

Probabilistic Reasoning

course slides to accompany syllabus

2023-2024

© L.C. van der Gaag, S. Renooij

UU – ICS Master Programmes:
Computing Science
Artificial Intelligence
Data Science



Universiteit Utrecht

Part One

Probabilistic reasoning

- Lecturer:** Silja Renooij (s.renooij@uu.nl)
Matthijs Vákár (m.i.l.vakar@uu.nl)
- Prerequisites:** probability theory & graph theory
- Literature:** syllabus, papers, slides & studymanual
- Additional info:** see course website:
<http://www.cs.uu.nl/docs/vakken/prob/>

Practical information

Course form:

- lectures
- exercises (formative self assessment)
(tip: discuss exercises together)

Grading:

- practical assignments (partially formative; qualitative)
(description + deadlines on Blackboard)
- written exam (summative)

Syllabus, Chapter 1:

Introduction

Reasoning under uncertainty

In numerous application areas of knowledge-based decision-support systems we have

- uncertainty concerning the general domain knowledge;
- problem-specific information that is often uncertain, incomplete and even contradictory.

A decision-support system should be capable of dealing with these types of knowledge.

Application of probability theory

Consider a joint probability distribution \Pr on a set of discrete random variables $\mathbf{V} = \{V_1, \dots, V_n\}$. Then, in general:

- representing \Pr requires **exponential space**
consider e.g. $n = 2$ binary-valued variables, or $n = 40$; what if they have 5 values each? (and how do you get the numbers?)
- calculating a probability from \Pr by conditioning and marginalisation requires **exponential time**
consider e.g. computing $\Pr(V_1 = \text{true})$ from $\Pr(\mathbf{V})$, or $\Pr(V_1 = \text{true} \mid V_2 = \text{true})$

This cannot be improved without additional **knowledge** about the probability distribution.

Diagnosis problem: pioneering in the 1960s

Let $H = \{h_1, \dots, h_n\}$, $n \geq 1$, be a set of hypotheses, and let $E = \{e_1, \dots, e_m\}$, $m \geq 1$, be a set of relevant findings (evidence).

Determine the 'best' diagnosis given findings $e \subseteq E$.

The approach: Compute for each $h \in H$ the probability

$$\Pr(h \mid e) = \frac{\Pr(e \mid h) \Pr(h)}{\Pr(e)}$$

Drawback: An exponential number of probabilities need to be computed; storage is also exponential.

Pioneering in the 1960s

Determine the diagnosis given findings $e \subseteq E$.

The approach: Assume $h_i \in H$ mutually exclusive, and collectively exhaustive: $\cup_{i=1}^n \{h_i\} = \Omega$.

Then, compute for each $h_i \in H$:

$$\Pr(h_i | e) = \frac{\Pr(e | h_i) \Pr(h_i)}{\Pr(e)} = \frac{\Pr(e | h_i) \Pr(h_i)}{\sum_{k=1}^n \Pr(e | h_k) \Pr(h_k)}$$

Drawback: We compute only $n - 1$ probabilities, but computation still requires an exponential number of probabilities.

Pioneering in the 1960s

Determine the diagnosis given findings $\mathbf{e} = \{e_p, \dots, e_q\}$,
 $1 \leq p, q \leq m$.

The approach: Assume *in addition* that all findings e_1, \dots, e_m are conditionally independent given $h_i, i = 1, \dots, n$. Then:

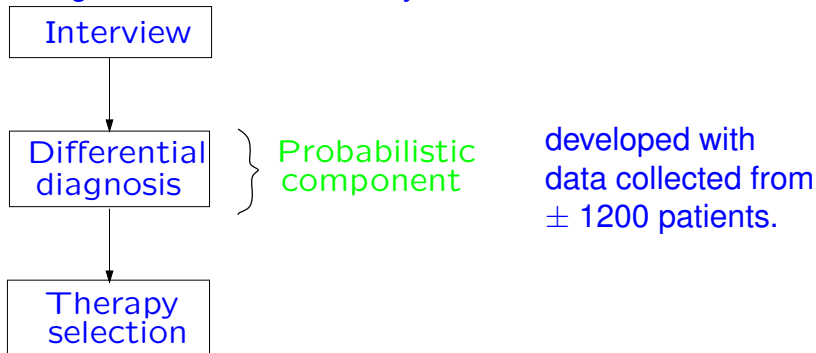
$$\begin{aligned}\Pr(h_i | \mathbf{e}) &= \frac{\Pr(e_p, \dots, e_q | h_i) \Pr(h_i)}{\sum_{k=1}^n \Pr(e_p, \dots, e_q | h_k) \Pr(h_k)} \\ &= \frac{\Pr(e_p | h_i) \cdot \dots \cdot \Pr(e_q | h_i) \Pr(h_i)}{\sum_{k=1}^n \Pr(e_p | h_k) \cdot \dots \cdot \Pr(e_q | h_k) \Pr(h_k)}\end{aligned}$$

Benefit: Only $m \cdot n$ conditional probabilities and $n - 1$ prior probabilities are required for the computation.

GLADYS

GLADYS (GLASGOW DYSPEPSIA SYSTEM) is a system for diagnosing dyspepsia.

The global structure of the system:



D.J. Spiegelhalter, R.P. Knill-Jones (1984). *Statistical and knowledge-based approaches to clinical decision-support systems with an application in gastroenterology*, Journal of the Royal Statistical Society (Series A), vol. 147, pp. 35-77.

Symptoms and diseases

Context: patients with an Ulcer. Question: which type?

		duodenal ulcer ($n = 248$)	gastric ulcer ($n = 43$)
Sex:	male	169	17
	female	79	26
Age:	< 26	43	1
	26 - 40	82	5
	41 - 55	87	19
	> 55	36	18
Daily pain:	yes	21	11
	no	214	27
Effect food on pain:	worsens	44	11
	no effect	82	9
	relieves	104	17
probability		0.85	0.15

The idea

Let \Pr be a joint distribution on the diagnosis search space including hypothesis h and observed findings e .

The **prior odds** for h , and **posterior odds** for h given e , are defined by

$$O(h) = \frac{\Pr(h)}{1 - \Pr(h)} = \frac{\Pr(h)}{\Pr(\neg h)}, \text{ and } O(h | e) = \frac{\Pr(h | e)}{\Pr(\neg h | e)}$$

Assume that all findings $e_i \in e$ are conditionally independent given h , then

$$O(h | e) = \frac{\Pr(e | h) \cdot \Pr(h)}{\Pr(e | \neg h) \cdot \Pr(\neg h)} = \prod_i \frac{\Pr(e_i | h)}{\Pr(e_i | \neg h)} \cdot O(h)$$

Now consider the following transformation: $10 \cdot \ln O(h | e) \dots$

The idea (cntd)

Applying the transformation $10 \cdot \ln$ to

$$O(h | e) = \prod_i \lambda_i \cdot O(h), \quad \text{where } \lambda_i = \frac{\Pr(e_i | h)}{\Pr(e_i | \neg h)}$$

results in a score s :

$$s = 10 \cdot \ln O(h | e) = 10 \cdot \ln O(h) + \sum_i 10 \cdot \ln \lambda_i = w_0 + \sum_i w_i$$

where w_i is a **weight** for finding e_i .

