN $\{windy = false\}$
IF $\{outlook = sunny, humidity = high\}$ THEN $\{play = no\}$
IF $\{outlook = sunny, play = no\}$ THEN $\{humidity = high\}$

2 IF $\{temp = cool, windy = false\}$ THEN $\{humidity = normal, play = yes\}$
IF $\{temp = cool, humidity = normal, windy = false\}$ THEN $\{play = yes\}$

Graphische Wahrscheinlichkeitsmodelle

Regen ← Jahreszeit

	$\mathbb{X}_1$ w	$\mathbb{X}_1$ f	$\mathbb{X}_1$ s	$\mathbb{X}_1$ h
$\mathbb{X}_2 = 0$	0.2	0.3	0.1	0.7
$\mathbb{X}_2 = 1$	0.8	0.7	0.9	0.3



Feucht ← Regen, Sprenger

	$\mathbb{X}_2\mathbb{X}_3$	$\bar{\mathbb{X}}_2\mathbb{X}_3$	$\mathbb{X}_2\bar{\mathbb{X}}_3$	$\bar{\mathbb{X}}_2\bar{\mathbb{X}}_3$
$\mathbb{X}_4 = 0$	0.1	0.3	0.4	0.8
$\mathbb{X}_4 = 1$	0.9	0.7	0.6	0.2

Wozu Graphische Modelle ?

- Visualisierung quantitativer Zusammenhänge
- Inferenz von Abhängigkeitsbeziehungen
- Berechnung kausaler Effekte
- Effiziente Auswertung multivariater Modelle

Jahreszeit ←

$\mathbb{X}_1 = w$	0.25
$\mathbb{X}_1 = f$	0.25
$\mathbb{X}_1 = s$	0.25
$\mathbb{X}_1 = h$	0.25

Glatt ← Feucht

	$\mathbb{X}_4$	$\bar{\mathbb{X}}_4$
$\mathbb{X}_5 = 0$	0.3	0.9
$\mathbb{X}_5 = 1$	0.7	0.1

Wahrscheinlichkeit und Graphstruktur

Datensatz

Wahrscheinlichkeits(dichte)werte

$$\mathrm{P} : \boldsymbol{\Omega} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_N \ \rightarrow \ \mathbb{R}_0^+$$

durch **Statistiken** des Datensatzes $\omega \subset \boldsymbol{\Omega}$ repräsentiert.

Modellformel

Saturiertes Modell $\prod_n L_n - 1 = 63$
Naives Modell $\sum_n L_n - N = 7$
Faktorisierung $3 + 4 + 4 + 4 + 2 = 17$

$$\mathrm{P}(\boldsymbol{x}) \ = \ \mathrm{P}(x_1) \cdot \mathrm{P}(x_2|x_1) \cdot \mathrm{P}(x_3|x_1) \cdot \mathrm{P}(x_4|x_2, x_3) \cdot \mathrm{P}(x_5|x_4)$$

Dependenzmodell

Menge aller **bedingten** Unabhängigkeiten zwischen **Mengen von** Zufallsvariablen

$$\Im(\mathbb{X}_2 \mid \mathbb{X}_1 \mid \mathbb{X}_3)$$

(*„Regen“* unabhängig von *„Rasensprenger“*, wenn *„Jahreszeit“* gegeben)

Graphisches Modell

- **Markovnetz**
  ungerichteter Graph
  „partielle Unabhängigkeit“
  $\mathbb{X}_i \not\leftrightarrow \mathbb{X}_j$
- **Bayesnetz**
  gerichteter azyklischer Graph
  „kausale Abhängigkeit“
  $\mathbb{X}_i \rightarrow \mathbb{X}_j$

Simpsons Paradoxon #1

Geschlechtsspezifische Diskriminierung an der Universität

Zweiwegetabelle: Geschlecht & Zulassungsquote

Geschlecht	*#Bewerbung*	*#Zulassung*	*%*
M	*600*	*350*	*58.3*
F	*600*	*250*	*41.6*

Frauen haben die geringeren Zulassungschancen!

„marginal abhängig“

$\Downarrow \leadsto$

„bedingt abhängig“

Zusatzvariable: Fakultätszugehörigkeit

Fakultät	*Geschlecht*	*#Bewerbung*	*#Zulassung*	*%*
TECH	*M*	*100*	*25*	*25*
TECH	*F*	*300*	*75*	*25*
PHIL	*M*	*200*	*100*	*50*
PHIL	*F*	*200*	*100*	*50*
THEO	*M*	*300*	*225*	*75*
THEO	*F*	*100*	*75*	*75*

Männer tendieren zu Fakultäten mit hoher Zulassungsquote!

Simpsons Paradoxon #2

Eine farbenfrohe Mordstatistik für den Bundesstaat Florida

Zweiwegetabelle: Hautfarbe & Strafmaß

Farbe$_{Mörder}$	*#Todesurteil*	*#Haftstrafe*	*% T.U.*
schwarz	*17*	*149*	*11.4*
weiß	*19*	*141*	*12.5*

kein Rassismus: ähnliche Todesurteilquote für Schwarz und Weiß

„marginal unabhängig“

$\Downarrow \leadsto$

„bedingt unabhängig“

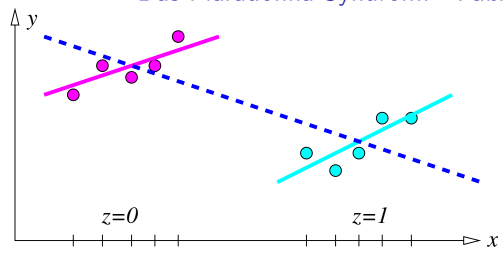
Zusatzvariable: Hautfarbe des Opfers

Farbe$_{Opfer}$	*Farbe$_{Mörder}$*	*#Tod*	*#Haft*	*% T.U.*
schwarz	*schwarz*	*6*	*97*	*5.8*
schwarz	*weiß*	*0*	*9*	*0.0*
weiß	*schwarz*	*11*	*52*	*17.5*
weiß	*weiß*	*19*	*132*	*12.6*

Der Mord an einem weißen Mitbürger kommt teurer zu stehen!

Simpsons Paradoxon #3

Das Maradonna-Syndrom: Fußballspielen ist ungesund!



Drei Attribute

$\mathbb{X}$ = *„Fußballaktivität“*
$\mathbb{Y}$ = *„Lebenserwartung“*
$\mathbb{Z}$ = *„Geschlecht“*

Bedingte Abhängigkeit

Weibliche• wie männliche• Regressionsgeraden

$$\mathbb{Y} = f(\mathbb{X}|0) \quad \text{bzw.} \quad \mathbb{Y} = f(\mathbb{X}|0)$$

besitzen **positive** Steigung.

Marginale Abhängigkeit

Geschlechtsneutrale Regressionsgerade

$$\mathbb{Y} = f(\mathbb{X})$$

besitzt **negative** Steigung.

Grund:

Frauen sind tendenziell **langlebig** und stehen eher auf **Volleyball+Ayurveda**.

Bedingte statistische Unabhängigkeit

von Mengen von Zufallsvariablen

Definition

Es seien A, B, Z drei paarweise disjunkte Teilmengen der Zufallsvariablen $\{\mathbb{X}_1, \ldots, \mathbb{X}_N\}$. Dann heißt A **bedingt statistisch unabhängig** von B bezogen auf Z genau dann, wenn gilt

$$\mathrm{P}(A \mid B, Z) \;=\; \mathrm{P}(A \mid Z)$$

und wir schreiben

$$\Im(A \mid Z \mid B)\ .$$

Ferner heißen A und B **bedingt faktorisierbar** bezogen auf Z, falls es zwei geeignete Funktionen (sic!) f und g gibt mit

$$\mathrm{P}(A, B, Z) \;=\; f(A, Z) \cdot g(B, Z)\ .$$

Marginale statistische Unabhängigkeit

Der Spezialfall *„gewöhnlicher"* statistischer Unabhängigkeit ergibt sich für UA-Postulate der Form $\Im(A \mid Z \mid B)$ mit $Z = \varnothing$.

Beweis.

- (1)⇒(2)⇒(3)⇒(1)

$$\mathrm{P}(b \mid a, z) \;=\; \frac{\mathrm{P}(a, b \mid z)}{\mathrm{P}(a|z)} \;=\; \frac{\mathrm{P}(a \mid b, z) \cdot \mathrm{P}(b|z)}{\mathrm{P}(a|z)} \;=\; \frac{\mathrm{P}(a|z) \cdot \mathrm{P}(b|z)}{\mathrm{P}(a|z)} \;=\; \mathrm{P}(b|z)$$

- (1)⇒(2)⇒(3)⇒(1)

$$\mathrm{P}(a, b, z) \;=\; \mathrm{P}(a, z) \cdot \mathrm{P}(b \mid a, z) \;=\; \mathrm{P}(a, z) \cdot \mathrm{P}(b|z) \;=:\; f(a, z) \cdot g(b, z)$$

- (1)⇒(2)⇒(3)⇒(1)

$$\mathrm{P}(a \mid b, z) \;=\; \frac{\mathrm{P}(a, b, z)}{\mathrm{P}(b, z)} \;=\; \frac{\mathrm{P}(a, b, z)}{\sum_a \mathrm{P}(a, b, z)} \;=\; \frac{f(a, z) \cdot g(b, z)}{\sum_a f(a, z) \cdot g(b, z)} \;=\; \frac{f(a, z)}{\sum_a f(a, z)}$$

Der letzte Ausdruck ist offenbar unabhängig von b.

□

Rechenregeln für bedingte Unabhängigkeiten

Lemma

Die folgenden Allaussagen über die Werte a, b und z dreier Zufallsvariablen $\mathbb{X}_a, \mathbb{X}_b, \mathbb{X}_z$ sind äquivalent:

1. $\mathrm{P}(a \mid b, z) \;=\; \mathrm{P}(a|z)$ *($a \nsim b$ wenn z bekannt)*
2. $\mathrm{P}(b \mid a, z) \;=\; \mathrm{P}(b|z)$ *($b \nsim a$ wenn z bekannt)*
3. $\mathrm{P}(a, b, z) \;=\; f(a, z) \cdot g(b, z)$ *(Faktorisierbarkeit)*

*Diese Äquivalenz gilt entsprechend für **Mengen** von Zufallsvariablen.*

Weitere äquivalente Formulierungen

für die bedingte statistische Unabhängigkeit zwischen drei Zufallsvariablen:

1. $\mathrm{P}(a, b, z) \;=\; \frac{\mathrm{P}(a,z) \cdot \mathrm{P}(b,z)}{\mathrm{P}(z)}$
2. $\mathrm{P}(a, b \mid z) \;=\; \mathrm{P}(a|z) \cdot \mathrm{P}(b|z)$
3. $\mathrm{P}(a, b, z) \;=\; \mathrm{P}(a|z) \cdot \mathrm{P}(b, z)$

Diskrete ZV

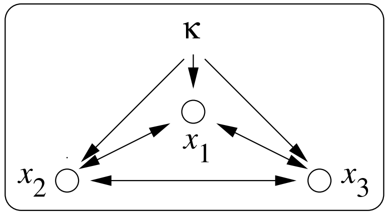
$\mathrm{P}(\mathbb{Y} = y \mid \mathbb{I} = i)$ ist konstant bzgl. i.

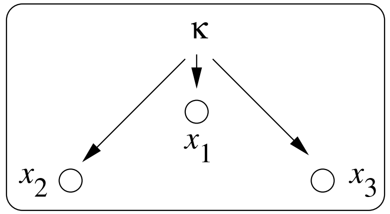
Stetige ZV

Lineare Regression $\mathbb{Y}|x \sim \mathcal{N}(a + bx, \sigma^2)$ mit $b = 0$.

Beispiel — Numerische Klassifikation

Normale und naive Bayesregel





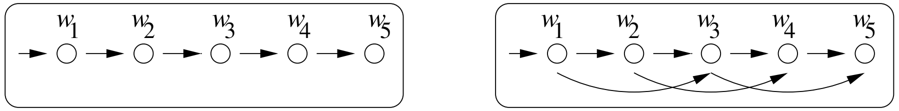
Datenerzeugungsmodell

für Merkmale $x_1, \ldots, x_N \in \mathbb{R}$ und Klassenvariable $y \in \{\Omega_1, \ldots, \Omega_K\}$:

$$f(\boldsymbol{x}, \Omega_\kappa) \;=\; \mathrm{P}(\Omega_\kappa) \cdot f(\boldsymbol{x}|\Omega_\kappa)$$

- **Multivariate Normalverteilungsdichte** (saturiertes Modell): $x_i \leftarrow \{x_j \mid j \neq i\}$ für alle i
- **Klassenbedingte Unabhängigkeit** (ausgedünntes Modell): $f(\boldsymbol{x}|\Omega_\kappa) = \prod_i \mathcal{N}(x_i \mid \mu_i, \sigma_i^2)$ ergibt $x_i \leftarrow \varnothing$ für alle i

Beispiel — N-Gramm-Grammatiken



Datenerzeugungsmodell

für eine Symbolfolge (Wortfolge) $\boldsymbol{w} = w_1 \dots w_M$ ist die **Kettenregel**

$$\mathrm{P}(\boldsymbol{w}) = \prod_{m=1}^{M} \mathrm{P}(w_m \mid w_1, \dots, w_{m-1}) \simeq \prod_{m=1}^{M} \mathrm{P}(w_m \mid w_{m-2}, w_{m-1})$$

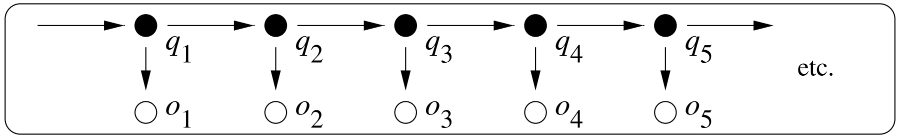
$$\simeq \prod_{m=1}^{M} \mathrm{P}(w_m \mid w_{m-1})$$

mit den statistischen Abhängigkeiten $\left\{ \begin{array}{c} w_m \leftarrow w_{m-1} \text{ (Bigramme)} \\ w_m \leftarrow \{w_{m-2}, w_{m-1}\} \text{ (Trigramme)} \end{array} \right\}$.

Unabhängigkeitspostulate der Bigramm-Grammatik

$\Im(\{w_m\} \mid \{w_{m-1}\} \mid \{w_1, \dots, w_{m-2}\})$ für alle $m = 2, \dots, M$.

Beispiel — Hidden Markov Modelle



Datenerzeugungsmodell

Beobachtbare Zeichenfolge $\boldsymbol{o} = o_1 \dots o_T$ mit $o_t \in \mathcal{O}$
Verborgene Zustandsfolge $\boldsymbol{q} = q_1 \dots q_T$ mit $q_t \in \mathcal{Q}$

$$\mathrm{P}(\boldsymbol{o}) = \mathrm{P}(\boldsymbol{o}|\boldsymbol{\lambda}) = \sum_{\boldsymbol{q} \in \mathcal{Q}^T} \mathrm{P}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\lambda}) = \sum_{\boldsymbol{q} \in \mathcal{Q}^T} \prod_{t=1}^{T} \mathrm{P}(q_t|q_{t-1}) \cdot \mathrm{P}(o_t|q_t)$$

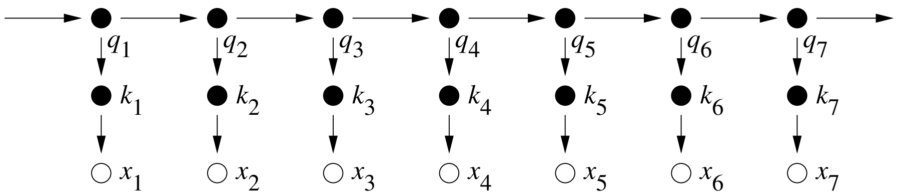
mit statistischen Abhängigkeiten $q_t \leftarrow q_{t-1}$ und $o_t \leftarrow q_t$.

Unabhängigkeitspostulate des HMM

$\Im(\{q_{t+1}\} \mid \{q_t\} \mid \{q_1, \dots, q_{t-1}; o_1, \dots, o_t\})$ und
$\Im(\{o_{t+1}\} \mid \{q_{t+1}\} \mid \{q_1, \dots, q_t; o_1, \dots, o_t\})$

Beispiel — (Semi-)kontinuierliches HMM

mit eindimensionalen Ausgabewerten



Datenerzeugungsmodell

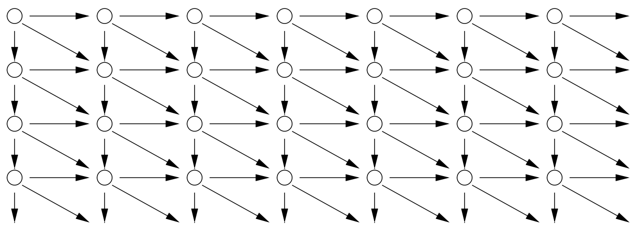
Beobachtbare Wertefolge $\boldsymbol{x} = x_1, \dots, x_T$ mit $x_t \in \mathbb{R}$
Verborgene Komponentenfolge $\boldsymbol{k} = k_1 \dots k_T$ mit $k_t \in \mathcal{K}$
Verborgene Zustandsfolge $\boldsymbol{q} = q_1 \dots q_T$ mit $q_t \in \mathcal{Q}$

$$\mathrm{P}(\boldsymbol{X}) = \mathrm{P}(\boldsymbol{X}|\boldsymbol{\lambda}) = \sum_{\boldsymbol{q} \in \mathcal{Q}^T} \sum_{\boldsymbol{k} \in \mathcal{K}^T} P(\boldsymbol{X}, \boldsymbol{k}, \boldsymbol{q} \mid \boldsymbol{\lambda})$$

mit statistischen Abhängigkeiten $q_t \leftarrow q_{t-1}$, $k_t \leftarrow q_t$ und $\boldsymbol{x}_t \leftarrow k_t$.

Beispiel — 2D Markov Random Field

Texturmodelle in der Grauwertbildanalyse



Datenerzeugungsmodell

Beobachtbare Zufallsvariablen $x_{i,j}$
auf dem **Ortsgitter** $i = 1, \dots, I$ und $j = 1, \dots, J$
mit statistischen Abhängigkeiten $x_{i,j} \leftarrow \{x_{i-1,j-1}, x_{i,j-1}, x_{i-1,j}\}$.

Unabhängigkeitspostulate des MRF

Für alle Gitterpunkte $(n, m) \in \mathbb{Z} \times \mathbb{Z}$ ist gefordert:
$\Im(\{\mathbb{X}_{n,m}\} \mid \{\mathbb{X}_{n,m-1}, \mathbb{X}_{n-1,m}, \mathbb{X}_{n-1,m-1}\} \mid \{\mathbb{X}_{i,j} \mid i < n, j < m\})$

Dependenzmodelle

Algebraische Charakterisierung von Abhängigkeitsstrukturen

Definition

Sei $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$ eine Menge von Zufallsvariablen und $\mathrm{P}(\cdot)$ eine Verteilung über V. Die Relation $\Im = \Im_{\mathrm{P}}$ mit

$$\Im : \mathfrak{P}X \times \mathfrak{P}X \times \mathfrak{P}X \;\to\; \{0,1\}$$

heißt **Dependenzmodell von** $\mathrm{P}(\cdot)$, wenn für alle (disjunkten) Variablenmengen $A, B, Z \subset V$ gilt:

$$\Im(A \mid Z \mid B) \quad\Leftrightarrow\quad \mathrm{P}(A \mid B, Z) = \mathrm{P}(A|Z)$$

Bemerkungen

1. Es gibt 4^N viele Variablenkombinationen A, B, Z.
   Es gibt 2^{4^N} viele dreistellige Mengenrelationen $\Im$ über V.
   Wieviele $\Im$ davon sind ein **valides Dependenzmodell** $\Im_{\mathrm{P}}$?
2. Simpsons Paradoxa: $\Im(A|Z|B) \not\Rightarrow \Im(A|\varnothing|B)$ und $\Im(A|Z|B) \not\Leftarrow \Im(A|\varnothing|B)$

Beweis.

SYM Symmetrie

$$\Im(A|Z|B) \Rightarrow \mathrm{P}(A,Z,B) = \underbrace{f(A,Z)\cdot g(B,Z)}_{\mathrm{P}(B,Z,A)} \Rightarrow \Im(B|Z|A)$$

DEC Dekomposition

$$\mathrm{P}(A,Z,B) \;=\; \sum_C \mathrm{P}(A,Z,B,C) \;=\; \sum_C f(A,Z)\cdot g(B,C,Z) \;=\; f(A,Z)\cdot\overbrace{\sum_C g(B,C,Z)}^{\tilde{g}(B,Z)}$$

beweist $\Im(A|Z|B)$; analoge Herleitung von $\Im(A|Z|C)$.

WUN Schwache Vereinigung

$$\begin{aligned}\Im(A \mid Z \mid B,C) \Rightarrow \mathrm{P}(A,Z,B,C) &= f(A,Z)\cdot g(B,C,Z)\\ &= \tilde{f}(A,Z,C)\cdot\tilde{g}(B,Z,C) \Rightarrow \Im(A \mid Z,C \mid B)\end{aligned}$$

CON Kontraktion

$$\mathrm{P}(A \mid Z,B,C) \;=\; \underbrace{\mathrm{P}(A \mid Z,B)}_{\Im(A|Z,B|C)} \;=\; \underbrace{\mathrm{P}(A \mid Z)}_{\Im(A|Z|B)} \Rightarrow \Im(A \mid Z \mid B,C)$$

INT Durchschnitt *(Beweis zu äquivalenter Formulierung INT* folgt)*

□

Pearlsche Dependenzaxiome

Axiomatische Charakterisierung aller *„erlaubten"* $\Im$-Relationen

Satz (Judea Pearl)

Es sei $\mathrm{P}(\cdot)$ eine Wahrscheinlichkeitsverteilung über $\mathbb{X}_1, \dots, \mathbb{X}_N$ und $\Im(\cdot|\cdot|\cdot)$ das zugehörige Dependenzmodell.

Dann gelten für alle (paarweise disjunkten) Variablenmengen A, B, C, Z die folgenden vier Aussagen:

SYM ***Symmetrie*** $\Im(A \mid Z \mid B) \Leftrightarrow \Im(B \mid Z \mid A)$

DEC ***Dekomposition*** $\Im(A \mid Z \mid B{\cup}C) \Rightarrow \Im(A \mid Z \mid B) \wedge \Im(A \mid Z \mid C)$

WUN ***Schwache Vereinigung*** $\Im(A \mid Z \mid B{\cup}C) \Rightarrow \Im(A \mid Z{\cup}C \mid B)$

CON ***Kontraktion*** $\Im(A \mid Z \mid B) \wedge \Im(A \mid Z{\cup}B \mid C) \Rightarrow \Im(A \mid Z \mid B{\cup}C)$

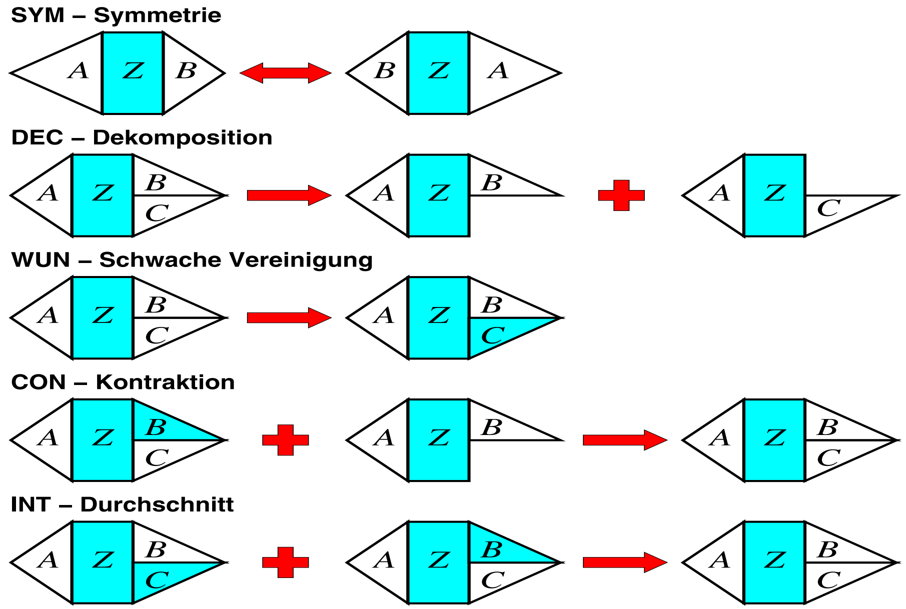
Falls $\mathrm{P}(\cdot)$ zudem streng positiv ($\forall \boldsymbol{x} \in \boldsymbol{\Omega} : \mathrm{P}(\boldsymbol{x}) > 0$) ist, gilt sogar:

INT ***Durchschnitt***

$\Im(A \mid Z{\cup}C \mid B) \wedge \Im(A \mid Z{\cup}B \mid C) \Rightarrow \Im(A \mid Z \mid B{\cup}C)$

Pearlsche Dependenzaxiome

Beweis durch angestrengtes Hingucken



Bemerkungen zu Pearls Axiomen

1. Die logische Umkehrung der Implikation CON folgt aus den Axiomen DEC und WUN.
2. Die logische Umkehrung von INT folgt mit zweimaliger Anwendung von WUN.
3. Die Axiomatisierung kann auf nichtdisjunkte ZV-Mengen ausgedehnt werden. Aus den obengenannten Axiomen sowie der zusätzlichen Forderung $\Im(A \mid Z \mid Z)$ beweist man die Aussage

$$\Im(A \mid Z \mid B) \Leftrightarrow \Im(A, Z \mid Z \mid B, Z)$$

4. Die fünf Axiome sind voneinander logisch unabhängig. Beweis durch Gegenbeispiele.
5. Das Axiom INT findet sich auch in der folgenden, äquivalenten Fassung INT* (Lauritzen, $Z = \varnothing$) wieder:

$$\Im(A \mid C \mid B) \wedge \Im(A \mid B \mid C) \Rightarrow \Im(A \mid \varnothing \mid B, C)$$

Durchschnittsaxiom INT

Garantiert ausschließlich für streng positive Verteilungen!

