
Einführung in Web- und Data-Science

Grundlagen der Stochastik

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

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: $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$
 - 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: $X(\{\omega_2, \omega_4, \omega_6\}) = \{2, 4, 6\}$

Laplace-Wahrscheinlichkeiten

- Betrachte die endliche Grundgesamtheit von Elementarereignissen $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$
- 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 $\omega_i, i = 1, \dots, n$ 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: $P(A) = \sum_{\omega \in A} P(\{\omega\})$
 $0 \leq P(\{\omega\}) \leq 1$ für alle ω und

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

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** Ereignisse $A_1, A_2, \dots, A_n \subset \Omega$
 - $P(A) \leq P(B)$ 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 $A, B \subset \Omega$
- ↪ Darstellung im Venn-Diagramm (John Venn [1834-1923])

Verbundwahrscheinlichkeit

- Betrachten wir zwei aufeinanderfolgende Würfelwürfe
- Der Ereignisraum (Ω, P) ist dann wie folgt definiert
 - $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\} \times \{\omega_1, \omega_2, \dots, \omega_6\}$
 - $P(A \times B)$ mit $A, B \subseteq \{\omega_1, \omega_2, \dots, \omega_6\}$ 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
$$\Omega' = \{\omega_1, \omega_2, \dots, \omega_6\} \times \{ \text{kreuz, pik, herz, karo} \}$$

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:
 - $P(A|B) = 1/2$
 - $P(A|B^c) = 2/4 = 1/2$

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 Unabhängigkeit

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

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

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

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 $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$, $M = \{1, 2, 3, 4, 5, 6\}$
- Sei X eine Zufallsvariable $X : \Omega \rightarrow M$
 - Beispiel: $X: \omega_2 \mapsto 2$
- Notation: $\text{dom}(X)$ 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)

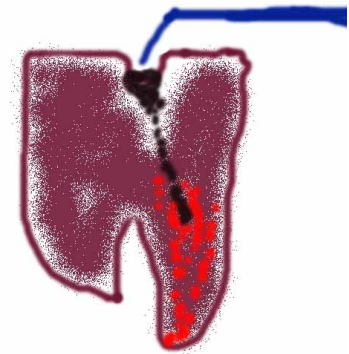
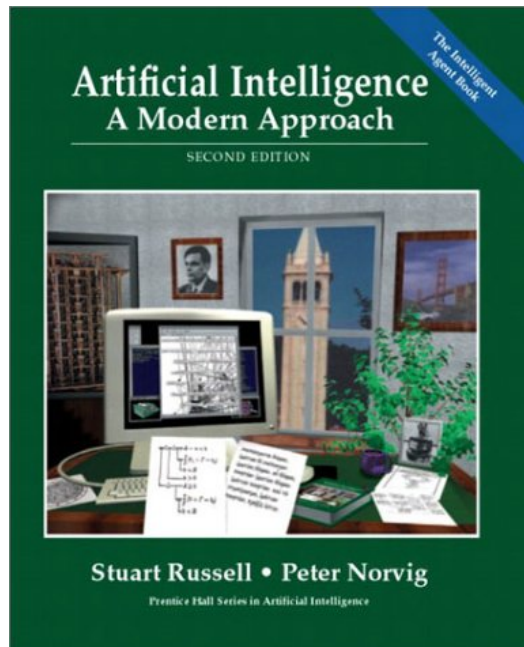
Verteilungsnotation

- Für eine diskrete Zufallsvariable X schreiben wir die **Verteilung** als $\mathbf{P}(X)$, wobei gilt
 $\mathbf{P}(X) = (P(x_1), \dots, P(x_n))^T$ für $x_1, x_2, \dots, x_n \in \text{dom}(X)$
- Auch im Verbund verwendet: $\mathbf{P}(X, Y)$
- $\mathbf{P}(X, Y) = \mathbf{P}(X | Y) \cdot \mathbf{P}(Y)$, wobei hier die Multiplikation komponentenweise erfolgt

Beispiel

Zahnarzt-Problem mit vier Variablen:

