

Applied Probability Theory

Why (Kolmogorov) Axioms?

If P models an *objectively* observable probability, these axioms are obviously reasonable.

However, why should an agent obey formal axioms when modeling degrees of (subjective) belief?

Objective vs. subjective probabilities

Axioms constrain the set of beliefs an agent can abide.

Finetti (1931) gave one of the most plausible arguments why subjective beliefs should respect axioms:

“When using contradictory beliefs, the agent will eventually fail.”

Unconditional Probabilities

$P(A)$ designates the *unconditioned* or *a priori* probability that $A \subseteq \Omega$ occurs if *no* other additional information is present. For example:

$$P(\text{cavity}) = 0.1$$

Note: Here, **cavity** is a proposition.

A formally different way to state the same would be via a binary random variable **Cavity**:

$$P(\text{Cavity} = \text{true}) = 0.1$$

A priori probabilities are derived from statistical surveys or general rules.

Unconditional Probabilities

In general a random variable can assume more than two values:

$$P(\text{Weather} = \text{sunny}) = 0.7$$

$$P(\text{Weather} = \text{rainy}) = 0.2$$

$$P(\text{Weather} = \text{cloudy}) = 0.02$$

$$P(\text{Weather} = \text{snowy}) = 0.08$$

$$P(\text{Headache} = \text{true}) = 0.1$$

$P(X)$ designates the vector of probabilities for the (ordered) domain of the random variable X :

$$P(\text{Weather}) = \langle 0.7, 0.2, 0.02, 0.08 \rangle$$

$$P(\text{Headache}) = \langle 0.1, 0.9 \rangle$$

Both vectors define the respective probability distributions of the two random variables.

Conditional Probabilities

New evidence can alter the probability of an event.

Example: The probability for cavity increases if information about a toothache arises.

With additional information present, the a priori knowledge must not be used!

$P(A | B)$ designates the *conditional* or *a posteriori* probability of A *given* the sole observation (*evidence*) B .

$$P(\text{cavity} | \text{toothache}) = 0.8$$

For random variables X and Y $P(X | Y)$ represents the set of conditional distributions for each possible value of Y .

Conditional Probabilities

$P(\text{Weather} \mid \text{Headache})$ consists of the following table:

	$h \hat{=} \text{Headache} = \text{true}$	$\neg h \hat{=} \text{Headache} = \text{false}$
Weather = sunny	$P(W = \text{sunny} \mid h)$	$P(W = \text{sunny} \mid \neg h)$
Weather = rainy	$P(W = \text{rainy} \mid h)$	$P(W = \text{rainy} \mid \neg h)$
Weather = cloudy	$P(W = \text{cloudy} \mid h)$	$P(W = \text{cloudy} \mid \neg h)$
Weather = snowy	$P(W = \text{snowy} \mid h)$	$P(W = \text{snowy} \mid \neg h)$

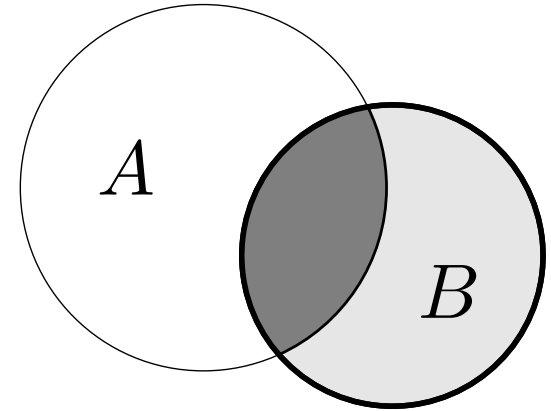
Note that we are dealing with *two* distributions now!
Therefore each column sums up to unity!

Formal definition:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} \quad \text{if } P(B) > 0$$

Conditional Probabilities

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



Product Rule: $P(A \wedge B) = P(A | B) \cdot P(B)$

Also: $P(A \wedge B) = P(B | A) \cdot P(A)$

A and B are *independent* iff

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

Equivalently, iff the following equation holds true:

$$P(A \wedge B) = P(A) \cdot P(B)$$

Interpretation of Conditional Probabilities

Caution! Common misinterpretation:

“ $P(A | B) = 0.8$ means, that $P(A) = 0.8$, given B holds.”

This statement is wrong due to (at least) two facts:

$P(A)$ is *always* the a-priori probability,
never the probability of A given that B holds!

$P(A | B) = 0.8$ is only applicable as long as no other evidence except B is present.
If C becomes known, $P(A | B \wedge C)$ has to be determined.