Herleitung für streng positive $\mathrm{P}(\cdot)$

Auf Grund der Prämissen von INT* gelten die Faktorisierungen

$$\mathrm{P}(a, b, c) \; = \; k(a, c) \cdot \ell(b, c) \; = \; g(a, b) \cdot h(b, c)$$

und für beliebige Werte c — also zum Beispiel für c_0 beliebig aber fest — gilt

$$g(a, b) \; = \; k(a, c_0) \cdot \frac{\ell(b, c_0)}{h(b, c_0)} \; =: \; \pi(a) \cdot \rho(b) \; .$$

Dann gilt die marginale Unabhängigkeit $\{a\} \not\sim \{b, c\}$ wegen der Faktorisierung

$$\mathrm{P}(a, b, c) \; = \; \pi(a) \cdot [\rho(b) \cdot h(b, c)] \; .$$

Gegenbeispiel

Die drei binärwertige Zufallsvariablen mit $\mathbb{X}_1 = \mathbb{X}_2 = \mathbb{X}_3$ und $\mathrm{P}(\mathbb{X}_i = 1) = 1/2$ für alle $i \in \{1, 2, 3\}$ sind nicht streng positiv (z.B. $\mathrm{P}(1, 1, 0) = 0$) und widerlegen INT:

$$\Im(\mathbb{X}_1 \mid \mathbb{X}_2 \mid \mathbb{X}_3) \; , \quad \Im(\mathbb{X}_1 \mid \mathbb{X}_3 \mid \mathbb{X}_2) \; , \quad \neg\Im(\mathbb{X}_1 \mid \varnothing \mid \mathbb{X}_2, \mathbb{X}_3)$$

Vollständigkeitsvermutung (Pearl & Paz, 1985)

Trügerische Hoffnung

Wenn $\Im$ die Axiome SYM, DEC, WUN & CON erfüllt, so heißt $(V, \Im)$ **Semigraphoid** und es gibt eine Wahrscheinlichkeitsverteilung $\mathrm{P}(\cdot)$ mit

$$P(A \mid B, Z) \; = \; P(A|Z) \quad \Leftrightarrow \quad \Im(A \mid Z \mid B) \; .$$

Wenn zusätzlich das Durchschnittsaxiom (INT) erfüllt ist, so kann für das **Graphoid** $(V, \Im)$ sogar ein streng positives $\mathrm{P}(\cdot)$ gefunden werden.

Satz (Studeny, 1992)

Weder für die Relationenmenge

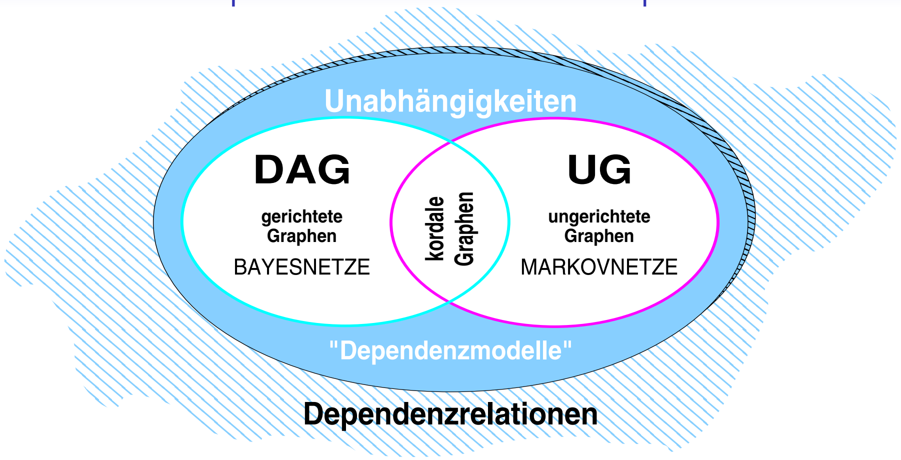
$$\{\Im_{\mathrm{P}} \mid \; \mathrm{P} \textit{ Wahrscheinlichkeitsverteilung} \}$$

noch für deren Teilmenge

$$\{\Im_{\mathrm{P}} \mid \; \mathrm{P} \textit{ streng positive Wahrscheinlichkeitsverteilung} \}$$

*gibt es ein korrektes und vollständiges **endliches Axiomensystem**.*

Dependenzmodelle und Graphen



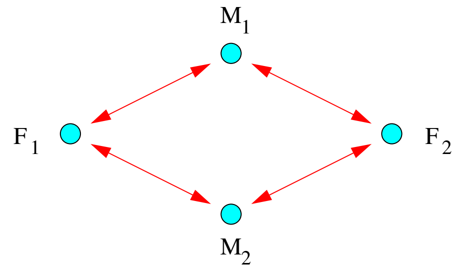
? Welche Dependenzmodelle sind durch UG charakterisierbar

? Welche Dependenzmodelle sind durch DAG charakterisierbar

? Welche Dependenzmodelle liegen gleichzeitig in beiden Klassen

? Welche Dependenzmodelle sind komplexer als jede Graphstruktur

Trennungsrelation im ungerichteten Graphen



Attributwerte: ±*infiziert*

Partnertauschmodell

Wegen

$$P(f_2 \mid f_1, m_1, m_2) = P(f_2 \mid m_1, m_2)$$

gilt $\Im(\mathbb{F}_1 \mid \mathbb{M}_1, \mathbb{M}_2 \mid \mathbb{F}_2)$.

Partnertauschgraph

$\mathrm{sep}\langle \mathbb{F}_1 \mid \mathbb{M}_1, \mathbb{M}_2 \mid \mathbb{F}_2 \rangle$ und
$\mathrm{sep}\langle \mathbb{M}_1 \mid \mathbb{F}_1, \mathbb{F}_2 \mid \mathbb{M}_2 \rangle$

Definition

Es sei $\mathcal{G} = (V, \mathcal{E})$ ein ungerichteter Graph und $A, B, Z \subset V$ disjunkte Knotenmengen. Die Menge Z **trennt** A **von** B genau dann, wenn alle Pfade zwischen Elementen $a \in A$ und $b \in B$ mindestens einen Knoten $z \in Z$ enthalten. Wir schreiben dafür:

$$\mathrm{sep}\langle A \mid Z \mid B \rangle$$

➪ *„Z blockiert alle Verbindungen zwischen Knoten aus A und B“*

Graphische Verteilungen und Dependenzmodelle

Überrepräsentation & Unterrepräsentation von $\Im(\cdot|\cdot|\cdot)$ durch $\mathrm{sep}\langle\cdot|\cdot|\cdot\rangle$

Definition

Es sei $P(\cdot)$ eine Wahrscheinlichkeitsverteilung auf V und $\Im$ ihr Dependenzmodell. Der ungerichtete Graph $\mathcal{G} = (V, \mathcal{E})$ heißt

- **Abhängigkeitsbild** von P gdw.
$$\Im(A \mid Z \mid B) \Rightarrow \mathrm{sep}\langle A \mid Z \mid B \rangle$$
- **Unabhängigkeitsbild** von P gdw.
$$\Im(A \mid Z \mid B) \Leftarrow \mathrm{sep}\langle A \mid Z \mid B \rangle$$
- **perfektes Bild** von P gdw.
$$\Im(A \mid Z \mid B) \Leftrightarrow \mathrm{sep}\langle A \mid Z \mid B \rangle$$

Die Verteilung $P(\cdot)$ (und das Modell $\Im$) heißen **graphisch**, wenn ein ungerichteter Graph existiert, der $\Im$ perfekt abbildet.

Über A-Bilder, U-Bilder und P-Bilder

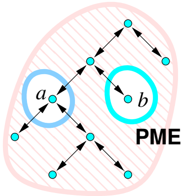
Bemerkungen

1. Die Trennungsrelation im UG ist **monoton** in der Barriere Z:
$$\mathrm{sep}\langle A|Z|B\rangle \text{ und } \tilde{Z} \supset Z \quad ➪ \quad \mathrm{sep}\langle A|\tilde{Z}|B\rangle$$
2. Es gilt die *„marginale Trennung“* $\mathrm{sep}\langle\{a\}|\varnothing|\{b\}\rangle$ genau dann, wenn $a, b \in V$ zu verschiedenen Zusammenhangskomponenten gehören.
3. A-Bild $\leadsto$ für adjazente Knoten gilt keinerlei Unabhängigkeit (der diskrete Graph ist A-Bild jedes P)
4. U-Bild $\leadsto$ für nichtadjazente Knoten gilt ≥ 1 Unabhängigkeit (der vollständige Graph ist U-Bild jedes P)
5. Nicht alle Dependenzmodelle $\Im$ besitzen ein perfektes Bild. Für das nichtmonotone Modell mit zwei Würfeln $\mathbb{W}_1$, $\mathbb{W}_2$ und die Signalglocke $\mathbb{G}$ für Pasch gilt nämlich
$$\Im(\mathbb{W}_1 \mid \varnothing \mid \mathbb{W}_2) \quad \text{und nicht} \quad \Im(\mathbb{W}_1 \mid \mathbb{G} \mid \mathbb{W}_2)\,.$$

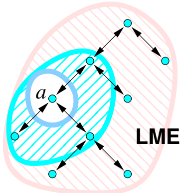
Die drei Markoveigenschaften

Definition

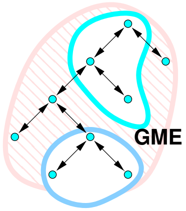
Es sei $\mathrm{P}(\cdot)$ eine Wahrscheinlichkeitsverteilung auf V und $\Im$ ihr Dependenzmodell. Der ungerichtete Graph $\mathcal{G} = (V, \mathcal{E})$ erfüllt die



- **paarweise Markoveigenschaft**
  gdw. für alle nichtadjazenten $a, b \in V$ gilt:
  $$\Im(a \mid V \setminus \{a, b\} \mid b)$$



- **lokale Markoveigenschaft**
  gdw. für alle jede Variable $a \in V$ gilt:
  $$\Im(a \mid \mathrm{bd}(a) \mid V \setminus \mathrm{cl}(a))$$



- **globale Markoveigenschaft**
  gdw. für alle $A, B, Z \subset V$ mit $\mathrm{sep}\langle A|Z|B\rangle$ gilt:
  $$\Im(A \mid Z \mid B)$$

Beweis.

- GME $\Rightarrow$ LME
  Es sei $a \in V$.
  Offensichtlich werden die beiden Mengen $\{a\}$ und $V \setminus \mathrm{cl}(a)$ durch den Rand $\mathrm{bd}(a)$ von a separiert.
  Damit folgt die Behauptung aus der Anwendung von GME.
- LME $\Rightarrow$ PME
  Zunächst gilt wegen der Voraussetzung LME die Aussage
  $$\Im(a \mid \mathrm{bd}(a) \mid V \setminus \mathrm{cl}(a))$$
  Wegen der Teilmengenbeziehung
  $$V \setminus \{a, b\} \;=\; \mathrm{bd}(a) \cup ((V \setminus \mathrm{cl}(a)) \setminus \{b\})$$
  kann mittels Axiom WUN
  $$\Im(a \mid V \setminus \{a, b\} \mid V \setminus \mathrm{cl}(a))$$
  gefolgert werden und mittels Axiom DEC wird verkürzt zu
  $$\Im(a \mid V \setminus \{a, b\} \mid b)\ .$$

□

Die Markoveigenschaften für „*Semigraphoide*"

Markovnetze $\hat{=}$ minimale Unabhängigkeitsbilder

Definition

Der Graph $\mathcal{G}$ heißt **Markovnetz** von $\mathrm{P}(\cdot)$, wenn er minimal mit der globalen Markoveigenschaft für $\mathrm{P}(\cdot)$ ist.

Das Markovnetz $\mathcal{G}$ ignoriert keine Abhängigkeiten, höchstens Unabhängigkeiten, aber davon so wenige wie möglich.

Satz

Sei $\mathcal{G} = (V, \mathcal{E})$ und $\mathrm{P}(\cdot)$ auf V gegeben. Dann gilt

globale ME ⇒ lokale ME ⇒ paarweise ME ,

aber es gilt im allgemeinen weder die Umkehrrichtung

paarweise ME ⇒ lokale ME

noch die Umkehrrichtung

lokale ME ⇒ globale ME

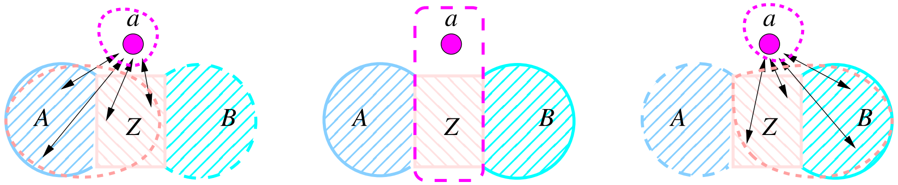
Beweis.

- LME $\not\Rightarrow$ GME (Gegenbeispiel)
  $$\mathbb{U} - \mathbb{W} - \mathbb{X} - \mathbb{Y} - \mathbb{Z} \qquad (\mathbb{U} = \mathbb{W}, \mathbb{Y} = \mathbb{Z}, \mathbb{X} = \mathbb{W} \cdot \mathbb{Y})$$
  mit binärwertigen, gleichverteilten Variablen.
  Es gilt zwar die lokale ME, aber $\Im(\mathbb{U}, \mathbb{W} \mid \mathbb{X} \mid \mathbb{Y}, \mathbb{Z})$ scheitert wegen
  $$\begin{aligned} \mathrm{P}(\mathbb{U} = \mathbb{W} = \mathbb{Y} = \mathbb{Z} = 1 \mid \mathbb{X} = 0) &= 0 \\ \mathrm{P}(\mathbb{U} = \mathbb{W} = 1 \mid \mathbb{X} = 0) \cdot \mathrm{P}(\mathbb{Y} = \mathbb{Z} = 1 \mid \mathbb{X} = 0) &\neq 0 \end{aligned}$$
- PME $\not\Rightarrow$ LME (Gegenbeispiel)
  $$\mathbb{X} \quad \mathbb{Y} - \mathbb{Z} \qquad (\mathbb{X} = \mathbb{Y} = \mathbb{Z})$$
  mit binärwertigen, gleichverteilten Variablen.
  Dann sind $\Im(X \mid Z \mid Y)$ und $\Im(X \mid Y \mid Z)$ trivialerweise erfüllt, überflüssigerweise sogar auch $\Im(Y \mid X \mid Z)$. Aber es gilt keineswegs
  $$\Im(X \mid \mathrm{bd}(X) \mid V \setminus \mathrm{cl}(X))$$
  denn $\mathrm{bd}(X) = \varnothing$ und $V \setminus \mathrm{cl}(X) = \{Y, Z\}$, und es ist $\mathbb{X}$ natürlich nicht marginal unabhängig von $\{\mathbb{Y}, \mathbb{Z}\}$.

□

Die Markoveigenschaften für „*Graphoide*“

Äquivalenz für strikt positive Wahrscheinlichkeitsverteilungen



$\text{sep}\langle\{a\} \mid A, Z \mid B\rangle$ $\text{sep}\langle A \mid Z\cup\{a\} \mid B\rangle$ $\text{sep}\langle\{a\} \mid B, Z \mid A\rangle$

Satz

Sei $\mathcal{G}$ ein UG. Erfüllt die Dependenzrelation $\Im$ von $\mathrm{P}(\cdot)$ für alle disjunkten Mengen $A, B, C, Z \subset V$ die Eigenschaft

- *INT* ***Durchschnitt***

$$\Im(A \mid Z\cup C \mid B) \wedge \Im(A \mid Z\cup B \mid C) \quad\Rightarrow\quad \Im(A \mid Z \mid B\cup C)\ ,$$

so gilt

globale ME ⇔ *lokale ME* ⇔ *paarweise ME .*

Beweis.

Es ist nur die Implikation PME ⇒ GME zu zeigen, die wir durch absteigende Induktion über die Größe $n = |Z|$ beweisen.

- Induktionsanfang:
  Für $n = N - 2$ liefert PME die Behauptung (o.B.d.A. sei $|A| = |B| = 1$).

  Induktionsschluß:
  Wir unterscheiden die beiden Fälle $A\cup B\cup Z = V$ und $A\cup B\cup Z \neq V$.
- Fall 1: Sei o.B.d.A. $|A| > 1$ und $a \in A$. Dann gelten nach WUN die beiden Trennungsaussagen

$$\text{sep}\langle A\setminus\{a\} \mid Z\cup\{a\} \mid B\rangle\ , \qquad \text{sep}\langle\{a\} \mid Z\cup A\setminus\{a\} \mid B\rangle\ .$$

  Nach I.V. übersetzen diese in die korrespondierenden Unabhängigkeiten und mit Axiom INT folgt $\Im(A \mid Z \mid B)$.
- Fall 2: Für jedes $a \in V\setminus(A\cup B\cup Z)$ gilt $\text{sep}\langle A \mid Z\cup\{a\} \mid B\rangle$ und mindestens eine der beiden Trennungsaussagen

$$\text{sep}\langle\{a\} \mid A, Z \mid B\rangle\ , \qquad \text{sep}\langle\{a\} \mid B, Z \mid A\rangle\ .$$

  Im ersten Fall folgt das Resultat $\Im(A \mid Z \mid B)$ nach den Axiomen INT, DEC und im zweiten Fall nach den Axiomen SYM, INT, DEC aus den übersetzten Trennungsaussagen (I.V.). □

Markovnetzkonstruktion

(1:1)-Abbildung aller partiellen (Un-)Abhängigkeiten

Lemma

Erfüllt die Dependenzrelation $\Im$ von $\mathrm{P}(\cdot)$ die Axiome SYM, DEC und INT, so gibt es ein ***eindeutiges Markovnetz*** *$\mathcal{G} = (V, \mathcal{E})$ zu $\Im$. Für alle Variablenpaare $a, b \in V$ gilt:*

$$\{a, b\} \notin \mathcal{E} \quad\Leftrightarrow\quad \Im(a \mid V\setminus\{a, b\} \mid b)$$

Satz (Pearl & Paz, 1985)

Die Dependenzrelation $\Im$ ist ***graphisch*** *genau dann, wenn sie die Axiome SYM, DEC, INT, SUN und TRA erfüllt.*

- *SUN* ***Starke Vereinigung***

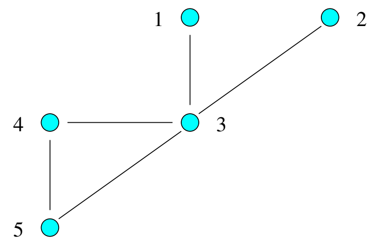
$$\Im(A \mid Z \mid B) \quad\Rightarrow\quad \Im(A \mid Z\cup C \mid B)$$

- *TRA* ***Transitivität***
  Für alle Variablen $x \in V$ gilt:

$$\Im(A \mid Z \mid B) \quad\Rightarrow\quad \Im(A \mid Z \mid \{x\}) \vee \Im(\{x\} \mid Z \mid B)$$

Beispiel — qualitative graphische Inferenz

„Vorhersage einer Reiseankunftszeit“



Uhrzeitwertige Zufallsvariable

Zwei Passanten — zwei Armbanduhren

$\mathbb{X}_1$	Zeit auf Armbanduhr I
$\mathbb{X}_2$	Zeit auf Armbanduhr II
$\mathbb{X}_3$	die wahre Uhrzeit
$\mathbb{X}_4$	die Fahrtzeit „*Jena–Weimar*“
$\mathbb{X}_5$	die Ankunftzeit in Weimar

Markovnetzerzeugung

Kantenlöschverfahren mit den Vorbehalten:

$$\left\{\begin{array}{l} \neg\,\Im(\mathbb{X}_1 \mid \mathbb{X}_2, \mathbb{X}_4, \mathbb{X}_5 \mid \mathbb{X}_3) \\ \neg\,\Im(\mathbb{X}_2 \mid \mathbb{X}_1, \mathbb{X}_4, \mathbb{X}_5 \mid \mathbb{X}_3) \end{array}\right. \quad , \quad \left\{\begin{array}{l} \neg\,\Im(\mathbb{X}_3 \mid \mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_5 \mid \mathbb{X}_4) \\ \neg\,\Im(\mathbb{X}_3 \mid \mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_4 \mid \mathbb{X}_5) \\ \neg\,\Im(\mathbb{X}_4 \mid \mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3 \mid \mathbb{X}_5) \end{array}\right.$$

Inferenz durch Ablesen von Trennungseigenschaften

Bedingte, aber nicht partielle Unabhängigkeiten: $\Im(\mathbb{X}_1, \mathbb{X}_2 \mid \mathbb{X}_3 \mid \mathbb{X}_5)$

Beispiel — Würfelpaar und Glocke

Nichtgraphische Verteilungen

Viele interessante Verteilungen liegen außerhalb der Klasse ungerichteter graphischer Modelle.

- Selbst ein streng positives $\mathrm{P}(\cdot)$ garantiert lediglich die Axiome SYM, DEC und INT, nicht aber SUN oder TRA.

Würfel-Glocken-Experiment

Es schlägt die starke Vereinigung (SUN) fehl:

$$\Im(\mathbb{W}_1 \mid \varnothing \mid \mathbb{W}_2) \quad \text{aber} \quad \neg\, \Im(\mathbb{W}_1 \mid \mathbb{G} \mid \mathbb{W}_2)$$

Bei **unfairen Würfeln** gilt auch keine Transitivität (TRA) mehr:

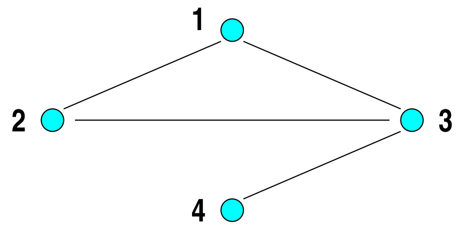
$$\Im(\mathbb{W}_1 \mid \varnothing \mid \mathbb{W}_2) \quad \text{aber weder} \quad \Im(\mathbb{W}_1 \mid \varnothing \mid \mathbb{G}) \quad \text{noch} \quad \Im(\mathbb{G} \mid \varnothing \mid \mathbb{W}_2)$$

Bemerkung

Die drei Axiome DEC, INT, SUN liefern eine beachtliche Äquivalenz:

$$\Im(A \mid Z \mid B) \quad \Leftrightarrow \quad \forall a \in A, b \in B: \ \Im(\{a\} \mid Z \mid \{b\})$$

Beispiel — eine Unverteilung mit Markovnetz



Eindeutiges Markovnetz

Dependenzstruktur

Gegeben sind die bedingten *„Unabhängigkeiten"*

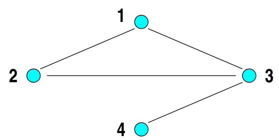
$$\Im(1 \mid 2 \mid 3) \quad \Im(1,2 \mid 3 \mid 4)$$
$$\Im(1 \mid 3 \mid 4) \quad \Im(1 \mid 2,3 \mid 4)$$
$$\Im(2 \mid 3 \mid 4) \quad \Im(2 \mid 1,3 \mid 4)$$

zuzüglich aller Symmetrien.

Eigenschaften

- $\Im$ erfüllt die Axiome SYM, DEC, WUN und INT.
- $\Im$ widerspricht dem Axiom CON, denn es gelten zwar $\Im(1 \mid 2 \mid 3)$ und $\Im(1 \mid 2,3 \mid 4)$, aber keineswegs $\Im(1 \mid 2 \mid 3,4)$.
- $\Im$ besitzt wegen ¬CON kein Wahrscheinlichkeitsmodell mit $\Im = \Im_{\mathrm{P}}$.
- $\Im$ besitzt aber wegen SYM, DEC, INT ein eindeutiges Markovnetz.

Beispiel — pathologische Verteilung ohne Markovnetz



Kantenlöschverfahren

Dependenzstruktur

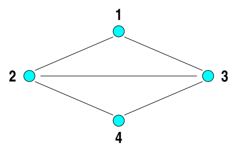
Gegeben sind bedingte Unabhängigkeiten

$$\Im(1 \mid 2,3 \mid 4) \qquad \Im(2 \mid 1,3 \mid 4)$$

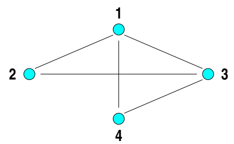
zuzüglich aller Symmetrien.

Eigenschaften

- $\Im$ erfüllt die Axiome SYM, DEC, WUN, CON.
- $\Im$ widerspricht dem Axiom INT, weil $\Im(1,2 \mid 3 \mid 4)$ fehlt.
- $\Im$ gehorcht einer Verteilung P, aber P ist wegen ¬INT nicht streng positiv!
- Das Kantenlöschverfahren ergibt kein Unabhängigkeitsbild, weil $\mathrm{sep}\langle 1|3|4 \rangle$ gilt, aber nicht $\Im(1|3|4)$.
- Es gibt kein eindeutiges Markovnetz!

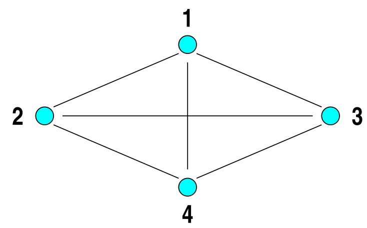


minimales U-Bild #1



minimales U-Bild #2

Beispiel — Unverteilung mit Monsternetz



vollständiger Graph

Dependenzstruktur

Gegeben sind die Postulate

$$\Im(1 \mid 3 \mid 4) \qquad \Im(2 \mid 3 \mid 4)$$
$$\Im(1,2 \mid 3 \mid 4) \qquad \Im(1 \mid 2 \mid 4)$$

zuzüglich aller Symmetrien.

Eigenschaften

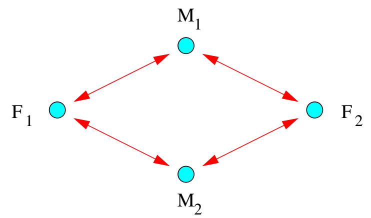
- $\Im$ erfüllt die Axiome SYM, DEC, INT.
- $\Im$ erfüllt nicht das Axiom WUN.
- Wegen SYM, DEC, INT gibt es ein eindeutiges Markovnetz $\mathcal{G}$.
- Der Graph $\mathcal{G}$ ist offenbar (keine Löschung) **vollständig**.
- Der Graph $\mathcal{G}$ *„hilft uns nicht sparen"* …

Faktorisierung von $\mathrm{P}(\cdot)$

über den Cliquen eines ungerichteten Graphen

Definition

Die Menge $C \subseteq V$ heißt **Clique** von $\mathcal{G} = (V, \mathcal{E})$, wenn $(C, \mathcal{E}|_C)$ einen maximalen zusammenhängenden Teilgraphen von $\mathcal{G}$ bildet.



