Einführung in Web- und Data-Science

Stochastische Grundlagen

Dr. Marcel Gehrke
Universität zu Lübeck
Institut für Informationssysteme

Stochastische Grundlagen

WAHRSCHEINLICHKEITEN

Erster Wahrscheinlichkeitsbegriff

- Grenzwert der relativen Häufigkeit des Auftreten eines Ereignisses
  - Beispiel: Würfeln einer geraden Zahl
- "Elementar-Ereignisse" besitzen gleiche Eintreffensw'keiten
  - Laplace'sches Prinzip
- Was sind Elementarereignisse eigentlich genau?
  - Elemente ω einer Grundgesamtheit Ω
  - Elementarereignisse sind abstrakt: $\Omega = \{\omega_1, \omega_2, ..., \omega_6\}$
    - Beispiel: ω_i steht für einen Würfelwurf
  - Abbildung von Ereignissen auf Merkmalswerte durch Zufallsvariablen
    - Zufallsvariable X für *Ergebnis* des Würfelwurfs ist eine Funktion
    - Ziel: Ereignis ω_i auf obenliegende Zahl i des geworfenen Würfels abbilden
    - $X : \Omega \rightarrow \{1, 2, 3, 4, 5, 6\}$

Ereignisse

- (Komplexe) Ereignisse sind Teilmengen von Ω
  - Beispiel: Würfelereignisse mit gerader Zahl
  - Definition der Zufallsvariablen entsprechend erweitert
  - $X: \mathcal{P}(\Omega) \rightarrow \mathcal{P}(M)$
    - Beispiel: $X(\{\omega_2, \omega_4, \omega_6\}) = \{2, 4, 6\}$

$\mathcal{P}(X)$ sei die Potenzmenge von X.

Laplace-Wahrscheinlichkeiten

- Betrachte die endliche Grundgesamtheit von Elementarereignissen $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$
- Für ein Ereignis $A \subseteq \Omega$ definiert man die Laplace-Wahrscheinlichkeit als die Zahl
  - $P(A) := |A| / |\Omega| = |A| / n$
    wobei $|A|$ die Anzahl der Elemente in A ist.
- Jedes Elementarereignis ω_i, $i = 1, \ldots, n$ hat also die Wahrscheinlichkeit $P(\{\omega_i\}) = 1/n$
- Wir sagen X hat Verteilung gekennzeichnet durch $P(\{\omega_i\}) = 1/n$
- Die Wahrscheinlichkeit von Ω ist $P(\Omega) = 1$

Bayessche Wahrscheinlichkeitstheorie

- Laplace-Verteilungen sind zu speziell!
- Beispiele:
  - Unfairer Würfel
  - Wahrscheinlichkeit für Knabengeburt
  - Auftreten von Kopf oder Zahl bei Euro Münzen
- Elementarereignisse nicht immer gleichwahrscheinlich!
- Konzept der A-priori-Wahrscheinlichkeit
  - Vorwissen und Grundannahmen des Beobachters in einer Wahrscheinlichkeitsverteilung zusammengefasst
  - ... und explizit im Modell ausgedrückt

Wahrscheinlichkeitsräume

Ein (diskreter) Wahrscheinlichkeitsraum ist definiert als ein Paar (Ω, P) wobei

- Ω eine (abzählbare) Grundgesamtheit ist und
- P ein Wahrscheinlichkeitsmaß, das jeder Teilmenge $A \subseteq \Omega$ eine Wahrscheinlichkeit $P(A)$ zuordnet.

P definiert man wieder über die Wahrscheinlichkeiten $P(\{\omega\})$ der Elementarereignisse $\omega \in A$:

wobei für $P(\{\omega\})$ gelten muss: $0 \leq P(\{\omega\}) \leq 1$ für alle ω und

$$P(A) = \sum_{\omega \in A} P(\{\omega\})$$

$$\sum_{\omega \in \Omega} P(\{\omega\}) = 1$$

Axiome von Kolmogorov [1903-1987]

- Wir betrachten eine beliebige (abzählbare) Grundgesamtheit Ω und eine Funktion P, die jedem Ereignis $A \subseteq \Omega$ eine Wahrscheinlichkeit zuordnet.
- Wir nennen P eine Wahrscheinlichkeitsverteilung auf Ω, wenn sie folgende Eigenschaften erfüllt:
  - Ax_1: $P(A) \geq 0$ für beliebige $A \subseteq \Omega$
  - Ax_2: $P(\Omega) = 1$
  - Ax_3: $P(A \cup B) = P(A) + P(B)$
    für disjunkte Ereignisse $A, B \subseteq \Omega$

Folgerungen

- $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$
  für paarweise disjunkte Eregnisse $A_1, A_2, \ldots, A_n \subset \Omega$
- $P(A) \leq P(B)$ falls $A \subseteq B$

- Definiere das Komplement von A: $\bar{A} = \Omega \setminus A$.
  Dann gilt: $P(\bar{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ für beliebige $A, B \subset \Omega$
  $\rightsquigarrow$ Darstellung im Venn-Diagramm

Verbundwahrscheinlichkeit

- Betrachten wir zwei aufeinanderfolgende Würfelwurfe
- Der Ereignisraum (Ω, P) ist dann wie folgt definiert
  - $\Omega = \{\omega_1,\omega_2,...,\omega_6\} \times \{\omega_1,\omega_2,...,\omega_6\}$
  - $P(A \times B)$ mit $A, B \subseteq \{\omega_1,\omega_2,...,\omega_6\}$ ist diskretes Wahrscheinlichkeitsmaß
- Wir sprechen von einer Verbundwahrscheinlichkeit und schreiben abkürzend
  - $P(A, B)$ für $P(A \times B)$ mit $A, B \subseteq \Omega$
- In einem Verbund können beliebige Grundgesamtheiten verknüpft werden
  - Beispiel: (Ω', P) mit
    $\Omega' = \{\omega_1,\omega_2,...,\omega_6\} \times \{$ kreuz, pik, herz, karo $\}$

Bedingte Wahrscheinlichkeiten

- Für Ereignisse $A, B \subseteq \Omega$ mit $P(B) > 0$ definiert man die bedingte Wahrscheinlichkeit von A gegeben B als die Zahl

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Beispiel: Würfelspiel
  - "Es wird eine gerade Zahl gewürfelt" $A := \{ \omega_2, \omega_4, \omega_6 \}$
    "Es wird eine Zahl > 4 gewürfelt" $B := \{ \omega_5, \omega_6 \}$
  - Dann:
    - $P(A|B) = 1/2$
    - $P(A|\bar{B}) = 2/4 = 1/2$

