

Marcel Gehrke

Data Understanding vs. Machine Training

Stochastical Basics

Stochastical Basics

Probabilities

Initial Concept of Probabilities

- Limit value of the relative frequency of occurrence of an event
 - Example: Rolling an even number
- "Elementary events" have equal probabilities of occurrence
 - Laplace's principle
- What exactly are **elementary events**?
 - Elements ω of a population Ω
 - Elementary events are abstract: $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$
 - Example: ω_i represents the roll of a die
 - Mapping of events to characteristic values using random variables
 - Random variable X for *the result* of the die roll is a function
 - Objective: Map event ω_i to the number i on the top of the rolled die
 - $X : \Omega \rightarrow \{1, 2, 3, 4, 5, 6\}$

Events

- (Complex) events are subsets of Ω
 - Example: Dice events with even numbers
 - Definition of random variables extended accordingly
 - $X: \mathcal{P}(\Omega) \rightarrow \mathcal{P}(M)$
 - Example: $X(\{\omega_2, \omega_4, \omega_6\}) = \{2, 4, 6\}$

$\mathcal{P}(X)$ is the power set of X .

Laplace Probabilities

- Consider the finite population of elementary events $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$
- For an event $A \subseteq \Omega$, we define the **Laplace probability** of an event $A \subseteq \Omega$ is defined as the number
 - $P(A) := |A| / |\Omega| = |A| / n$, where $|A|$ is the number of elements in A .
- Each elementary event $\omega_i, i = 1, \dots, n$ therefore has the probability $P(\{\omega_i\}) = 1/n$
- We say that X has a **distribution** characterised by $P(\{\omega_i\}) = 1/n$
- The probability of Ω is $P(\Omega) = 1$

Bayesian Probability Theory

- Laplace distributions are too specific!
- Examples:
 - Unfair dice
 - Probability of giving birth to a boy
 - Occurrence of heads or tails on Euro coins
- Elementary events are not always equally probable!
- Concept of a **priori probability**
 - **Prior knowledge and basic assumptions of an observer** summarised in a probability distribution
 - ... and explicitly expressed in the model

Probability Spaces

A (discrete) probability space is defined as a pair (Ω, P) where

- Ω is a (countable) population and
- P is a probability measure that assigns a probability $P(A)$ to each subset $A \subseteq \Omega$.

P is again defined via the probabilities $P(\{\omega\})$ of the elementary events $\omega \in A$:

where the following must apply for $P(\{\omega\})$: $P(A) = \sum_{\omega \in A} P(\{\omega\})$

$0 \leq P(\{\omega\}) \leq 1$ for all ω and $\sum_{\omega \in \Omega} P(\{\omega\}) = 1$

Axioms of Kolmogorov [1903-1987]

- We consider any (countable) population Ω and a function P that assigns a probability to each event $A \subseteq \Omega$.
- We call P a probability distribution on Ω if it satisfies the following properties:
 - Ax_1 : $P(A) \geq 0$ for any $A \subseteq \Omega$
 - Ax_2 : $P(\Omega) = 1$
 - Ax_3 : $P(A \cup B) = P(A) + P(B)$ for disjoint events $A, B \subseteq \Omega$

Implications

- $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$
for pairwise disjoint events $A_1, A_2, \dots, A_n \subset \Omega$
- $P(A) \leq P(B)$ if $A \subseteq B$
- Define the complement of A : $\bar{A} = \Omega \setminus A$.
Then, the following applies: $P(\bar{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any $A, B \subset \Omega$
 \rightsquigarrow representation in the Venn diagram

Joint Probability

- Let us consider two consecutive dice rolls
- The event space (Ω, P) is then defined as follows
 - $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\} \times \{\omega_1, \omega_2, \dots, \omega_6\}$
 - $P(A \times B)$ with $A, B \subseteq \{\omega_1, \omega_2, \dots, \omega_6\}$ is a discrete probability measure
- We refer to this as a **joint probability** and write
 - $P(A, B)$ for $P(A \times B)$ with $A, B \subseteq \Omega$
- **Any populations** can be linked in a joint probability
 - Example: (Ω', P) with $\Omega' = \{\omega_1, \omega_2, \dots, \omega_6\} \times \{\text{clubs, spades, hearts, diamonds}\}$