Der diamantene Graph

besitzt genau vier Cliquen:

$$C_1 = \{\mathbb{M}_1, \mathbb{F}_1\}, \quad C_2 = \{\mathbb{M}_1, \mathbb{F}_2\},$$

$$C_3 = \{\mathbb{M}_2, \mathbb{F}_1\}, \quad C_4 = \{\mathbb{M}_2, \mathbb{F}_2\}$$

Gibbs-Verteilung des PT-Modells über dem Diamanten

$$\mathrm{P}(m_1, m_2, f_1, f_2) \;=\; \frac{1}{z} \cdot \phi_1(m_1, f_1) \cdot \phi_2(m_1, f_2) \cdot \phi_3(m_2, f_1) \cdot \phi_4(m_2, f_2)$$

mit den **Kompatibilitäts**- oder **Kernfunktionen** *(keine Wahrscheinlichkeiten!)*

$$\phi_i(\xi_{i_1}, \xi_{i_2}) \;=\; \begin{cases} \alpha_i & \xi_{i_1} = \xi_{i_2} \quad \text{gleicher Gesundheitszustand} \\ \beta_i & \xi_{i_1} \neq \xi_{i_2} \quad \text{genau ein Partner infiziert} \end{cases}$$

Faktorisierungs- und Markoveigenschaften

Lemma

Für alle ungerichteten Graphen $\mathcal{G} = (V, \mathcal{E})$ und für alle Wahrscheinlichkeitsmodelle $\mathrm{P} : \mathcal{X}_V \to \mathbb{R}$ gilt:

$$\boxed{\mathit{FAK} \Rightarrow \mathit{GME}} \Rightarrow \mathit{LME} \Rightarrow \mathit{PME}$$

Satz (Hammersley & Clifford, 1971)

Für jede streng positive Wahrscheinlichkeitsverteilung $\mathrm{P}(\cdot)$ und jeden ungerichteten Graphen $\mathcal{G}$ gilt:

$$\boxed{\mathit{FAK} \Leftrightarrow \mathit{GME}} \Leftrightarrow \mathit{LME} \Leftrightarrow \mathit{PME}$$

Bemerkung

Im Falle numerischer Zufallsvariablen ist als Voraussetzung des HC-Satzes auch die **Existenz und Stetigkeit** der Dichtefunktion $f : \mathcal{X}_V \to \mathbb{R}$ zu fordern.

FAK — die Faktorisierungeigenschaft

Definition

Die Wahrscheinlichkeitsverteilung $\mathrm{P}(\cdot)$ **zerfällt über dem Graphen** $\mathcal{G} = (V, \mathcal{E})$, wenn es für jede vollständige Menge $A \subset V$ eine nichtnegative **Kernfunktion**

$$\phi_A : \bigotimes_{a \in A} \mathcal{X}_a \;\to\; \mathbb{R}_0^+$$

über dem kartesischen Produkt aller A-Wertebereiche gibt mit

$$\mathrm{P}(\boldsymbol{x}) \;=\; \prod_{A \text{ vollständig}} \phi_A(\boldsymbol{x}_A)$$

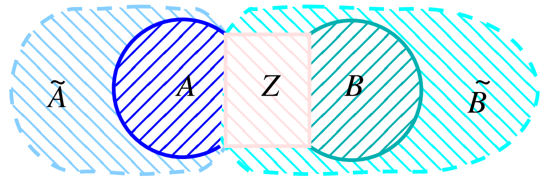
O.B.d.A. können wir diese Faktorisierungseigenschaft (FAK) aber auch unter Beschränkung auf die Menge $\mathcal{C}(\mathcal{G})$ der **Cliquen** von $\mathcal{G}$ definieren:

$$\mathrm{P}(\boldsymbol{x}) \;=\; \prod_{A \in \mathcal{C}(\mathcal{G})} \phi_A(\boldsymbol{x}_A)$$

Beweis.

FAK $\Rightarrow$ GME

Es seien $A, B, Z \subset V$ disjunkt mit $\mathrm{sep}\langle A \mid Z \mid B\rangle$.
Sei $\tilde{A}$ die Zusammenhangshülle von A in $\mathcal{G}_{V \setminus Z}$ und sei $\tilde{B} = V \setminus (\tilde{A} \cup Z)$.



A, B gehören sicherlich zu verschiedenen Zusammenhangskomponenten im Restgraphen $\mathcal{G}_{V \setminus Z}$, also gilt für jede Clique $C \subset V$ genau eine der beiden Bedingungen

$$C \;\subseteq\; \tilde{A} \cup Z \qquad \text{oder} \qquad C \;\subseteq\; \tilde{B} \cup Z\,.$$

Die (garantierte: FAK) Faktorisierung gewinnt damit das folgendes Aussehen:

$$\mathrm{P}(\boldsymbol{x}) \;=\; \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{x}_C) \;=\; \prod_{C \in \mathcal{C}_A} \phi_C(\boldsymbol{x}_C) \cdot \prod_{C \in \mathcal{C}_B} \phi_C(\boldsymbol{x}_C) \;=\; g(\boldsymbol{x}_{\tilde{A} \cup Z}) \cdot h(\boldsymbol{x}_{\tilde{B} \cup Z})$$

Nach Definition der bedingten Unabhängigkeit folgt daraus $\Im(\tilde{A} \mid Z \mid \tilde{B})$ und nach zweimaliger Anwendung des Axioms DEC auch die GME-Behauptung $\Im(A \mid Z \mid B)$. □

Beweis.

GME $\Rightarrow$ FAK

Aus völlig trivialen Gründen (auch $V \subseteq V$) gibt es eine Mammut-Faktorisierung à la

$$\mathrm{P}(\boldsymbol{x}) \;=\; \prod_{A \subset V} \phi_A(\boldsymbol{x}_A) \; .$$

Wegen der Eigenschaft $\mathrm{P}(\boldsymbol{x}) > 0$ strenger Positivität läßt sich diese Darstellung schmerzfrei logarithmieren:

$$\log \mathrm{P}(\boldsymbol{x}) \;=\; \sum_{A \subset V} \log \phi_A(\boldsymbol{x}_A)$$

Nach einer sogenannten „*Möbius-Inversion*" (sehr schwierig!) lassen sich in obigem Ausdruck durch Faktorisierung nach partiellen Unabhängigkeiten Zug um Zug alle Nicht-Cliquen-Summanden eliminieren. □

FAK $\not\Leftrightarrow$ GME für pathologische $\mathrm{P}(\cdot)$

Moussouris (1974)

Gegenbeispiel

Betrachte $V = \{\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3, \mathbb{X}_4\}$ und die Verteilung

$$\mathrm{P}(x) \;=\; \begin{cases} 1/8 & x \in \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \\ 0 & \text{sonst} \end{cases}$$

Für alle (x_2, x_4) ist entweder $\mathrm{P}(\mathbb{X}_1 \mid x_2, x_4)$ oder $\mathrm{P}(\mathbb{X}_3 \mid x_2, x_4)$ eine degenerierte Abbildung, also besteht trivialerweise keinerlei Abhängigkeit von $\mathbb{X}_3$ bzw. $\mathbb{X}_1$. Gleiches gilt auch für alle (x_1, x_3), also gilt insgesamt

$$\Im(\mathbb{X}_1 \mid \mathbb{X}_2, \mathbb{X}_4 \mid \mathbb{X}_3) \quad \wedge \quad \Im(\mathbb{X}_2 \mid \mathbb{X}_1, \mathbb{X}_3 \mid \mathbb{X}_4)$$

➯ Der **Diamant** besitzt die GME, aber $\mathrm{P}(\cdot)$ ist nicht $\Diamond$-faktorisierbar:

$$\begin{aligned}
0 \neq 1/8 &= \mathrm{P}(0,0,0,0) = \phi_{1,2}(0,0) \cdot \phi_{2,3}(0,0) \cdot \phi_{3,4}(0,0) \cdot \phi_{4,1}(0,0) \\
0 &= \mathrm{P}(0,0,1,0) = \phi_{1,2}(0,0) \cdot \phi_{2,3}(0,1) \cdot \phi_{3,4}(1,0) \cdot \phi_{4,1}(0,0) \\
0 \neq 1/8 &= \mathrm{P}(0,0,1,1) = \phi_{1,2}(0,0) \cdot \phi_{2,3}(0,1) \cdot \phi_{3,4}(1,1) \cdot \phi_{4,1}(1,0) \\
0 \neq 1/8 &= \mathrm{P}(1,1,1,0) = \phi_{1,2}(1,1) \cdot \phi_{2,3}(1,1) \cdot \phi_{3,4}(1,0) \cdot \phi_{4,1}(0,1)
\end{aligned}$$

Zerlegbare graphische Modelle

Gibbs-Verteilungen

Verteilungen in Cliquenproduktform:

$$\mathrm{P}(\boldsymbol{x}) \;=\; \mathrm{P}(\boldsymbol{x}_V) \;=\; \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{x}_C) \Bigg/ \sum_{\boldsymbol{x} \in \boldsymbol{\Omega}} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{x}_C)$$

Die Potentialfunktionen $\phi_C(\cdot)$ sind i.a. **keine** (⚡) Wahrscheinlichkeiten.

Zerlegbarkeit

Wann zerfällt $\mathrm{P}(\boldsymbol{x})$ in ein Produkt **bedingter Randverteilungen** ?

- Wenn die Cliquen des Modellgraphen als Baum angeordnet sind!
- Die Baumstruktur regelt die Abhängigkeitsrichtungen.

Es besteht Freiheit in der Wahl, welche **Außencliquen** ein Blatt und welche eine Wurzel werden.

Beispiel — Markovketten I

Faktorisierung mit unterschiedlicher Variablenordnung

$$\mathbb{X}_1 \longleftrightarrow \mathbb{X}_2 \longleftrightarrow \mathbb{X}_3 \longleftrightarrow \mathbb{X}_4$$

Faktorisierung = Kettenregel + Unabhängigkeiten

$$\mathrm{P}(x_1, x_2, x_3, x_4) \;=\; \mathrm{P}(x_1) \cdot \mathrm{P}(x_2 | x_1) \cdot \underbrace{\mathrm{P}(x_3 \mid x_1, x_2)}_{\mathrm{P}(x_3|x_2)} \cdot \underbrace{\mathrm{P}(x_4 \mid x_1, x_2, x_3)}_{\mathrm{P}(x_4|x_3)}$$

Jede Variable kann als **Baumwurzel** nominiert werden — so auch $\mathbb{X}_3$:

$$\mathrm{P}(x_3, x_2, x_4, x_1) \;=\; \mathrm{P}(x_3) \cdot \mathrm{P}(x_2 | x_3) \cdot \underbrace{\mathrm{P}(x_4 \mid x_3, x_2)}_{\mathrm{P}(x_4|x_3)} \cdot \underbrace{\mathrm{P}(x_1 \mid x_3, x_2, x_4)}_{\mathrm{P}(x_1|x_2)}$$

Aber **nicht jede Variablenfolge** ist mit der Baumstruktur verträglich:

$$\mathrm{P}(x_1, x_4, x_2, x_3) \;=\; \mathrm{P}(x_1) \cdot \underbrace{\mathrm{P}(x_4 | x_1)}_{⚡} \cdot \underbrace{\mathrm{P}(x_2 \mid x_1, x_4)}_{⚡} \cdot \underbrace{\mathrm{P}(x_3 \mid x_1, x_4, x_2)}_{⚡}$$

Beispiel — Markovketten II

Faktorisierung mit unterschiedlichen Cliquenbäumen

$$(\mathbb{X}_1, \mathbb{X}_2) \longleftrightarrow (\mathbb{X}_2, \mathbb{X}_3) \longleftrightarrow (\mathbb{X}_3, \mathbb{X}_4)$$

Faktorisierung = Cliquen + Baum + Wurzelauswahl

$$\mathrm{P}(x_1, x_2, x_3, x_4) = f(x_1, x_2) \cdot g(x_2, x_3) \cdot h(x_3, x_4)$$

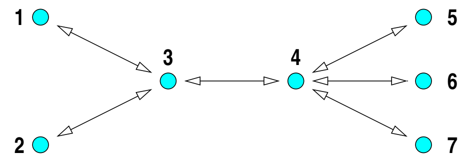
$$\mathrm{P}(x_1, x_2) \cdot \mathrm{P}(x_3|x_2) \cdot \mathrm{P}(x_4|x_3)$$

$$\mathrm{P}(x_1|x_2) \cdot \mathrm{P}(x_2, x_3) \cdot \mathrm{P}(x_4|x_3)$$

$$\mathrm{P}(x_1|x_2) \cdot \mathrm{P}(x_2|x_3) \cdot \mathrm{P}(x_3, x_4)$$

➯ Jede Wurzelnominierung definiert eine **valide** Modellformel.

Beispiel — Markovbäume I



Fakt

Ist $\mathcal{G}$ ein Baum, so sind alle Cliquen zweielementig.

Die $N-1$ Cliquen bilden selbst wieder einen Baum.

Faktorisierung im Beispiel

Kettenregel & Variablenbaumtraversierung

$$\mathrm{P}(\boldsymbol{x}) = \mathrm{P}(3) \cdot \mathrm{P}(1|3) \cdot \mathrm{P}(2|3) \cdot \mathrm{P}(4|3) \cdot \mathrm{P}(5|4) \cdot \mathrm{P}(6|4) \cdot \mathrm{P}(7|4)$$

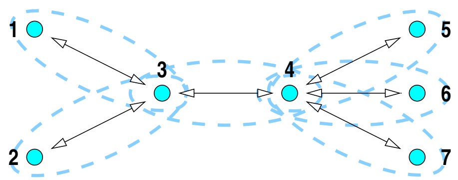
Faktorisierung allgemein

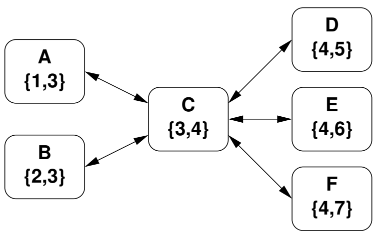
Traversieren ⇝ konsistente Variablenordnung ⇝ Einfachbedingungen

$$\mathrm{P}(\boldsymbol{x}) = \prod_n \mathrm{P}(x_n|\cdot) = \prod_n \mathrm{P}(x_n|x_{\pi(n)})$$

Denn für alle $\mathbb{X}_n \in V$ gilt: $\mathrm{sep}\langle \mathbb{X}_n \mid \mathbb{X}_{\pi(n)} \mid V \setminus \mathrm{off}(\mathbb{X}_n)\rangle$

Beispiel — Markovbäume II





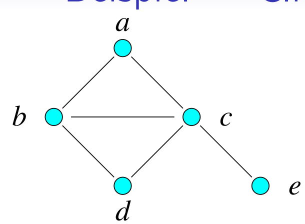
$$\mathrm{P}(\boldsymbol{x}) = \frac{\text{Cliquenwahrscheinlichkeit}}{\text{Cliquenschnittwahrscheinlichkeit}}$$

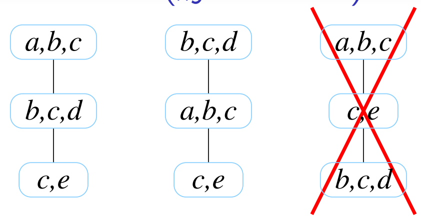
$$= \frac{\mathrm{P}(A) \cdot \mathrm{P}(B) \cdot \mathrm{P}(C) \cdot \mathrm{P}(D) \cdot \mathrm{P}(E) \cdot \mathrm{P}(F)}{\mathrm{P}(A \cap C) \cdot \mathrm{P}(B \cap C) \cdot \mathrm{P}(C \cap D) \cdot \mathrm{P}(C \cap E) \cdot \mathrm{P}(C \cap F)}$$

$$= \frac{\mathrm{P}(1,3) \cdot \mathrm{P}(2,3) \cdot \mathrm{P}(3,4) \cdot \mathrm{P}(4,5) \cdot \mathrm{P}(4,6) \cdot \mathrm{P}(4,7)}{\mathrm{P}(3) \cdot \mathrm{P}(3) \cdot \mathrm{P}(4) \cdot \mathrm{P}(4) \cdot \mathrm{P}(4)}$$

$$= \mathrm{P}(3) \cdot \mathrm{P}(1|3) \cdot \mathrm{P}(2|3) \cdot \mathrm{P}(4|3) \cdot \mathrm{P}(5|4) \cdot \mathrm{P}(6|4) \cdot \mathrm{P}(7|4)$$

Beispiel — Cliquenverbundbaum (*„join tree“*)





Drei Cliquen — aber welche Baumstruktur?

Die Cliquen $C_1 = \{a, b, c\}$, $C_2 = \{b, c, d\}$, $C_3 = \{c, e\}$ bilden paarweise einen nichtleeren Schnitt.

⇧ $\Im(C_1|C_2|C_3) \rightsquigarrow C_1 - C_2 - C_3$ ist U-Bild von $\mathrm{P}(\cdot)$
$\Im(C_2|C_1|C_3) \rightsquigarrow C_2 - C_1 - C_3$ ist U-Bild von $\mathrm{P}(\cdot)$

⇩ $C_1 - C_3 - C_2$ ist **nicht** U-Bild von $\mathrm{P}(\cdot)$, da $\neg\Im(C_1|C_3|C_2)$

$\Im(C_1|C_2|C_3)$ und $\Im(C_2|C_1|C_3) \rightsquigarrow \Im(C_1, C_2|\varnothing|C_3)$ (INT) verletzt, also ≥ 2 minimale U-Bilder.

Zerlegung

Beide konsistenten Verbundbäume ergeben nach Traversierung:

$$\mathrm{P}(a, b, c, d, e) = \mathrm{P}(a) \cdot \mathrm{P}(b|a) \cdot \mathrm{P}(c|a, b) \cdot \mathrm{P}(d|b, c) \cdot \mathrm{P}(e|c)$$

Kordalität und Zerlegbarkeit

Äquivalente Eigenschaften ungerichteter Graphen

Definition

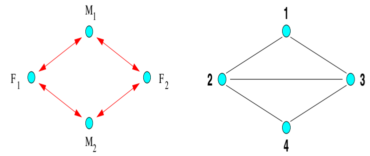
Das Mengentripel (A, Z, B) heißt **Zerlegung des ungerichteten Graphen** $\mathcal{G} = (V, \mathcal{E})$, falls gilt:

Partition	Trennung	Vollständigkeit
$A \uplus Z \uplus B = V$	$\text{sep}\langle A \mid Z \mid B\rangle$	$\mathcal{G}_Z$ ist vollständig

Der Graph $\mathcal{G}$ selbst heißt **zerlegbar**, wenn er vollständig ist oder aber eine Zerlegung mit zerlegbaren Teilgraphen $\mathcal{G}_{A\cup Z}$ und $\mathcal{G}_{B\cup Z}$ besitzt.

Definition

Ein ungerichteter Graph $\mathcal{G} = (V, \mathcal{E})$ heißt **kordal** oder **trianguliert** genau dann, wenn jeder Zyklus der Länge ≥ 4 mindestens eine **Sehne** besitzt.

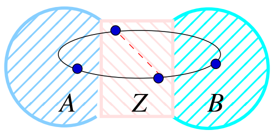


nicht kordal kordal

Beweis.

Wir zeigen die Implikation *„zerlegbar ⇒ kordal"*

- Induktionsanfang:
  $|V| \leq 3$ impliziert trivialerweise die Kordalität.
- Induktionsschritt: Sei also (A, Z, B) eine Zerlegung von $\mathcal{G}$. Nach Induktionsvoraussetzung sind dann $\mathcal{G}_{A\cup Z}$ und $\mathcal{G}_{B\cup Z}$ kordal.



Angenommen, $\mathcal{G}$ besitzt einen Zyklus ≥ 4 ohne Sehne. Dieser muß wegen der I.V. Knoten in A und auch in B haben, passiert also mindestens $2\times$ die Menge Z und teilt deshalb ≥ 2 Knoten mit Z. Diese sind aber wegen der Vollständigkeit von Z verbunden — ⚡

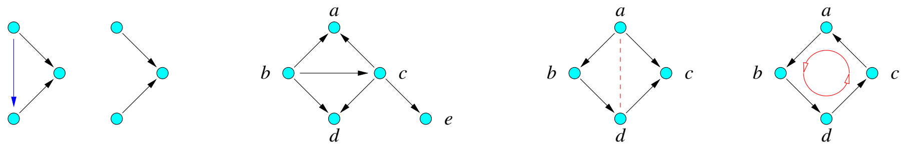
□

Moralische Graphen

„Alle Elternpaare sind (miteinander!) verheiratet."

Definition

Ein gerichteter Graph heißt **moralisch**, wenn jedes konvergierende Kantenpaar aus zwei adjazenten Knoten entspringt.



Satz

Für einen ungerichteten Graphen $\mathcal{G}$ sind die Eigenschaften äquivalent:

1. *$\mathcal{G}$ ist zerlegbar.*
2. *$\mathcal{G}$ ist kordal.*
3. *$\mathcal{G}$ läßt sich azyklisch und moralisch richten.*
4. *$\mathcal{G}$ besitzt die Cliqueneliminationseigenschaft.*
5. *Es gibt einen verträglichen Verbundbaum für $\mathcal{G}$.*

Cliqueneliminationseigenschaft

Definition

Der ungerichtete Graph $\mathcal{G}$ besitzt die **Cliqueneliminationseigenschaft**, wenn alle Knoten aller seiner Cliquen durch wiederholte Anwendung folgender Operationen eliminiert werden können:

- **Unikatknoten**
  Lösche einen Knoten, der nur in einer einzigen Clique auftaucht.
- **Dominierte Mengen**
  Lösche eine Clique, die Teilmenge einer anderen Clique ist.

Der Schlüsselgraph besitzt die CEP

Schachmatt in sieben Zügen:

$\{a, b, c\}$	$\{b, c, d\}$	$\{c, e\}$	die 3 Cliquen des Verbundbaumbeispiels
$\{b, c\}$	$\{b, c\}$	$\{c\}$	Unikate a, d und e gelöscht
$\{b, c\}$	•	•	zwei dominierte Cliquen gelöscht
•	•	•	Unikate b und c gelöscht

Verträgliche Verbundbäume

Definition

Sei $\mathcal{G} = (V, \mathcal{E})$ ein ungerichteter Graph. Der Graph $\mathcal{G}^*$ ist ein **mit** $\mathcal{G}$ **verträglicher Verbundbaum**, falls gilt:

1. Die Knoten $\mathcal{G}^*$ sind genau die Cliquen $\mathcal{C}(\mathcal{G})$.
2. $\mathcal{G}^*$ ist zusammenhängend und zyklenfrei.
3. Für jeden Knoten $a \in V$ gilt:

   Je zwei a enthaltende Cliquen besitzen einen Verbindungspfad, der ausschließlich Cliquen C mit $a \in C$ enthält.

Für den Schlüsselgraphen ist VB #3 nicht verträglich

Im dritten Verbundbaum

$$\underbrace{\{a,b,c\}}_{C_1} \;—\; \underbrace{\{c,e\}}_{C_3} \;—\; \underbrace{\{b,c,d\}}_{C_2}$$

gilt $b \in C_1$ und $b \in C_2$, aber es ist $b \notin C_3$, obwohl C_3 auf dem einzigen verfügbaren Pfad von C_1 nach C_2 liegt.

Beweis.

GME $\Rightarrow$ FAK

(die Umkehrung gilt ja sowieso)

Induktion über die Zerlegungshierarchie von $\mathcal{G}$:

Sei $\operatorname{sep}\langle A|Z|B\rangle$ eine Zerlegung von $\mathcal{G}$. Dann gilt

$$\begin{aligned} \mathrm{P}(\boldsymbol{x}_V) &= \mathrm{P}(\boldsymbol{x}_{A\cup Z}) \cdot \mathrm{P}(\boldsymbol{x}_B \mid \boldsymbol{x}_{A\cup Z}) \\ &= \mathrm{P}(\boldsymbol{x}_{A\cup Z}) \cdot \mathrm{P}(\boldsymbol{x}_B \mid \boldsymbol{x}_Z) \\ &= \frac{\mathrm{P}(\boldsymbol{x}_{A\cup Z}) \cdot \mathrm{P}(\boldsymbol{x}_{B\cup Z})}{\mathrm{P}(\boldsymbol{x}_Z)} \end{aligned}$$

wegen $\Im(A|Z|B)$ nach GME.

Der Nenner $\mathrm{P}(\boldsymbol{x}_Z)$ ist bereits ein Cliquenfaktor, weil Z vollständig ist.

Die beiden Zählerterme sind nach Induktionsvoraussetzung über $\mathcal{G}_{A\cup Z}$ bzw. $\mathcal{G}_{B\cup Z}$ faktorisierbar, bestehen also ausschließlich aus Cliquentermen.

Die behauptete Faktorisierung ergibt sich durch Zusammenfassen und Umgruppieren nach $\mathcal{G}$-Cliquen. □

Zerlegbarkeit und Faktorisierung

Lemma (Cliquenschnittformel)

Sei $\mathcal{G}$ ein zerlegbarer ungerichteter Graph. Dann gilt für alle $\mathrm{P}(\cdot)$

$$\textit{FAK} \Leftrightarrow \textit{GME}$$

und diese Faktorisierung besteht aus cliquenbezogenen Randverteilungen:

$$\mathrm{P}(\boldsymbol{x}) = \prod_{C\in\mathcal{C}} \frac{\mathrm{P}(\boldsymbol{x}_C)}{\mathrm{P}(\boldsymbol{x}_{C\cap\pi(C)})}$$

Dabei bezeichnet $\pi(C)$ die eindeutig bestimmte Vorgängerclique von C im (festen, aber beliebigen) verträglichen Verbundbaum.

Die beiden CSF für den Schlüsselgraphen

Die Verbundbäume $\{a,b,c\}—\{b,c,d\}—\{c,e\}$ und $\{b,c,d\}—\{a,b,c\}—\{c,e\}$ liefern die äquivalenten Faktorisierungen

$$\mathrm{P}(\cdot) = \frac{\mathrm{P}(a,b,c)\cdot\mathrm{P}(b,c,d)\cdot\mathrm{P}(c,e)}{\mathrm{P}(b,c)\cdot\mathrm{P}(c)} \quad \text{und} \quad \mathrm{P}(\cdot) = \frac{\mathrm{P}(b,c,d)\cdot\mathrm{P}(a,b,c)\cdot\mathrm{P}(c,e)}{\mathrm{P}(b,c)\cdot\mathrm{P}(c)}\,.$$

Beweis.