In general we have:

$$P(A | B \wedge C) \neq P(A | B)$$

E. g. $C \rightarrow A$ might apply.

Joint Probabilities

Let X_1, \dots, X_n be random variables over the same frame of discernment Ω and event algebra \mathcal{E} . Then $\vec{X} = (X_1, \dots, X_n)$ is called a *random vector* with

$$\vec{X}(\omega) = (X_1(\omega), \dots, X_n(\omega))$$

Shorthand notation:

$$P(\vec{X} = (x_1, \dots, x_n)) = P(X_1 = x_1, \dots, X_n = x_n) = P(x_1, \dots, x_n)$$

Definition:

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n) &= P\left(\left\{ \omega \in \Omega \mid \bigwedge_{i=1}^n X_i(\omega) = x_i \right\}\right) \\ &= P\left(\bigcap_{i=1}^n \{X_i = x_i\}\right) \end{aligned}$$

Joint Probabilities

Example: $P(\text{Headache}, \text{Weather})$ is the *joint probability distribution* of both random variables and consists of the following table:

	$h \hat{=} \text{Headache} = \text{true}$	$\neg h \hat{=} \text{Headache} = \text{false}$
Weather = sunny	$P(W = \text{sunny} \wedge h)$	$P(W = \text{sunny} \wedge \neg h)$
Weather = rainy	$P(W = \text{rainy} \wedge h)$	$P(W = \text{rainy} \wedge \neg h)$
Weather = cloudy	$P(W = \text{cloudy} \wedge h)$	$P(W = \text{cloudy} \wedge \neg h)$
Weather = snowy	$P(W = \text{snowy} \wedge h)$	$P(W = \text{snowy} \wedge \neg h)$

All table cells sum up to unity.

Calculating with Joint Probabilities

All desired probabilities can be computed from a joint probability distribution.

	toothache	\neg toothache
cavity	0.04	0.06
\neg cavity	0.01	0.89

$$\begin{aligned} \text{Example: } P(\text{cavity} \vee \text{toothache}) &= P(\text{cavity} \wedge \text{toothache}) \\ &\quad + P(\neg\text{cavity} \wedge \text{toothache}) \\ &\quad + P(\text{cavity} \wedge \neg\text{toothache}) = 0.11 \end{aligned}$$

$$\begin{aligned} \text{Marginalizations: } P(\text{cavity}) &= P(\text{cavity} \wedge \text{toothache}) \\ &\quad + P(\text{cavity} \wedge \neg\text{toothache}) = 0.10 \end{aligned}$$

Conditioning:

$$P(\text{cavity} \mid \text{toothache}) = \frac{P(\text{cavity} \wedge \text{toothache})}{P(\text{toothache})} = \frac{0.04}{0.04 + 0.01} = 0.80$$

Problems

Easiness of computing all desired probabilities comes at an unaffordable price:

Given n random variables with k possible values each, the joint probability distribution contains k^n entries which is infeasible in practical applications.

Hard to handle.

Hard to estimate.

Therefore:

1. Is there a more *dense* representation of joint probability distributions?
2. Is there a more *efficient* way of processing this representation?

The answer is *no* for the general case, however, certain dependencies and independencies can be exploited to reduce the number of parameters to a practical size.

Stochastic Independence

Two events A and B are *stochastically independent* iff

$$\begin{aligned} P(A \wedge B) &= P(A) \cdot P(B) \\ &\Leftrightarrow \\ P(A \mid B) &= P(A) = P(A \mid \overline{B}) \end{aligned}$$

Two random variables X and Y are *stochastically independent* iff

$$\begin{aligned} \forall x \in \text{dom}(X) : \forall y \in \text{dom}(Y) : \quad &P(X = x, Y = y) = P(X = x) \cdot P(Y = y) \\ &\Leftrightarrow \\ \forall x \in \text{dom}(X) : \forall y \in \text{dom}(Y) : \quad &P(X = x \mid Y = y) = P(X = x) \end{aligned}$$

Shorthand notation: $P(X, Y) = P(X) \cdot P(Y)$.

Note the formal difference between $P(A) \in [0, 1]$ and $P(X) \in [0, 1]^{|\text{dom}(X)|}$.