Conditional Probabilities

- For events $A, B \subseteq \Omega$ with $P(B) > 0$, the conditional probability of A given B is defined as the number

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

- Example: Dice game
 - "An even number is rolled" $A := \{ \omega_2, \omega_4, \omega_6 \}$
 - "A number > 4 is rolled" $B := \{ \omega_5, \omega_6 \}$
 - Then:
 - $P(A|B) = 1/2$
 - $P(A|\bar{B}) = 2/4 = 1/2$

Bayes' Theorem

- Thomas Bayes [1701-1761]
- This theorem is based on the asymmetry of the definition of conditional probabilities:

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

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

- Analogous for joint probabilities

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

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

Stochastical Basics

Independencies

Stochastic Independence

- When are two events **A** and **B** independent?
- Motivation via conditional probabilities:
- Two events **A** and **B** are **independent** if
 - $P(A | B) = \frac{P(A \cap B)}{P(B)} = P(A) \quad P(B) > 0$
 - or $P(B | A) = \frac{P(A \cap B)}{P(A)} = P(B) \quad P(A) > 0$
 - or $P(A, B) = P(A) \cdot P(B)$ applies.
 - The conditions $P(B) > 0$ and $P(A) > 0$ are not necessary here.

Example: Rolling the Die Twice

- A fair die is thrown twice in succession.
 - **A** stands for "a six on the first roll"
 - **B** stands for "a six on the second roll"
- For each roll of the die, the population is $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$ with $\omega_i =$ "The number rolled is i "
- According to Laplace, $P(A) = P(B) = 1/6$.
- For an "independent" roll, the following therefore applies
 $P(A, B) = P(A) \cdot P(B) = 1/36$

Conditional Independence

- Let C be any event with $P(C) > 0$.

Two events A and B are called **conditionally independent** given C if the following applies:

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

- In other words:

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

Notation

- Let $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$, $M = \{1, 2, 3, 4, 5, 6\}$
- Let X be a random variable $X : \Omega \rightarrow M$
 - Example: $X: \omega_2 \mapsto 2$
- Notation: $\text{dom}(X)$ stands for M
- Notation: $X = 2$ stands for **elementary event** $\{\omega_2\}$
 - $P(X=2)$ stands for $P(\{\omega_2\})$
- Notation: $X = 2 \vee X = 4 \vee X = 6$ stands for **complex event** $\{\omega_2, \omega_4, \omega_6\}$
 - $P(X = 2 \vee X = 4 \vee X = 6)$ stands for $P(\{\omega_2, \omega_4, \omega_6\})$
- Notation: $X = 2 \wedge Y = 4$ stands for compound event $\{\omega_2\} \times \{\omega_4\}$ (different variables)

Distribution Notation

- For a discrete random variable X , we write the **distribution** as $\mathbf{P}(X)$, where
$$\mathbf{P}(X) = (P(x_1), \dots, P(x_n))^T \text{ for } x_1, x_2, \dots, x_n \in \text{dom}(X)$$
- Also used in combination: $\mathbf{P}(X, Y)$
- $\mathbf{P}(X, Y) = \mathbf{P}(X | Y) \cdot \mathbf{P}(Y)$, where the multiplication is performed component-wise
- $\langle P(x_1), \dots, P(x_n) \rangle$ stands for $(P(x_1), \dots, P(x_n))^T$