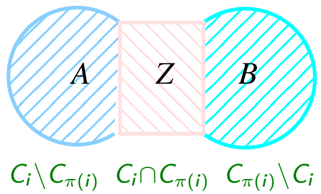
Cliquenschnittformel

Es sei $C_1, \ldots, C_M$ eine mit der Nachfolgerrelation eines Cliquenverbundbaums von $\mathcal{G}$ verträgliche Cliquenordnung.

Für jede Clique C_i bezeichne $C_{\pi(i)}$ die eindeutig bestimmte Elterclique ($\pi(i) < i$). Dann gilt (für alle i) die Trennungsrelation

$$\operatorname{sep}\langle C_i \mid C_{\pi(i)} \mid C_1, \ldots, C_{i-1}\rangle$$

und wegen GME auch die entsprechende bedingte Unabhängigkeit.



$$\begin{aligned} \mathrm{P}(\boldsymbol{x}) = \mathrm{P}(x_1, \ldots, x_N) &= \prod_{i=1}^{M} \mathrm{P}(\boldsymbol{x}_{C_i} \mid \boldsymbol{x}_{C_1}, \ldots, \boldsymbol{x}_{C_{i-1}}) \\ &= \prod_{i=1}^{M} \mathrm{P}(\boldsymbol{x}_{C_i} \mid \boldsymbol{x}_{C_{\pi(i)}}) \\ &= \prod_{i=1}^{M} \mathrm{P}(\boldsymbol{x}_{C_i} \mid \boldsymbol{x}_{C_i\cap C_{\pi(i)}}) \\ &= \prod_{i=1}^{M} \frac{\mathrm{P}(\boldsymbol{x}_{C_i})}{\mathrm{P}(\boldsymbol{x}_{C_i\cap C_{\pi(i)}})} \end{aligned}$$

□

Graphtriangulierung & Verbundbaumkonstruktion

(Algorithmus)

1. KNOTENORDNUNG
   Ordne Knoten nach maximalem Rang; setze sukzessiv:
   $$v_{i+1} \stackrel{\text{def}}{=} \operatorname*{argmax}_{v \notin V(i)} |\{v' \in V \mid (v, v') \in \mathcal{E},\ v' \in V(i)\}|$$
2. KANTENERZEUGUNG
   Für $i = N, \ldots, 1$
   $$\mathcal{E} \leftarrow \mathcal{E} \cup \{v', v''\}$$
   falls $v', v'' \in V(i-1)$ und falls $\{v_i, v'\}, \{v_i, v''\} \in \mathcal{E}$.
3. CLIQUENORDNUNG
   Fixiere Reihenfolge $C_1, \ldots, C_M$ nach dem maximalen Knotenrang.
4. KANTENERZEUGUNG
   Für $i = 2, \ldots, M$ erzeuge neue Kante $C_{\pi(i)} \to C_i$ mit
   $$\pi(i) < i \quad \text{und } |C_{\pi(i)} \cap C_i| \text{ ist maximal.}$$

Beispiel

Knotenfolge:
$a^0 b^1 c^2 d^2 e^1$

Neue Kanten:
(keine)

Cliquenfolge:
C_1 : abc^{012}
C_2 : bcd^{122}
C_3 : ce^{21}
(beliebig)

VB-Kanten:
$1 \to 2$,
$1 \to 3$

Zwischenbilanz

für ungerichtete graphische Modelle

1. Nicht jede Verteilung ist graphisch.
2. Streng positive Verteilungen erlauben aber, mit dem Kantenlöschverfahren ein Markovnetz (minimales U-Bild) zu erzeugen.
3. Markovnetze faktorisieren gemäß ihrer Cliquenstruktur, aber nicht zwingend in Wahrscheinlichkeiten.
4. Durch Triangulieren des Markovnetzes werden einige Unabhängigkeiten außer Gefecht gesetzt, aber dafür gewinnen wir eine Kettenregel (CSF).

Ursache und Wirkung

Gerichtete azyklische Graphen

Kausalrichtung

Drei Attribute · zwei Interaktionen · drei Wirkkonfugurationen:

kaskadierend	*„Wetter“ → „Ernte“ → „Preis“*	$\Im(\mathbb{X}_1 \mid \mathbb{X}_2 \mid \mathbb{X}_3)$
divergierend	*„Größe“ ← „Alter“ → „Lesefähigkeit“*	$\Im(\mathbb{X}_1 \mid \mathbb{X}_2 \mid \mathbb{X}_3)$
konvergierend	*„Würfel$_1$“ → „Glocke“ ← „Würfel$_2$“*	$\neg\,\Im(\mathbb{X}_1 \mid \mathbb{X}_2 \mid \mathbb{X}_3)$

Modelle kausaler Beziehungen: $\left\{\begin{array}{c}\textbf{erklärende}\\ \textbf{vermittelnde}\\ \textbf{diagnostische}\end{array}\right\}$ Variablen.

Lemma (Erinnerung)

*Ein gerichteter Graph $\mathcal{G} = (V, \mathcal{E})$, $\mathcal{E} \subseteq V \times V$, ist **azyklisch** genau dann, wenn es eine **kantenverträgliche Knotenordnung** $V = \{v_1, \ldots, v_N\}$ gibt:*

$$(v_i, v_j) \in \mathcal{E} \quad ➪ \quad i < j \qquad \text{für alle } i, j \in \{1, \ldots, N\}$$

*M.a.W.: Ein DAG („directed acyclic graph“) besitzt keine gerichteten Zyklen (**Pfade**); ungerichtete Zyklen (**Ketten**) sind hingegen erlaubt.*

δ-Trennungsrelation

für gerichtete azyklische Graphen

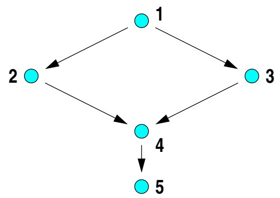
Definition

Es sei $\mathcal{G} = (V, \mathcal{E})$ ein gerichteter azyklischer Graph und $A, B, Z \subset V$ disjunkte Knotenmengen. Eine **Kette** zwischen den Knoten a und b heißt **blockiert von** Z

- wenn sie einen nichtkonvergierenden Knoten $c \in Z$ enthält
- oder wenn sie einen konvergierenden Knoten $c \notin Z$ enthält, der auch keinen Nachfolger in Z besitzt.

Die Menge Z **trennt** A **von** B genau dann, wenn alle Ketten zwischen Elementen $a \in A$ und $b \in B$ von Elementen aus Z blockiert werden. Wir schreiben dafür:

$$\mathrm{sep}_\delta \langle A \mid Z \mid B \rangle$$



Beispiel

Es gilt $\mathrm{sep}_\delta \langle 2 \mid 1 \mid 3 \rangle$, denn:

Die Kette $2 \leftarrow 1 \rightarrow 3$ ist von $\mathbb{X}_1$ blockiert wegen $1 \in Z = \{1\}$

Die Kette $2 \rightarrow 4 \leftarrow 3$ ist von $\mathbb{X}_4$ blockiert wg. $4, 5 \notin Z = \{1\}$

Ein Trennungskriterium

Satz

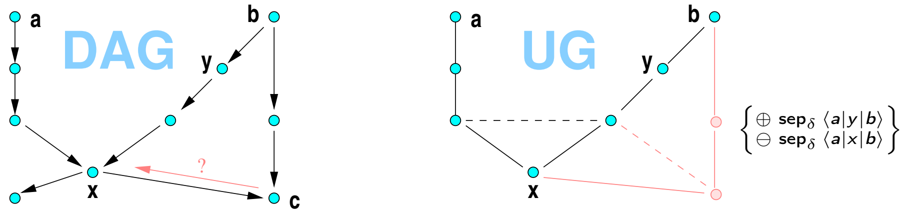
In einem DAG $\mathcal{G} = (V, \mathcal{E})$ gilt für alle disjunkten Mengen $A, B, Z \subset V$:

$$\mathrm{sep}_\delta \langle A \mid Z \mid B \rangle_{\mathcal{G}} \quad \Leftrightarrow \quad \mathrm{sep} \langle A \mid Z \mid B \rangle_{\mathcal{G}^*}$$

*Dabei bezeichne W die Vorgängerhülle von $A \cup Z \cup B$ und $\mathcal{G}^*$ sei der **moralische** Graph $(\mathcal{G}_W)^{\mathrm{m}}$ von $\mathcal{G}_W$.*

Beispiel

- **obere Kette:** y blockiert, aber x blockiert nicht!
- **untere Kette:** x blockiert und c blockiert.



Gilt $\mathrm{sep}_\delta \langle a \mid \{x, y\} \mid b \rangle$? Na klar: $\mathrm{sep} \langle a \mid \{x, y\} \mid b \rangle$

Kausale Verteilungen und Dependenzmodelle

Überrepräsentation & Unterrepräsentation von $\Im(\cdot \mid \cdot \mid \cdot)$ durch $\mathrm{sep}_\delta \langle \cdot \mid \cdot \mid \cdot \rangle$

Definition

Es sei $\mathrm{P}(\cdot)$ eine Wahrscheinlichkeitsverteilung auf V und $\Im$ ihr Dependenzmodell. Der gerichtete azyklische Graph $\mathcal{G} = (V, \mathcal{E})$ heißt

- **Abhängigkeitsbild** von P gdw.

$$\Im(A \mid Z \mid B) \Rightarrow \mathrm{sep}_\delta \langle A \mid Z \mid B \rangle$$

- **Unabhängigkeitsbild** von P gdw.

$$\Im(A \mid Z \mid B) \Leftarrow \mathrm{sep}_\delta \langle A \mid Z \mid B \rangle$$

- **perfektes Bild** von P gdw.

$$\Im(A \mid Z \mid B) \Leftrightarrow \mathrm{sep}_\delta \langle A \mid Z \mid B \rangle$$

Die Verteilung $\mathrm{P}(\cdot)$ (und das Modell $\Im$) heißen **kausal**, wenn ein gerichteter azyklischer Graph existiert, der $\Im$ perfekt abbildet.

Rekursive Faktorisierung

Definition

Die Wahrscheinlichkeitsverteilung $\mathrm{P}(\cdot)$ **zerfällt rekursiv** über dem gerichteten azyklischen Graphen $\mathcal{G} = (V, \mathcal{E})$, wenn es für jede Variable $a \in V$ eine nichtnegative **Kernfunktion**

$$\phi_a : \mathcal{X}_a \times \bigotimes_{v \in \mathrm{pa}(a)} \mathcal{X}_v \rightarrow \mathbb{R}_0^+$$

gibt mit

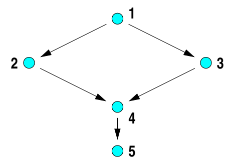
$$\mathrm{P}(\boldsymbol{x}) = \prod_{a \in V} \phi_a(x_a, \boldsymbol{x}_{\mathrm{pa}(a)})$$

Es bezeichnet $\mathrm{pa}(a)$ die **Eltermenge** $\{v \mid (v, a) \in \mathcal{E}\}$ von a.

Beispiel

Im Rasensprengergraphen zerfällt die Verteilung, falls es Potentialfunktionen gibt mit:

$$\mathrm{P}(\boldsymbol{x}) = \phi_1(x_1) \cdot \phi_2(x_1, x_2) \cdot \phi_3(x_1, x_3) \cdot \phi_4(x_2, x_3, x_4) \cdot \phi_5(x_4, x_5)$$



Die reduzierte Kettenregel

Lemma

Wenn $\mathrm{P}(\cdot)$ über $\mathcal{G}$ rekursiv zerfällt, können die Kernfunktionen ϕ_a o.B.d.A. gemäß

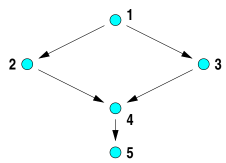
$$\phi_a(x_a, \boldsymbol{x}_{pa(a)}) \;=\; \mathrm{P}_{a|pa(a)}(x_a \mid \boldsymbol{x}_{pa(a)})$$

als bedingte Einzelwertwahrscheinlichkeiten gestaltet werden und es gilt — bei kantenverträglicher Variablenordnung — die ***reduzierte Kettenregel****:*

$$\mathrm{P}(\boldsymbol{x}) \;=\; \prod_{i=1}^{N} \mathrm{P}(x_i \mid \boldsymbol{x}_{pa(\mathbb{X}_i)})$$

Beispiel

Im Rasensprengergraphen kann die Faktorisierung wie folgt gewählt werden:



$$\mathrm{P}(\boldsymbol{x}) \;=\; \mathrm{P}(x_1)\cdot\mathrm{P}(x_2|x_1)\cdot\mathrm{P}(x_3|x_1)\cdot\mathrm{P}(x_4 \mid x_2, x_3)\cdot\mathrm{P}(x_5|x_4)$$

Beweis.

Wir vereinbaren eine verträgliche Ordnung $V = \{\mathbb{X}_1, \ldots, \mathbb{X}_N\}$ und wir wissen, daß nun die Kausalitätsbeziehung gilt:

$$\mathbb{X}_j \in \mathrm{pa}(\mathbb{X}_i) \quad\Rightarrow\quad j < i$$

Wir berechnen nun die Randverteilung der ersten n Variablen:

$$\begin{aligned}\mathrm{P}(\boldsymbol{x}_{1..n}) \;&=\; \sum_{x_{n+1}} \cdots \sum_{x_N} \mathrm{P}(\boldsymbol{x}) \;=\; \sum_{x_{n+1}} \cdots \sum_{x_N} \prod_{i=1}^{N} \phi_i(x_i, \boldsymbol{x}_{\mathrm{pa}(\mathbb{X}_i)}) \\ &=\; \prod_{i=1}^{n} \phi_i(x_i, \boldsymbol{x}_{\mathrm{pa}(\mathbb{X}_i)}) \cdot \prod_{i=n+1}^{N} \underbrace{\left(\sum_{x_i \in \mathcal{X}_i} \phi_i(x_i, \boldsymbol{x}_{\mathrm{pa}(\mathbb{X}_i)}) \right)}_{\sigma_i}\end{aligned}$$

Daraus folgt für die bedingte Wahrscheinlichkeit $\mathrm{P}(x_n \mid \boldsymbol{x}_{1..n-1})$:

$$\ldots \;=\; \frac{\mathrm{P}(\boldsymbol{x}_{1..n})}{\mathrm{P}(\boldsymbol{x}_{1..n-1})} \;=\; \frac{\prod_{i=1}^{n} \phi_i(x_i, \boldsymbol{x}_{\mathrm{pa}(\mathbb{X}_i)}) \cdot \prod_{i=n+1}^{N} \sigma_i}{\prod_{i=1}^{n-1} \phi_i(x_i, \boldsymbol{x}_{\mathrm{pa}(\mathbb{X}_i)}) \cdot \prod_{i=n}^{N} \sigma_i} \;=\; \frac{\phi_n(x_n, \boldsymbol{x}_{\mathrm{pa}(\mathbb{X}_n)})}{\sigma_n}$$

Wenn wir also normierte Faktoren verwenden ($\sigma_n \equiv 1$), entsprechen die ϕ_n gerade den klassischen Kettenregelgliedern $\mathrm{P}(x_n|\cdot)$. Daß diese tatsächlich nur von $\mathbb{X}_n$ und deren Eltervariablen abhängen, ergibt sich aus der Argumentstruktur von ϕ_n. □

Faktorisierung und Markoveigenschaft

Satz

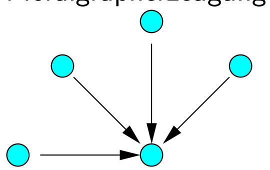
Wenn die Verteilung $\mathrm{P}(\cdot)$ über dem DAG $\mathcal{G}$ rekursiv zerfällt, dann zerfällt $\mathrm{P}(\cdot)$ auch über dem moralischen Graphen $(\mathcal{G})^{\mathrm{m}}$ von $\mathcal{G}$.

$\mathcal{G}$ ist dann sicherlich ein Unabhängigkeitsbild von $\mathrm{P}(\cdot)$, d.h. es gilt:

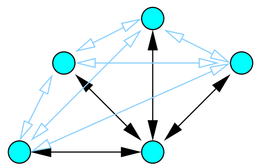
*FAK**	⇒	*GME**
⇓		⇑
FAK	⇒	*GME*

Beweisidee

Moralgrapherzeugung



Knoten mit allen Eltern in $\mathcal{G}$



„just married“ (Clique in $(\mathcal{G})^{\mathrm{m}}$)

Beweis.

- FME* ⇒ FME
  (moralische Faktorisierung)

  Für jede Variable $a \in V$ ist die Menge $\{a\} \cup \mathrm{pa}(a)$ im moralischen Graphen $(\mathcal{G})^{\mathrm{m}}$ von $\mathcal{G}$ vollständig, denn a ist mit jedem Elter adjazent und alle Eltern wurden miteinander verheiratet.
  Damit bilden die Potentialfunktionen ϕ_a auch eine Cliquenfaktorisierung auf $(\mathcal{G})^{\mathrm{m}}$.
- FME ⇒ GME
  (für $(\mathcal{G})^{\mathrm{m}}$; gilt immer)
- GME ⇒ GME*

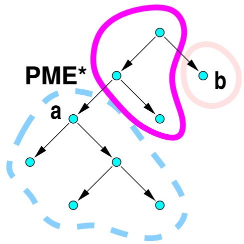
  Gilt nun $\mathrm{sep}_\delta \langle A \mid Z \mid B \rangle$ in $\mathcal{G}$, so ist auch $\mathrm{sep}\langle A \mid Z \mid B \rangle$ in $(\mathcal{G})^{\mathrm{m}}$.
  Es besitzt $(\mathcal{G})^{\mathrm{m}}$ die globale ME für $\mathrm{P}(\cdot)$, also ist auch $\Im(A \mid Z \mid B)$.

□

Die drei Markoveigenschaften

Definition

Es sei $\mathrm{P}(\cdot)$ eine Wahrscheinlichkeitsverteilung auf V und $\Im$ ihr Dependenzmodell. Der gerichtete azyklische Graph $\mathcal{G} = (V, \mathcal{E})$ erfüllt die

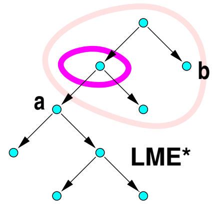


- **paarweise Markoveigenschaft**
  gdw. für alle nichtadjazenten $a, b \in V$ gilt:

$$\Im(a \mid V \setminus \mathrm{off}(a) \setminus \{b\} \mid b)$$

- **lokale Markoveigenschaft**
  gdw. für jede Variable $a \in V$ gilt:

$$\Im(a \mid \mathrm{pa}(a) \mid V \setminus \mathrm{off}(a))$$



- **globale Markoveigenschaft**
  gdw. für alle $A, B, Z \subset V$ mit $\mathrm{sep}_\delta \langle A|Z|B \rangle$ gilt:

$$\Im(A \mid Z \mid B)$$

Beweis.

PME* $\not\Rightarrow$ LME* (alle anderen Richtungen nur im alten Vorlesungsskriptum)

Als Gegenbeispiel betrachte die vier binärwertigen, uniform verteilten Zufallsvariablen $\mathbb{X} = \mathbb{Y} = \mathbb{Z}$ und $\mathbb{W}$ und den DAG mit den Kanten

$$\mathbb{Z} \to \mathbb{W} \to \mathbb{X} \quad \text{und} \quad \mathbb{Z} \to \mathbb{Y} \to \mathbb{W} .$$

Der Graph besitzt die paarweise ME, denn von den insgesamt vier nichtadjazenten Variablenpaaren erfüllen nur $(\mathbb{X}, \mathbb{Y})$ und $(\mathbb{X}, \mathbb{Z})$ die Nachkommenbedingung. Damit sind

$$\Im(\mathbb{X} \mid \mathbb{W}, \mathbb{Z} \mid \mathbb{Y}) \quad \text{und} \quad \Im(\mathbb{X} \mid \mathbb{W}, \mathbb{Y} \mid \mathbb{Z})$$

zu überprüfen — die Faktorzerlegung ergibt sich aber wie folgt:

$$\mathrm{P}(x, y \mid z, w) \;=\; \begin{cases} 1 & x = y = z \\ 0 & \text{sonst} \end{cases} \;=\; \delta_{xz} \cdot \delta_{yz}$$

Ganz analog ergibt sich auch $\mathrm{P}(x, z \mid y, w) = \delta_{xy} \cdot \delta_{zy}$. Der Graph besitzt aber nicht die lokale ME, denn die Unabhängigkeit

$$\Im(\mathbb{X} \mid \underbrace{\mathrm{pa}(\mathbb{X})}_{\mathbb{W}} \mid \underbrace{V \setminus \mathrm{off}(\mathbb{X})}_{\{\mathbb{W},\mathbb{Y},\mathbb{Z}\}})$$

bedingt nach Axiom DEC auch $\Im(\mathbb{X} \mid \mathbb{W} \mid \mathbb{Y}, \mathbb{Z})$, was die Verteilung $\mathrm{P}(\cdot)$ offensichtlich nicht hergibt. □

Die Markoveigenschaften für Semi-/Graphoide

Bayesnetze $\hat{=}$ minimale Unabhängigkeitsbilder

Definition

Der Graph $\mathcal{G}$ heißt **Bayesnetz** von $\mathrm{P}(\cdot)$, wenn er minimal mit der globalen Markoveigenschaft für $\mathrm{P}(\cdot)$ ist.

Das Bayesnetz $\mathcal{G}$ ignoriert keine Abhängigkeiten, höchstens Unabhängigkeiten, aber davon so wenige wie möglich.

Satz

Sei $\mathcal{G} = (V, \mathcal{E})$ und $\mathrm{P}(\cdot)$ auf V gegeben. Dann gilt

*FAK** ⇔ *GME** ⇔ *LME** ⇒ *paarweise ME* ,

aber es gilt im allgemeinen nicht die Umkehrrichtung

*PME** ⇒ *LME** .

Für streng positive Verteilungen $\mathrm{P}(\cdot)$ gilt sogar die Äquivalenz

*LME** ⇔ *PME** .

Axiomatisierung kausaler Dependenzmodelle ?

Satz

Ist das Dependenzmodell $\Im$ kausal, so gelten die folgenden sieben unabhängigen Axiome:

SYM ***Symmetrie*** $\quad \Im(A \mid Z \mid B) \Leftrightarrow \Im(B \mid Z \mid A)$

C/D ***Komposition/Dekomposition***

$$\Im(A \mid Z \mid B \cup C) \Leftrightarrow \Im(A \mid Z \mid B) \wedge \Im(A \mid Z \mid C)$$

INT ***Durchschnitt***

$$\Im(A \mid Z \cup C \mid B) \wedge \Im(A \mid Z \cup B \mid C) \Rightarrow \Im(A \mid Z \mid B \cup C)$$

WUN ***Schwache Vereinigung*** $\quad \Im(A \mid Z \mid B \cup C) \Rightarrow \Im(A \mid Z \cup C \mid B)$

CON ***Kontraktion*** $\quad \Im(A \mid Z \mid B) \wedge \Im(A \mid Z \cup B \mid C) \Rightarrow \Im(A \mid Z \mid B \cup C)$

WTR ***Schwache Transitivität***

$$\Im(A \mid Z \mid B) \wedge \Im(A \mid Z \cup \{x\} \mid B) \Rightarrow \Im(A \mid Z \mid \{x\}) \vee \Im(\{x\} \mid Z \mid B)$$

CHO ***Kordalität***

$$\Im(a \mid c, d \mid b) \wedge \Im(c \mid a, b \mid d) \Rightarrow \Im(a \mid c \mid b) \vee \Im(a \mid d \mid b)$$

Markovdecken und Markovgrenzen

Definition

Sei $\Im$ ein Dependenzmodell auf V und $V = \{\mathbb{X}_1, \ldots, \mathbb{X}_N\}$ eine **Variablenordnung**.

- Eine Menge $B \subset V$ heißt **Markovdecke** von $c \in V$ bezüglich $A \subset V$ genau dann, wenn gilt:

$$B \subseteq A \quad \wedge \quad \Im(\{c\} \mid B \mid A \setminus B)$$

- Ist B minimal mit dieser Eigenschaft, so heißt B eine **Markovgrenze**.
- Die Folge $B_1, \ldots, B_N$ heißt **Grenzensystem** von $\Im$ bezüglich Variablenordnung $\mathbb{X}_1, \ldots, \mathbb{X}_N$ genau dann, wenn jede Menge B_n eine Markovgrenze von $\mathbb{X}_n$ bezüglich $V_n = \{\mathbb{X}_1, \ldots, \mathbb{X}_{n-1}\}$ ist.
- Ein gerichteter azyklischer Graph $\mathcal{G}$, dessen Eltermengen $\mathrm{pa}(\mathbb{X}_n)$ ein Grenzensystem von $\Im$ bilden, heißt **Grenzengraph** von $\Im$.

Markovdecken einer Markovkette: $\Im(\mathbb{X}_n \mid \{\mathbb{X}_{n-1}, \mathbb{X}_{n+1}\} \mid V \setminus \{\mathbb{X}_{n-1}, \mathbb{X}_n, \mathbb{X}_{n+1}\})$

Bayesnetzkonstruktion

Lemma (Verma 1986)

Ist $\Im$ ein Semigraphoid, so ist jeder Grenzengraph von $\Im$ ein Bayesnetz von $\Im$.

Ist $\Im$ ein Graphoid, so ist der Grenzengraph von $\Im$ ***bei gegebener Variablenordnung*** *eindeutig.*

$\mathcal{G}$ ist ein Bayesnetz für die Verteilung $\mathrm{P}(\cdot)$ genau dann, wenn er die LME für $\Im = \Im_{\mathrm{P}}$ besitzt und die Eltermengen $pa(\mathbb{X}_n)$ minimal mit dieser Eigenschaft sind (Markovgrenzen von $\mathbb{X}_n$ bzgl. $V \setminus off(\mathbb{X}_n)$).*

(Algorithmus)

1. Wähle eine Variablenordnung $V = \{\mathbb{X}_1, \ldots, \mathbb{X}_N\}$ aus.
2. Wähle $\mathbb{X}_1$ als Wurzel und ordne die Randverteilung $P_1(x_1)$ zu.
3. Für alle $i \geq 2$ berechne ein minimales B_i mit

$$B_i \subseteq \{\mathbb{X}_1, \ldots, \mathbb{X}_{i-1}\} \quad \text{und} \quad P(x_i \mid x_1, \ldots, x_{i-1}) = P(x_i \mid \boldsymbol{x}_{B_i})$$

und kreiere Knoten $\mathbb{X}_i$ mit der Vorgängermenge $\mathrm{pa}(\mathbb{X}_i) = B_i$ und der lokalen Verteilung $P_i(x_i \mid \boldsymbol{x}_{B_i})$.