Der Satz von Bayes

- Thomas Bayes [1701-1761]
- Dieser Satz beruht auf der Asymmetrie der Definition von bedingten Wahrscheinlichkeiten:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \Rightarrow \quad P(A \cap B) = P(A|B)P(B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \Rightarrow \quad P(A \cap B) = P(B|A)P(A)$$

- Analog für Verbundwahrscheinlichkeiten

$$P(A|B) = \frac{P(A\ ,\ B)}{P(B)} \quad \Rightarrow \quad P(A\ ,\ B) = P(A|B)P(B)$$

$$P(B|A) = \frac{P(A\ ,\ B)}{P(A)} \quad \Rightarrow \quad P(A\ ,\ B) = P(B|A)P(A)$$

Stochastische Grundlagen

UNABHÄNIGKEITEN

Stochastische Unabhängigkeit

- Wann sind 2 Ereignisse A, B unabhängig?
- Motivation über bedingte Wahrscheinlichkeiten:
- Zwei Ereignisse A, B sind unabhängig, wenn

$$\underbrace{P(A|B)}_{\frac{P(A\cap B)}{P(B)}} = P(A) \qquad P(B) > 0$$

$$\text{bzw. } \underbrace{P(B|A)}_{\frac{P(A\cap B)}{P(A)}} = P(B) \qquad P(A) > 0$$

bzw. $P(A, B) = P(A) \cdot P(B)$ gilt.

» Voraussetzung $P(B) > 0$ und $P(A) > 0$ ist hier nicht nötig

Beispiel: Zweimaliges Würfeln

- Ein fairer Würfel wird zweimal hintereinander geworfen.
  - A stehe für “Beim 1. Würfelwurf eine Sechs”
  - B stehe für “Beim 2. Würfelwurf eine Sechs”
- Bei jeden Würfelwurf ist die Grundgesamtheit
  $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$
  mit $\omega_i =$ "Gewürfelte Zahl ist i"
- Nach Laplace gilt $P(A) = P(B) = 1/6$.
- Bei “unabhängigem” Werfen gilt somit
  $P(A, B) = P(A) \cdot P(B) = 1/36$

Bedingte Unabhängigkeit

- Sei C ein beliebiges Ereignis mit $P(C) > 0$.
  Zwei Ereignisse A und B nennt man
  bedingt unabhängig gegeben C, wenn gilt:

$$P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

- Anders geschrieben:

$$P(A \mid B, C) = P(A \mid C)$$

Notation

- Sei $\Omega = \{\omega_1,\omega_2,...,\omega_6\}$, $M = \{1,2,3,4,5,6\}$
- Sei X eine Zufallsvariable $X : \Omega \rightarrow M$
  - Beispiel: $X: \omega_2 \longmapsto 2$
- Notation: $dom(X)$ steht für M
- Notation: $X = 2$ steht für Elementarereignis $\{\omega_2\}$
  - $P(X=2)$ steht für $P(\{\omega_2\})$
- Notation:
  $X=2 \vee X=4 \vee X=6$ steht für komplexes Ereignis $\{\omega_2 , \omega_4 , \omega_6\}$
  - $P(X=2 \vee X=4 \vee X=6)$ steht für $P(\{\omega_2 , \omega_4 , \omega_6\})$
- Notation: $X=2 \wedge Y=4$ steht für Verbundereignis $\{\omega_2\} \times \{\omega_4'\}$ (verschiedene Variablen)

Verteilungsnotation

- Für eine diskrete Zufallsvariable X schreiben wir die Verteilung als $\mathbf{P}(X)$, wobei gilt
  $\mathbf{P}(X) = (P(x_1), ..., P(x_n))^T$ für $x_1, x_2, ..., x_n \in \text{dom}(X)$

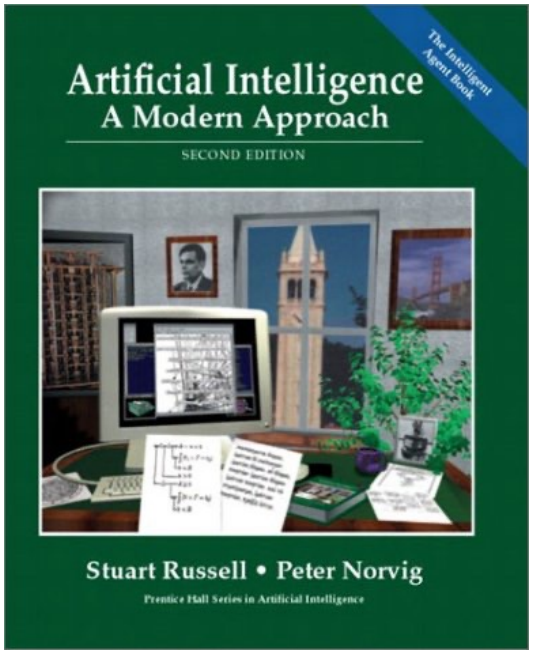
- Auch im Verbund verwendet: $\mathbf{P}(X, Y)$

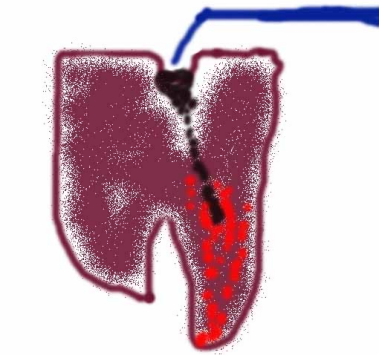
- $\mathbf{P}(X, Y) = \mathbf{P}(X \mid Y) \cdot \mathbf{P}(Y)$, wobei hier die Multiplikation komponentenweise erfolgt

Beispiel

Zahnarzt-Problem mit vier Variablen:

- Zahnschmerzen (Sind besagte Schmerzen wirklich Zahnschmerzen?)
- Loch (Es könnte ein Loch sein?)
- Fang (Stahlinstrument erzeugt Testschmerz?)
- Wetter (Wetter: sonnig, regnerisch, bewölkt, schneit)





Nachfolgende Präsentationen enthalten Material aus Kapitel 14 (Sektion 1 and 2)

Prior Wahrscheinlichkeit

- Prior oder bedingte Wahrscheinlichkeit von Propositionen

  e.g., P(*Loch* = wahr) = 0.1 und P(*Wetter* = sonnig) = 0.72 entsprechen den Annahmen bevor wir (neue) Beobachtungen erhalten