- **Toothache** (Sind besagte Schmerzen wirklich Zahnschmerzen?)
- **Cavity** (Es könnte ein Loch sein?)
- **Catch** (Stahlinstrument erzeugt Testschmerz?)
- **Weather** (Wetter: sunny,rainy,cloudy,snow)



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

Prior probability

- Prior or unconditional probabilities of propositions
e.g., $P(\text{Cavity} = \text{true}) = 0.1$ and $P(\text{Weather} = \text{sunny}) = 0.72$ correspond to belief prior to arrival of any (new) evidence
- Probability distribution
gives values for all possible assignments:
 $P(\text{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$
(normalized, i.e., sums to 1 because one must be the case)

Full joint probability distribution

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables

$P(\text{Weather}, \text{Cavity})$ is a 4×2 matrix of values:

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

- Full joint probability distribution: all random variables involved
 - $P(\text{Toothache}, \text{Catch}, \text{Cavity}, \text{Weather})$
- Every query about a domain can be answered by the full joint distribution

Discrete random variables: Notation

- $\text{Dom}(\text{Weather}) = \{\text{sunny}, \text{rainy}, \text{cloudy}, \text{snow}\}$ and $\text{Dom}(\text{Weather})$ disjoint from domain of other random variables:
 - Atomic event $\text{Weather}=\text{rainy}$ often written as rainy
 - Example: $P(\text{rainy})$, the random variable Weather is implicitly defined by the value rainy
- Boolean variable Cavity
 - Atomic event $\text{Cavity}=\text{true}$ written as cavity
 - Atomic event $\text{Cavity}=\text{false}$ written as $\neg\text{cavity}$
 - Examples: $P(\text{cavity})$ or $P(\neg\text{cavity})$

Conditional probability

- Conditional or posterior probabilities
e.g., $P(\text{cavity} \mid \text{toothache}) = 0.8$
or: $\langle 0.8 \rangle$
i.e., given that *toothache* is all I know
- (Notation for conditional distributions:
 $\mathbf{P}(\text{Cavity} \mid \text{Toothache})$ is a 2-element vector of 2-element vectors
- If we know more, e.g., *cavity* is also given, then we have
 $P(\text{cavity} \mid \text{toothache}, \text{cavity}) = 1$
- New evidence may be irrelevant, allowing simplification, e.g.,
 $P(\text{cavity} \mid \text{toothache}, \text{sunny}) = P(\text{cavity} \mid \text{toothache}) = 0.8$
- This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

- A general version holds for whole distributions, e.g.,

$$\mathbf{P}(\text{Weather}, \text{Cavity}) = \mathbf{P}(\text{Weather} \mid \text{Cavity}) \mathbf{P}(\text{Cavity})$$

$$\mathbf{P}(\text{Cavity}, \text{Weather}) = \mathbf{P}(\text{Cavity}) \mathbf{P}(\text{Weather} \mid \text{Cavity})$$

View as a set of 4×2 equations, **not** matrix mult.

$$(1,1) \mathbf{P}(\text{Weather}=\text{sunny} \mid \text{Cavity}=\text{true}) \mathbf{P}(\text{Cavity}=\text{true})$$

$$(1,2) \mathbf{P}(\text{Weather}=\text{sunny} \mid \text{Cavity}=\text{false}) \mathbf{P}(\text{Cavity}=\text{false}), \dots$$

- Chain rule** is derived by successive application of product rule:

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

Inference by enumeration

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
$\neg$ <i>cavity</i>	.016	.064	.144	.576

- For any proposition ϕ , sum the probability where it is true: $P(\phi) = \sum_{\omega: \omega \models \phi} P(\omega)$
- $P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$
- Unconditional or **marginal probability** of toothache
- Process is called marginalization or summing out

$$\text{Ax}_3: P(A \cup B) = P(A) + P(B)$$

für disjunkte Ereignisse $A, B \subseteq \Omega$

Marginalization and conditioning

- Let Y, Z be sequences of random variables s.th.
 $Y \cup Z$ denotes all random variables describing the world
- Marginalization
 - $P(Y) = \sum_{z \in Z} P(Y, z)$
- Conditioning
 - $P(Y) = \sum_{z \in Z} P(Y|z)P(z)$