The probability $\Pr(h | e)$ is now computed from

$$\Pr(h | e) = \frac{O(h | e)}{1 + O(h | e)} = \frac{e^{\frac{s}{10}}}{1 + e^{\frac{s}{10}}} = \frac{1}{1 + e^{-\frac{s}{10}}}$$

A scoring system

	h : duodenal ulcer (du) ($n = 248$)	$\neg h$: gastric ulcer (gu) ($n = 43$)
male (m)	169	17
female (f)	79	26

Calculation of probabilities, likelihood ratios and weights:

$$\Pr(m \mid \text{du}) = \frac{169}{248} \sim 0.68, \Pr(m \mid \text{gu}) \sim 0.40 \Rightarrow$$

$$\lambda_m = \frac{\Pr(m \mid \text{du})}{\Pr(m \mid \text{gu})} = \frac{0.68}{0.40} \sim 1.7 \implies w_m = 10 \cdot \ln \lambda_m \sim 5$$

$$\Pr(f \mid \text{du}) = \frac{79}{248} \sim 0.32, \Pr(f \mid \text{gu}) \sim 0.60 \Rightarrow$$

$$\lambda_f = \frac{\Pr(f \mid \text{du})}{\Pr(f \mid \text{gu})} = \frac{0.32}{0.60} \sim 0.53 \implies w_f = 10 \cdot \ln \lambda_f \sim -6$$

Symptoms and their weights

		duodenal ulcer (<i>n</i> = 248)	gastric ulcer (<i>n</i> = 43)	weight
Sex:	male	169	17	5
	female	79	26	-6
Age:	< 26	43	1	18
	26 - 40	82	5	10
	41 - 55	87	19	-2
	> 55	36	18	-10
Daily pain:	yes	21	11	-12
	no	214	27	3
Effect food on pain:	worsens	44	11	-4
	no effect	82	9	4
	relieves	104	17	0
prior		0.85	0.15	17

An example diagnosis

A 30 year old woman reports to the clinic. She has pain in the abdominal area, but not on a daily basis; the pain worsens as soon as she eats.

Calculation of the score:

- the initial score: +17
 - the patient is female: - 6
 - her age is 30: +10
 - she is in pain, but not every day: + 3
 - food intake worsens the pain: - 4
-
- +20

Given that the patient has one of the two diseases, duodenal ulcer and gastric ulcer, she has with probability

$$(1 + e^{-\frac{20}{10}})^{-1} \approx 1.14^{-1} \approx 0.88$$

a duodenal ulcer and a gastric ulcer with probability 0.12.

Reviewing ‘Idiot’s Bayes’

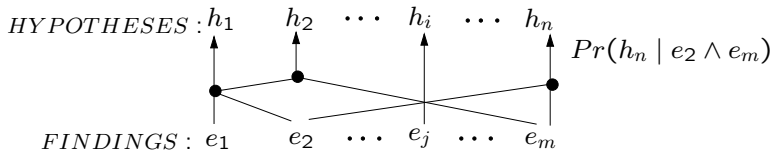
The naive Bayes approach is

- mathematically **correct**, and
- computationally **easy**.

However

- underlying assumptions usually **unacceptable**;
- and, *at the time*, for larger applications
 - number of hypotheses often large → **undoable** to compute each $\Pr(h_i | e)$;
 - often **not enough** information for reliable probability assessments.

History: diagnosis in the 1970s



The most likely hypothesis given observed findings is determined as follows:

- prune the search space using **heuristic rules**;
- **approximate** the missing probabilities required, for example with:

$$\Pr(e_i \wedge e_j) = \min\{\Pr(e_i), \Pr(e_j)\};$$

- select the hypothesis with the highest probability.

Reviewing the quasi-probabilistic models

The quasi-probabilistic models are

- computationally **easy**, and
- easy to **use**,

even for larger applications.

However, these models are

- mathematically **incorrect**, and
- even as an approximation model **not convincing**.

The rehabilitation of probability theory in the 1980s

J. Pearl introduces what we now call probabilistic graphical models (PGMs) :

- a graphical model to represent the knowledge in a complex multi-variate domain
 - graph encodes probabilistic independences
 - joint probability distribution is factorized into smaller functions
- knowledge representation is separated from reasoning
 - ⇒ allows generic algorithms for
 - inference (computing probabilities)
 - learning
 - ...