- Verbundsahrscheinlichkeitsverteilung

  Gibt Werte für alle möglichen Zuweisungen an:

  P(*Wetter*) = <0.72, 0.1, 0.08, 0.1>

  (normalisiert, i.e., summiert sich zu 1 auf)

Vollständige Verbundswahrscheinlichkeitsverteilung

- Verbundswahrscheinlichkeitsverteilung gibt für eine Menge von Zufallsvariablen die Wahrscheinlichkeit jedes atomaren Ereignisses an
- Für **P**(Wetter, Loch):

Wetter =	sonnig	regnerisch	bewölkt	schneit
Loch = wahr	0.144	0.02	0.016	0.02
Loch = falsch	0.576	0.08	0.064	0.08

- Vollständige Verbundswahrscheinlichkeitsverteilung : umfasst alle Zufallsvariablen
  - **P**(Zahnschmerzen, Fang, Loch, Wetter)
- Jede Anfrage zu einer Domain kann durch die vollständige Verbundswahrscheinlichkeitsverteilung beantwortet werden

Diskrete Zufallsvariablen: Notation

- Dom(Wetter) = {sonnig, regnerisch, bewölkt, schneit} und Dom(Wetter) disjunkt von der Domäne anderer Zufallsvariablen:
  - Atomares Ereignis Wetter=regnerisch wird oft geschrieben als regnerisch
  - Beispiel: P(regnerisch), Die Zufallsvariable Wetter ist implizit definiert durch den Wert regnerisch
- Boolean Variable Loch
  - Atomares Ereignis Loch = wahr geschrieben als loch
  - Atomares Ereignis Loch = falsch geschrieben als ¬ loch
  - Beispiele: P(loch) oder P(¬ loch)

Bedingte Wahrscheinlichkeit

- Bedingte oder posterior Wahrscheinlichkeiten
  e.g., P(*loch* | *zahnschmerzen*) = 0.8
  oder: <0.8>
  i.e., gegeben *zahnschmerzen* sind Beobachtungen, die bekannt sind

- Notation für bedingte Verteilungen:
  P(*Loch* | *Zahnschmerzen*) ist ein 2-Element-Vektor von 2-Element-Vektoren

- Wenn wir mehr wissen, e.g., *loch* auch gegeben ist, dann haben wir
  P(*loch* | *zahnschmerzen, loch*) = 1

- Neue Beobachtungen können irrelevant sein und eine Vereinfachung ermöglichen, e.g.,
  P(*loch* | *zahnschmerzen, sonnig*) = P(*loch* | *zahnschmerzen*) = 0.8

- Diese Art von Schlussfolgerung, ermöglicht durch Domänenwissen, ist wichtig

Bedingte Wahrscheinlichkeit

- Im Allgemeinen für Verteilungen von Zufallsvariablen gilt, z.B.:

$$\boldsymbol{P}(Wetter, Loch) = \boldsymbol{P}(Wetter \mid Loch)\, \boldsymbol{P}(Loch)$$
$$\boldsymbol{P}(Loch, Wetter) = \boldsymbol{P}(Loch)\boldsymbol{P}(Wetter \mid Loch)$$

Daraus ergeben sich 4 × 2 Gleichungen, nicht Matrixmultiplikation

(1,1) $P(Wetter = sonnig \mid Loch = wahr)\ P(Loch = wahr)$

(1,2) $P(Wetter = sonnig \mid Loch = falsch)\ P(Loch = false), \ldots.$

- Der Multplikationssatz wird von der mehrmaligen Anwendung der Produktregel abgeleitet:

$$\begin{aligned} P(X_1, \ldots, X_n) &= P(X_1,\ldots,X_{n-1})\, P(X_n \mid X_1,\ldots,X_{n-1}) \\ &= P(X_1,\ldots,X_{n-2})\, P(X_{n-1} \mid X_1,\ldots,X_{n-2})\, P(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \ldots \\ &= \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}) \end{aligned}$$

Inferenz durch Enumeration

- Beginnen Sie mit der vollständigen Verbundswahrscheinlichkeitsverteilung:

	zahnschmerzen		¬zahnschmerzen	
	fang	¬fang	*fang*	*¬fang*
loch	.108	.012	.072	.008
¬loch	.016	.064	.144	.576

- Für jede Anfrage φ, summieren Sie die Wahrscheinlichkeiten wo φ wahr ist: $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

- $P(zahnschmerzen) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

- Unbedingte oder marginale Wahrscheinlichkeit von zahnschmerzen
- Prozess wird als Marginalisierung oder Aussummieren bezeichnet

Ax_3: $P(A \cup B) = P(A) + P(B)$
für disjunkte Ereignisse $A, B \subseteq \Omega$

Marginalisierung und Konditionierung

- Seien Y, Z Sequenzen von Zufallsvariablen, so dass $Y \cup Z$ alle Zufallsvariablen beschreibt

- Marginalisierung
  - $\mathbf{P}(\mathbf{Y}) = \Sigma_{\mathbf{z} \in \mathbf{Z}} \mathbf{P}(\mathbf{Y},\mathbf{z})$

- Konditionierung
  - $\mathbf{P}(\mathbf{Y}) = \Sigma_{\mathbf{z} \in \mathbf{Z}} \mathbf{P}(\mathbf{Y}|\mathbf{z})P(\mathbf{z})$

Inferenz durch Enumeration

- Beginnen Sie mit der vollständigen Verbundswahrscheinlichkeitsverteilung:

	$zahnschmerzen$		¬zahnschmerzen	
	$fang$	¬fang	$fang$	$\neg fang$
$loch$	.108	.012	.072	.008
$\neg loch$	.016	.064	.144	.576

Für jede Anfrage φ, summieren Sie die Wahrscheinlichkeiten wo φ wahr ist: $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

- P(loch ∨ *zahnschmerzen*) = 0.108 + 0.012 + 0.072 + 0.008+ 0.016 + 0.064 = 0.28

(P(loch ∨ *zahnschmerzen*) = P(loch) + P(*zahnschmerzen*) – P(loch ∧ *zahnschmerzen*))

Inferenz durch Enumeration

- Beginnen Sie mit der vollständigen Verbundswahrscheinlichkeitsverteilung:

	zahnschmerzen		¬zahnschmerzen	
	fang	¬fang	*fang*	*¬fang*
loch	.108	.012	.072	.008
¬loch	.016	.064	.144	.576