Example

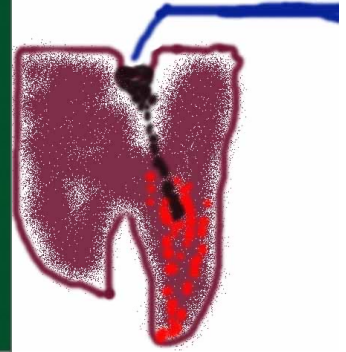
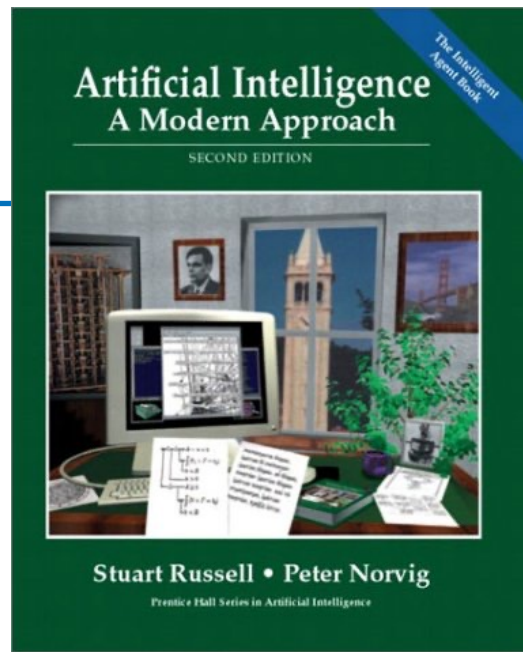
Dentist problem with four variables:

Toothache (Is the pain really toothache?)

Cavity (Could it be a hole?)

Catch (Is the steel instrument causing test pain?)

Weather (sunny, rainy, cloudy, snow)



The following presentations contain material from Chapter 14 (Sections 1 and 2)

Prior Probability

- **Prior or conditional probability** of propositions

e.g., $P(\text{hole} = \text{true}) = 0.1$ and $P(\text{weather} = \text{sunny}) = 0.72$ correspond to the assumptions before we receive (new) observations

- **Joint probability distribution**

Provides values for all possible assignments:

$P(\text{weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$

(**normalised**, i.e., adds up to 1)

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

- $\text{Dom}(\text{Weather}) = \{\text{sunny, rainy, cloudy, snow}\}$ 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: $P(\text{Cavity} \mid \text{Toothache})$ 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., $P(\text{Weather}, \text{Cavity}) = P(\text{Weather} | \text{Cavity}) P(\text{Cavity})$
 $P(\text{Cavity}, \text{Weather}) = P(\text{Cavity})P(\text{Weather} | \text{Cavity})$

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

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

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

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

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

Inference by Enumeration

- Start with the complete joint probability distribution:

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

- For each query ϕ , sum the probabilities, where ϕ 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

- This process is called marginalisation or summation

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

for disjoint events $A, B \subseteq \Omega$

Marginalisation and Conditioning

- Let **Y** and **Z** be sequences of random variables such that **Y U Z** describes all random variables
- Marginalisation
 - $P(\mathbf{Y}) = \sum_{z \in Z} P(\mathbf{Y}, z)$
- Conditioning
 - $P(\mathbf{Y}) = \sum_{z \in Z} P(\mathbf{Y}|z)P(z)$

Inference by Enumeration

- Start with the joint probability distribution:

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

- Can also compute conditional probabilities:

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

Normalisation

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

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

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

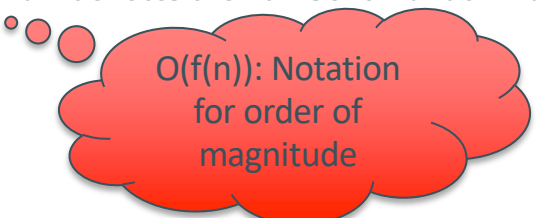
General idea: compute distribution on query variable by fixing **evidence variables** (toothache) and summing over **hidden variables** (catch)

Inference by Enumeration Contd.

Typically, we are interested in the posterior joint distribution of the **query variables** Y given specific values e for the **evidence variables** E (X are all variables of the modelled world)

Let the **hidden variables** be $H = X - Y - E$ then the required summation of joint entries is done by summing out the hidden variables: $P(Y | E = e) = \alpha P(Y, E = e) = \alpha \sum_h P(Y, E = e, H = h)$

- The terms in the summation are joint entries because Y , E and H together exhaust the set of random variables (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?