The Probabilistic Graphical Model framework

Probabilistic Graphical Model: a compact representation of a joint probability distribution $P_{\mathbf{r}}$ over a set of random variables, comprised of:

- **qualitative** knowledge of $P_{\mathbf{r}}$: a graph representation of the independences between the variables involved;
- **quantitative** knowledge of $P_{\mathbf{r}}$: functions that capture part of $P_{\mathbf{r}}$ 'locally' per group of variables.

Algorithms associated with the framework are often tailored to

- the type of graphical model: directed or undirected
- the type of random variables: discrete and/or continuous

Probabilistic Graphical Models

J. Pearl introduced PGMs based on

- **undirected graphs: Markov networks** (Markov Random Fields)
 - Gibbs random field: joint distribution is strictly positive
 - Ising/Potts model (Physics): pairwise MRF with discrete variables
 - Applications in image processing, computer vision, . . .
- **directed (acyclic) graphs: Bayesian networks (BNs)**
 - Naive Bayes: restricted topology, discrete or continuous (Gaussian) variables
 - Hidden Markov model (HMM): 'Dynamic' BN with restricted topology, discrete variables
 - Particle/Kalman Filter: HMM with continuous/Gaussian variables
 - Applications in medicine, biology, genetics, speech recognition, spamfiltering, . . .

Focus on Bayesian networks

PGMs are considered to be interpretable models. When used as a modelling tool, directed models are often preferred.

Judea Pearl introduced several algorithms for inferring 'beliefs' from those represented in a Bayesian network:

- first for trees and polytrees (singly connected graphs)
- then for multiply-connected graphs
- for the latter, the algorithm by Steffen Lauritzen & David Spiegelhalter was the first to find wide-spread use.

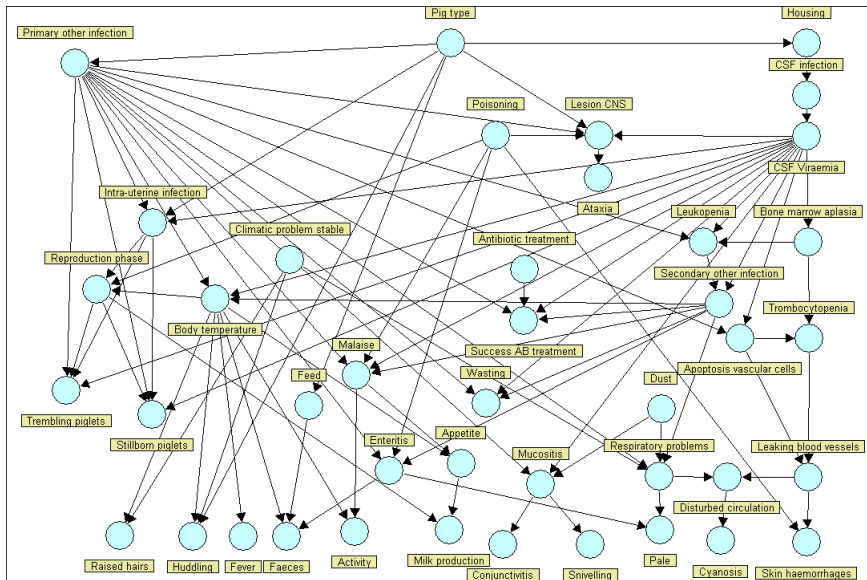
Also see “Inference in Bayesian Networks: a Historical Perspective”, by Adnan Darwiche

An example: Classical Swine Fever (CSF)