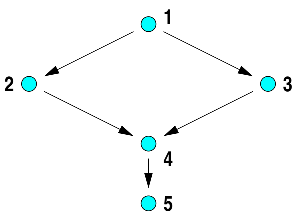
Markovdecken gegen den Rest der Welt

Fragestellung

In ungerichteten Graphen fallen die beiden folgenden Fragestellungen zusammen:

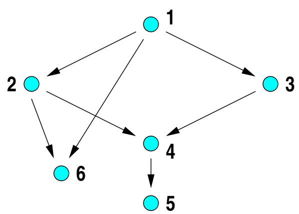
LME *Lokale Markoveigenschaft:* Gegen welche Variablen wird $a \in V$ durch seine unmittelbaren Nachbarn $\mathrm{bd}(a)$ abgeschirmt ?

AME *Allgemeine Markoveigenschaft:* Durch welche Menge wird $a \in V$ gegen den „*Rest der Welt*“ abgeschirmt ?

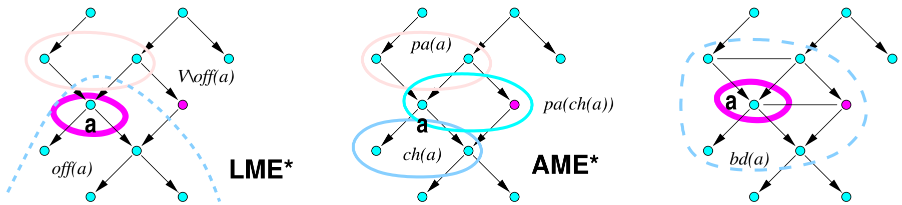


$\Im(1 \mid 2,3 \mid \text{„Rest“})$
$\Im(2 \mid 1,3,4 \mid \text{„Rest“})$
$\Im(3 \mid 1,2,4 \mid \text{„Rest“})$
$\Im(4 \mid 2,3,5 \mid \text{„Rest“})$
$\Im(5 \mid 4 \mid \text{„Rest“})$

$\Im(1 \mid 2,3,6 \mid \text{„Rest“})$
$\Im(2 \mid 1,3,4,6 \mid \text{„Rest“})$
$\Im(3 \mid 1,2,4 \mid \text{„Rest“})$
$\Im(4 \mid 2,3,5 \mid \text{„Rest“})$
$\Im(5 \mid 4 \mid \text{„Rest“})$
$\Im(6 \mid 1,2 \mid \text{„Rest“})$



Allgemeine Markoveigenschaft



Lemma (AME*)

Es sei $\mathcal{G}$ ein Bayesnetz für $\Im$. Für jedes $a \in V$ bildet die Vereinigung der folgenden Variablenmengen eine Markovdecke bzgl. V:

1. *die Menge $pa(a)$ der direkten Vorfahren von a,*
2. *die Menge $ch(a)$ der direkten Nachkommen von a,*
3. *die Menge der direkten Vorfahren der direkten Nachkommen von a.*

Mit anderen Worten:

$$\Im(a \mid pa(a) \cup ch(a) \cup pa(ch(a)) \setminus \{a\} \mid \text{„Rest“})$$

Beweis.

- $\mathcal{G}$ ist ein Bayesnetz von $\Im$, also insbesondere ein Unabhängigkeitsbild; folglich gilt die GME*.
- Wir haben also nur die Trennungseigenschaft

$$\mathrm{sep}_\delta \,\langle a \mid B_a \mid \text{„Rest"} \rangle_{\mathcal{G}} \ , \qquad B_a \overset{\text{def}}{=} \mathrm{pa}(a) \cup \mathrm{ch}(a) \cup \mathrm{pa}(\mathrm{ch}(a)) \setminus \{a\}$$

  zu zeigen.
- Die Trennungseigenschaft beweisen wir im Moralgraphen $(\mathcal{G})^{\mathrm{m}}$.

  Dort hat Knoten $a \in V$ als Nachbarn genau alle ehemaligen Eltern und Kinder des DAG sowie zusätzlich all jene Knoten, zu denen gemeinsame Kinder in $\mathcal{G}$ existieren, mit anderen Worten gilt:

$$\mathrm{bd}_{(\mathcal{G})^{\mathrm{m}}}(a) \;=\; B_a$$

- Selbstverständlich wird a im Moralgraphen $(\mathcal{G})^{\mathrm{m}}$ — wie in jedem UG wegen der LME — durch seinen Rand $\mathrm{bd}_{(\mathcal{G})^{\mathrm{m}}}(a)$ von allen Restknoten getrennt:

$$\mathrm{sep}_\delta \,\langle a \mid B_a \mid \text{„Rest"} \rangle_{(\mathcal{G})^{\mathrm{m}}}$$

  Damit ist die Behauptung gezeigt.

□

Graphische versus kausale Verteilungen

Lemma

1. *Es gibt graphische Verteilungen, die nicht kausal sind.*
2. *Es gibt kausale Verteilungen, die nicht graphisch sind.*
3. *Es gibt Verteilungen, die weder graphisch noch kausal sind.*
4. *Ist* $\mathrm{P} : \mathcal{X} \to \mathbb{R}$ *sowohl graphisch als auch kausal, so ist jedes Markovnetz von* $\Im_{\mathrm{P}}$ *kordal/trianguliert.*
5. *Ist* $\mathrm{P} : \mathcal{X} \to \mathbb{R}$ *sowohl graphisch als auch kausal, so ist jedes Bayesnetz von* $\Im_{\mathrm{P}}$ *moralisch.*

Beweisidee

$$\begin{array}{ccccc}
 & & \Im(\{a\} \mid \text{„Rest"} \mid \{b\})_{\mathrm{P}(\cdot)} & & \\
 & & \Uparrow & & \\
\mathrm{sep}\langle A|Z|B\rangle_{\mathcal{G}_{UG}} & \Leftrightarrow & \Im(A|Z|B)_{\mathrm{P}(\cdot)} & \Leftrightarrow & \mathrm{sep}_\delta \,\langle A|Z|B\rangle_{\mathcal{G}_{DAG}} \\
 & & & & \Updownarrow \\
 & & & & \mathrm{sep}\langle A|Z|B\rangle_{(\mathcal{G}_{DAG})^{\mathrm{m}}}
\end{array}$$

Beweis.

1. Jedes P mit dem Diamant-UG $f_1 \overset{m_1}{\underset{m_2}{\Diamond}} f_2$ als perfektem Bild.
2. Jedes P mit dem Konvergenz-DAG $w_1 \to g \leftarrow w_2$ als perfektem Bild.
3. Jede nichtkausale loglineare Verteilung mit der Modellformel

$$\mathrm{P}(a,b,c) \;=\; \phi_1(b,c) \cdot \phi_2(a,c) \cdot \phi_3(a,b)$$

   denn der **vollständige UG** ist das eindeutige Markovnetz zu P, enthält aber die $\{b,c\}$, $\{a,c\}$, $\{a,b\}$ nicht als Cliquen.
4. Wegen des Spezialfalls partieller Unabhängigkeiten besitzen $\mathcal{G}_{UG}$ und $(\mathcal{G}_{DAG})^{\mathrm{m}}$ identische Kanten, das Markovnetz ist also der Moralgraph des Bayesnetzes.

   $\mathcal{G}_{UG}$ muß dann aber auch kordal sein, denn jeder Kreis ≥ 4 muß im (azyklischen) Bayesnetz einen konvergierenden Knoten besitzen, folglich (aus Gründen der Moral) auch eine Sehne.
5. Im Falle der Unmoral gäbe es $a \to z \leftarrow b$, aber weder $a \to b$ noch $a \leftarrow b$. Für die *„historischen Abschlüsse"* Z von $\{z\}$ und W von $\{a,b\}$ gilt dann aber

$$\mathrm{sep}\langle \{a\} \mid W \backslash \{a,b\} \mid \{b\} \rangle_{(\mathcal{G}_W)^{\mathrm{m}}} \ ,$$

   aber nicht

$$\mathrm{sep}\langle \{a\} \mid Z \backslash \{a,b\} \mid \{b\} \rangle_{(\mathcal{G}_Z)^{\mathrm{m}}} \ ,$$

   ein eklatanter ⚡ zum SUN-Axiom (P graphisch!), da $W \backslash \{a,b\} \subset Z \backslash \{a,b\}$ gilt.

□

Beispiele

Markovnetze mit 3, 4, 5 oder 6 Variablen

• • •	$\mathrm{P}(x) \cdot \mathrm{P}(y) \cdot \mathrm{P}(z)$	3-diskret	⊕
•—• •	$\mathrm{P}(x,y) \cdot \mathrm{P}(z)$	2+1-diskret	⊕
•—•—•	$\mathrm{P}(x,y) \cdot \mathrm{P}(y,z) / \ \mathrm{P}(y)$	kaskadiert	⊕
$\triangle$	$\mathrm{P}(x,y,z)$	saturiert	⊕
$\Diamond$	$\phi(x,y) \cdot \phi(y,z) \cdot \phi(z,w) \cdot \phi(w,x)$	Diamant	⊖
$\lhd\rhd$	$\mathrm{P}(x,y,z) \cdot \mathrm{P}(y,z,w) / \ \mathrm{P}(y,z)$	3/3-Cliquen	⊕
$\rhd\lhd$	$\mathrm{P}(x,y,z) \cdot \mathrm{P}(v,w,z) / \ \mathrm{P}(z)$	3/3-Cliquen	⊕
$\triangle \equiv \triangle$	$\phi(x_1,x_2,x_3) \cdot \phi(y_1,y_2,y_3) \cdot \phi(x_1,y_1) \cdot \phi(x_2,y_2) \cdot \phi(x_3,y_3)$	Toblerone	⊖

Beispiele

Bayesnetze mit 3 oder 4 Variablen

• • •	$P(x) \cdot P(y) \cdot P(z)$	3-diskret	⊕
•→• •	$P(x) \cdot P(y \mid x) \cdot P(z)$	2+1-diskret	⊕
•→•→•	$P(x) \cdot P(y \mid x) \cdot P(z \mid y)$	kaskadiert	⊕
•←•→•	$P(x \mid y) \cdot P(y) \cdot P(z \mid y)$	divergent	⊕
•→•←•	$P(x) \cdot P(y \mid x, z) \cdot P(z)$	konvergent	⊖
△	$P(x) \cdot P(y \mid x) \cdot P(z \mid x, y)$	saturiert	⊕
◁▷	$P(x) \cdot P(y \mid x) \cdot P(z \mid x, y) \cdot P(w \mid y, z)$	3-3-Cliquen	⊕
◁▷	$P(x) \cdot P(y \mid x) \cdot P(w \mid y) \cdot P(z \mid x, y, w)$	unmoralisch!	⊖

Berechnung bedingter Wahrscheinlichkeiten

Verbundverteilung

Gemeinsame Verteilung $P(x_1, \ldots, x_n)$ **aller** Variablen in **Produktform**.

Randverteilungen

Gemeinsame Verteilung für eine Teilmenge $A \subset V$:

$$\begin{aligned}
P(V \setminus \{x_i\}) &= \sum_{x_i} P(x_1, \ldots, x_n) \\
P(V \setminus \{x_i, x_j\}) &= \sum_{x_i} \sum_{x_j} P(x_1, \ldots, x_n) \\
P(V \setminus \{x_{i_1}, \ldots, x_{i_m}\}) &= \sum_{x_{i_1}} \cdots \sum_{x_{i_m}} P(x_1, \ldots, x_n)
\end{aligned}$$

Bedingte Verteilungen

Einfluß einer Zufallsvariablen $\mathbb{X}_j$ auf eine andere $\mathbb{X}_i$:

$$P(x_i \mid x_j) = \frac{P(x_i, x_j)}{P(x_j)} = \frac{\sum \cdots \sum P(x_1, \ldots, x_n)}{\sum_{x_i} \sum \cdots \sum P(x_1, \ldots, x_n)}$$

Warum Bayesnetze ?

Weil sie in Wahrscheinlichkeiten faktorisieren !

Was ist Inferenz ?

Logik	Axiome, Schlußregeln ➪	neue Sätze
Arithmetik	Parameterwerte, Operationen ➪	Funktionswerte
Stochastik	Observablen, W-Modell ➪	bedingte W'keiten

A posteriori Verteilungen

Evidenz $E = \{e_1, \ldots, e_m\}$ („ *instanziierte*" Variablen)

$$P(x_i \mid E) = P(x_i = \xi \mid e_1 = \eta_1, \ldots, e_m = \eta_m)$$

Rand- und Rückschlußverteilungen sind aufwendig zu berechnen!

- Eliminiere Variablen in ökonomischer Reihenfolge — gemäß Dependenzstruktur bzw. Modellformel.
- Propagationsalgorithmen, Marker-Passing, Sampling ...

Notation der Rechengrößen

für baumförmige Bayesnetze

Wahrscheinlichkeitsparametermatrix

Jeder Knoten y im DAG hat **genau einen** Elterknoten x.

$$\begin{aligned} \boldsymbol{M}_{y|x} &= \mathrm{P}(y|x) = [\mathrm{P}(y=\eta_j \mid x=\xi_i)]_{ij} \\ &= \begin{Bmatrix} \mathrm{P}(y=\eta_1 \mid x=\xi_1) & \cdots & \mathrm{P}(y=\eta_k \mid x=\xi_1) \\ \vdots & & \vdots \\ \mathrm{P}(y=\eta_1 \mid x=\xi_m) & \cdots & \mathrm{P}(y=\eta_k \mid x=\xi_m) \end{Bmatrix} \end{aligned}$$

Evidenz

Instanziierte Variablen $e \in V$ bzw. $E \subseteq V$.

Belief-Funktion

Subjektive Einschätzung von x auf Grundlage von E (Wahr'keitsfeld):

$$\begin{aligned} \mathrm{bel}(x) &\overset{\text{def}}{=} \mathrm{P}(x|E) \\ \mathrm{bel}(x) &= \mathrm{P}(x \mid z=\zeta) = (\mathrm{P}(x=\xi_1 \mid z=\zeta), \ldots, \mathrm{P}(x=\xi_\ell \mid z=\zeta))^\top \end{aligned}$$

Unidirektionale Fortpflanzung in Ketten

Zwei Variablen

Beispiel: $x \to y$, Evidenz $\{y = \eta\}$

Nach der Bayesformel gilt:

$$\mathrm{bel}(x) = \mathrm{P}(x \mid y=\eta) = \frac{\mathrm{P}(x)\cdot\mathrm{P}(y=\eta \mid x)}{\mathrm{P}(y=\eta)} \propto \mathrm{P}(x)\cdot\boldsymbol{\lambda}(x)$$

mit der a priori Wahrsch'keit $\mathrm{P}(x)$ und dem **diagnostischen Vektor**

$$\boldsymbol{\lambda}(x) = \mathrm{P}(y=\eta \mid x) \qquad (\eta\text{-te Spalte der Matrix } \boldsymbol{M}_{y|x}).$$

$\mathrm{P}(x)\cdot\boldsymbol{\lambda}(x)$ bezeichnet das komponentenweise Produkt.

Unidirektionale Fortpflanzung in Ketten

Drei Variablen

Beispiel: $x \to y \to z$, Evidenz $\{z = \zeta\}$

Nach der Bayesformel gilt wiederum:

$$\mathrm{bel}(x) = \mathrm{P}(x \mid z=\zeta) = \frac{\mathrm{P}(x)\cdot\mathrm{P}(z=\zeta \mid x)}{\mathrm{P}(z=\zeta)} \propto \mathrm{P}(x)\cdot\boldsymbol{\lambda}(x)$$

Der **diagnostische Vektor** lautet nunmehr

$$\begin{aligned} \boldsymbol{\lambda}(x) = \mathrm{P}(z=\zeta \mid x) &= \sum_y \mathrm{P}(z=\zeta, y \mid x) \\ &= \sum_y \mathrm{P}(z=\zeta \mid y)\cdot\mathrm{P}(y|x) \\ &= \boldsymbol{M}_{y|x} \bullet \boldsymbol{\lambda}(y) \end{aligned}$$

$\boldsymbol{M}_{y|x} \bullet \boldsymbol{\lambda}(y)$ bezeichnet das Vektor-Matrix-Produkt über die Variable y.

Unidirektionale Fortpflanzung in Ketten

Mehr als drei Variablen

Beispiel: $x_1 \to x_2 \to \ldots \to x_n$, Evidenz $\{x_n = \xi\}$

Nach der Bayesformel gilt wiederum:

$$\mathrm{bel}(x_1) = \mathrm{P}(x_1 \mid x_n=\xi) = \frac{\mathrm{P}(x_1)\cdot\mathrm{P}(x_n=\xi \mid x_1)}{\mathrm{P}(x_n=\xi)} \propto \mathrm{P}(x_1)\cdot\boldsymbol{\lambda}(x_1)$$

Der **diagnostische Vektor** gehorcht der Rekursion:

$$\begin{aligned} \boldsymbol{\lambda}(x_1) &= \boldsymbol{M}_{x_2|x_1} \bullet \boldsymbol{\lambda}(x_2) \\ &= \boldsymbol{M}_{x_2|x_1} \bullet \boldsymbol{M}_{x_3|x_2} \bullet \boldsymbol{\lambda}(x_3) \\ &= \boldsymbol{M}_{x_2|x_1} \bullet \boldsymbol{M}_{x_3|x_2} \bullet \boldsymbol{M}_{x_4|x_3} \bullet \boldsymbol{\lambda}(x_4) \\ &= \boldsymbol{M}_{x_2|x_1} \bullet \boldsymbol{M}_{x_3|x_2} \bullet \ldots \bullet \boldsymbol{M}_{x_{n-1}|x_{n-2}} \bullet \underbrace{\mathrm{P}(x_n=\xi \mid x_{n-1})}_{\boldsymbol{M}_{\xi|x_{n-1}}} \end{aligned}$$

Bidirektionale Fortpflanzung in Ketten

Beispiel: $e^+ \to v \to w \to x \to y \to z \to e^-$

A posteriori Wahrscheinlichkeiten nach Bayesformel:

$$\begin{aligned} \mathrm{bel}(x) \;=\; \mathrm{P}(x \mid e^+, e^-) &\;\propto\; \mathrm{P}(e^- \mid x, e^+) \cdot \mathrm{P}(x|e^+) \\ &\;=\; \mathrm{P}(e^-|x) \cdot \mathrm{P}(x|e^+) \;=\; \boldsymbol{\lambda}(x) \cdot \boldsymbol{\pi}(x) \end{aligned}$$

Diagnostische Evidenz $\boldsymbol{\lambda}(x) = \mathrm{P}(e^-|x)$

Kausale Evidenz $\boldsymbol{\pi}(x) = \mathrm{P}(x|e^+)$

Fortpflanzung rückwärts

$$\begin{aligned} \boldsymbol{\pi}(x) &= \mathrm{P}(x|e^+) \\ &= \sum_w \mathrm{P}(x \mid w, e^+) \cdot \mathrm{P}(w|e^+) \\ &= \sum_w \mathrm{P}(x|w) \cdot \mathrm{P}(w|e^+) \\ &= \boldsymbol{\pi}(w) \bullet \boldsymbol{M}_{x|w} \end{aligned}$$

Fortpflanzung vorwärts

$$\begin{aligned} \boldsymbol{\lambda}(x) &= \mathrm{P}(e^-|x) \\ &= \sum_y \mathrm{P}(e^-, y \mid x) \\ &= \sum_y \mathrm{P}(e^-|y) \cdot \mathrm{P}(y|x) \\ &= \boldsymbol{M}_{y|x} \bullet \boldsymbol{\lambda}(y) \end{aligned}$$

Bidirektionale Fortpflanzung in Bäumen

Zerlegung der Belief-Funktion

Zerlegung der Evidenz

Für $x \in V$ unterscheiden wir zwei Quellgebiete:

$$E \;=\; E_x^+ \uplus E_x^- \qquad \text{mit} \quad \begin{cases} E_x^- \subset \mathrm{ch}(x) & \text{„flußabwärts“} \\ E_x^+ \subset V \backslash \mathrm{ch}(x) & \text{„flußaufwärts“} \end{cases}$$

Belief-Funktion

Nach Kettenregel und $\mathrm{sep}_\delta \langle E_x^- \mid \{x\} \mid E_x^+ \rangle$ folgt:

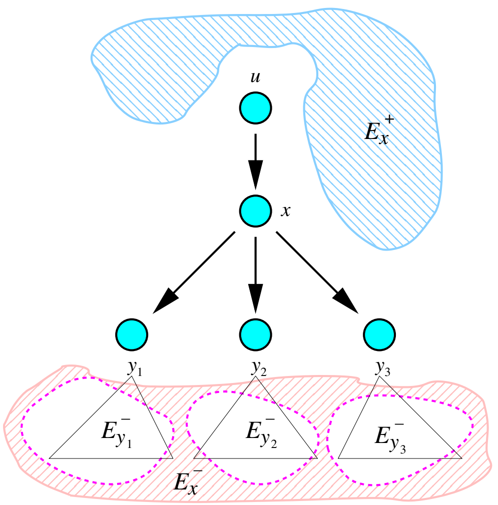
$$\begin{aligned} \mathrm{bel}(x) &= \mathrm{P}(x \mid E_x^+, E_x^-) \\ &\propto \mathrm{P}(E_x^-, x \mid E_x^+) \\ &= \mathrm{P}(E_x^- \mid x, E_x^+) \cdot \mathrm{P}(x|E_x^+) \;=\; \boldsymbol{\lambda}(x) \cdot \boldsymbol{\pi}(x) \end{aligned}$$

$\boldsymbol{\pi}(x) =$ **kausale** Unterstützung von x durch die Vorgänger

$\boldsymbol{\lambda}(x) =$ **diagnostische** Unterstützung von x durch die Nachfolger

Bidirektionale Fortpflanzung in Bäumen

Zerlegung der Evidenz



Vertikale Zerlegung

kausal/diagnostisch:

$$E_x \;=\; E_x^+ \uplus E_x^-$$

Horizontale Zerlegung

des diagnostischen Teils

$$E_x^- \;=\; \biguplus_{y_\ell \in \mathrm{ch}(x)} E_{y_\ell}^-$$

Horizontale Zerlegung

des kausalen Teils

$$E_{y_\ell}^+ \;=\; E_x^+ \uplus \biguplus_{k \neq \ell} E_{y_k}^-$$

Bidirektionale Fortpflanzung in Bäumen

Diagnostische und prädiktive Wahrscheinlichkeiten

Diagnostische Komponente

Seien $u_1, \ldots, u_r$ die Nachfolger von x:

$$\boldsymbol{\lambda}(x) \;=\; \mathrm{P}(E_x^-|x) \;=\; \mathrm{P}(E_{u_1}^-, \ldots, E_{u_r}^- \mid x) \;=\; \prod_{s=1}^{r} \underbrace{\mathrm{P}(E_{u_s}^-|x)}_{\boldsymbol{\lambda}_{u_s}(x)}$$

Falls $\{x = \xi\}$ selbst instanziiert, so erzeuge Dummyknoten d mit $\boldsymbol{\lambda}_d(x) = \mathbf{I}_{x=\xi}$.

Prädiktive Komponente

Sei $u \in V$ der Vater (die Mutter) von x:

$$\begin{aligned} \boldsymbol{\pi}(x) \;=\; \mathrm{P}(x|E_x^+) &= \sum_u \mathrm{P}(x, u \mid E_x^+) \\ &= \sum_u \mathrm{P}(x|u) \cdot \mathrm{P}(u|E_x^+) \;=:\; \boldsymbol{M}_{x|u} \bullet \boldsymbol{\pi}_x(u) \end{aligned}$$

Bidirektionale Fortpflanzung in Bäumen

Variablenversetzte diagnostische und prädiktive Komponenten

Berechnung von $\boldsymbol{\lambda}_x(u)$

für $u \to x$

$$\begin{aligned}\boldsymbol{\lambda}_x(u) &= \sum_x \mathrm{P}(E_x^- \mid u, x) \cdot \mathrm{P}(x|u)\\ &= \sum_x \mathrm{P}(E_x^- | x) \cdot \mathrm{P}(x|u)\\ &= \sum_x \boldsymbol{\lambda}(x) \cdot \mathrm{P}(x|u)\\ &= \boldsymbol{M}_{x|u} \bullet \boldsymbol{\lambda}(x)\end{aligned}$$

Berechnung von $\boldsymbol{\pi}_y(x)$

für $u \to x$ und $y \leftarrow x \to z$

$$\begin{aligned}\boldsymbol{\pi}_y(x) &= \mathrm{P}(x|E_y^+)\\ &= \mathrm{P}(x \mid E_x^+, E_z^-)\\ &\propto \mathrm{P}(E_z^- \mid x, E_x^+) \cdot \mathrm{P}(x|E_x^+)\\ &= \boldsymbol{\lambda}_z(x) \cdot \boldsymbol{\pi}(x)\\ &= \boldsymbol{\lambda}_z(x) \cdot \boldsymbol{M}_{x|u} \bullet \boldsymbol{\pi}_x(u)\end{aligned}$$

Spezialfall: $x = \xi$ evident

$$\boldsymbol{\lambda}_x(u) = \mathrm{P}(x = \xi \mid u)$$

(ξ-te Spalte von Matrix $\boldsymbol{M}_{x|u}$)

Allgemeinfall: ≥ 3 Kinder

$$\boldsymbol{\pi}_y(x) = \boldsymbol{\pi}(x) \cdot \sum_{z \neq y} \boldsymbol{\lambda}_z(x)$$

Vorwärts-Rückwärts-Algorithmus

in baumförmigen Bayesnetzen

Start

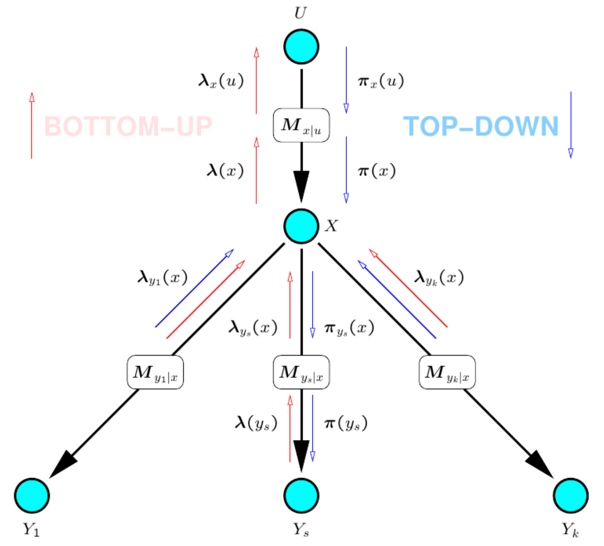
$\boldsymbol{\pi}(x_0) = \boldsymbol{M}_{x_0|\cdot}$ (Wurzel)
$\boldsymbol{\lambda}(x_\ell) = \mathbf{1}$ (Blatt)
$\boldsymbol{\lambda}(x_e) = \boldsymbol{e}^{(\zeta)}$ (Evidenz)

Bottom-up

$\boldsymbol{\lambda}(y_k)$ (I.V.)
$\boldsymbol{\lambda}_{y_k}(x) = \boldsymbol{M}_{y_k|x} \bullet \boldsymbol{\lambda}(y_k)$
$\boldsymbol{\lambda}(x) = \prod_k \boldsymbol{\lambda}_{y_k}(x)$

Top-down

$\boldsymbol{\pi}(x)$ (I.V.)
$\boldsymbol{\pi}_{y_k}(x) \propto \boldsymbol{\pi}(x) \cdot \prod_{\ell \neq k} \boldsymbol{\lambda}_{y_\ell}(x)$
$\boldsymbol{\pi}(y_k) = \boldsymbol{M}_{y_k|x} \bullet \boldsymbol{\pi}_{y_k}(x)$



Inferenz in moralischen Bayesnetzen

Vorwärts-Rückwärts-Algorithmus über Variablenkomplexen

(Algorithmus)

1. ENTFERNE ALLE KANTENRICHTUNGEN
   $\leadsto$ äquivalentes kordales Markovnetz
2. BILDE VETRÄGLICHEN VERBUNDBAUM
   mit Cliquensequenz $\mathcal{C} = \{C_1, \ldots, C_K\}$
3. KONSTRUIERE VARIABLENKOMPLEXE
   $\mathbb{Y}_k := \bigotimes_{\mathbb{X}_j \in B_k} \mathbb{X}_j$ mit $B_k := C_k \setminus C_{k-1}$
4. EXPANDIERE VERTEILUNGSPARAMETER
   $\boldsymbol{M}_{k|\ell} = (\mathrm{P}(\mathbb{Y}_k = \boldsymbol{\eta} | \mathbb{Y}_\ell = \boldsymbol{\zeta}) \mid \boldsymbol{\eta} \in \mathcal{Y}_k, \boldsymbol{\zeta} \in \mathcal{Y}_\ell)$
5. EXEKUTIERE VR-ALGORITHMUS AUF VB

Moralische Bayesnetze

Verbundbaum:
spezielle $\boldsymbol{x}/y$-Kombinationen

Markovnetz:
nur Imputation!