- Kann auch bedingte Wahrscheinlichkeiten berechnen:

$$P(\neg \text{ loch} \mid \text{zahnschmerzen}) = \frac{P(\neg \text{ loch} \wedge \text{zahnschmerzen})}{P(\text{zahnschmerzen})}$$

Produktregel

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.08/0.2 = 0.4$$

$$P(\text{loch} \mid \text{zahnschmerzen}) = (0.108+0.012)/0.2 = 0.6$$

Normalisierung

	zahnschmerzen		¬zahnschmerzen	
	fang	¬fang	*fang*	*¬fang*
loch	.108	.012	.072	.008
¬loch	.016	.064	.144	.576

- Nenner **P**(**z**) (or P(zahnschmerz) im vorherigen Beispiel) kann als Normalisierungskonstante α angesehen werden

P(Loch | zahnschmerz) = α **P**(Loch, zahnschmerz)
= α [**P**(Loch, zahnschmerz,fang) + **P**(Loch, zahnschmerz,¬ fang)]
= α [<0.108,0.016> + <0.012,0.064>]
= α <0.12,0.08> = <0.6,0.4>

Allgemeine Idee: Berechnen Sie die Verteilung auf die Abfragevariable durch Fixierung von Evidenzvariablen (Zahnschmerzen) und Summe über die latente Variablen (Fang)

Inferenz durch Enumeration, contd.

Typischerweise interessieren wir uns für
die posterior Verbundswahrscheinlichkeit von den anfrage Variablen **Y**
gegebene spezifische Werte **e** für die beobachteten Variablen **E**
(**X** sind alle Variablen der modellierten Welt)

Lassen Sie die latenten Variablen $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$ sein, dann erfolgt die erforderliche Summierung der gemeinsamen Einträge durch das Summieren über die latenten Variablen:

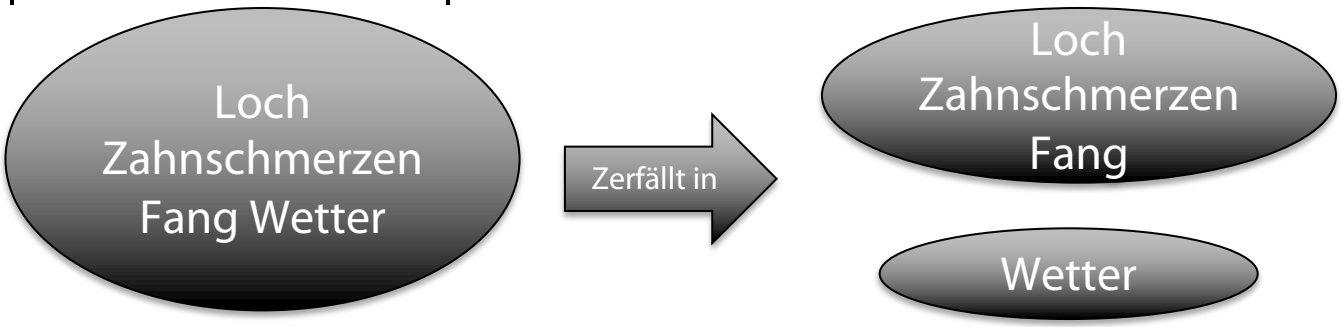
$$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \Sigma_{\mathbf{h}} \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$$

- Die Terms in der Summation sind verbundene Einträge, da **Y**, **E** und **H** schöpfen zusammen die Menge der Zufallsvariablen aus (**X**)

- Probleme:
  1. Worst-Case Zeitkomplexität $O(d^n)$ wo d die größte Arität ist und n bezeichnet die Anzahl der Zufallsvariablen
  2. Raumkomplexität $O(d^n)$ zur Speicherung der Verbundsverteilung
  3. Wie können wir Zahlen für $O(d^n)$ Einträge finden?

Unabhängigkeiten

- A und B sind unabhängig, genau dann wenn
  $\mathbf{P}(A|B) = \mathbf{P}(A)$ oder $\mathbf{P}(B|A) = \mathbf{P}(B)$ oder $\mathbf{P}(A, B) = \mathbf{P}(A)\,\mathbf{P}(B)$



$\mathbf{P}$(Zahnschmerzen, Fang, Loch, Wetter)
$= \mathbf{P}$(Zahnschmerzen, Fang, Loch) $\mathbf{P}$(Wetter)

- 32 Einträge reduziert zu 12 Einträgen;
- Absolute Unabhängigkeit ist sehr mächtig, aber auch selten
- Die Zahnmedizin ist ein großes Feld mit Hunderten von Variablen, von denen keine unabhängig ist. Was können wir da tun?

Bedingte Unabhängigkeit

- **P**(Zahnschmerzen, Loch, Fang) hat $2^3 - 1 = 7$ unabhängige Einträge
- Wenn ich ein Loch habe, dann ist die Wahrscheinlichlichkeit, dass der Fang das Loch findet unabhängig davon, ob ich Zahnschmerzen habe:
  (1) **P**(fang | zahnschmerzen, loch) = **P**(fang | loch)
- Die gleiche Unabhängigkeit gilt auch, wenn ich kein Loch habe:
  (2) **P**(fang | zahnschmerzen,¬ loch) = **P**(fang | ¬ loch)
- Fang ist bedingt Unabhängig von Zahnschmerzen gegeben Loch:
  P(Fang | Zahnschmerzen, Loch) = **P**(Fang | Loch)
- Äquivalente Aussagen:
  P(Zahnschmerzen | Fang, Loch) = **P**(Zahnschmerzen | Loch)

  P(Zahnschmerzen, Fang | Loch) = **P**(Zahnschmerzen | Loch) **P**(Fang | Loch)

Bedingte Unabhängigkeit contd.

- Bestimme die vollständige Verbundswahrscheinlichkeit mit Hilfe des Multplikationssatzes:

  P(Zahnschmerzen, Fang, Loch)

  = **P**(Zahnschmerzen | Fang, Loch) **P**(Fang, Loch)

  = **P**(Zahnschmerzen | Fang, Loch) **P**(Fang | Loch) **P**(Loch)

  bedingte Unabhängigkeit

  = **P**(Zahnschmerzen | Loch) **P**(Fang | Loch) **P**(Loch)

  i.e., 2 + 2 + 1 = 5 unabhängige Einträge
- In den meisten Fällen verringert die Anwendung der bedingten Unabhängigkeit die größe der Verbundsserteilung von exponentiell in n zu linear in n.
- Bedingte Unabhängigkeit ist unsere grundlegendste und robusteste Form des Wissens über Unsicherheiten.