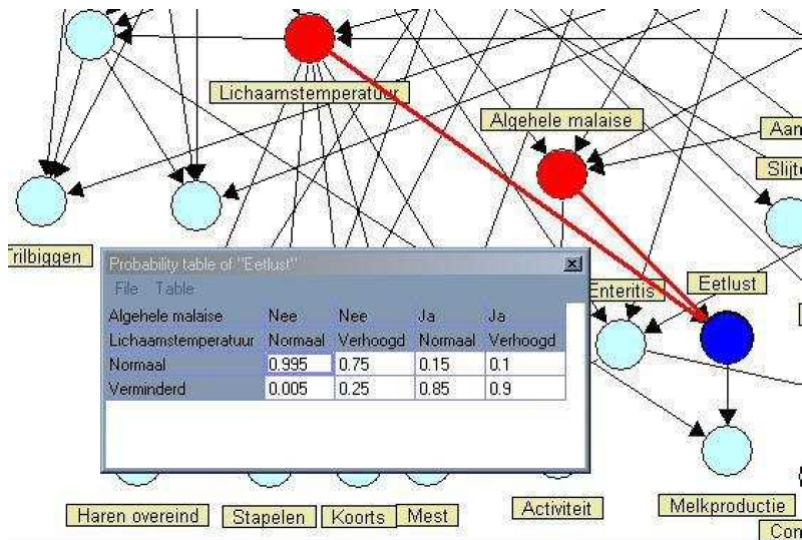
The classical swine fever network is a decision-support system for the early detection of classical swine fever (varkenspest).

- early detection of CSF is important, but hard;
- the network has been developed in cooperation with 2 veterinarians of the Central Veterinary Institute of Wageningen UR;
- part of european EPIZONE project;
- veterinarians all over the country collected data with PDAs

The Classical swine fever network: initial graphical structure



The Classical swine fever network: probability tables



$$\Pr(\text{Appetite} \mid \text{BodyTemp} \wedge \text{Malaise})$$

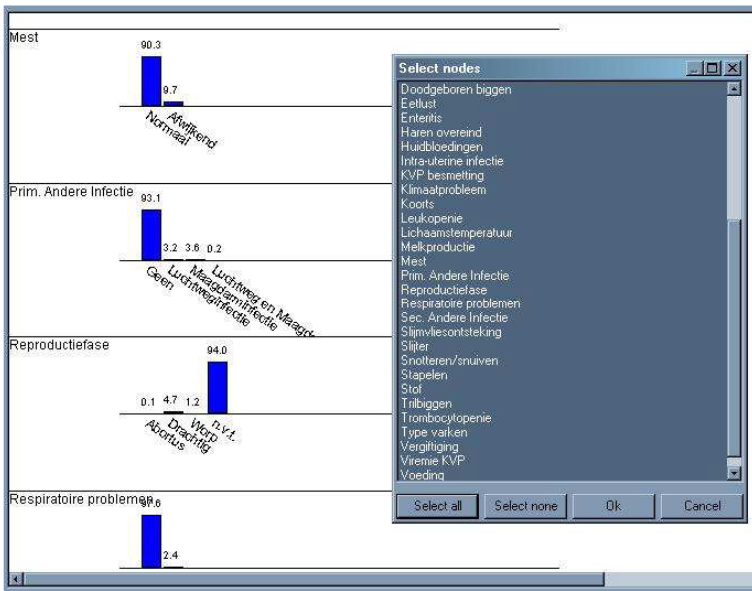
Classical swine fever: prior probabilities