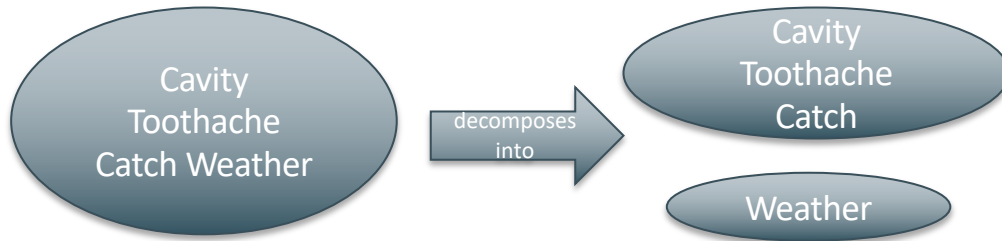


$O(f(n))$: Notation
for order of
magnitude

Independence

- A and B are independent iff

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B) \quad \text{or} \quad P(A, B) = P(A) P(B)$$



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

- 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

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

Conditional Independence Contd.

- Write out full joint distribution using chain rule:

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

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

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

conditional independence

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

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.

Stochastical Basics

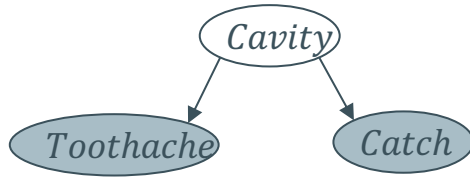
Bayesian Models

Naïve Bayes Model

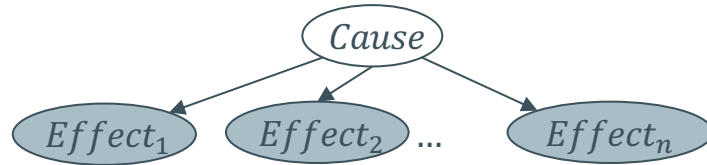
$$\begin{aligned} P(\text{Cavity} \mid \text{toothache} \wedge \text{catch}) &= \alpha P(\text{toothache} \wedge \text{catch} \mid \text{Cavity}) P(\text{Cavity}) \\ &= \alpha P(\text{toothache} \mid \text{Cavity}) P(\text{catch} \mid \text{Cavity}) P(\text{Cavity}) \end{aligned}$$

Is an example of a **Naïve Bayes** model

$$P(\text{Cause}, \text{Effect}_1, \dots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i \mid \text{Cause})$$



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



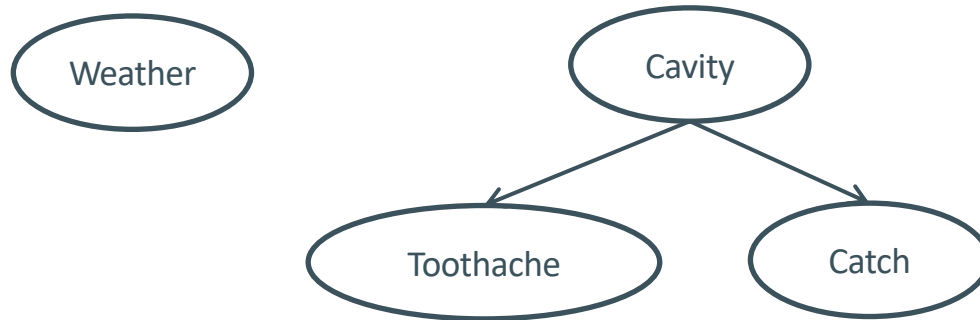
Usually, the assumption that effects are independent is wrong, but works well in practice

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: $P(X_i | \text{Parents}(X_i))$
- 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:

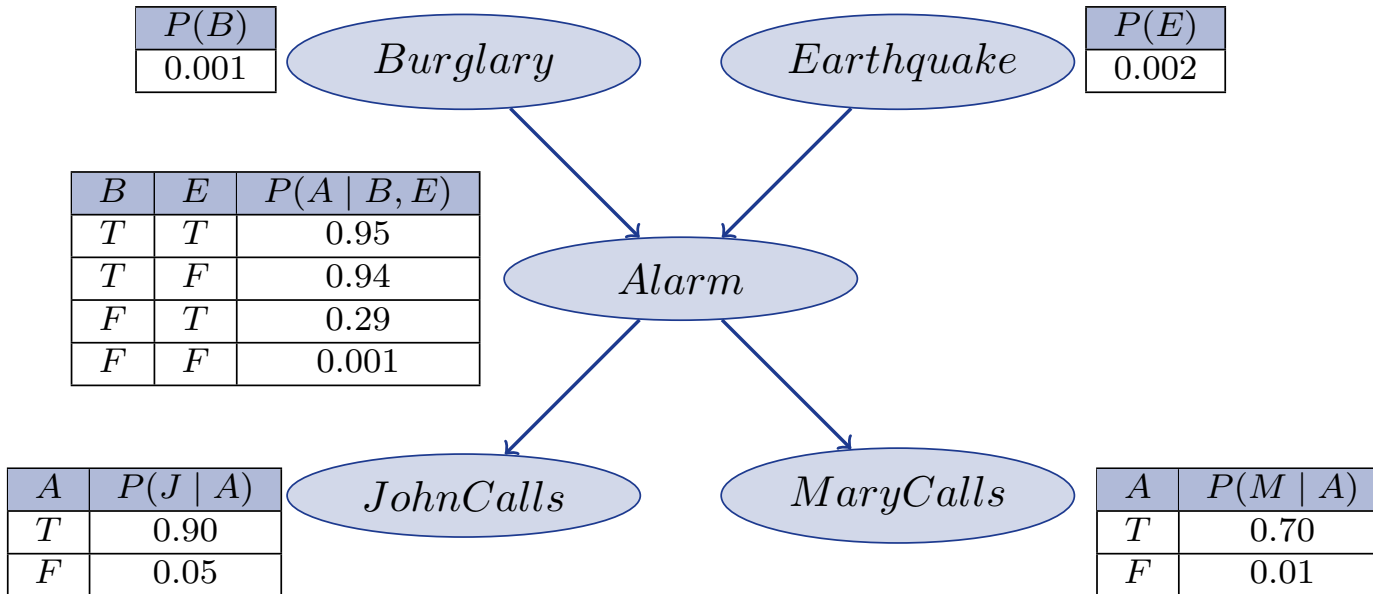


- Weather* is independent of the other variables
- Toothache* and *Catch* are conditionally independent given *Cavity*

Example

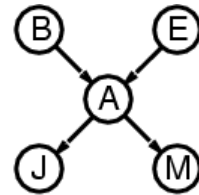
- I'm at work, neighbour John calls to say my alarm is ringing, but neighbour Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Network topology can reflect "causal" knowledge:
 - A burglar can 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. $2^5 - 1 = 31$)

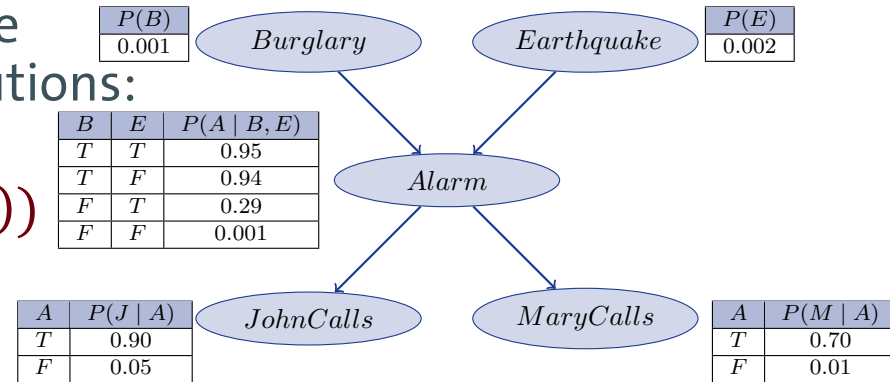


Semantics

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

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

- e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$
 $= P(j \mid a) \cdot P(m \mid a) \cdot P(a \mid \neg b, \neg e) \cdot P(\neg b) \cdot P(\neg e)$
 $= 0.90 \cdot 0.70 \cdot 0.001 \cdot 0.999 \cdot 0.998$
 ≈ 0.00063