Inference by enumeration

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
$\neg$ <i>cavity</i>	.016	.064	.144	.576

For any proposition ϕ , sum the atomic events where it is true:

$$P(\phi) = \sum_{\omega: \omega \models \phi} P(\omega)$$

- $P(\text{cavity} \vee \text{toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

$$(P(\text{cavity} \vee \text{toothache}) = P(\text{cavity}) + P(\text{toothache}) - P(\text{cavity} \wedge \text{toothache}))$$

Inference by enumeration

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
$\neg$ <i>cavity</i>	.016	.064	.144	.576

- Can also compute conditional probabilities:

$$\begin{aligned} P(\neg \text{cavity} \mid \text{toothache}) &= \frac{P(\neg \text{cavity} \wedge \text{toothache})}{P(\text{toothache})} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} \\ &= 0.08 / 0.2 = 0.4 \end{aligned}$$

Product rule

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

Normalization

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
$\neg$ <i>cavity</i>	.016	.064	.144	.576

- Denominator $\mathbf{P(\mathbf{z})}$ (or $P(\text{toothache})$ in the example before) can be viewed as a **normalization constant** α

$$\begin{aligned} P(\text{Cavity} \mid \text{toothache}) &= \alpha P(\text{Cavity}, \text{toothache}) \\ &= \alpha [P(\text{Cavity}, \text{toothache}, \text{catch}) + P(\text{Cavity}, \text{toothache}, \neg \text{catch})] \\ &= \alpha [<0.108, 0.016> + <0.012, 0.064>] \\ &= \alpha <0.12, 0.08> = <0.6, 0.4> \end{aligned}$$

General idea: compute distribution on query variable by fixing **evidence variables** (Toothache) and summing over **hidden variables** (Catch)

Inference by enumeration, contd.

Typically, we are interested in
the posterior joint distribution of the **query variables** $\mathbf{Y}$
given specific values $\mathbf{e}$ for the **evidence variables** $\mathbf{E}$
($\mathbf{X}$ are all variables of the modeled world)

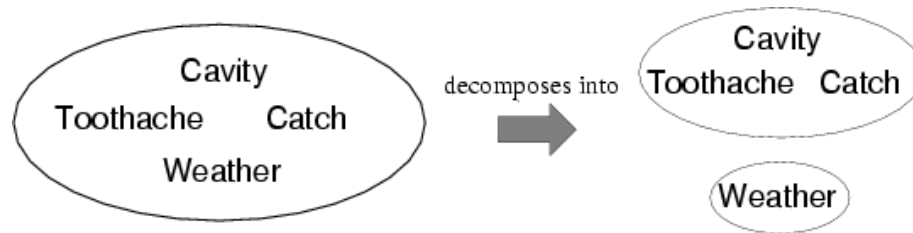
Let the **hidden variables** be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$ then the required summation of joint entries is done by summing out the hidden variables:

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

- The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$ and $\mathbf{H}$ together exhaust the set of random variables ($\mathbf{X}$)
- Obvious problems:
 1. Worst-case time complexity $O(d^n)$ where d is the largest arity and n denotes the number of random variables
 2. Space complexity $O(d^n)$ to store the joint distribution
 3. How to find the numbers for $O(d^n)$ entries?

Independence

- A and B are independent iff
 $\mathbf{P(A|B) = P(A)}$ or $\mathbf{P(B|A) = P(B)}$ or $\mathbf{P(A, B) = P(A) P(B)}$



$$\mathbf{P(\text{Toothache, Catch, Cavity, Weather})}$$
$$= \mathbf{P(\text{Toothache, Catch, Cavity}) P(\text{Weather})}$$

- 32 entries reduced to 12;
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