Conditional Independence

Let X , Y and Z be three random variables. We call X and Y *conditionally independent given Z* , iff the following condition holds:

$$\forall x \in \text{dom}(X) : \forall y \in \text{dom}(Y) : \forall z \in \text{dom}(Z) : \\ P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z) \cdot P(Y = y \mid Z = z)$$

Shorthand notation: $X \perp\!\!\!\perp_P Y \mid Z$

Let $\mathbf{X} = \{A_1, \dots, A_k\}$, $\mathbf{Y} = \{B_1, \dots, B_l\}$ and $\mathbf{Z} = \{C_1, \dots, C_m\}$ be three disjoint sets of random variables. We call \mathbf{X} and \mathbf{Y} *conditionally independent given \mathbf{Z}* , iff

$$P(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}) \cdot P(\mathbf{Y} \mid \mathbf{Z}) \Leftrightarrow P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})$$

Shorthand notation: $\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$

Conditional Independence

The complete condition for $\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$ would read as follows:

$$\forall a_1 \in \text{dom}(A_1) : \dots \forall a_k \in \text{dom}(A_k) :$$

$$\forall b_1 \in \text{dom}(B_1) : \dots \forall b_l \in \text{dom}(B_l) :$$

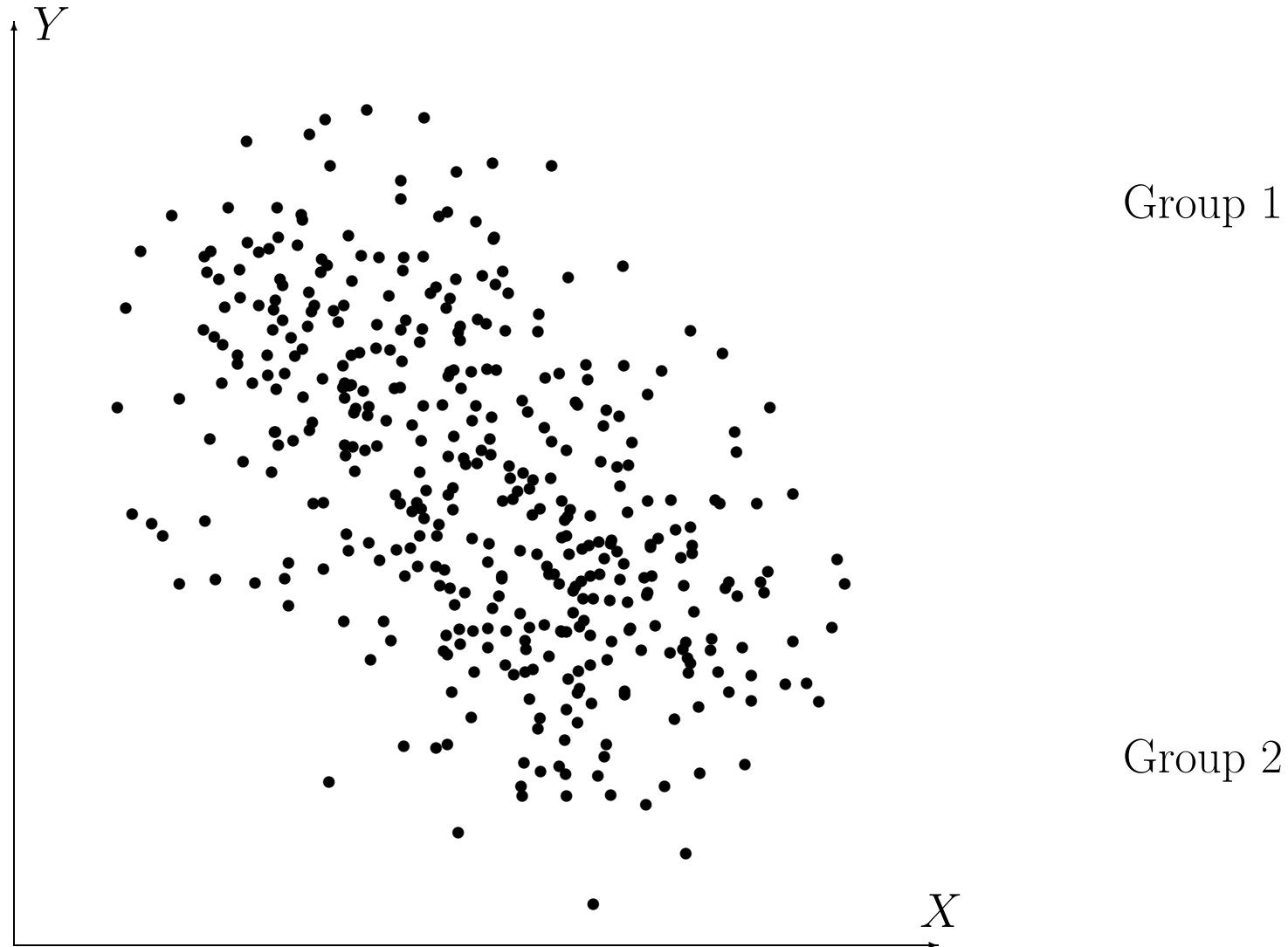
$$\forall c_1 \in \text{dom}(C_1) : \dots \forall c_m \in \text{dom}(C_m) :$$