Stochastische Grundlagen

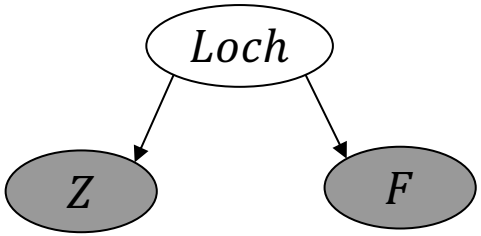
BAYESSCHE MODELLE

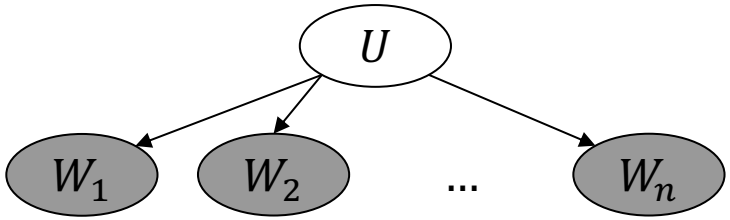
Naïve Bayes Modell

$$\boldsymbol{P}(\text{Loch} \mid zahnschmerzen \wedge fang)$$
$$= \boldsymbol{\alpha}\, \boldsymbol{P}(zahnschmerzen \wedge fang \mid \text{Loch})\boldsymbol{P}(\text{Loch})$$
$$= \boldsymbol{\alpha}\, \boldsymbol{P}(zahnschmerzen \mid \text{Loch})\, \boldsymbol{P}(fang \mid \text{Loch})\boldsymbol{P}(\text{Loch})$$

Ist ein Beispiel für ein naïve Bayes Modell

$$\boldsymbol{P}(Ursache, Wirkung_1, \ldots, Wirkung_n) =$$
$$\boldsymbol{P}(Ursache) \prod_i \boldsymbol{P}(Wirkung_i \mid Ursache)$$





Die Anzahl der Parameter ist linear in n

Normalerweise ist die Annahme, dass Wirkungen/Effekte unabhängig sind, falsch, funktioniert aber in der Praxis gut.

Bayesian Netzwerk

- Eine einfache grafische Notation für bedingte Unabhängigkeitsbehauptungen und damit für die kompakte Spezifikation vollständiger Verbundsverteilungen

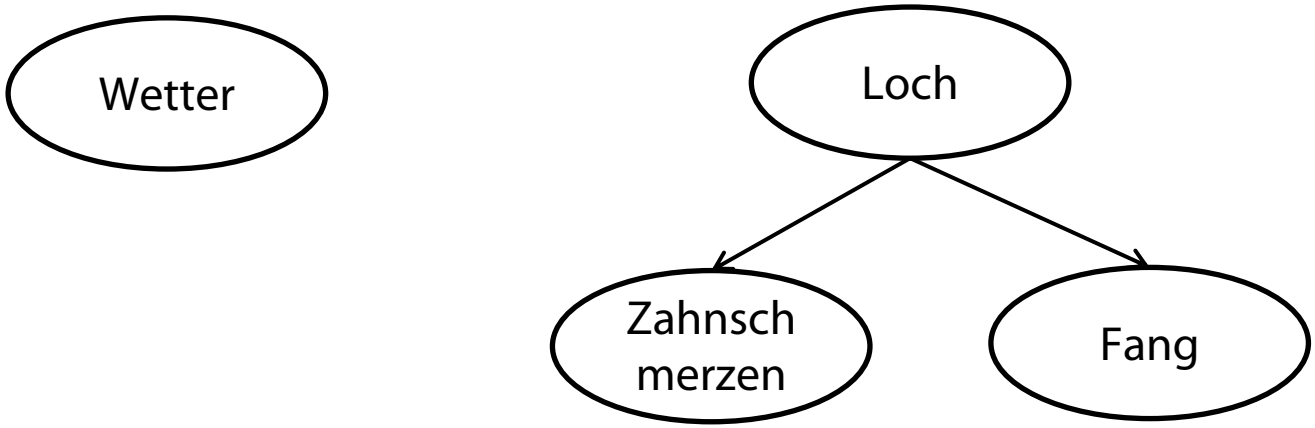
- Syntax:
  - eine Menge von Knoten, einer pro Zufallsvariable
  - Ein gerichteter azyklischer Graph (Verbindung ≈ "direkter Einflüsse")
  - eine bedingte Verteilung für jeden Knoten gegeben seinen Eltern:

$$\mathbf{P}\,(X_i \mid \text{Eltern}\,(X_i))$$

- Die Bedingteverteilung wird oftmals als Bedingte Wahrscheinlichkeitstabelle (conditional probability table (CPT)), gegeben Werte für die Elternknoten, dargestellt

Beispiel

- Die Topologie des Netzwerks enkodiert durch die bedingte Unabhängigkeitsannahmen:



- Das *Wetter* ist unabhängig von den anderen Variablen
- *Zahnschmerzen* und *Fang* sind bedingt unabhängig gegeben *Loch*