- $\mathbf{P}(\text{Toothache}, \text{Cavity}, \text{Catch})$ has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
(1) $\mathbf{P}(\text{catch} \mid \text{toothache}, \text{cavity}) = \mathbf{P}(\text{catch} \mid \text{cavity})$
- The same independence holds if I haven't got a cavity:
(2) $\mathbf{P}(\text{catch} \mid \text{toothache}, \neg \text{cavity}) = \mathbf{P}(\text{catch} \mid \neg \text{cavity})$
- Catch is **conditionally independent** of Toothache given Cavity:
 $\mathbf{P}(\text{Catch} \mid \text{Toothache}, \text{Cavity}) = \mathbf{P}(\text{Catch} \mid \text{Cavity})$
- Equivalent statements:
 $\mathbf{P}(\text{Toothache} \mid \text{Catch}, \text{Cavity}) = \mathbf{P}(\text{Toothache} \mid \text{Cavity})$
 $\mathbf{P}(\text{Toothache}, \text{Catch} \mid \text{Cavity}) = \mathbf{P}(\text{Toothache} \mid \text{Cavity}) \mathbf{P}(\text{Catch} \mid \text{Cavity})$

Conditional independence contd.

- Write out full joint distribution using chain rule:

$$\mathbf{P}(\text{Toothache}, \text{Catch}, \text{Cavity})$$

$$= \mathbf{P}(\text{Toothache} \mid \text{Catch}, \text{Cavity}) \mathbf{P}(\text{Catch}, \text{Cavity})$$

$$= \mathbf{P}(\text{Toothache} \mid \text{Catch}, \text{Cavity}) \mathbf{P}(\text{Catch} \mid \text{Cavity}) \mathbf{P}(\text{Cavity})$$

conditional independence

$$= \mathbf{P}(\text{Toothache} \mid \text{Cavity}) \mathbf{P}(\text{Catch} \mid \text{Cavity}) \mathbf{P}(\text{Cavity})$$

i.e., $2 + 2 + 1 = 5$ independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n .
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Naïve Bayes Model

$$\begin{aligned} & \mathbf{P}(Cavity|toothache \wedge catch) \\ &= \alpha \mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity) \\ &= \alpha \mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity) \end{aligned}$$

This is an example of a **naive Bayes** model:

$$\mathbf{P}(Cause, Effect_1, \dots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i|Cause)$$



Total number of parameters is **linear** in n

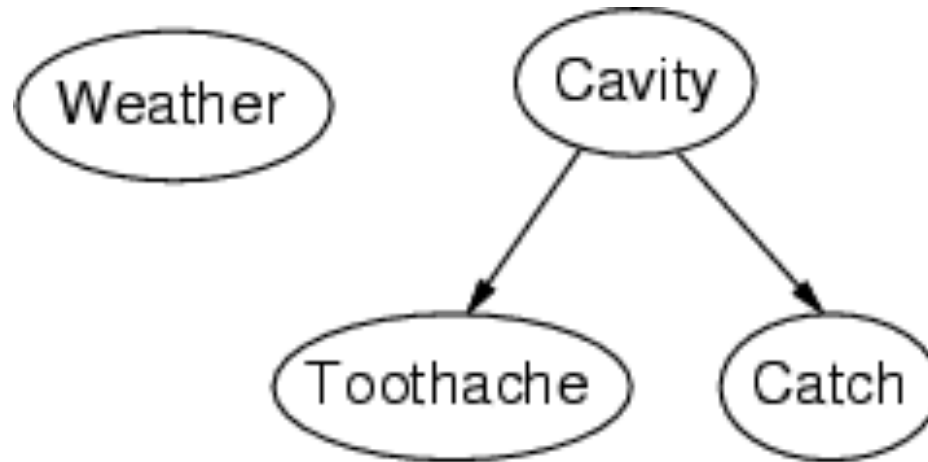
Usually, the assumption that effects are independent is wrong,
but works well in practice

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- Syntax:
 - a set of nodes, one per variable
 - a directed, acyclic graph (link $\approx$ "directly influences")
 - a conditional distribution for each node given its parents:
$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$
- In the simplest case, conditional distribution represented as a **conditional probability table** (CPT) giving the distribution over X_i for each combination of parent values

Example

- Topology of network encodes conditional independence assertions:

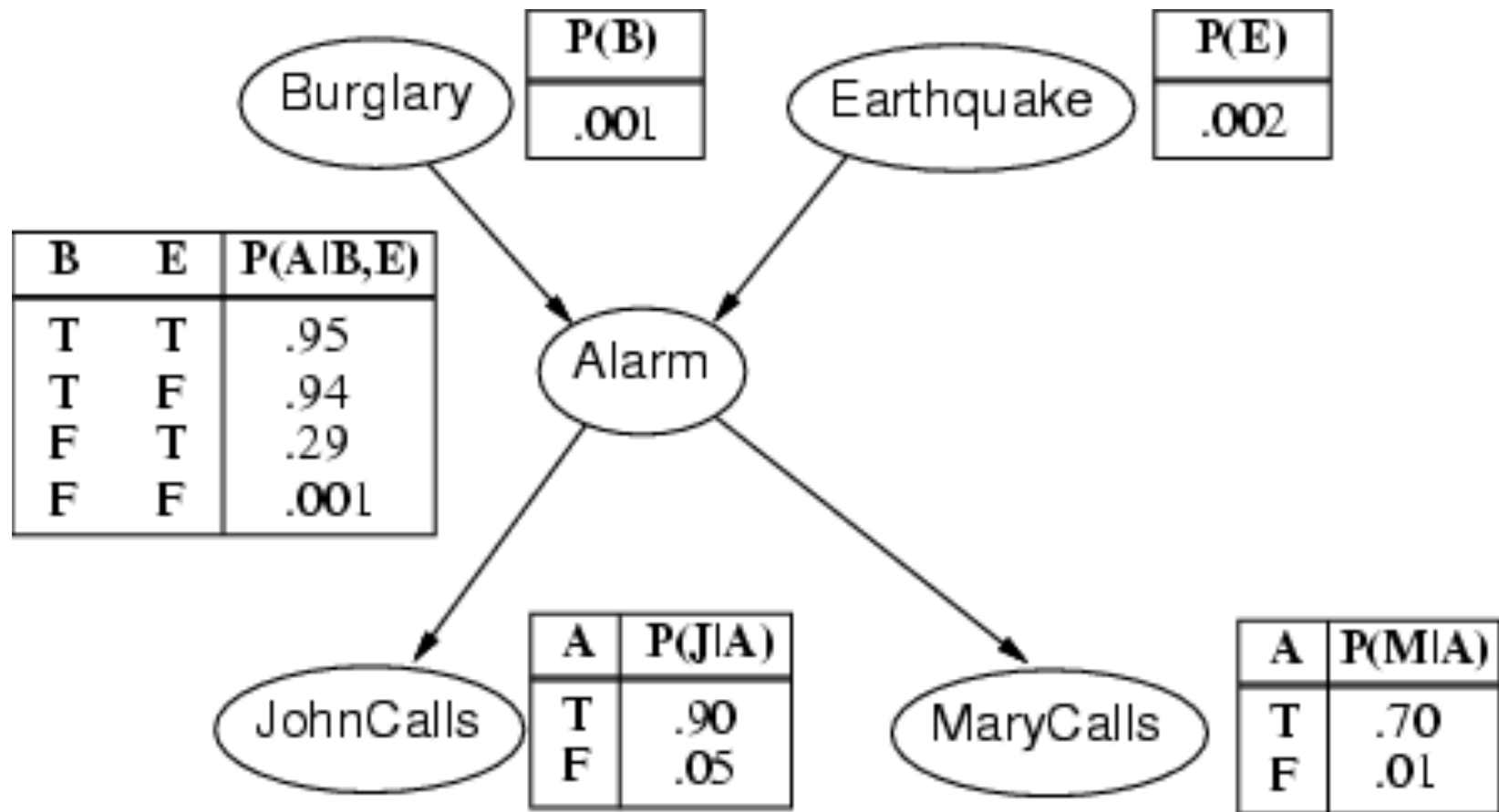


- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Example

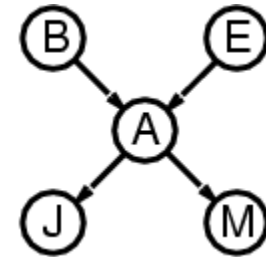
- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: *Burglary, Earthquake, Alarm, JohnCalls, MaryCalls*
- Network topology can reflect "causal" knowledge:
 - A burglar can set the alarm off
 - An earthquake can set the alarm off
 - The alarm can cause Mary to call
 - The alarm can cause John to call

Example contd.



Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values
- Each row requires one number p for $X_i = \text{true}$ (the number for $X_i = \text{false}$ is just $1-p$)
- If each variable has no more than k parents, the complete network requires $n \cdot 2^k$ numbers
- i.e., grows linearly with n , vs. 2^n for the full joint distribution
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)



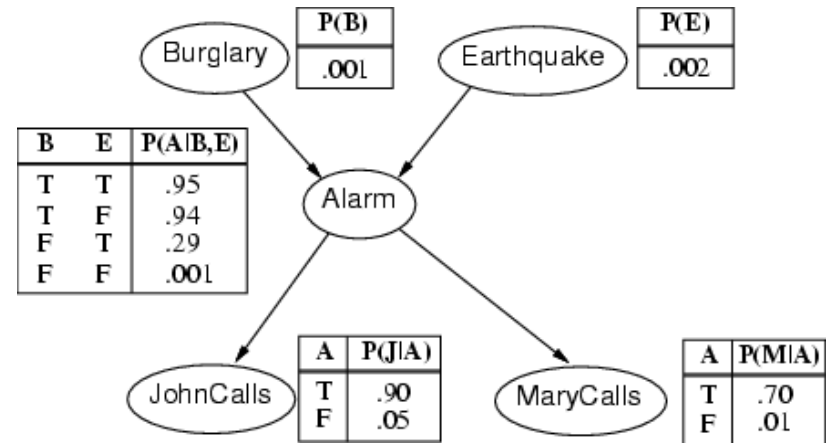
Semantics

The full joint distribution is defined as the product of the local conditional distributions:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{Parents}(X_i))$$

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

$$\begin{aligned} &= P(j | a) P(m | a) P(a | \neg b, \neg e) P(\neg b) P(\neg e) \\ &= 0.90 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

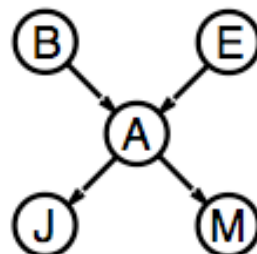


Inference by enumeration

Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation

Simple query on the burglary network:

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

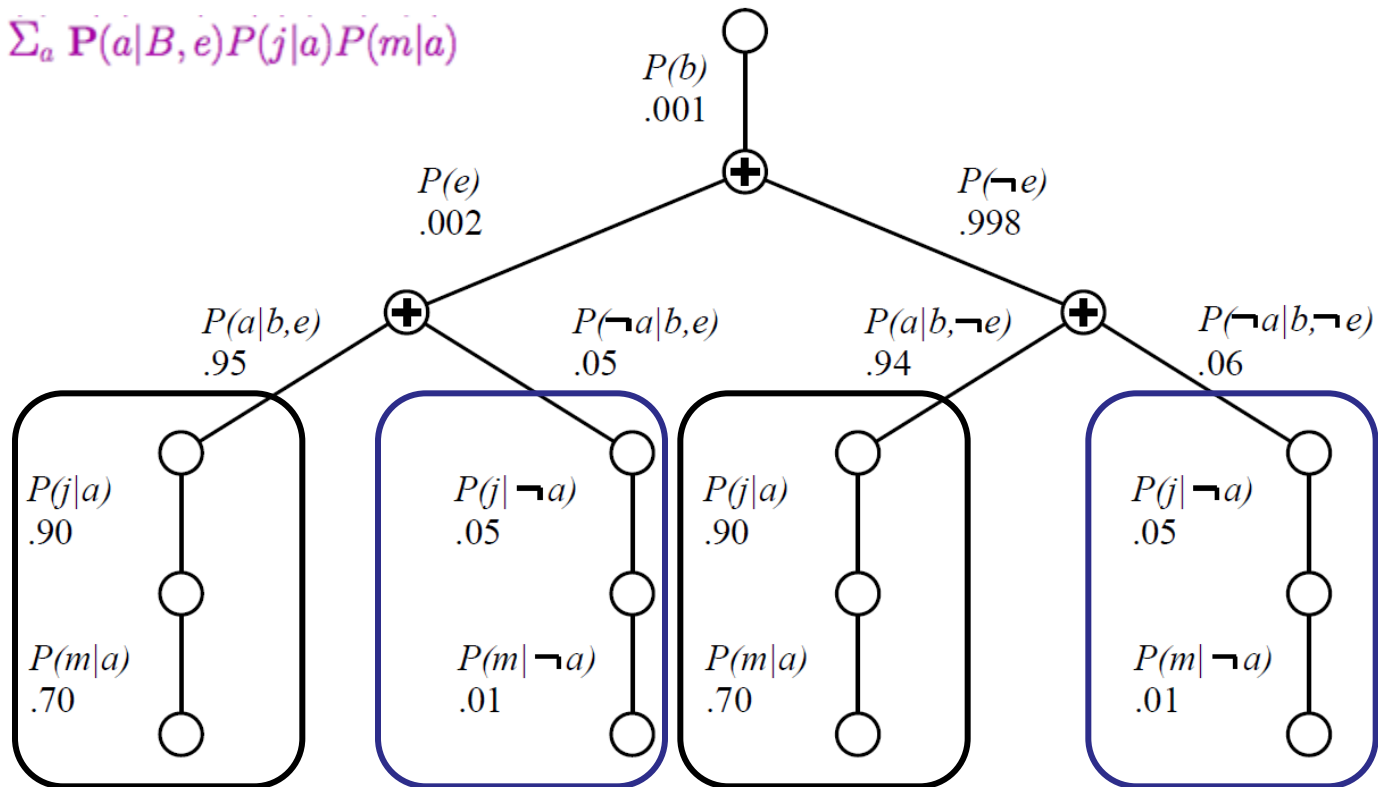


Rewrite full joint entries using product of CPT entries:

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

Evaluation Tree

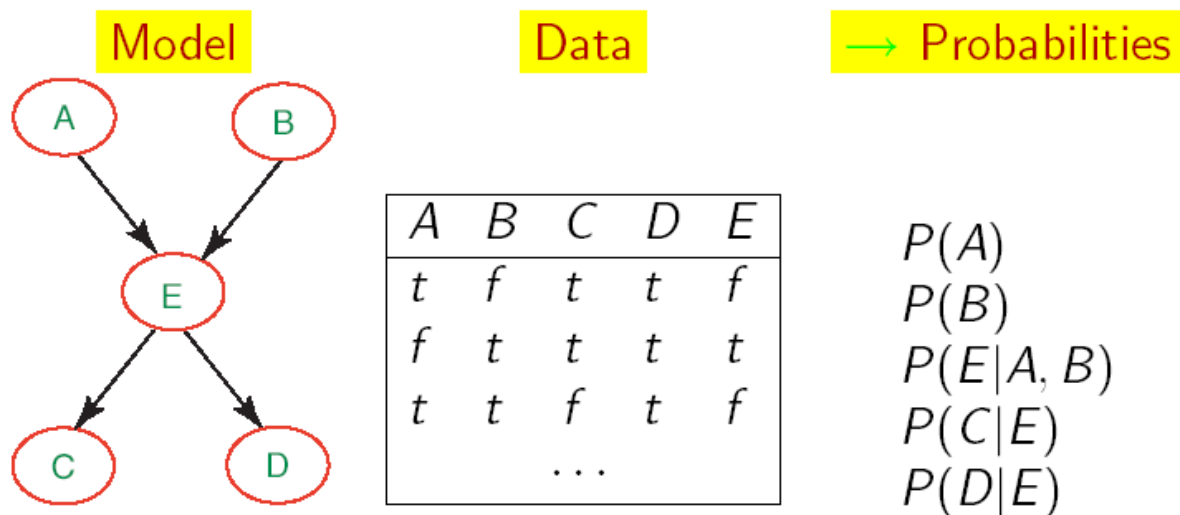
$$P(B) \sum_e P(e) \sum_a P(a|B, e) P(j|a) P(m|a)$$