$$\begin{aligned} & P(A_1 = a_1, \dots, A_k = a_k, B_1 = b_1, \dots, B_l = b_l \mid C_1 = c_1, \dots, C_m = c_m) \\ & = P(A_1 = a_1, \dots, A_k = a_k \mid C_1 = c_1, \dots, C_m = c_m) \\ & \quad \cdot P(B_1 = b_1, \dots, B_l = b_l \mid C_1 = c_1, \dots, C_m = c_m) \end{aligned}$$

Remarks:

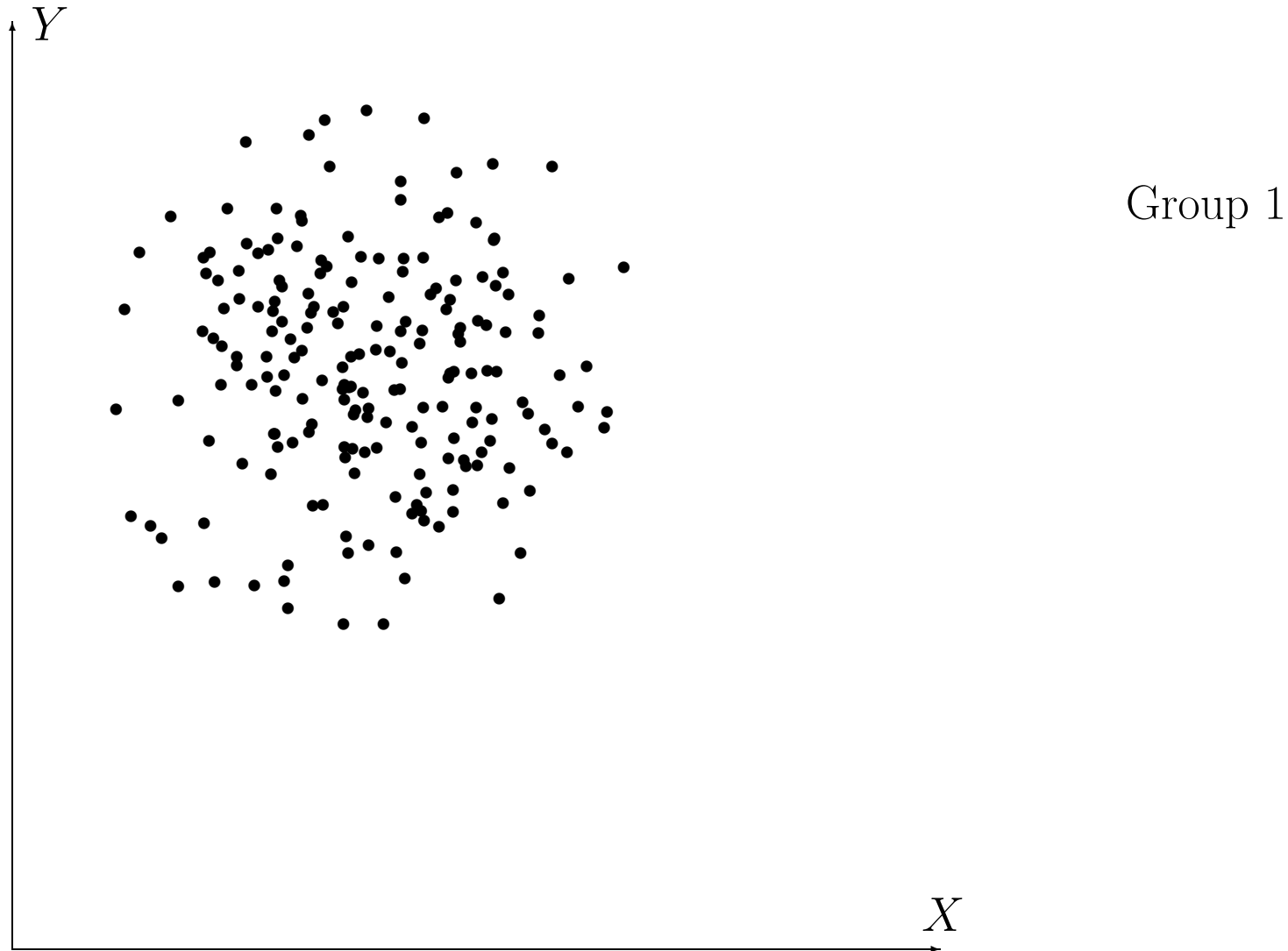
1. If $\mathbf{Z} = \emptyset$ we get (unconditional) independence.
2. We do not use curly braces ($\{\}$) for the sets if the context is clear. Likewise, we use X instead of \mathbf{X} to denote sets.