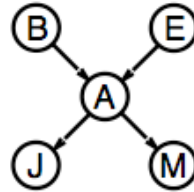


Inference by Enumeration

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

Simple query on the burglary network:

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



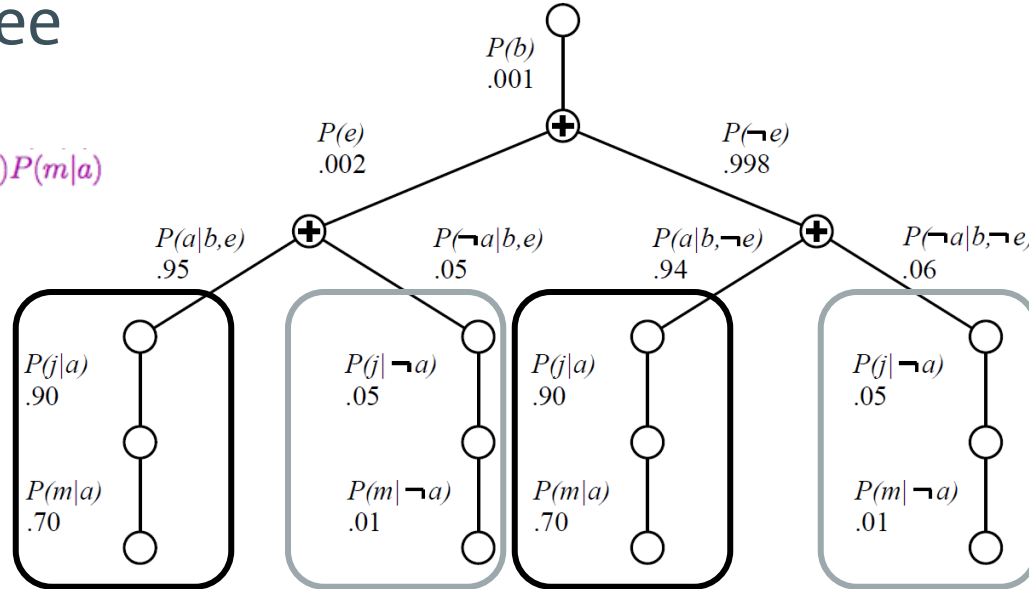
Rewrite full joint entries using product of CPT entries:

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

Recursive depth-first enumeration: $O(n)$ space, $O(d^n)$ time

Evaluation Tree

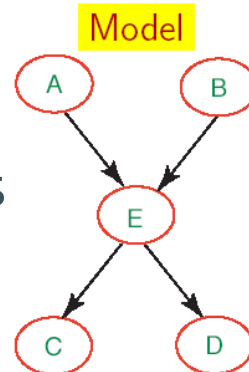
$$P(B) \sum_e P(e) \sum_a P(a|B, e) P(j|a) P(m|a)$$



Enumeration is inefficient: recurring calculations
 e.g. calculations of $P(j | a)P(m | a)$ for the different values of e

Learning BNs: Data Science with Complete Data

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



Data

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

→ Probabilities

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

Maximum Likelihood Parameter Estimation

- Assume that the structure of a Bayesian network is known
- Objective: Estimate Bayesian network parameters θ
 - Entries in CPTs, $P(X | Parents(X))$
- A parameterisation θ is good if it is likely to generate the observed data:

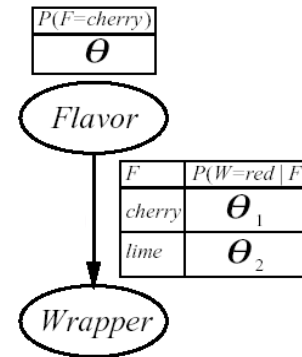
$$P(D | \theta) = \prod_m P(x[m] | \theta)$$

- Maximum likelihood estimation (MLE) principle:
Choose θ^* such that $P(D | \theta^*)$ is maximised