Enumeration is inefficient: repeated computation
 e.g., computes $P(j|a)P(m|a)$ for each value of e

Learning BNs: Data Science w/ Complete Data

- We will start by applying ML to the simplest type of BNets learning:
 - known structure
 - Data containing observations for all variables
 - ✓ All variables are observable, no missing data
- The only thing that we need to learn are the network's parameters



Maximum-Likelihood-Parameterschätzung

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

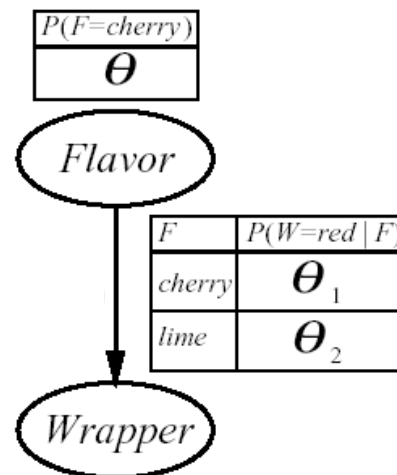
$$P(D \mid \theta) = \prod_m P(x[m] \mid \theta)$$

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

Gleichverteilte,
unabhängige
Stichproblem
(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

➤ $P(W=\text{green}, F = \text{cherry} | h_{\theta\theta_1\theta_2}) = (*)$

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

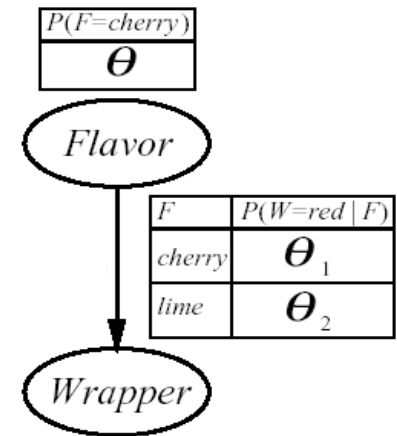
$$= \theta (1 - \theta_1)$$

➤ 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 (*)

➤ $P(\mathbf{d} | h_{\theta\theta_1\theta_2})$

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



Anwendungsbeispiel Bonbonfabrik

➤ Maximierung des Logarithmus der Zielfunktion

- $L = c \log \theta + \ell \log(1 - \theta) + r^c \log \theta_1 + g^c \log(1 - \theta_1) + r^\ell \log \theta_2 + g^\ell \log(1 - \theta_2)$

➤ Bestimmung der Ableitungen bzgl. $\theta, \theta_1, \theta_2$

- Ausdrücke ohne Term, nach dem abgeleitet wird, verschwinden

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

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

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

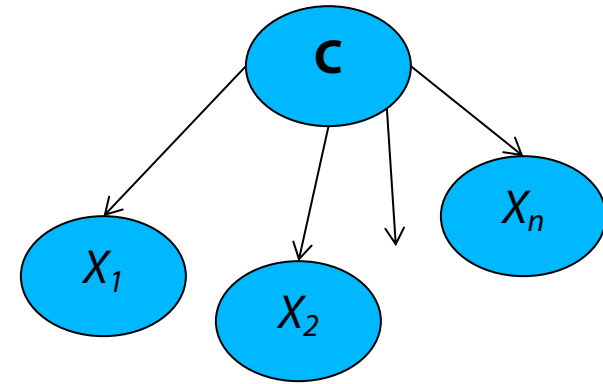
Maximum-Likelihood-Parameterschätzung

- 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

$$P(C|x_1, x_2, \dots, x_n) = \frac{P(C, x_1, x_2, \dots, x_n)}{P(x_1, x_2, \dots, x_n)} = \alpha P(C) \prod_i P(x_i | C)$$

➤ Deterministische Vorhersagen können durch Wahl der wahrscheinlichsten Klasse erreicht werden

➤ Skalierung auf realen Daten sehr gut:

- $2n + 1$ Parameter benötigt

Anwendung: Diagnose

Useful for assessing **diagnostic** probability from **causal** probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

E.g., let M be meningitis, S be stiff neck:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!