Unmoralische Bayesnetze

Schummeln:
Verheiraten aller Elternpaare

Monte Carlo:
Iteratives Auswürfeln und Neuschätzen

Spezialfall Imputation

Vorhersage eines Attributwertes aus allen anderen

Belief-Funktion

mit Zielvariable x_k und Evidenzvariablen $E = V \setminus \{x_k\}$:

$$\mathrm{bel}(x_k)_\xi = \mathrm{P}(x_k = \xi \mid \boldsymbol{x}_E) = \frac{\mathrm{P}(x_k = \xi, \boldsymbol{x}_E)}{\mathrm{P}(\boldsymbol{x}_E)} = \frac{\mathrm{P}(\boldsymbol{x}_{|x_k=\xi})}{\mathrm{P}(\boldsymbol{x}_E)}$$

$\mathbb{X}_k$ ist diskretes Attribut

1. Für alle $\xi_\ell \in \mathcal{X}_k$ berechne $q_\ell = \mathrm{P}(\boldsymbol{x}_{|x_k=\xi_\ell})$ mit
   $$\boldsymbol{x}_{|x_k=\xi_\ell} = (x_1, \ldots, x_{k-1}, x_k = \xi_\ell, x_{k+1}, \ldots, x_n)^\top \in \mathbb{R}^n\,.$$
2. Setze $\mathrm{bel}(x_k)_\ell = q_\ell \,/\, \sum_i q_i\,.$

$\mathbb{X}_k$ ist stetiges Attribut

Effiziente Lösungsmöglichkeit trotz $|\mathcal{X}_k| = \infty$?

Spezialfall Imputation

Vorhersage eines normalverteilten Attributwertes

Die Geheimfunktion

Es ist $\mathrm{P}(x_k = \xi \mid \boldsymbol{x}_E) = \mathcal{N}(\xi \mid \mu, \sigma^2)$, also erhalten wir Resultate der Form

$$\mathrm{P}(\boldsymbol{x}_{|x_k=\xi}) \;=\; \mathrm{P}(\boldsymbol{x}_E)\cdot\mathrm{P}(\xi|\boldsymbol{x}_E) \;=\; \underbrace{c\cdot\mathcal{N}(\xi \mid \mu, \sigma^2)}_{g_{c,\mu,\sigma}(\xi)}$$

durch Auswertung des Bayesnetzes an der Stelle $\xi \in \mathbb{R}$.

Lemma

Die unbekannten Parameter $c > 0$, $\mu \in \mathbb{R}$ und $\sigma > 0$ der skalierten univariaten Gaußdichte

$$g_{c,\mu,\sigma}(\xi) \;\stackrel{def}{=}\; c\cdot\mathcal{N}(\xi \mid \mu, \sigma^2)$$

können aus den Funktionswerten von $g(\cdot)$ an vier reellen Stützstellen bestimmt werden.

Diese Entschlüsselungstechnik läßt sich auf **multivariate** Gaußdichten verallgemeinern.

Beweis.

Wir definieren die *lograt*-Funktion $\ell(x,y) = -2\cdot\log(g(x)/g(y))$ und folgern die Identität

$$\ell(x,y) \;=\; 1/\sigma^2\cdot(x^2 - y^2 - 2\mu\cdot(x-y)) \;.$$

Wir definieren nun die Differentiale

$$\ell_h^-(x) \;\stackrel{\text{def}}{=}\; \ell(x, x-h) \;=\; 1/\sigma^2\cdot(+2hx - h^2 - 2h\mu)$$
$$\ell_h^+(x) \;\stackrel{\text{def}}{=}\; \ell(x, x+h) \;=\; 1/\sigma^2\cdot(-2hx - h^2 + 2h\mu)$$

für $h > 0$ und finden nach deren Addition einen Lösungsausdruck

$$\hat{\sigma}^2 \;=\; -2\cdot\frac{h^2}{\ell_h^+(x) + \ell_h^-(x)}$$

für die gesuchte Varianz. Anschließend können wir aus jeder der Differentialformeln den Erwartungswert berechnen, z.B.:

$$\hat{\mu} \;=\; \frac{\hat{\sigma}^2\cdot\ell_h^+(x) + 2hx + h^2}{2h} \;=\; \frac{\hat{\sigma}^2}{2h}\cdot\ell_h^+(x) + x + \frac{h}{2}$$

Schließlich bestimmen wir noch den Skalierungsfaktor c; die numerisch stabilste Methode besteht in einer weiteren Auswertung der Geheimfunktion $g(\cdot)$, und zwar am Dichtegipfel:

$$\hat{c} \;=\; \frac{g(x)}{\mathcal{N}(x \mid \hat{\mu}, \hat{\sigma}^2)} \;=\; \frac{g(\hat{\mu})}{\mathcal{N}(\hat{\mu} \mid \hat{\mu}, \hat{\sigma}^2)} \;=\; \frac{g(\hat{\mu})}{\mathcal{N}(0 \mid 0, \hat{\sigma}^2)} \;=\; \sqrt{2\pi}\cdot\hat{\sigma}\cdot g(\hat{\mu})$$

□

Imputation in (nichtkordalen) Markovnetzen

Effiziente Berechnung als $\mathrm{bd}(x_k)$-ausgedünntes Cliquenprodukt

Bedingte Wahrscheinlichkeit nach Faktorisierung

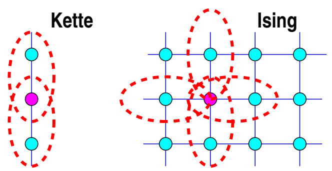
$$\mathrm{P}(\mathbb{X}_k = x_k \mid \mathbb{X}_{V\setminus k} = \boldsymbol{x}') \;=\; \frac{\mathrm{P}(x_k, \boldsymbol{x}')}{\mathrm{P}(\boldsymbol{x}')} \;=\; \frac{\prod_{C\in\mathcal{C}} \phi_C(x_k, \boldsymbol{x}')}{\sum_{\xi\in\mathcal{X}_k}\prod_{C\in\mathcal{C}} \phi_C(x_k, \boldsymbol{x}')}$$

Ausklammern & Kürzen aller Gibbspotenziale ϕ_C mit $x_k \notin C$

Reduzierte Faktorisierung über $\mathcal{C}_{(k)} := \{C \mid x_k \in C\} = \{C \mid C \subseteq \mathrm{cl}(x_k)\}$ (wegen $\Im(x_k \mid \mathrm{bd}(x_k) \mid$ *„Rest“*$)$ nicht ganz unerwartet!)

Binäres Zielattribut $|\mathcal{X}_k| = 2$

$$\log\mathrm{odds}(\boldsymbol{x}') = \log\frac{\mathrm{P}(1|\boldsymbol{x}')}{\mathrm{P}(0|\boldsymbol{x}')} = \sum_{C\in\mathcal{C}_{(k)}} \log\frac{\phi_C(\boldsymbol{x}_{C\setminus k}, 1)}{\phi_C(\boldsymbol{x}_{C\setminus k}, 0)}$$



Diskrete loglineare Modelle

Spezialfall: drei Variablen ($N = 3$)

Dreiwegetabellen

Drei diskrete Zufallsvariablen $V = \{\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3\} = \{a, b, c\}$

- Endliche Wertebereiche $\mathcal{X}_a, \mathcal{X}_b, \mathcal{X}_c$
- Endlich viele Zellen $(j, k, l) \in \mathcal{X}_a \times \mathcal{X}_b \times \mathcal{X}_c$
- Würfel $\{p_{jkl}\}$ von Wahrscheinlichkeiten $\sum_j \sum_k \sum_l p_{jkl} = 1$
- Würfel $\{n_{jkl}\}$ von (absoluten) Häufigkeiten $\sum_j \sum_k \sum_l n_{jkl} = T$

Loglineares Verteilungsmodell

Produktform $\Delta \subset \mathfrak{P}V$

$$p_{jkl} = \prod_{A \in \Delta} \underbrace{\phi_A(\boldsymbol{x}_A)}_{z^A_{jkl}}$$

Summenform

$$\log p_{jkl} = \sum_{A \in \Delta} \underbrace{\log \phi_A(\boldsymbol{x}_A)}_{u^A_{jkl}}$$

Beispiele — Dreiwegemodelle

Menge der (maximalen) Interaktionsterme · „Generatoren“

Unabhängiges Modell

$\boxed{a, b, c}$

$$\log p_{jkl} = u + u^a_j + u^b_k + u^c_l$$
$$p_{jkl} = \mathrm{P}(a = \alpha_j) \cdot \mathrm{P}(b = \beta_k) \cdot \mathrm{P}(c = \chi_l) = p_{j..} \cdot p_{.k.} \cdot p_{..l}$$

Kettenförmiges Modell

$\boxed{ab, ac}$

$$\log p_{jkl} = u + u^a_j + u^b_k + u^c_l + u^{ab}_{jk} + u^{ac}_{jl}$$
$$p_{jkl} = \frac{p_{jk.} \cdot p_{j.l}}{p_{j..}} \quad \text{bzw.} \quad \frac{p_{jkl}}{p_{j..}} = \frac{p_{jk.}}{p_{j..}} \cdot \frac{p_{j.l}}{p_{j..}}$$

Saturiertes Modell

$\boxed{abc}$

$$\log p_{jkl} = u + u^a_j + u^b_k + u^c_l + u^{ab}_{jk} + u^{ac}_{jl} + u^{bc}_{kl} + u^{abc}_{jkl}$$

Schätzung der kanonischen Modellparameter

Normierungseigenschaft

$$1 \stackrel{!}{=} \sum_{jkl} p_{jkl} = \sum_{jkl} \exp\left\{\sum_{A \in \Delta} u^A_{jkl}\right\} = e^u \cdot \sum_{jkl} \exp\left\{\sum_{A \neq \varnothing} u^A_{jkl}\right\}$$

Multinomial gezogene Stichprobe

$$P(\boldsymbol{n}|\boldsymbol{p}) = P(\{n_{jkl}\} \mid \{p_{jkl}\}) = \frac{T!}{\prod_{j,k,l} n_{jkl}!} \cdot \prod_{j,k,l} p_{jkl}^{n_{jkl}}$$

Logarithmierte Likelihood-Funktion

$$\ell_{\mathrm{ML}}(\boldsymbol{n}|\boldsymbol{p}) = \log \frac{T!}{\prod_{j,k,l} n_{jkl}!} + \sum_{j,k,l} n_{jkl} \log p_{jkl}$$

Maximum-Likelihood-Schätzwerte

Kanonische Verteilungsparameter für das saturierte Modell

$$\hat{\boldsymbol{p}} = \operatorname*{argmax}_{\boldsymbol{p}} \ell_{\mathrm{ML}}(\boldsymbol{n}|\boldsymbol{p}) \quad \Rightarrow \quad \hat{p}_{jkl} = \frac{n_{jkl}}{T}$$

Diskrete Loglinearmodelle

Definition

Die Familie diskreter Wahrscheinlichkeitsfunktionen $\{p_x\}_{x \in \Omega}$ der Gestalt

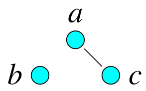
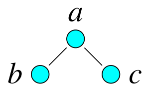
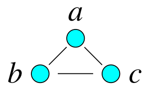
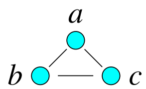
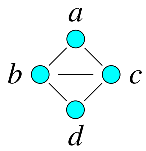
$$\log p_x = \sum_{A \in \Delta} u^A_x, \qquad \boldsymbol{x} \in \boldsymbol{\Omega}, \qquad \Delta \subseteq \mathfrak{P}V$$

heißt **Loglinearmodell** mit der Menge Δ von **Interaktionstermen**.

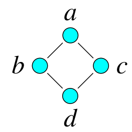
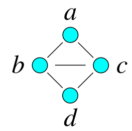
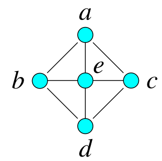
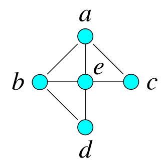
1. Das Loglinearmodell heißt **hierarchisch**, falls gilt:
   $$A \subseteq B \quad \text{und} \quad B \in \Delta \quad \Rightarrow \quad A \in \Delta$$
2. Das Loglinearmodell heißt **graphisch**, wenn gilt:
   $$C \in \Delta \quad \Leftrightarrow \quad \forall a, b \in C : \{a, b\} \in \Delta$$
3. Ein graphisches LLM heißt **zerlegbar**, wenn sein Graph kordal ist.

Die maximalen Interaktionsterme eines hierarchischen Loglinearmodells heißen **Generatoren**. Die Generatorenmenge wird auch als **Modellformel** bezeichnet.

Beispiele — Loglinearmodelle I

	GRAPHISCH	ZERLEGBAR	*b, ac*
	GRAPHISCH	ZERLEGBAR	*ab, ac*
	□	□	*ab, ac, bc*
	GRAPHISCH	ZERLEGBAR	*abc*
	GRAPHISCH	ZERLEGBAR	*abc, bcd*

Beispiele — Loglinearmodelle II

	GRAPHISCH	□	*ab, ac, bd, cd*
	□	□	*ab, bcd, ac*
	GRAPHISCH	□	*abe, ace, bde, cde*
	GRAPHISCH	□	*ab, ac, bde, ce*

Elementare und marginale Ereignisse

Häufigkeit und charakteristische Funktion

Definition

Es sei $\{n_x\}_{x\in\Omega}$ die Tafel elementarer Ereignishäufigkeiten über V.
Das Zahlenfeld $\{n_{x_A}\}_{x_A\in\Omega_A}$ für eine Variablenmenge $A \subseteq V$ mit Einträgen

$$n_{x_A} \stackrel{\text{def}}{=} \sum_{x_{A'}\in\Omega_{A'}} n_x \,, \qquad A' = V \setminus A$$

heißt **marginale Tafel** oder Tabelle für A.

Definition

Sei $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$, $A \subseteq V$ und $\boldsymbol{x}_A \in \boldsymbol{\Omega}_A$ ein marginales Ereignis.
Die zweiwertige Abbildung

$$\varphi_{x_A} : \begin{cases} \boldsymbol{\Omega} & \to \quad \{1, 0\} \\ \boldsymbol{\xi} & \mapsto \quad \begin{cases} 1 & \xi_j = x_j \text{ für alle } j \in A \\ 0 & \text{sonst} \end{cases} \end{cases}$$

heißt **charakteristische Funktion** von $\boldsymbol{x}_A$.

Maximum-Entropie-Prinzip

Edwin Thompson Jaynes, 1957

Satz (ML$\hat{=}$ME)

Es sei ein hierarchisches loglineares Modell

$$\log p_x = \sum_{A\in\Delta} u_x^A$$

gegeben sowie die Häufigkeitstafel $\{n_x\}_{x\in\Omega}$ der Daten $\omega \subset \boldsymbol{\Omega}$.

1. *Die **Maximum-Likelihood**-Parameter $\left\{u_x^A\right\}_{x_A}$ des Modells erfüllen die Bedingungsgleichungen (*)*

$$\mathcal{E}[\varphi_{x_A}(\mathbb{X}) \mid \boldsymbol{u}] = \frac{n_{x_A}}{T} \,, \qquad A \in \Delta,\ \boldsymbol{x} \in \boldsymbol{\Omega} \,.$$

2. *Unter allem Wahrscheinlichkeitsverteilungen, die das Gleichungssystem (*) erfüllen, hat obiges loglineare Modell mit Parametern $\left\{u_x^A\right\}_{x_A}$ die **maximale Entropie**.*

„Unter allen Zuständen eines physikalischen Systems, die kompatibel mit dem vorhandenen Wissen sind, ist der zu wählen, welcher die Entropie maximiert. “

Lernen der Loglinearparameter

GIS-Algorithmus — Generalized Iterative Scaling

Satz (Deming & Stephan, 1940)

Mit der abkürzenden Schreibweise $z_x^A = \exp(u_x^A)$ *gilt:*
Das Iterationsverfahren

$$z_x^A \leftarrow z_x^A \cdot \left(\frac{n_{x_A}/T}{\mathcal{E}[\varphi_{x_A}(\mathbb{X})]} \right)^{1/|\Delta|} = z_x^A \cdot \left(\frac{\sum_{y\in\Omega} \varphi_{x_A}(\boldsymbol{y}) \cdot \frac{n_y}{T}}{\sum_{y\in\Omega} \varphi_{x_A}(\boldsymbol{y}) \cdot \prod_{B\in\Delta} z_y^B} \right)^{1/|\Delta|}$$

mit den Startwerten $z_x^A \equiv 1$ *konvergiert gegen die Maximum-Likelihood-Schätzwerte des loglinearen Modells.*

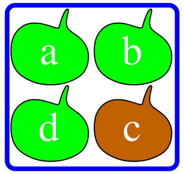
Bemerkung

Die Gleichung für $\varnothing \in \Delta$ garantiert $\sum p_x = 1$.
Das Bedingungssystem ist konsistent: alle $C_y = \sum \varphi_{x_A}(\boldsymbol{y})$ sind gleich $|\Delta|$.
Beweis $\rightsquigarrow$ Skriptum „*Stochastische Grammatikmodelle*"

Beispiel — Tafelobst im Tetrapack

Diamantenes Verteilungsmodell

für die vier frischen/faulen Äpfel:

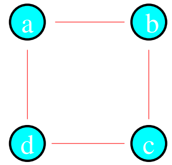


$$\mathrm{P}: \begin{cases} \{0,1\}^4 & \rightarrow \quad [0,1] \\ (\alpha,\beta,\zeta,\delta) & \mapsto \quad z \cdot z^{ab}_{\alpha\beta} \cdot z^{bc}_{\beta\zeta} \cdot z^{cd}_{\zeta\delta} \cdot z^{da}_{\delta\alpha} \end{cases}$$

Datensammlung und Statistiken

Absolute Häufigkeiten

$$\{n_{\alpha\beta\zeta\delta} \mid \alpha,\beta,\zeta,\delta \in \{0,1\}\}$$



Minimale suffiziente Statistiken, z.B. für $ab \in \Delta$:

$$n^{ab}_{\alpha\beta} = \sum_{a,b,c,d} \delta_{a=\alpha} \cdot \delta_{b=\beta} \cdot n_{abcd}$$

Symmetrie I

$$n^{ab}_{\xi\eta} = n^{bc}_{\xi\eta} = n^{cd}_{\xi\eta} = n^{da}_{\xi\eta} = n_{\xi\eta}$$

Symmetrie II

$$n_{01} = n_{10}$$

Beispiel — Tafelobst im Tetrapack

Daten = 100 Obstkörbe

	0:4	1:3	2:2	3:1	4:0	
n_{00}	60	2	1	0	0	63
n_{01}	0	1	2	1	0	4
n_{10}	0	1	2	1	0	4
n_{11}	0	0	1	2	26	29
	60	4	6	4	26	100

Iterationsanfang

$$z_{00} = z_{01} = z_{11} = 1$$

Iterationsschritt

$$z_{\xi\eta} \leftarrow z_{\xi\eta} \cdot \left(\frac{n_{\xi\eta}/100}{\mathcal{E}[\varphi_{\xi\eta}(\mathbb{X})]} \right)^{1/5}$$

Generalized Iterative Scaling

	Loglinearparameter				Wahrscheinlichkeiten in Promille				
i	z_{00}	z_{01}	z_{11}	$1/z$	$\binom{00}{00}$	$\binom{10}{00}$	$\binom{11}{00}$	$\binom{11}{10}$	$\binom{11}{11}$
0	1	1	1	16	62.5	62.5	62.5	62.5	62.5
1	1.2	0.693	1.03	10.9	192	63.9	54.7	46.8	103
2	1.33	0.509	1.06	9.05	351	51.1	40.5	32.2	139
3	1.4	0.402	1.1	8.39	460	37.8	29.5	23.1	172
4	1.43	0.339	1.13	8.15	517	28.9	22.8	18	201
6	1.46	0.279	1.17	8.03	561	20.6	16.6	13.4	235
9	1.47	0.252	1.19	8.01	579	17.1	13.9	11.3	251
12	1.47	0.245	1.19	8	584	16.3	13.2	10.7	255
16	1.47	0.244	1.2	8	585	16	13	10.6	256
20	1.47	0.243	1.2	8	585	16	13	10.6	256
saturiertes Modell:					600	10	10	10	260

MLS $\hat{=}$ relative Ereignishäufigkeiten

Happy End — für alle kausalen Verteilungen

Zerlegbare Loglinearmodelle

Cliquen $\mathcal{C} = \{C_1, \ldots, C_M\}$

$$\mathrm{P}(\boldsymbol{x}) = \prod_{C\in\mathcal{C}} z_x^C = \prod_{C\in\mathcal{C}} \frac{\mathrm{P}(\boldsymbol{x}_C)}{\mathrm{P}(\boldsymbol{x}_{C\cap\pi(C)})}$$

Bayesnetze

Ordnung $V = \{V_1, \ldots, V_N\}$

$$\mathrm{P}(\boldsymbol{x}) = \prod_{n=1}^{N} \mathrm{P}(x_n \mid \boldsymbol{x}_{\mathrm{pa}(n)}) = \prod_{n=1}^{N} \boldsymbol{M}_{x_n|\mathrm{pa}(n)}(\boldsymbol{x})$$

Maximum-Likelihood

$$\hat{z}_x^C = \frac{n_{x_C}}{n_{x_{C\cap\pi(C)}}}$$

Maximum-Likelihood

$$\hat{\boldsymbol{M}}_{x_n|\mathrm{pa}(n)}(\boldsymbol{x}) = \frac{n_{x_{\{x_n\}\cup\mathrm{pa}(n)}}}{n_{x_{\mathrm{pa}(n)}}}$$

Welches ist die beste Modellstruktur ?

Wahrscheinlichste Kombination aus Struktur und Parametern

Gegeben

Datenprobe ω aus der Objektmenge $\mathbf{\Omega}$ über den Variablen V

$$\mathbf{\Omega} \;=\; \bigotimes_{a \in V} \mathcal{X}_a$$

Gesucht

Das bestpassende graphische/kausale/kordale/loglineare Modell

$$\hat{\Delta} \;=\; \underset{\Delta \subset \mathfrak{P} V}{\operatorname{argmax}} \, \mathrm{J}_\omega(\Delta) \;=\; \underset{\Delta \subset \mathfrak{P} V}{\operatorname{argmax}} \frac{f_{\text{prior}}(\Delta) \cdot \mathrm{P}(\omega|\Delta)}{\mathrm{P}(\omega)}$$

$$\mathrm{P}(\omega|\Delta) \;=\; \sum_{\theta \in \mathcal{M}(\Delta)} f_{\text{prior}}(\theta|\Delta) \cdot \mathrm{P}(\omega \mid \Delta, \theta)$$

Markovnetze
$\binom{N}{2}$ Kanten & insgesamt
$2^{\binom{N}{2}}$ ungerichtete Graphen

Bayesnetze
$N!$ Ordnungen & jeweils
$2^{\binom{N}{2}}$ zyklenfreie Graphen

Loglinear
2^N Terme
2^{2^N} Modelle

Gütemaße für Modellstrukturen

Definition

Mit den **Maximum-Likelihood-Parametern**

$$\hat{\theta}_\Delta(\omega) \;=\; \underset{\theta}{\operatorname{argmax}} \, \ell_\omega(\Delta, \theta) \;=\; \underset{\theta}{\operatorname{argmax}} \sum_{\boldsymbol{x} \in \omega} \log \mathrm{P}(\boldsymbol{x} \mid \Delta, \theta)$$

und der ML-bezogenen Bewertung $\hat{\ell}_\omega(\Delta) = \ell_\omega(\Delta, \hat{\theta}_\Delta(\omega))$ heißt

$$\mathrm{dev}(\Delta) \;\overset{\text{def}}{=}\; 2 \cdot \left(\hat{\ell}(\mathfrak{P} V) - \hat{\ell}(\Delta)\right), \qquad \mathfrak{P} V = \{V\} = \text{saturiertes Modell}$$

die **Devianz** der Modellstruktur Δ für die Daten ω.