Equally distributed,
independent samples
(i.i.d. samples)

Application Example: Candy Factory

- A manufacturer chooses the colour of the sweet wrapper with a certain probability depending on the flavour, whereby the corresponding distribution is unknown
 - If flavour = cherry, choose red paper with probability θ_1
 - If flavour = lime, choose red paper with probability θ_2
- The Bayesian network contains three parameters to be learned
 - $\theta, \theta_1, \theta_2$

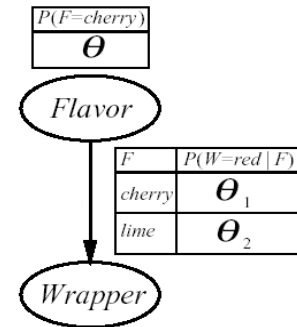


Application Example: Candy Factory

- $$\begin{aligned}
 P(W = \text{green}, F = \text{cherry} | h_{\theta, \theta_1, \theta_2}) &= (*) \\
 &= P(W = \text{green} | F = \text{cherry}, h_{\theta, \theta_1, \theta_2}) P(F = \text{cherry} | h_{\theta, \theta_1, \theta_2}) = (1 - \Theta_1) \Theta
 \end{aligned}$$

- We unwrap N sweets

- c are cherry and l are lime
 - r^c cherry with red paper, g^c cherry with green paper
 - r^l lime with red paper, g^l lime with green paper
 - Each attempt yields a combination of paper and flavour as in (*)



- $$P(\mathbf{d} | h_{\theta, \theta_1, \theta_2}) = \prod_j P(d_j | h_{\theta, \theta_1, \theta_2}) = \Theta^c \cdot (1 - \Theta)^l \cdot (\Theta_1)^{r^c} \cdot (1 - \Theta_1)^{g^c} \cdot (\Theta_2)^{r^l} \cdot (1 - \Theta_2)^{g^l}$$

Application Example: Candy Factory

- Maximising the logarithm of the objective function

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

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

$$\frac{\partial L}{\partial \theta_1} = \frac{r^c}{\theta_1} - \frac{g^c}{1 - \theta_1} = 0 \Rightarrow \theta_1 = \frac{r^c}{r^c + g^c}$$

- Determination of the derivatives with respect to $\theta, \theta_1, \theta_2$

- Expressions without a term to be differentiated disappear

$$\frac{\partial L}{\partial \theta_2} = \frac{r^l}{\theta_2} - \frac{g^l}{1 - \theta_2} = 0 \Rightarrow \theta_2 = \frac{r^l}{r^l + g^l}$$

Maximum Likelihood Parameter Estimation

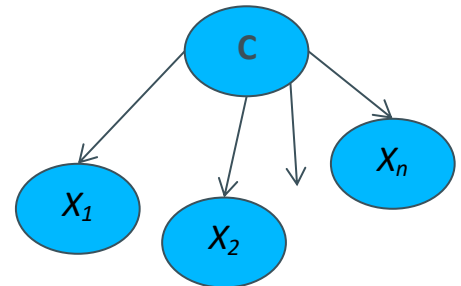
- Estimation by forming relative frequencies
- This process is applicable to any fully observable Bayesian network
- With complete data and maximum likelihood parameter estimation:
 - Parameter learning breaks down into separate learning problems for each parameter (CPT) through logarithmisation
 - Each parameter is determined by the relative frequency of a node value given the values of the parent nodes

Popular Application: Naïve Bayes Model

- Naïve Bayes model: Very simple Bayesian network for classification
 - *Class variable* C (to be predicted) forms root
 - Attribute variables X_i (observations) are leaves
- Naïve because it is assumed that the attribute values are conditionally independent when the class is given

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

- Deterministic predictions can be achieved by selecting the most probable class
- Scales very well to real data:
 - $2n + 1$ parameters required



Application: Diagnosis

Useful for estimating **diagnoses** Probabilities of **causal** dependencies

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

Let M be meningitis and S be a stiff neck:

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

Note: The conditional probability of meningitis is still very small!