Conditional Independence — Example 1



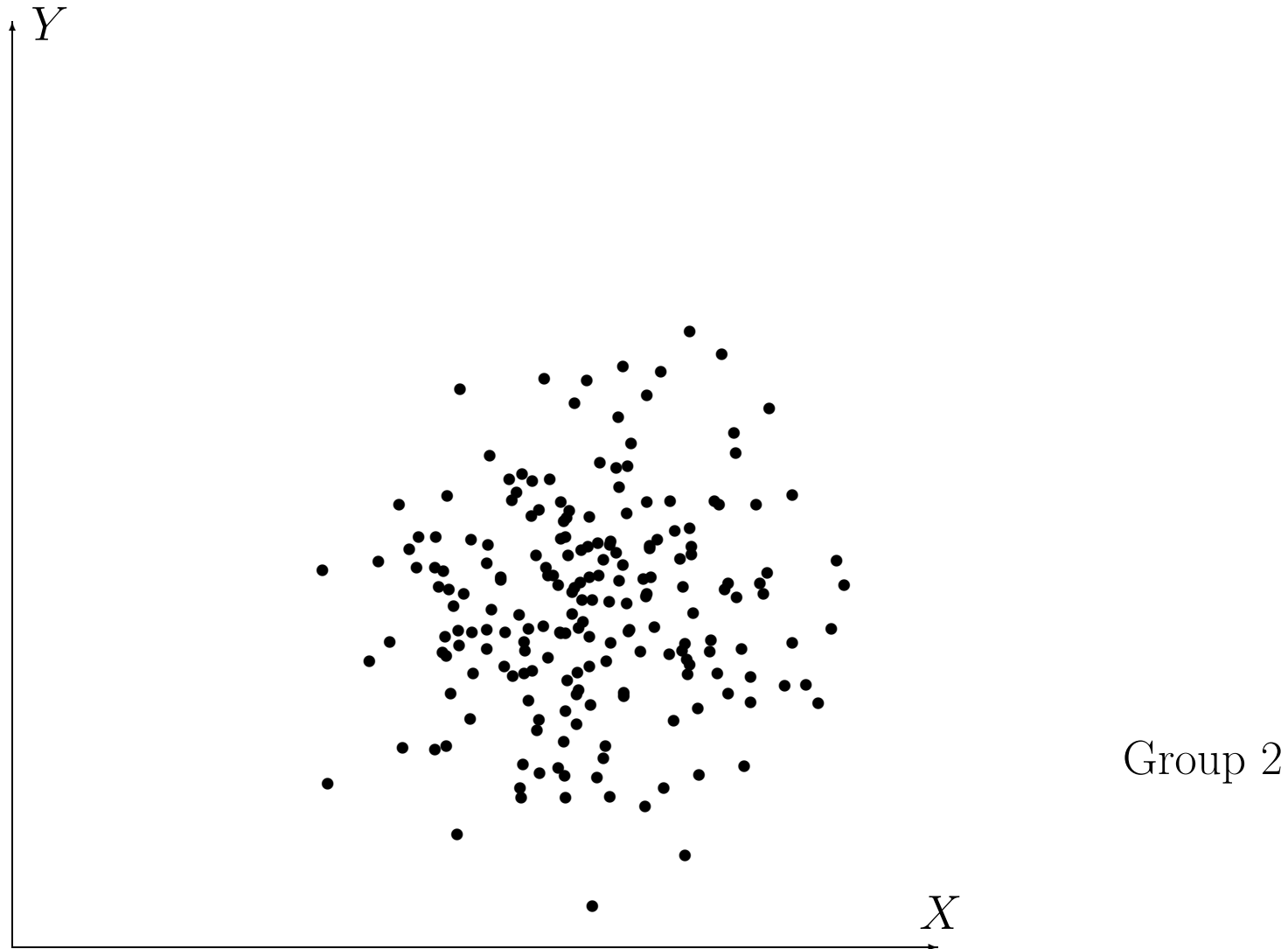
(Weak) Dependence in the entire dataset: X and Y dependent.