Lemma

Das Devianzmaß besitzt die folgenden Eigenschaften:

1. *Gilt $\omega \sim \mathrm{P}(\cdot|\Delta)$, so ist die Devianz asymptotisch χ^2_d-verteilt, wobei d die Differenz der Freiheitsgrade von Δ und saturiertem Modell bezeichne.*
2. *Es gilt $\mathrm{dev}(\mathfrak{P} V) = 0$ und $\mathcal{E}[\mathrm{dev}(\Delta)] = d$.*

Einige regularisierte Gütemaße

Kreuzvalidierung

Datenpartition $\omega = \omega_a \uplus \omega_b$

$$\mathrm{J}(\Delta) \;=\; \ell_{\omega_b}(\Delta, \hat{\theta}_\Delta(\omega_a))$$

Rotationsvalidierung (L^1O)

„leave-one-out“ $\omega^{(\boldsymbol{x})} = \omega \backslash \{\boldsymbol{x}\}$

$$\mathrm{J}(\Delta) \;=\; \sum_{\boldsymbol{x} \in \omega} \ell_{\{\boldsymbol{x}\}}(\Delta, \hat{\theta}_\Delta(\omega^{(\boldsymbol{x})}))$$

ML-Bewertung + Strafterm

$$\mathrm{J}(\Delta) \;=\; \hat{\ell}_\omega(\Delta) \;-\; \psi(N) \cdot |\theta_\Delta|$$

AIC ➪ $\psi(N) \equiv 1$
„Akaike Information Criterion“

BIC ➪ $\psi(N) = \frac{1}{2} \log N$
„Bayesian Information Criterion“

Entropie

Bedingte Entropien $\mathcal{H}(\mathbb{X}_n|\boldsymbol{x}) = -\sum_{\xi \in \mathcal{X}_n} \mathrm{P}(\xi|\boldsymbol{x}) \cdot \log \mathrm{P}(\xi|\boldsymbol{x})$

$$\mathrm{J}(\Delta) \;=\; \mathcal{H}(\Delta) \;=\; \sum_{n=1}^{N} \sum_{x \in \mathcal{X}_{\mathsf{pa}(n)}} \mathrm{P}(\boldsymbol{x}) \cdot \mathcal{H}(\mathbb{X}_n \mid \mathbb{X}_{\mathsf{pa}(n)} = \boldsymbol{x})$$

Die $K2$-Metrik für Bayesnetze

Cooper & Herskovitz, 1991

Fakt

Eine perfekte Gütefunktion wäre die a posteriori Wahrscheinlichkeit $\mathrm{P}(\Delta|\omega)$ *der Modellstruktur auf Basis der Datenprobe.*

Gleich- und Dirichletverteilungsannahme

für Bayesnetzstruktur Δ und -parameter $\boldsymbol{M}_{n|\mathrm{pa}(n)}$:

$$\mathrm{P}(\Delta|\omega) \;\propto\; \mathrm{P}(\omega|\Delta) \;=\; \int \underbrace{\mathcal{D}(\theta|\Psi)\cdot \mathrm{P}(\omega \mid \theta, \Delta)}_{\mathrm{P}(\omega^{(\Psi)}|\theta,\Delta)}\, d\theta$$

K2-Metrik

Geschlossene Darstellung der a posteriori Wahrscheinlichkeit:

$$\mathrm{J}(\Delta) \;=\; \prod_{n=1}^{N} \prod_{\boldsymbol{x}\in\mathcal{X}_{\mathbf{pa}(n)}} \frac{(L_n-1)!}{(n_{\boldsymbol{x}}^{\mathrm{pa}(n)} + L_n - 1)!} \cdot \prod_{\xi\in\mathcal{X}_n} n_{\boldsymbol{x},\xi}^{\mathrm{pa}(n),\{n\}}$$

Suchverfahren

Wer findet die Stecknadel im Heuhaufen vor Anbruch des jüngsten Tages ?

Ungerichtete Graphen — Markovnetze

Gesucht ist eine J-optimale Teilmenge von

$$\mathfrak{E}_V = \{\{a,b\} \mid a,b \in V, a \neq b\}$$

➪ Jedes $\mathcal{E} \subseteq \mathfrak{E}_V$ ist „*erlaubt*"!

Kombinatorische Merkmalauswahl

Alle „*wrapper*"-Verfahren sind sinngemäß anwendbar:

- $\left\{\begin{matrix}\textbf{backward}\\ \textbf{forward}\end{matrix}\right\}$ **selection**: sukzessiv Kanten $\left\{\begin{matrix}\text{entfernen}\\ \text{einfügen}\end{matrix}\right\}$
- pulsierende Suche · geordnete Suche · evolutionäre Suche

SFS — Sequential Forward Selection

Gierige bottom-up Suche (Whitney 1971 · Buntine 1991)

(Algorithmus)

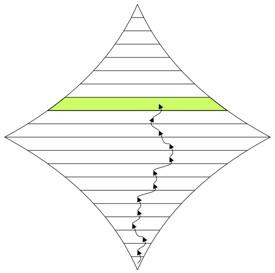
1. INITIALISIERUNG

$$\mathcal{G} \;=\; (V, \varnothing)$$

2. AUSWAHL
einer nützlichsten neuen Kante

$$\mathfrak{e}^* \;=\; \operatorname{argmax}\{\mathrm{J}(E,\mathfrak{e}) \mid \mathfrak{e} \in \mathfrak{E}_V \backslash E\}$$

3. TERMINIERUNG
Wenn $\mathrm{J}(E,\mathfrak{e}^*) \leq \mathrm{J}(E)$
dann ⇝ ENDE
sonst $E \leftarrow E \cup \{\mathfrak{e}^*\}$ und ⇝ 2.



Bemerkung
SFS trifft voreilige Entscheidungen (Horizont=1) und verfehlt i.a. die Optimallösung.

$E^{(1)} \subset E^{(2)} \subset E^{(3)} \subset \ldots$

SBE — Sequential Backward Elimination

Gierige top-down Suche (Marill & Green 1963 · Edwards/MIM 1995)

(Algorithmus)

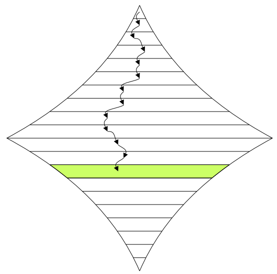
1. INITIALISIERUNG

$$\mathcal{G} \;=\; (V, \mathfrak{E}_V)$$

2. AUSWAHL
einer nutzlosesten alten Kante

$$\mathfrak{e}^* \;=\; \operatorname{argmax}\{\mathrm{J}(E\backslash\mathfrak{e}) \mid \mathfrak{e} \in E\}$$

3. TERMINIERUNG
Wenn $\mathrm{J}(E\backslash\mathfrak{e}^*) \leq \mathrm{J}(E)$
dann ⇝ ENDE
sonst $E \leftarrow E \backslash \{\mathfrak{e}^*\}$ und ⇝ 2.



Bemerkung
SBE aufwändiger als SFS:

Start mit umfangreicheren E

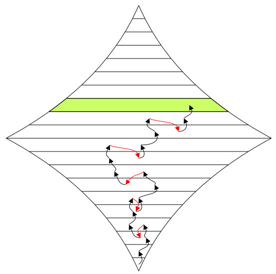
Längerer Weg zum Ziel

FBS — Forward/Backward Selection

Gierige bidirektionale Suche (Wahba 1988)

⟨Algorithmus⟩

1. INITIALISIERUNG
   $\mathcal{G} \; = \; (V, \varnothing)$
2. KANTENAUSWAHL
   $$\begin{aligned} \mathfrak{e}^F &= \operatorname{argmax}\{\mathrm{J}(E, \mathfrak{e}) \mid \mathfrak{e} \in \mathfrak{E}_V \backslash E\} \\ \mathfrak{e}^B &= \operatorname{argmax}\{\mathrm{J}(E \backslash \mathfrak{e}) \mid \mathfrak{e} \in E\} \end{aligned}$$
3. WENN MÖGLICH LÖSCHEN
   Wenn $\mathrm{J}(E \backslash \mathfrak{e}^B) > \mathrm{J}(E)$
   dann $E \leftarrow E \backslash \{\mathfrak{e}^B\}$ und $\rightsquigarrow$ 2.
4. TERMINIERUNG
   Wenn $\mathrm{J}(E, \mathfrak{e}^F) > \mathrm{J}(E)$
   dann $E \leftarrow E \cup \{\mathfrak{e}^F\}$ und $\rightsquigarrow$ 2.
   Sonst $\rightsquigarrow$ ENDE.



Bemerkung
Redundant gewordene Merkmale können jetzt wieder eliminiert werden.

Bewertungsgesteuert oder immer löschen.

Extrem riskantes „*Hillclimbing*"

SFFS — Sequential Forward Floating Search

Pulsierende Suche mit $q = \binom{N}{2}$ Schubladen (Pudil 1994)

⟨Algorithmus⟩

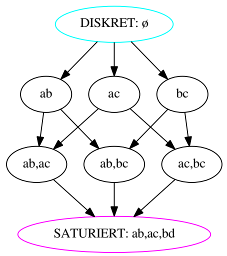
1. INITIALISIERUNG
   $$\mathcal{G} \; = \; (V, \varnothing)\,, \qquad n = 0\,, \qquad \iota_0 = \mathrm{J}(\varnothing)$$
2. VORWÄRTSAUSWAHL
   $$\mathfrak{e}^* \; = \; \operatorname{argmax}\{\mathrm{J}(E, \mathfrak{e}) \mid \mathfrak{e} \in \mathfrak{E}_V \backslash E\}$$
   Setze $E \leftarrow E \cup \{\mathfrak{e}^*\}$, $n \leftarrow n+1$ und $\iota_n = \mathrm{J}(E)$ und $\rightsquigarrow$ 2.
3. RÜCKWÄRTSAUSWAHL
   $$\mathfrak{e}^* \; = \; \operatorname{argmax}\{\mathrm{J}(E \backslash \mathfrak{e}) \mid \mathfrak{e} \in \mathfrak{E}_V \backslash E\}$$
   Wenn $\mathrm{J}(E \backslash \mathfrak{e}^*) \le \iota_{n-1}$ dann $\rightsquigarrow$ 2.
   Sonst setze $E \leftarrow E \backslash \{\mathfrak{e}^*\}$, $n \leftarrow n-1$ und $\iota_n = \mathrm{J}(E)$ und $\rightsquigarrow$ 3.
4. TERMINIERUNG
   Wenn Zielkardinalität $n = n_0$ erreicht dann $\rightsquigarrow$ ENDE.

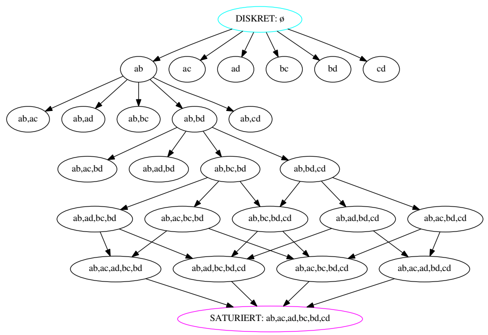
Geordnete Suche

Branch & Bound (Narendra/Fukunaga 1977)

Suchraum

Zustände	Kantenmengen $E \subset \mathfrak{E}_V$
Startzustand	$E = \varnothing$
Übergänge	Zusatzkante $\{a, b\}$
Kosten	Devianzabbau $\frac{\partial}{\partial a}\mathrm{dev}(E)$





Geordnete Suche

$\mathrm{dev}(E)$ monoton ⇨ Kosten ≥ 0

Branch&Bound-Algorithmus

A* mit trivialer Restschätzung

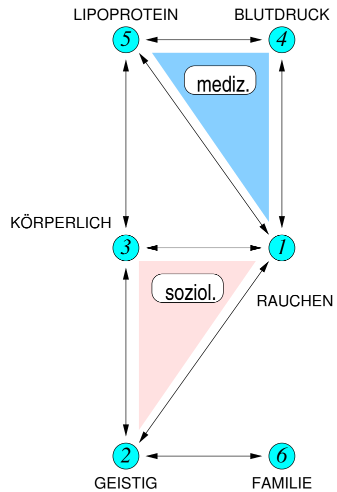
$$f(E) \; = \; \underbrace{g(E)}_{\mathrm{dev}(E)} + \underbrace{h(E)}_{\equiv 0}$$

Beispiel — Koronare Herzschwäche

P. Stuyvesant (1978)

Sechs binäre Attribute

$\mathbb{X}_1$	Versuchsperson ist Raucher ?
$\mathbb{X}_2$	Streß durch geistige Arbeit
$\mathbb{X}_3$	Streß durch körperliche Arbeit
$\mathbb{X}_4$	systolischer Blutdruck $\le 140mm$
$\mathbb{X}_5$	Lipoproteinquotient $\beta/\alpha \le 3$
$\mathbb{X}_6$	famil. Befund koronarer Schwäche



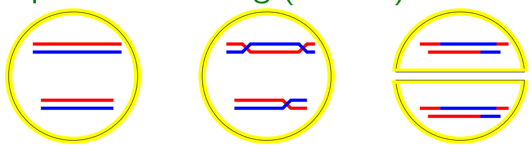
Datenerhebung

- Befundung bei $T = 1841$ Automobilarbeitern in Detroit
- Statistik mit $2^6 = 64$ Zellen
- Inkrem. Löschen partieller Abhängigkeiten
- χ^2-Kriterium: Abweichung vom saturierten Modell
- Kordale Graphen (zerlegbare Modelle!)

Beispiel — Chromosomensequenzierung

Wir sortieren das Erbgut von „*barley powdery mildew fungus*“

Haploide Vererbung (Meiose)



nach der Verschmelzung Crossover-Operation nach der Zellteilung

Erhebung

70 Geschwisterindividuen

6 binäre phänotypische Attribute

$\mathbb{X}_i \stackrel{1:1}{=}$ unbekannter Genlocus

Hypothese über Genloci $\mathbb{X}_1, \ldots, \mathbb{X}_6$

- unterschiedliche Chromosomen $\rightsquigarrow$ unabhängig
- gleiches Chromosom $\rightsquigarrow$ distanzabhängig korelliert
- Sequenz von Genen g_1, g_2, g_3 ⇨ $\Im(g_1 \mid g_2 \mid g_3)$

Resultat

$$d \longleftrightarrow a \longleftrightarrow b \longleftrightarrow e \longleftrightarrow c \longleftrightarrow f$$

Suchverfahren

Die Stecknadel piekst jetzt nur noch auf einer Seite !

Gerichtete azyklische Graphen — Bayesnetze

Optimale Teilmenge von $\mathfrak{E}_{\{x_i\}} = \{(a,b) \mid a \neq b\}$

- UG-Kantenselektion — Test auf Kordalität
- DAG-Kantenselektion — Test auf Zyklen
  DAG-Kantenselektion — Test auf Zyklen und Moralität

Optimale Teilmengen von $\mathfrak{E}_{(x_i)} = \{(x_i, x_j) \mid 1 \leq i < j \leq N\}$

- Lineare Variablenordnung $V = \{\mathbb{X}_1, \ldots, \mathbb{X}_n\}$ vorlegen
  Optimale Eltermenge $B_n \subseteq V_n$ für jedes $\mathbb{X}_n$ berechnen
  (zulässiges Verfahren sofern $\mathrm{J}(\cdot)$ in „*Familienterme*“ zerfällt)

Exakte Suche für eingeschränkte Netzstrukturen

- Bäume & Fallschirme
  Minimaler Spannbaum (Cormen, Leiserson, Rivest 1990)

K2-Algorithmus

Elternsuchverfahren (Cooper & Hershkovits, 1992)

(Algorithmus)

1. INITIALISIERUNG
   Eine Variablenordnung ist a priori vorzugeben:
   $$V = \{x_1, \ldots, x_N\}\ , \qquad n = 2$$
2. ELTERNAUSWAHL
   Triff eine Vorwärtsauswahl (SFS) bezüglich K2-Bewertung:
   $$\mathrm{pa}(x_n) = \mathrm{argmax}\,\{\mathrm{J}(A) \mid A \subseteq \{x_1, \ldots, x_{n-1}\}\}$$
3. TERMINIERUNG
   Wenn $n < N$ dann $n \leftarrow n+1$ und $\rightsquigarrow$ 2 sonst $\rightsquigarrow$ ENDE.

(Algorithmus)

Tetrad III Algorithmus

UG-Kantenselektion (Scheines 1996)

(Algorithmus)

1. KORRELATIONSTEST
   Lösche Verbindungskanten für marginal unabhängige ZV
   $$\Im(\mathbb{X}_i \mid \varnothing \mid \mathbb{X}_j)$$
2. PARTIELLE UNABHÄNGIGKEIT
   Lösche Verbindungskanten für bedingt unabhängige ZV
   $$\Im(\mathbb{X}_i \mid \mathrm{bd}(\mathbb{X}_i) \cup \mathrm{bd}(\mathbb{X}_j) \mid \mathbb{X}_j)$$
3. Teste m-elementige Teilmengen von $C_{ij} = \mathrm{bd}(\mathbb{X}_i) \cup \mathrm{bd}(\mathbb{X}_j)$.
4. ORIENTIERUNGSPHASE
   1. Wähle eine Variablenordnung $V = \{\mathbb{X}_1, \ldots, \mathbb{X}_N\}$.
   2. Orientiere alle Kanten gemäß Ordnungsindex.
   3. Ergänze Kanten für „*unshielded collider*“ und orientiere sie.

(Algorithmus)

Bayesian Network SFFS

Pulsierende DAG-Kantenselektion für Bayesnetze (Blanco & Inza, 2002)

(Algorithmus)

1 INITIALISIERUNG

$$\mathcal{G} = (V, \varnothing)\ , \qquad n = 0\ , \qquad \iota_0 = \mathrm{J}(\varnothing)$$

2 VORWÄRTSAUSWAHL

$$\mathfrak{e}^* \;=\; \operatorname{argmax}\left\{\mathrm{J}(E,\mathfrak{e}) \mid \mathfrak{e} \in \mathfrak{E}_{\{x_i\}}\backslash E \wedge \mathrm{DAG}\,(E,\mathfrak{e})\right\}$$

Setze $E \leftarrow E \cup \{\mathfrak{e}^*\}$, $n \leftarrow |n| + 1$ und $\iota_n = \mathrm{J}(E)$ und dann $\rightsquigarrow$ 2.

3 RÜCKWÄRTSAUSWAHL

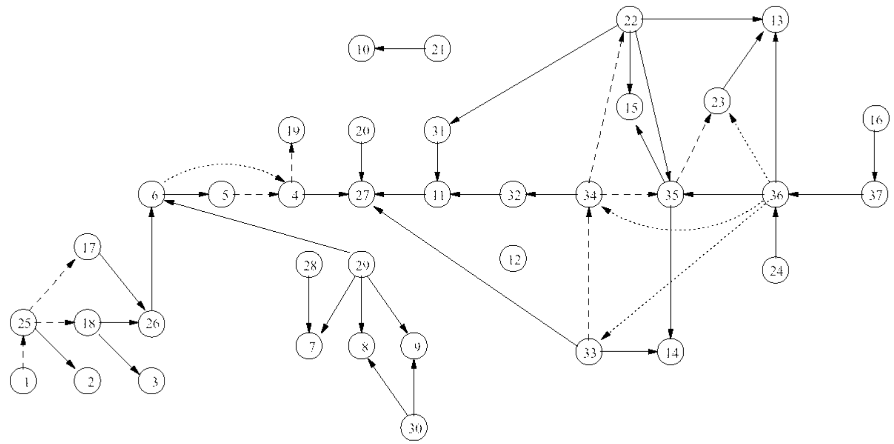
$$\mathfrak{e}^* \;=\; \operatorname{argmax}\left\{\mathrm{J}(E\backslash\mathfrak{e}) \mid \mathfrak{e} \in \mathfrak{E}_{\{x_i\}} \setminus E\right\}$$

Wenn $\mathrm{J}(E\backslash\mathfrak{e}^*) \leq \iota_{n-1}$ dann $\rightsquigarrow$ 2.
Sonst setze $E \leftarrow E \setminus \{\mathfrak{e}^*\}$, $n \leftarrow n - 1$ und $\iota_n = \mathrm{J}(E)$ und $\rightsquigarrow$ 3.

4 TERMINIERUNG
Wenn Zielkardinalität $n = n_0$ erreicht dann $\rightsquigarrow$ ENDE.

Beispiel — Alarmkette

37 Attribute · 46 → 45 Kanten · 370 Lernbeispiele



Gelernt: 45 Kanten

TBN — Baumförmige Bayesnetze

Erinnerung: moralische Bayesnetze $\hat{=}$ zerlegbare Markovnetze

Modellformel für ein TBN mit Wurzel $\mathbb{X}_{i_0}$

$V = \{x_1, \ldots, x_N\}$ und $\pi : V\backslash\{i_0\} \to V$ mit $\mathrm{pa}(x_j) = \{x_{\pi_j}\}$ für $j \neq i_0$:

$$\mathrm{P}(\boldsymbol{x}) \;=\; \mathrm{P}(x_{i_0}) \cdot \prod_{j \neq i_0} \mathrm{P}(x_j | x_{\pi_j}) \;=\; \prod_{i=1}^{n} \mathrm{P}(x_i) \cdot \prod_{j \neq i_0} \underbrace{\frac{\mathrm{P}(x_j, x_{\pi_j})}{\mathrm{P}(x_j) \cdot \mathrm{P}(x_{\pi_j})}}_{\exp\left(\Im(x_j; x_{\pi_j})\right)}$$

Nur die **punktweisen Transinformationen** sind abhängig von der Baumstruktur!

Relevanter Anteil der logarithmierten Likelihood-Zielgröße

für Lerndatenprobe ω und Baumkantenmenge $E = \{\{j, \pi_j\} \mid j \neq i_0\}$:

$$\ell_{\mathrm{ML}}(\omega | E) \;=\; \sum_{(i,j) \in E} \underbrace{\left\{\mathcal{H}(\omega, \mathrm{P}_{\mathbb{X}_i}) + \mathcal{H}(\omega, \mathrm{P}_{\mathbb{X}_j}) - \mathcal{H}(\omega, \mathrm{P}_{\mathbb{X}_i \mathbb{X}_j})\right\}}_{\Im_\omega(\mathbb{X}_i; \mathbb{X}_j)}$$

⇨ Berechnung aller **empirischen Transinformationen**

Kausalpfadanalyse mit baumförmigen Bayesnetzen

Suche nach dem minimalen Spannbaum (Chow & Liu 1968)

(Algorithmus)

1 INITIALISIERUNG
Berechne alle Transinformationswerte $(i, j = 1, \ldots, N)$:

$$\mathrm{TI}(\mathbb{X}_i, \mathbb{X}_j) \;\stackrel{\text{def}}{=}\; \sum_{x_i \in \mathcal{X}_i} \sum_{x_j \in \mathcal{X}_j} \mathrm{P}(x_i, x_j) \cdot \log \frac{\mathrm{P}(x_i, x_j)}{\mathrm{P}(x_i) \cdot \mathrm{P}(x_j)}$$

2 BEWERTETER GRAPH
Erzeuge $\tilde{\mathcal{G}} = (V, V^2, \beta)$ mit der Kantengewichtung

$$\beta : \begin{cases} V^2 & \to \quad \mathbb{R} \\ \{x_i, x_j\} & \mapsto \quad -\mathrm{TI}(\mathbb{X}_i, \mathbb{X}_j) \end{cases}$$

3 SPANNBAUM $(\mathrm{O}(N^2 \log N))$ SLAC — *„single-linkage agglomerative clustering“*
Konstruiere den minimalen spannenden Baum $\mathcal{G} \subset \tilde{\mathcal{G}}$.

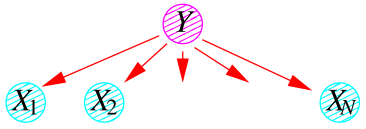
4 ORIENTIERUNG VON $\mathcal{G}$
Wähle eine beliebige Wurzelvariable $v_0 \in V$.
Alle Kanten von $\mathcal{G}$ werden *„wurzelauswärts“* gerichtet.