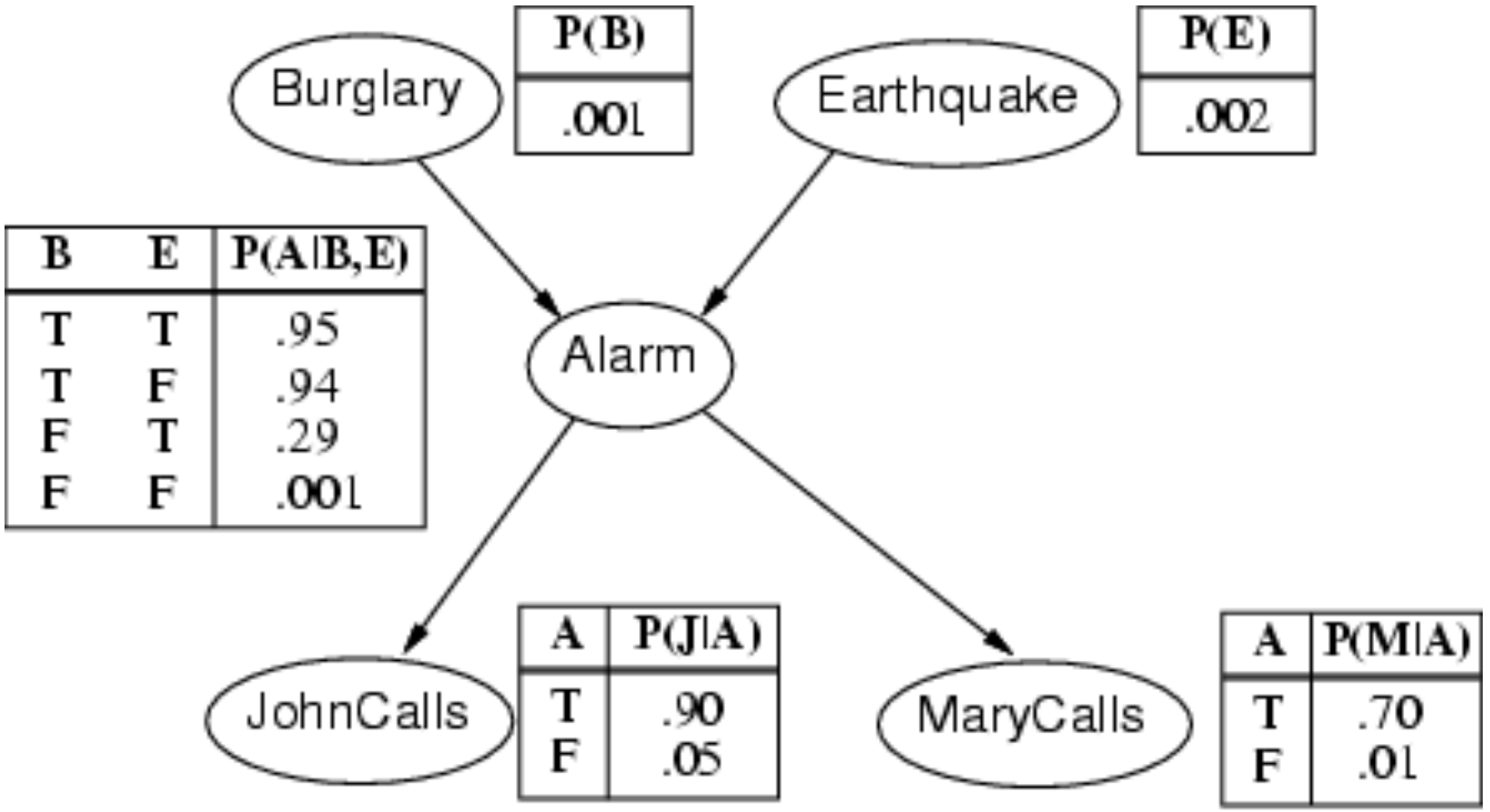
Beispiel

- Ich bin bei der Arbeit, mein Nachbar John ruft an, um zu sagen, dass mein Alarm klingelt, aber Nachbarin Mary ruft nicht an. Manchmal wird der Alarm durch kleinere Erdbeben ausgelöst. Gibt es einen Einbrecher?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

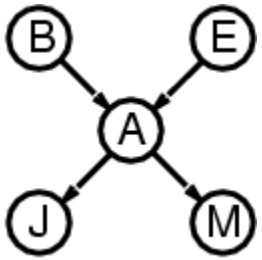
- Die Netzwerktopologie kann "kausales" Wissen widerspiegeln:
  - Ein Einbrecher kann den Alarm auslösen
  - Ein Erdbeben kann den Alarm auslösen
  - Der Alarm kann dazu führen, dass Mary anruft
  - Der Alarm kann dazu führen, dass John anruft

Example contd.



Kompaktheit

- Eine CPT für eine Boolean ZV X_i mit k Boolean Eltern hat 2^k Zeilen für die Kombinationen von Elternknoten



- Jede Zeile hat eine Nummer p für $X_i = wahr$ (der Wert für $X_i = falsch$ ist $1\text{-}p$)

- Wenn jede Variable nicht mehr als k Eltern hat, dann braucht das komplette Netzwerk $n \cdot 2^k$ Werte

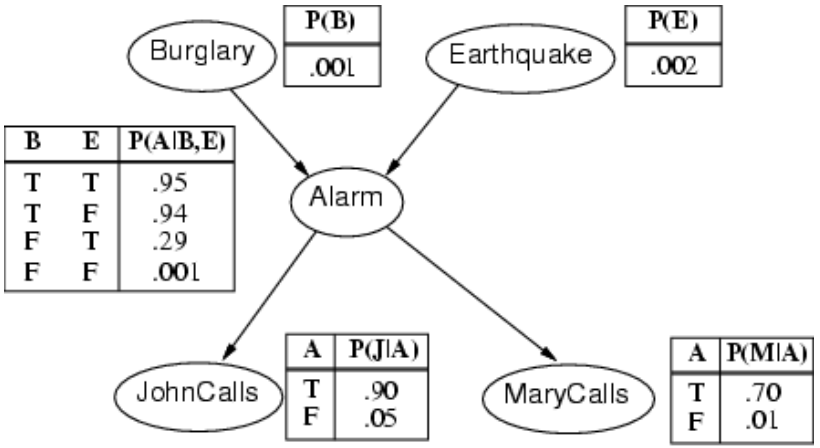
- Also das Netzwerk wächst linear in n, vs. 2^n für die vollständige Verbundswahrscheinlichkeitsverteilung

- Für dieses Beispiel also, $1 + 1 + 4 + 2 + 2 = 10$ Werte (vs. $2^5\text{-}1 = 31$)

Semantik

Die vollständige Verbundswahrscheinlichkeitsverteilung ist definiert als das Produkt der lokalen bedingten Verteilungen:

$$P(X_1, \ldots, X_n) = \pi_{i=1}^{n} P(X_i \mid Eltern(X_i))$$



e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$= P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b, \neg e)\, P(\neg b)\, P(\neg e)$

$= 0.90 \times 0.7 \times 0.001 \times 0.999 \times 0.998$

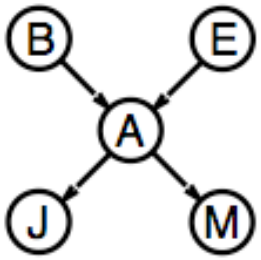
≈ 0.00063

Inferenz durch Enumeration

Man kann Anfragen, mit Hilfe der CPTs, beantworten, ohne die vollständige Verbundswahrscheinlichkeitsverteilung zu bestimmen