Conditional Independence — Example 1



No Dependence in Group 1: X and Y conditionally independent given Group 1.

Conditional Independence — Example 1



No Dependence in Group 2: X and Y conditionally independent given Group 2.

Conditional Independence — Example 2

- $\text{dom}(G) = \{\text{mal}, \text{fem}\}$ Geschlecht (gender)
- $\text{dom}(S) = \{\text{sm}, \overline{\text{sm}}\}$ Raucher (smoker)
- $\text{dom}(M) = \{\text{mar}, \overline{\text{mar}}\}$ Verheiratet (married)
- $\text{dom}(P) = \{\text{preg}, \overline{\text{preg}}\}$ Schwanger (pregnant)

p_{GSMP}		G = mal		G = fem	
		S = sm	S = $\overline{\text{sm}}$	S = sm	S = $\overline{\text{sm}}$
M = mar	P = preg	0	0	0.01	0.05
	P = $\overline{\text{preg}}$	0.04	0.16	0.02	0.12
M = $\overline{\text{mar}}$	P = preg	0	0	0.01	0.01
	P = $\overline{\text{preg}}$	0.10	0.20	0.07	0.21

Conditional Independence — Example 2

$$P(\mathbf{G}=\text{fem}) = P(\mathbf{G}=\text{mal}) = 0.5$$

$$P(\mathbf{S}=\text{sm}) = 0.25$$

$$P(\mathbf{P}=\text{preg}) = 0.08$$

$$P(\mathbf{M}=\text{mar}) = 0.4$$

Gender and Smoker are not independent:

$$P(\mathbf{G}=\text{fem} \mid \mathbf{S}=\text{sm}) = 0.44 \neq 0.5 = P(\mathbf{G}=\text{fem})$$

Gender and Marriage are marginally independent but conditionally dependent given Pregnancy:

$$P(\text{fem}, \text{mar} \mid \overline{\text{preg}}) \approx 0.152 \neq 0.169 \approx P(\text{fem} \mid \overline{\text{preg}}) \cdot P(\text{mar} \mid \overline{\text{preg}})$$

Bayes Theorem

Product Rule (for events A and B):

$$P(A \cap B) = P(A | B)P(B) \quad \text{and} \quad P(A \cap B) = P(B | A)P(A)$$

Equating the right-hand sides:

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

For random variables X and Y :

$$\forall x \forall y : \quad P(Y = y | X = x) = \frac{P(X = x | Y = y)P(Y = y)}{P(X = x)}$$

Generalization concerning background knowledge/evidence E :

$$P(Y | X, E) = \frac{P(X | Y, E)P(Y | E)}{P(X | E)}$$

Bayes Theorem — Application

$$P(\text{toothache} \mid \text{cavity}) = 0.4$$

$$P(\text{cavity}) = 0.1$$

$$P(\text{toothache}) = 0.05$$

$$P(\text{cavity} \mid \text{toothache}) = \frac{0.4 \cdot 0.1}{0.05} = 0.8$$

Why not estimate $P(\text{cavity} \mid \text{toothache})$ right from the start?

Causal knowledge like $P(\text{toothache} \mid \text{cavity})$ is more robust than diagnostic knowledge $P(\text{cavity} \mid \text{toothache})$.

The causality $P(\text{toothache} \mid \text{cavity})$ is independent of the a priori probabilities $P(\text{toothache})$ and $P(\text{cavity})$.

If $P(\text{cavity})$ rose in a caries epidemic, the causality $P(\text{toothache} \mid \text{cavity})$ would remain constant whereas both $P(\text{cavity} \mid \text{toothache})$ and $P(\text{toothache})$ would increase according to $P(\text{cavity})$.

A physician, after having estimated $P(\text{cavity} \mid \text{toothache})$, would not know a rule for updating.

Relative Probabilities

Assumption:

We would like to consider the probability of the diagnosis **GumDisease** as well.

$$\begin{aligned}P(\text{toothache} \mid \text{gumdisease}) &= 0.7 \\P(\text{gumdisease}) &= 0.02\end{aligned}$$

Which diagnosis is more probable?