Klassifizieren mit Bayesnetzen

$\mathbb{X}_1, \ldots, \mathbb{X}_N \Rightarrow \mathbb{Y} \in \{1, 2, \ldots, K\}$

Naives Bayesnetz

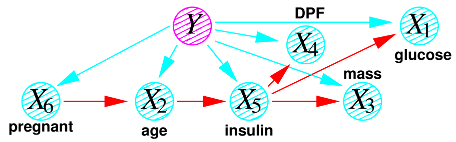
$$\mathrm{P}(y) \cdot \prod_n \mathrm{P}(x_n | y)$$



naive Bayesregel

Tree Augmented Bayesnet

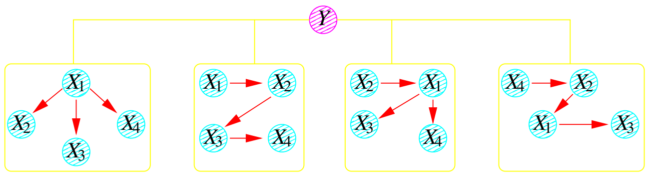
$$\mathrm{P}(x_1 | y) \cdot \prod_{n=2}^{N} \mathrm{P}(x_n \mid y, x_{\mathrm{pa}(n)})$$



TABN (Friedman, Geiger, Goldszmidt 1998)

Bayes-Multinetz

$$\mathrm{P}(y) \cdot \mathrm{P}(\boldsymbol{x} \mid E_y, \theta_y)$$



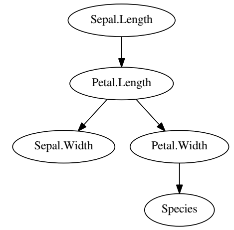
„Bayes wood" (Heckerman 1991)

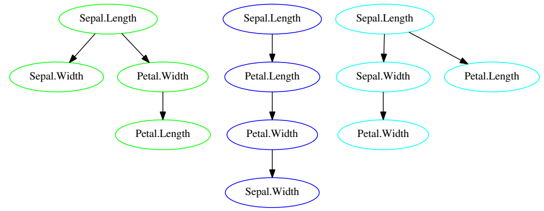
Beispiel — Fishers Irisdatensatz

5 Attribute ($\mathbb{R}^4 \times \{1, 2, 3\}$) · 150 Objekte (50 je Spezies)

Transinformationsmatrix

		$\mathbb{X}_1$	$\mathbb{X}_2$	$\mathbb{X}_3$	$\mathbb{X}_4$	$\mathbb{X}_5$
ℓ_{sepal}	$\mathbb{X}_1$	0	1.33	2.04	1.88	0.69
ω_{sepal}	$\mathbb{X}_2$	1.33	0	1.43	1.40	0.38
ℓ_{petal}	$\mathbb{X}_3$	2.04	1.43	0	2.64	1.37
ω_{petal}	$\mathbb{X}_4$	1.88	1.40	2.64	0	1.43
species	$\mathbb{X}_5$	0.69	0.38	1.37	1.43	0





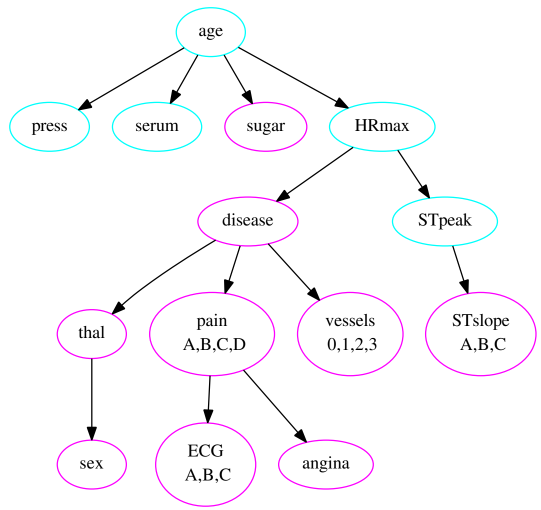
Setosa Versicolor Virginica

Bayesbaum ⇧
für alle fünf Attribute

⇦ Bayeswald
ein Baum je Spezies

Beispiel — Statlog Herzdatensammlung

13 Attribute · 270 Objekte · Klassifikation: *„disease"*



$\Im(\mathbb{X}_{14} \mid \mathbb{X}_3, \mathbb{X}_8, \mathbb{X}_{12}, \mathbb{X}_{13} \mid$ *„Rest"*$)$

1. age ($\mathbb{R}$)
2. sex $\{male, female\}$
3. chest pain $\{A, B, C, D\}$
4. blood pressure ($\mathbb{R}$)
5. serum cholestoral ($\mathbb{R}$)
6. fasting blood sugar $\{T, F\}$
7. resting ECG results $[0 : 2]$
8. maximum heart rate achieved ($\mathbb{R}$)
9. exercise induced angina $\{T, F\}$
10. ST depression (exercise:rest) ($\mathbb{R}$)
11. slope of peak exercise ST $\{A, B, C\}$
12. vessels colored by flourosopy $[0 : 3]$
13. thal $\{normal, fixed, defect\}$
14. heart disease $\{T, F\}$

Stetige Loglinearmodelle

Motivation: multivariate Normalverteilungsdichte

Definition

Es sei die N-dimensionale multivariate Normalverteilungsdichte

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{S}) \;=\; |2\pi \boldsymbol{S}|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

mit dem Mittelwertvektor $\boldsymbol{\mu}$ und der Kovarianzmatrix $\boldsymbol{S}$ gegeben.
Die Werte α, β_i, κ_{ij} der Darstellung

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{S}) \;=\; \exp\left(\alpha + \sum_i \beta_i \cdot x_i + \sum_{i,j} \kappa_{ij} \cdot x_i x_j\right)$$

heißen **kanonische Parameter** der exponentiellen Familie; die Matrix $\boldsymbol{K} = [\kappa_{ij}]$ heißt **Konzentrationsmatrix** oder **Präzisionsmatrix**.

Kanonische Parameter & Standardparameter

$\alpha = -\frac{1}{2} \cdot (\log |2\pi \boldsymbol{S}| + \boldsymbol{\mu}^\top \boldsymbol{S}^{-1}\boldsymbol{\mu})$, $\quad \boldsymbol{\beta} \;=\; \boldsymbol{S}^{-1}\boldsymbol{\mu}$, $\quad \boldsymbol{K} \;=\; \boldsymbol{S}^{-1}$.

Loglinearmodelle

Die Kovarianzterme in Δ sind die nichtnegativen Koeffizienten der Summationsterme

$$u^{(i_1,\dots,i_N)} \cdot x_1^{i_1} x_2^{i_2} x_3^{i_3} \dots x_n^{i_n} \dots x_{N-1}^{i_{N-1}} x_N^{i_N}$$

des Dichtefunktionsexponenten. Insbesondere fällt dem Term

$$u^{(0,\dots,0)} \cdot x_1^0 x_2^0 \dots x_N^0 \;=\; u^{(0,\dots,0)} \cdot 1 \;=\; u^{(0,\dots,0)}$$

wieder die Rolle des Normierungsfaktors zu.

Die Vektoren $\boldsymbol{i}$ können wir auch als *Multimengen* von Zufallsvariablen auffassen.

Gaußsche graphische Modelle

Hier werden ausschließlich Kovarianzterme $\boldsymbol{i} \in \Delta$ zugelsaaen mit

$$\sum_{n=1}^{N} i_n \;=\; i_1 + i_2 + i_3 + \dots + i_N \;\leq\; 2\,.$$

Beweis.

Das Modell ist auch graphisch, denn es gibt grundsätzlich keinerlei Interaktion zwischen mehr als zwei Variablen. Es gilt $\Im(A \mid Z \mid B)$ genau dann, wenn Δ ausschließlich $A \cup Z$-Terme und $B \cup Z$-Terme enthält, also wenn es keine $\{a, b\}$-Terme mit $a \in A$ und $b \in B$ gibt. □

Stetige Loglinearmodelle

Definition

Sei $\Delta \subset \mathbb{N}^N$ eine (endliche) Menge von Exponenten-N-Tupeln.
Die Familie stetiger Wahrscheinlichkeitsdichtefunktionen der Gestalt

$$\log f_\Delta(\boldsymbol{x}) \;=\; \sum_{i \in \Delta} u^i \cdot \prod_{n=1}^{N} x_n^{i_n}\,, \qquad \boldsymbol{x} \in \mathbb{R}^N$$

heißt **stetiges Loglinearmodell** über V mit der Menge Δ von **Kovarianztermen**.

Lemma

Für Loglinearmodelle Δ, die Normalverteilungen sind, gilt:

$$\Delta \text{ hierarchisch} \quad \Rightarrow \quad \Delta \text{ graphisch}$$

*Wir nennen diese Familien **Gaußsche Graphische Modelle** oder **Kovarianzselektionsmodelle**.*

Wissenswertes über multivariate Normalverteilungsdichten

Lemma

Für normalverteilte Zufallsvariablen $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$ gelten die folgenden Aussagen:

1. *Summenbildung:* $\quad \mathbb{X} \;=\; \mathbb{X}' + \mathbb{X}'' \;\sim\; \mathcal{N}(\boldsymbol{\mu}' + \boldsymbol{\mu}'', \boldsymbol{S}' + \boldsymbol{S}'')$
2. *Affine Abbildung:* $\quad \boldsymbol{A}\mathbb{X} + \boldsymbol{b} \;\sim\; \mathcal{N}(\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{ASA}^\top)$
3. *Marginalisierung:* $\quad \mathbb{X}_C \;\sim\; \mathcal{N}(\boldsymbol{\mu}_C, \boldsymbol{S}_{CC})$
4. *Konditionierung:* $\quad \mathbb{X}_{A|\boldsymbol{x_B}} \;\sim\; \mathcal{N}(\boldsymbol{\mu}_{A|\boldsymbol{x_B}}, \boldsymbol{S}_{A|\boldsymbol{x_B}})$

   Dabei gelte $A \uplus B = V$ und es sind definiert:

$$\begin{aligned} \boldsymbol{\mu}_{A|\boldsymbol{x_B}} &= \boldsymbol{\mu}_A + \boldsymbol{S}_{AB} \cdot \boldsymbol{S}_{BB}^{-1} \cdot (\boldsymbol{x}_B - \boldsymbol{\mu}_B) & \qquad \boldsymbol{\mu} &= (\boldsymbol{\mu}_A^\top, \boldsymbol{\mu}_B^\top)^\top \\ \boldsymbol{S}_{A|\boldsymbol{x_B}} &= \boldsymbol{S}_{AA} - \boldsymbol{S}_{AB} \cdot \boldsymbol{S}_{BB}^{-1} \cdot \boldsymbol{S}_{BA} & \qquad \boldsymbol{S} &= \begin{pmatrix} \boldsymbol{S}_{AA} & \boldsymbol{S}_{AB} \\ \boldsymbol{S}_{BA} & \boldsymbol{S}_{BB} \end{pmatrix} \end{aligned}$$

5. *Für die bedingte Kreuzkovarianzmatrix gilt der Zusammenhang:*

$$\left(\boldsymbol{S}_{A|\boldsymbol{x_B}}\right)^{-1} \;=\; \boldsymbol{K}_{AA} \;=\; \left(\boldsymbol{S}^{-1}\right)_{AA}$$

Marginalisierung

Ähnlich wie schon zuvor für Vektoren definieren wir Matrixausschnitte durch

$$\boldsymbol{M}_{A,B} \overset{\text{def}}{=} (M_{ab} \mid a \in A, b \in B) \ .$$

Die Matrix $\boldsymbol{S}_{CC}$ insbesondere enthält also alle Varianzen von und Kovarianzen zwischen Variablen aus C.
Die Matrix $\boldsymbol{S}_{AB}$ heißt übrigens auch „Kreuzkovarianzmatrix" der Variablenmengen A und B.

Konditionierung

Bei geeigneter Variablennummerierung gilt in der Situation $A \uplus B = V$:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{AA} & \boldsymbol{S}_{AB} \\ \boldsymbol{S}_{BA} & \boldsymbol{S}_{BB} \end{pmatrix} = \begin{pmatrix} \boldsymbol{S}_{AA} & \boldsymbol{S}_{AB} \\ \boldsymbol{S}_{AB}^\top & \boldsymbol{S}_{BB} \end{pmatrix} , \qquad \boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_{AA} & \boldsymbol{K}_{AB} \\ \boldsymbol{K}_{BA} & \boldsymbol{K}_{BB} \end{pmatrix}$$

Die Matrixgleichung ergibt sich aus der (unschönen!) Formel zur Blockmatrixinvertierung.

Beweis.

Die partielle Unabhängigkeit, d.h. die Frage nach einer Kante oder keiner Kante zwischen zwei Variablen im Markovnetz, läßt sich ganz einfach aus der inversen Kovarianzmatrix $\boldsymbol{K}$ ablesen.

- Beweisidee 1:
  Betrachte die bedingte Verteilung mit $A = \{a, b\}$ und $B = V \setminus \{a, b\}$.

  Das Variablenpaar $(\mathbb{X}_a, \mathbb{X}_b)$ ist, bei gegebenem $\boldsymbol{x}_B$, mit der Kovarianzmatrix $\boldsymbol{S}_{ab|\cdot}$ normalverteilt.

  $\mathbb{X}_a, \mathbb{X}_b$ sind unabhängig genau dann, wenn $\boldsymbol{S}_{ab|\cdot}$ eine Diagonalmatrix ist; dies aber ist genau dann der Fall, wenn ihre Inverse, also $\boldsymbol{K}_{\{a,b\}}$ diagonal ist, also falls $\kappa_{ab} = \kappa_{ba} = 0$ ist.
- Beweisidee 2:
  Die Normalverteilungsdichte ist faktorisierbar in Gibbs-Komponenten mit maximal zwei Variablen.

  Sie läßt sich also in zwei Faktoren $g_{V\setminus\{a\}}$ und $h_{V\setminus\{b\}}$ genau dann zerlegen, wenn mindestens die Gibbs-Komponente für $\{a, b\}$ fehlt.

$\square$

Kovarianz und Konzentration

Nulleinträge $\hat{=}$ marginale & partielle Unabhängigkeiten

Lemma (Wermuth 1976)

Für normalverteilte Variablen $a, b \in V \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$ mit $a \neq b$ gilt:

- ***Marginale Unabhängigkeit**:* $\quad \Im(a \mid \varnothing \mid b) \Leftrightarrow s_{ab} = 0$
- ***Partielle Unabhängigkeit**:* $\quad \Im(a \mid Rest \mid b) \Leftrightarrow \kappa_{ab} = 0$

Gaußscher Diamant

$$\begin{pmatrix} \kappa_{11} & \kappa_{12} & 0 & \kappa_{14} \\ \kappa_{21} & \kappa_{22} & \kappa_{23} & 0 \\ 0 & \kappa_{32} & \kappa_{33} & \kappa_{34} \\ \kappa_{41} & 0 & \kappa_{43} & \kappa_{44} \end{pmatrix}$$

$$\Updownarrow$$

$$\begin{array}{ccc} \mathbb{X}_1 & - & \mathbb{X}_2 \\ | & & | \\ \mathbb{X}_4 & - & \mathbb{X}_3 \end{array}$$

Gaußscher Schlüssel

$$\begin{pmatrix} \kappa_{11} & \kappa_{12} & 0 & 0 & 0 \\ \kappa_{21} & \kappa_{22} & \kappa_{23} & \kappa_{23} & 0 \\ 0 & \kappa_{32} & \kappa_{33} & \kappa_{34} & \kappa_{35} \\ 0 & \kappa_{42} & \kappa_{43} & \kappa_{44} & \kappa_{45} \\ 0 & 0 & \kappa_{53} & \kappa_{54} & \kappa_{55} \end{pmatrix}$$

$$\Updownarrow$$

$$\mathbb{X}_1 - \mathbb{X}_2 \lhd \begin{array}{c} \mathbb{X}_4 \\ \mathbb{X}_3 \end{array} \rhd \mathbb{X}_5$$

Charakterisierung bedingter Unabhängigkeiten

Lemma

Für Variablenmengen A, B, Z mit $V = A \uplus B \uplus Z$ und die bedingte Kreuzkovarianzmatrix

$$\boldsymbol{S}_{AB|Z} = [s_{ab}]_{a\in A}^{b\in B} , \qquad s_{ab} \overset{\text{def}}{=} \mathrm{Cov}[\mathbb{X}_a, \mathbb{X}_b \mid \mathbb{X}_Z = \boldsymbol{x}_Z]$$

gilt die Beziehung:

$$\boldsymbol{S}_{AB|Z} = \boldsymbol{S}_{AB} - \boldsymbol{S}_{AZ} \cdot (\boldsymbol{S}_{ZZ})^{-1} \cdot \boldsymbol{S}_{ZB}$$

Satz (Speed & Kiiveri 1986)

Für normalverteilte Variablen $V = A \uplus B \uplus Z$ mit Kovarianzmatrix $\boldsymbol{S}$ sind äquivalent:

1. $\boldsymbol{S}_{AB} = \boldsymbol{S}_{AZ} \cdot (\boldsymbol{S}_{ZZ})^{-1} \cdot \boldsymbol{S}_{ZB}$
2. $(\boldsymbol{S}^{-1})_{AB} = \boldsymbol{0}$ *beziehungsweise* $\boldsymbol{K}_{AB} = \boldsymbol{0}$
3. $\Im(A \mid Z \mid B)$

Maximum-Likelihood-Schätzung

für Gaußsche Graphische Modelle

Satz (Dempster 1972)

Es sei $\mathcal{G} = (V, \mathcal{E})$ ein Gaußsches Graphisches Modell mit der Generatorenmenge $\mathcal{C} \subset \mathfrak{P}V$ und sei $\omega \subset \mathbb{R}^N$ ein Datensatz mit den Statistiken $\boldsymbol{m}$ und $\boldsymbol{\Sigma}$. Dann bilden $\boldsymbol{m}$ und $\{\boldsymbol{\Sigma}_{CC} \mid C \in \mathcal{C}\}$ eine minimale suffiziente Statistik des Modells für ω.

Die Maximum-Likelihood-Parameter $\boldsymbol{\mu}$ und $\boldsymbol{S}$ bzw. $\boldsymbol{K}$ gehorchen den Bedingungen

$$\begin{array}{rcll} \boldsymbol{\mu} & = & \boldsymbol{m} & \\ s_{ab} & = & \sigma_{ab} & \{a,b\} \in \mathcal{E} \vee a = b \\ \kappa_{ab} & = & 0 & \{a,b\} \notin \mathcal{E} \wedge a \neq b \end{array}$$

Bemerkung

Da $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ das saturierte Modell ist, beträgt die **Devianz**:

$$\mathrm{dev}(\mathcal{C}) \;=\; 2 \cdot (\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \ell(\boldsymbol{\mu}, \boldsymbol{S})) \;=\; T \cdot \log(\det \boldsymbol{S} / \det \boldsymbol{\Sigma})$$

Maximum-Likelihood-Schätzung

Existenz & Eindeutigkeit

Datenkovarianzmatrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 3.023 & 1.258 & 1.004 \\ 1.258 & 1.709 & 0.842 \\ 1.004 & 0.842 & 1.116 \end{pmatrix}$$

ML-Kovarianzmatrix

$$\boldsymbol{S} = \begin{pmatrix} 3.023 & 1.258 & 0.620 \\ 1.258 & 1.709 & 0.842 \\ 0.620 & 0.842 & 1.116 \end{pmatrix}$$

ML-Konzentrationsmatrix

$$\boldsymbol{K} = \begin{pmatrix} 0.477 & -0.351 & 0.000 \\ -0.351 & 1.190 & -0.703 \\ 0.000 & -0.703 & 1.426 \end{pmatrix}$$

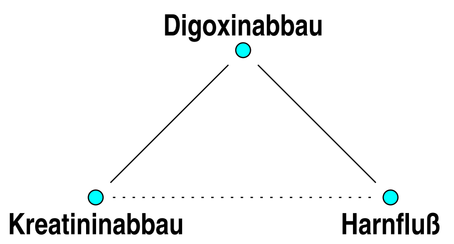
Satz (Dempster 1972)

Es seien $\boldsymbol{A}$, $\boldsymbol{B}$ zwei symmetrische, positiv-definite $(N \times N)$-Matrizen. Ferner sei $\mathcal{E} \subseteq \{1, \ldots, N\} \times \{1, \ldots, N\}$ symmetrisch mit $(i,i) \in \mathcal{E}$ für alle i. Dann gibt es eine symmetrische, positiv-definite Matrix $\boldsymbol{S}$ mit

$$\begin{array}{rcll} s_{ij} & = & a_{ij} & \forall (i,j) \in \mathcal{E} \\ (\boldsymbol{S}^{-1})_{ij} & = & b_{ij} & \forall (i,j) \notin \mathcal{E} \end{array}$$

und $\boldsymbol{S}$ ist eindeutig mit diesen Eigenschaften.

Beispiel — Digoxin-Abbau



Datensammlung

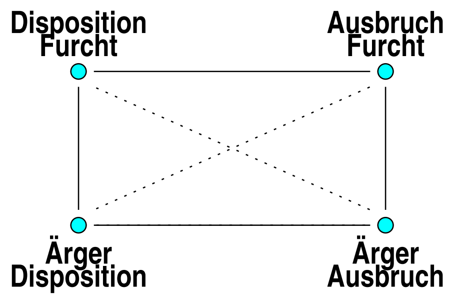
$\omega \subset \mathbb{R}^3$

$|\omega| = 35$ Patienten

- $\mathbb{X}$ = Abbau von Kreatinin
- $\mathbb{Y}$ = Abbau von Digoxin
- $\mathbb{Z}$ = Harnflußrate

χ^2-Test ➪ $\{\mathbb{X}, \mathbb{Z}\} \notin \mathcal{E}$

Beispiel — Furcht versus Ärger



Datensammlung

$\omega \subset \mathbb{R}^4$

$|\omega| = 684$ Versuchspersonen

- **Augenblickszustand:**

  $\mathbb{X}$ = Furcht

  $\mathbb{W}$ = Ärger

- **mentale Prägung:**

  $\mathbb{Z}$ = Furcht

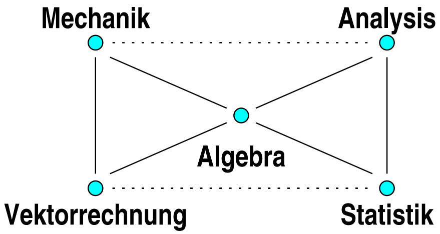
  $\mathbb{Y}$ = Ärger

χ^2-Test ergibt:

$$\Im(\mathbb{X} \mid \mathbb{W}, \mathbb{Z} \mid \mathbb{Y})$$

$$\Im(\mathbb{W} \mid \mathbb{X}, \mathbb{Y} \mid \mathbb{Z})$$

Beispiel — Punktezahl in Übungsserien



Datensammlung

$\omega \subset \{0, 1, 2, \ldots, 100\}^5$
$|\omega| = 88$ Studierende

- 5 Übungsgruppen in 5 Fächern
- je 100 Punkte erzielbar

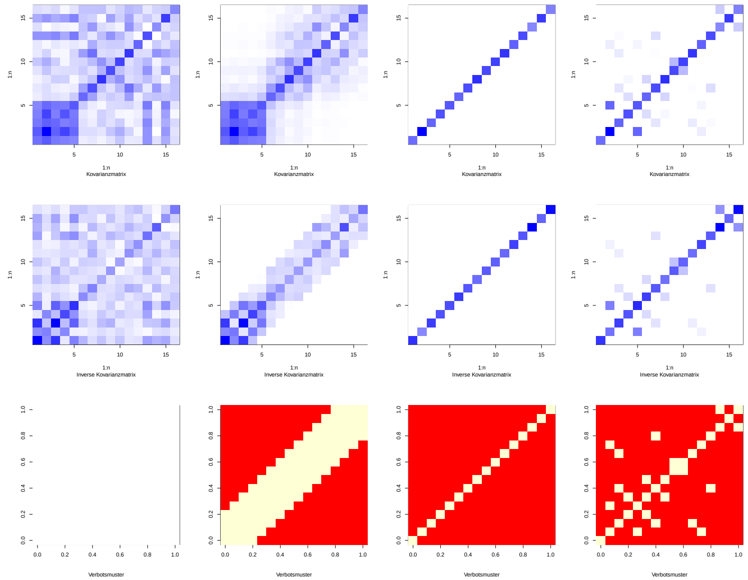
Inferenz

$\Im(\text{Stat} \mid \text{Alg}, \text{Analysis} \mid \cdot)$

$\Im(\text{Mech} \mid \text{Alg}, \text{Vektor} \mid \cdot)$

Zentrale Befähigung: Algebra

Beispiel — Schriftzeichenklassifikation



Beispiel

Datensatz
`letter`
16 Merkmale
alle Klassen

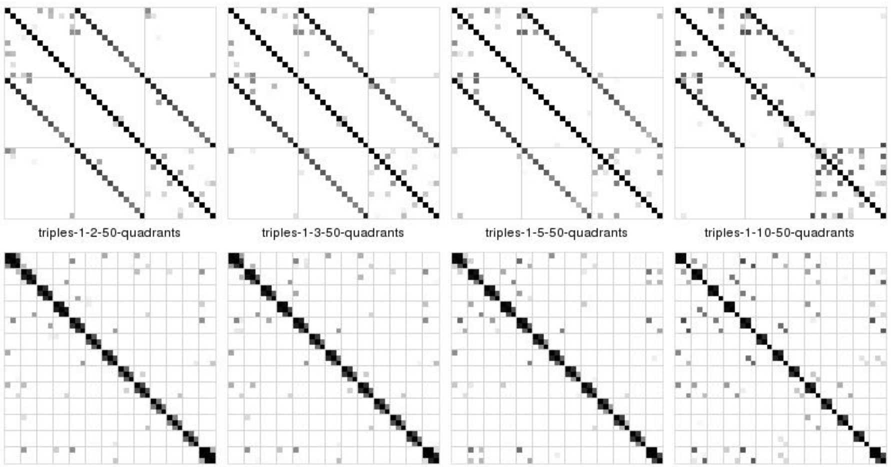
oben:
Kovarianz
$\hat{\boldsymbol{S}} = \boldsymbol{C}^{-1}$

Mitte:
Konzentration
$\boldsymbol{C}$ erfüllt $\boldsymbol{A}$

unten:
Adjazenz $\boldsymbol{A}$
Abhängigkeits-muster
(gegeben)

Beispiel — Sprachsignalparameter

12 MFCC-Parameter · drei Zeitpunkte



triples-1-2-50-quadrants | triples-1-3-50-quadrants | triples-1-5-50-quadrants | triples-1-10-50-quadrants

triples-1-2-50-reordered | triples-1-3-50-reordered | triples-1-5-50-reordered | triples-1-10-50-reordered

Dünne Abhängigkeitsstruktur

Die Vergangenheit wird durch unmittelbare Vorgänger „*maskiert*“.

Zusammenfassung (6)

1. **Kovarianz** und **Korrelation** sind **quantitative** Charakterisierungen der **linearen** Aspekte („*Regression*") statistischer Abhängigkeit.
2. Die **Transinformation** quantifiziert statistische Abhängigkeit in allgemeiner Form, setzt aber die Kenntnis der **wahren Verteilung** voraus.
3. Die **Warenkorbanalyse** sucht **Assoziationsregeln** mit gleichermaßen hohen Werten für **Support**, **Konfidenz** und **Relevanz** (z.B. **Apriori-Algorithmus**).
4. Das **Dependenzmodell** charakterisiert die **bedingten Unabhängigkeiten** $\Im(A \mid Z \mid B)$ je dreier Attributmengen einer Verteilung.
5. Verteilungen heißen **graphisch** (**kausal**), wenn ihr DM durch die (δ-)**Separation** eines **UG** (**DAG**) gegeben ist.
6. Kausale Modelle faktorisieren **attributweise** in **bedingte Wahrsch'keiten**, graphische Modelle faktorisieren **cliquenweise** in **Gibbspotenziale**.
7. **Kordale Modelle** besitzen einen **triangulierten** UG, einen **moralischen** DAG und eine **Cliquenschnitt**faktorisierung.
8. Statistische **Inferenz** ist nur für **Ketten** und (Verbund-)**Bäume** effizient.
9. Die **ML-Schätzung** der **Modellparameter** aus Daten beruht auf **relativen Häufigkeiten** (DAG) oder dem **Maximum-Entropie-Prinzip** (UG).
10. Die Aufdeckung der **Modellstruktur** basiert auf **Unabhängigkeitstests** (Kantenelimination/Grenzengraph) oder **gieriger Suche** mit **Strafterm**.