Naïve Anfrage an das Netzwerk:

$$\begin{aligned}&\mathbf{P}(B|j,m)\\&=\mathbf{P}(B,j,m)/P(j,m)\\&=\alpha\mathbf{P}(B,j,m)\\&=\alpha\ \Sigma_e\ \Sigma_a\ \mathbf{P}(B,e,a,j,m)\end{aligned}$$



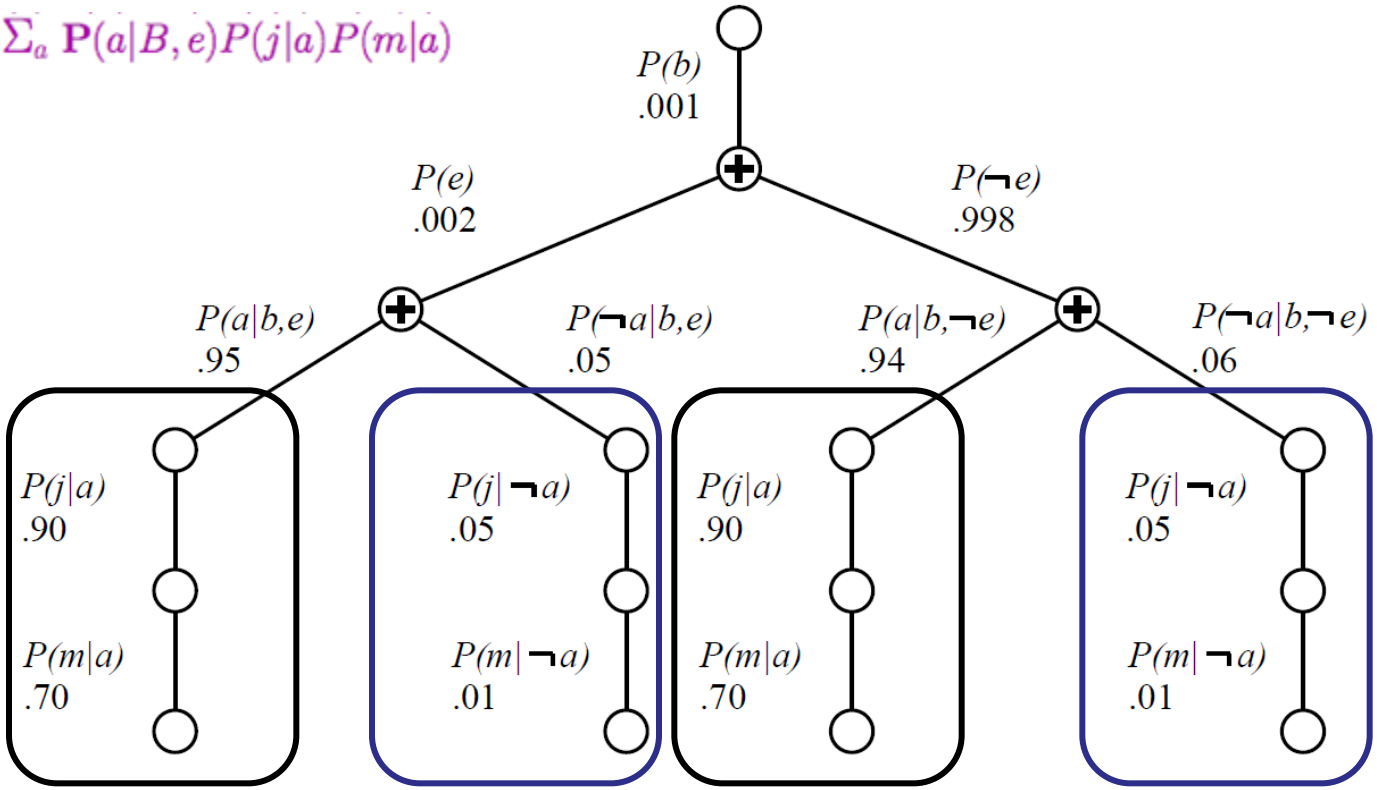
Anfrage an das Netzwerk mit Hilfe der CPTs:

$$\begin{aligned}&\mathbf{P}(B|j,m)\\&=\alpha\ \Sigma_e\ \Sigma_a\ \mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)\\&=\alpha\mathbf{P}(B)\ \Sigma_e\ P(e)\ \Sigma_a\ \mathbf{P}(a|B,e)P(j|a)P(m|a)\end{aligned}$$

Rekursive tiefen Enumeration: $O(n)$ space, $O(d^n)$ time

Evaluationsbaum

$$\mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) P(m|a)$$

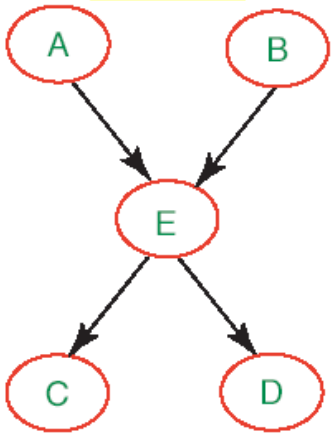


Enumeration ist ineffizient: wiederkehrende Berechnungen
z.B. Berechnungen von $P(j \mid a)P(m \mid a)$ für die verschiedenen Werte von e

Lernen von Bayesschen Netzwerken

- Wir beginnen mit der Anwendung von ML auf die einfachste Art des Bayesschen Netzwerken-Lernens:
- Bekannte Struktur
- Daten, enthalten Beobachtungen für alle Variablen
  - ✓ Alle Variablen sind beobachtbar, keine fehlenden Daten
- Das einzige, was wir lernen müssen, sind die Parameter des Netzwerks

Model Data → Probabilities



A	B	C	D	E
t	f	t	t	f
f	t	t	t	t
t	t	f	t	f
		...		