If we are interested in *relative probabilities* only (which may be sufficient for some decisions), $P(\text{toothache})$ needs not to be estimated:

$$\begin{aligned}\frac{P(C \mid T)}{P(G \mid T)} &= \frac{P(T \mid C)P(C)}{P(T)} \cdot \frac{P(T)}{P(T \mid G)P(G)} \\&= \frac{P(T \mid C)P(C)}{P(T \mid G)P(G)} = \frac{0.4 \cdot 0.1}{0.7 \cdot 0.02} \\&= 28.57\end{aligned}$$

Normalization

If we are interested in the absolute probability of $P(C | T)$ but do not know $P(T)$, we may conduct a complete case analysis (according C) and exploit the fact that $P(C | T) + P(\neg C | T) = 1$.

$$P(C | T) = \frac{P(T | C)P(C)}{P(T)}$$

$$P(\neg C | T) = \frac{P(T | \neg C)P(\neg C)}{P(T)}$$

$$1 = P(C | T) + P(\neg C | T) = \frac{P(T | C)P(C)}{P(T)} + \frac{P(T | \neg C)P(\neg C)}{P(T)}$$

$$P(T) = P(T | C)P(C) + P(T | \neg C)P(\neg C)$$

Normalization

Plugging into the equation for $P(C | T)$ yields:

$$P(C | T) = \frac{P(T | C)P(C)}{P(T | C)P(C) + P(T | \neg C)P(\neg C)}$$

For general random variables, the equation reads:

$$P(Y = y | X = x) = \frac{P(X = x | Y = y)P(Y = y)}{\sum_{\forall y' \in \text{dom}(Y)} P(X = x | Y = y')P(Y = y')}$$

Note the “loop variable” y' . Do not confuse with y .

Multiple Evidences

The patient complains about a toothache. From this first evidence the dentist infers:

$$P(\text{cavity} \mid \text{toothache}) = 0.8$$

The dentist palpates the tooth with a metal probe which catches into a fracture:

$$P(\text{cavity} \mid \text{fracture}) = 0.95$$

Both conclusions might be inferred via Bayes rule. But what does the combined evidence yield? Using Bayes rule further, the dentist might want to determine:

$$P(\text{cavity} \mid \text{toothache} \wedge \text{fracture}) = \frac{P(\text{toothache} \wedge \text{fracture} \mid \text{cavity}) \cdot P(\text{cavity})}{P(\text{toothache} \wedge \text{fracture})}$$

Multiple Evidences

Problem:

He needs $P(\text{toothache} \wedge \text{catch} \mid \text{cavity})$, i. e. diagnostics knowledge for all combinations of symptoms in general. Better incorporate evidences step-by-step:

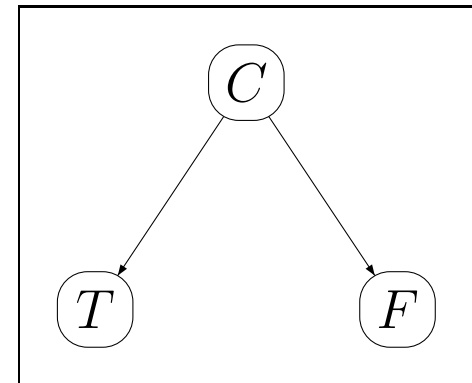
$$P(Y \mid X, E) = \frac{P(X \mid Y, E)P(Y \mid E)}{P(X \mid E)}$$

Abbreviations:

C — cavity

T — toothache

F — fracture



Objective:

Computing $P(C \mid T, F)$ with just causal statements of the form $P(\cdot \mid C)$ and under exploitation of independence relations among the variables.

Multiple Evidences

A priori: $P(C)$

Evidence toothache: $P(C | T) = P(C) \frac{P(T | C)}{P(T)}$

Evidence fracture: $P(C | T, F) = P(C | T) \frac{P(F | C, T)}{P(F | T)}$

$$T \perp\!\!\!\perp F | C \quad \Leftrightarrow \quad P(F | C, T) = P(F | C)$$

$$P(C | T, F) = P(C) \frac{P(T | C)}{P(T)} \frac{P(F | C)}{P(F | T)}$$

Seems that we still have to cope with symptom inter-dependencies?!

Multiple Evidences

Compound equation from last slide:

$$\begin{aligned} P(C | T, F) &= P(C) \frac{P(T | C) P(F | C)}{P(T) P(F | T)} \\ &= P(C) \frac{P(T | C) P(F | C)}{P(F, T)} \end{aligned}$$

$P(F, T)$ is a normalizing constant and can be computed if $P(F | \neg C)$ and $P(T | \neg C)$ are known:

$$P(F, T) = \underbrace{P(F, T | C)}_{P(F|C)P(T|C)} P(C) + \underbrace{P(F, T | \neg C)}_{P(F|\neg C)P(T|\neg C)} P(\neg C)$$

Therefore, we finally arrive at the following solution...