$P(A)$
$P(B)$
$P(E|A, B)$
$P(C|E)$
$P(D|E)$

Maximum-Likelihood-Parameterschätzung

- Nehme an, die Struktur eines BNs sei bekannt
- Ziel: Schätze BN-Parameter θ
  - Einträge in CPTs, P(X | Parents(X))
- Eine Parametrierung θ ist gut, falls hierdurch die beobachteten Daten wahrscheinlich generiert werden:

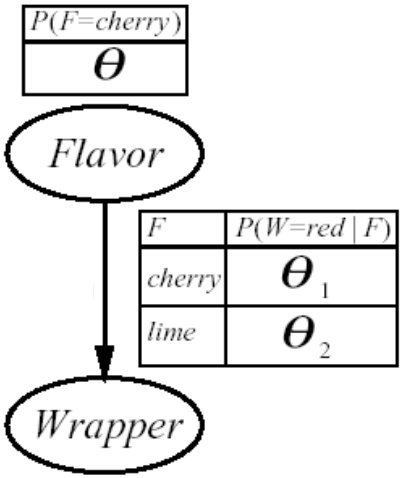
$$P(D \mid \theta) = \prod_m P(x[m] \mid \theta)$$

- Maximum Likelihood Estimation (MLE) Prinzip: Wähle θ^* so, dass $P(D/\theta^*)$ maximiert wird

Gleichverteilte, unabhängige Stichprobem (i.i.d. samples)

Anwendungsbeispiel Bonbonfabrik

- Ein Hersteller wählt die Farbe des Bonbonpapiers mit einer bestimmten Wahrscheinlichkeit je nach Geschmack, wobei die entsprechende Verteilung nicht bekannt sei
  - Wenn Geschmack=cherry, wähle rotes Papier mit W'keit θ_1
  - Wenn Geschmack=lime, wähle rotes Papier mit W'keit θ_2
- Das Bayessche Netzwerk enthält drei zu lernende Parameter
  - $\theta\,\theta_1\,\theta_2$

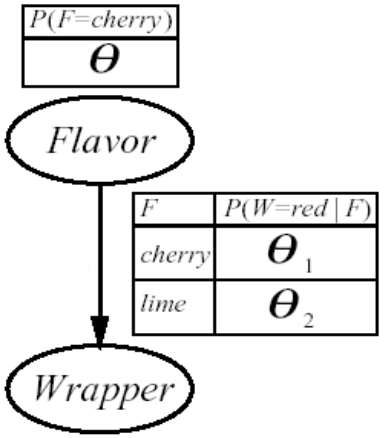


Anwendungsbeispiel Bonbonfabrik

- $P(W=green, F=cherry \mid h_{\theta\theta_1\theta_2}) = (*)$

  $= P(W=green \mid F=cherry, h_{\theta\theta_1\theta_2})\, P(F=cherry \mid h_{\theta\theta_1\theta_2})$

  $= (1-\theta_1)\,\theta$



- Wir packen N Bonbons aus
  - *c* sind cherry und *ℓ* sind lime
  - r^c cherry mit rotem Papier, g^c cherry mit grünem Papier
  - r^l lime mit rotem Papier, g^ℓ lime mit grünem Papier
  - Jeder Versuch liefert eine Kombination aus Papier und Geschmack wie bei (*)

- $P(\mathbf{d} \mid h_{\theta\theta_1\theta_2})$

  $= \prod_j P(d_j \mid h_{\theta\theta_1\theta_2}) = \theta^c (1-\theta)^\ell (\theta_1)^{r^c} (1-\theta_1)^{g^c} (\theta_2)^{r^\ell} (1-\theta_2)^{g^\ell}$

Anwendungsbeispiel Bonbonfabrik

- Maximierung des Logarithmus der Zielfunktion
  - $L = c \log\theta + \ell \log(1-\theta) + r^c \log\theta_1 + g^c \log(1-\theta_1) + r^l \log\theta_2 + g^\ell \log(1-\theta_2)$
- Bestimmung der Ableitungen bzgl. $\theta, \theta_1, \theta_2$
  - Ausdrücke ohne Term, nach dem abgeleitet wird, verschwinden

$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell}$$

$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \quad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c+g_c}$$

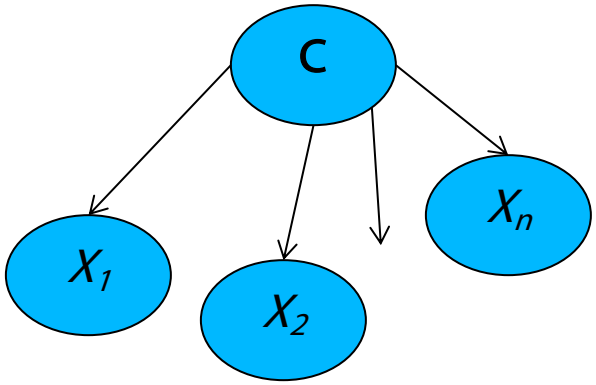
$$\frac{\partial L}{\partial \theta_2} = \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 \quad \Rightarrow \quad \theta_2 = \frac{r_\ell}{r_\ell+g_\ell}$$

Maximum-Likelihood-Parameterschätzung

- Schätzung durch Bildung relativer Häufigkeiten
- Dieser Prozess ist auf jedes voll beobachtbare BN anwendbar
- Mit vollständigen Daten und Maximum-Likelihood-Parameterschätzung:
  - Parameterlernen zerfällt in separate Lernprobleme für jeden Parameter (CPT) durch Logarithmierung
  - Jeder Parameter wird durch die relative Häufigkeit eines Knotenwertes bei gegebenen Werten der Elternknoten bestimmt

Beliebte Anwendung: Naives Bayes-Modell

- Naïve Bayes-Modell: Sehr einfaches Bayessches Netzwerk zur Klassifikation
  - *Klassenvariable C* (vorherzusagen) bildet Wurzel
  - Attributvariablen X_i (Beobachtungen) sind Blätter



- Naiv, weil angenommen wird, dass die Attributwerte bedingt unabhängig sind, wenn die Klasse gegeben ist

$$\boldsymbol{P}(C|x_1,x_2,..,x_n) = \frac{\boldsymbol{P}(C,x_1,x_2,..,x_n)}{\boldsymbol{P}(x_1,x_2,..,x_n)} = \alpha \boldsymbol{P}(C)\prod_i \boldsymbol{P}(x_n \mid C)$$

- Deterministische Vorhersagen können durch Wahl der wahrscheinlichsten Klasse erreicht werden
- Skalierung auf realen Daten sehr gut:
  - 2n + 1 Parameter benötigt

Anwendung: Diagnose

Nützlich für das Abschätzen von Diagnosen

Wahrscheinlichkeiten von kausalen Abhängigkeiten

$$P(Ursache \mid Wirkung) = \frac{P(Wirkung \mid Ursache)P(Ursache)}{P(Wirkung)}$$

Lassen Sie M Meningitis sein und S einen steifen Nacken:

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0{,}8 \cdot 0{,}0001}{0{,}1} = 0{,}0008$$

Bemerkung: Die bedingte Wahrscheinlichkeit von Meningitis ist immer noch sehr klein!