Multiple Evidences

$$P(C | F, T) = \frac{\boxed{P(C)} \boxed{P(T | C)} \boxed{P(F | C)}}{\boxed{P(F | C)} \boxed{P(T | C)} \boxed{P(C)} + \boxed{P(F | \neg C)} \boxed{P(T | \neg C)} \boxed{P(\neg C)}}$$

Note that we only use causal probabilities $P(\cdot | C)$ together with the a priori (marginal) probabilities $P(C)$ and $P(\neg C)$.

Multiple Evidences — Summary

Multiple evidences can be treated by reduction on
a priori probabilities
(causal) conditional probabilities for the evidence
under assumption of conditional independence

General rule:

$$P(Z | X, Y) = \alpha P(Z) P(X | Z) P(Y | Z)$$

for X and Y conditionally independent given Z and with normalizing constant α .

Monty Hall Puzzle

Marylin Vos Savant in her riddle column in the New York Times:

You are a candidate in a game show and have to choose between three doors. Behind one of them is a Porsche, whereas behind the other two there are goats. After you chose a door, the host Monty Hall (who knows what is behind each door) opens another (not your chosen one) door with a goat. Now you have the choice between keeping your chosen door or choose the remaining one.

Which decision yields the best chance of winning the Porsche?

Monty Hall Puzzle

G You win the Porsche.

R You revise your decision.

A Behind your initially chosen door is (and remains) the Porsche.

$$\begin{aligned}P(G \mid R) &= P(G, A \mid R) + P(G, \bar{A} \mid R) \\&= P(G \mid A, R)P(A \mid R) + P(G \mid \bar{A}, R)P(\bar{A} \mid R) \\&= 0 \cdot P(A \mid R) + 1 \cdot P(\bar{A} \mid R) \\&= P(\bar{A} \mid R) = P(\bar{A}) = \frac{2}{3}\end{aligned}$$

$$\begin{aligned}P(G \mid \bar{R}) &= P(G, A \mid \bar{R}) + P(G, \bar{A} \mid \bar{R}) \\&= P(G \mid A, \bar{R})P(A \mid \bar{R}) + P(G \mid \bar{A}, \bar{R})P(\bar{A} \mid \bar{R}) \\&= 1 \cdot P(A \mid \bar{R}) + 0 \cdot P(\bar{A} \mid \bar{R}) \\&= P(A \mid \bar{R}) = P(A) = \frac{1}{3}\end{aligned}$$

Simpson's Paradox

Example: C = Patient takes medication, E = patient recovers

	E	$\neg E$	Σ	Recovery rate
C	20	20	40	50%
$\neg C$	16	24	40	40%
Σ	36	44	80	

Men	E	$\neg E$	Σ	Rec.rate	Women	E	$\neg E$	Σ	Rec.rate
C	18	12	30	60%	C	2	8	10	20%
$\neg C$	7	3	10	70%	$\neg C$	9	21	30	30%
	25	15	40			11	29	40	

$$P(E | C) > P(E | \neg C)$$

but

$$P(E | C, M) < P(E | \neg C, M)$$

$$P(E | C, W) < P(E | \neg C, W)$$

Probabilistic Reasoning

Probabilistic reasoning is difficult and may be problematic:

- $P(A \wedge B)$ is not determined simply by $P(A)$ and $P(B)$:
 $P(A) = P(B) = 0.5 \Rightarrow P(A \wedge B) \in [0, 0.5]$
- $P(C | A) = x, P(C | B) = y \Rightarrow P(C | A \wedge B) \in [0, 1]$
Probabilistic logic is *not truth functional!*

Central problem: How does additional information affect the current knowledge?
I. e., if $P(B | A)$ is known, what can be said about $P(B | A \wedge C)$?

High complexity: n propositions $\rightarrow 2^n$ full conjunctives

Hard to specify these probabilities.

Summary

Uncertainty is inevitable in complex and dynamic scenarios that force agents to cope with ignorance.

Probabilities express the agent's inability to vote for a definitive decision. They model the degree of belief.

If an agent violates the axioms of probability, it may exhibit irrational behavior in certain circumstances.

The Bayes rule is used to derive unknown probabilities from present knowledge and new evidence.

Multiple evidences can be effectively included into computations exploiting conditional independencies.