Faeces

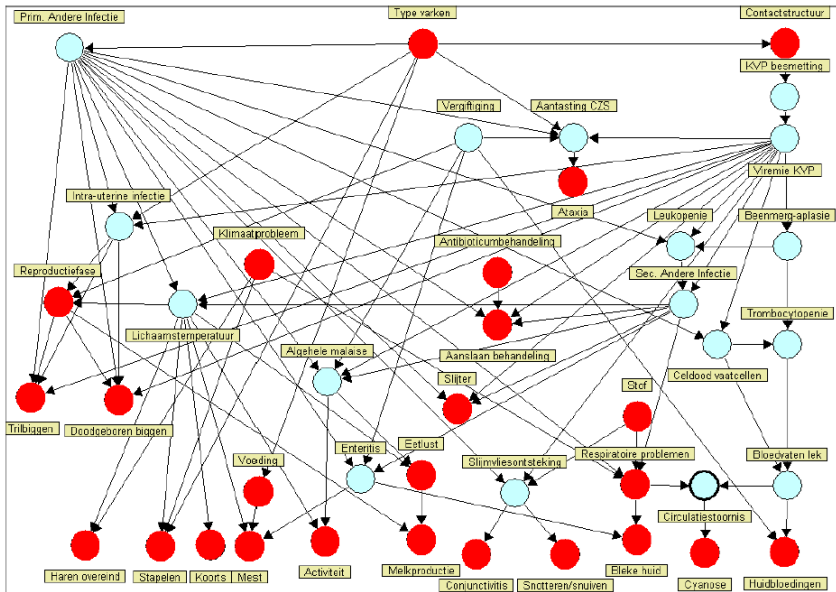
Prim. Other
Infection

Reproduction
phase

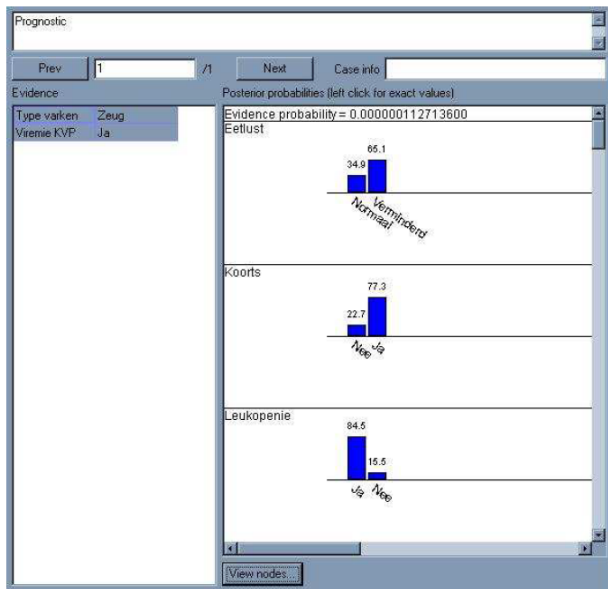
Respiratory
problems



Classical swine fever: diagnostic reasoning



Classical swine fever: prognostic reasoning



A Bayesian network: necessary ingredients

Definition:

A Bayesian network is a pair $\mathcal{B} = (G, \Gamma)$ such that

- G is an *acyclic directed graph* with nodes representing a set of *random variables* \mathbf{V} ;
- $\Gamma = \{\gamma_{V_i} \mid V_i \in \mathbf{V}\}$ is a set of *assessment functions*.

Property:

$$\Pr(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} \gamma_{V_i}(V_i \mid \rho(V_i))$$

defines a *joint probability distribution* \Pr on \mathbf{V} such that G is a *directed I-map* for the *independence relation* I_{\Pr} of \Pr .

About this course ...

The following subjects will be addressed in this course:

- the **syntactics** and **semantics** of PGMs;
- for BNs and probabilistic models in general (latter through Probabilistic Programming):
 - algorithms for **probabilistic inference** (exact and approximate);
 - methods for **constructing** a probabilistic model for a domain of application;
- methods for **evaluating** a discrete Bayesian network's performance and behaviour;
- (methods for **controlling** and **explaining** reasoning).

Syllabus, Chapter 2:

Preliminaries

(Discrete) Random variables

Let $\mathcal{V} = \{V_1, \dots, V_n\}$, $n \geq 1$, be a set of random variables. Each variable $V_i \in \mathcal{V}$ can take on one of $m \geq 2$ values.

For ease of exposition we mostly consider ‘binary’ variables:

- $V_i = \text{true}$, denoted by v_i ;
- $V_i = \text{false}$, denoted by $\neg v_i$ (or by $\overline{v_i}$).

The set \mathcal{V} spans a Boolean Algebra of logical propositions \mathcal{V} :

- T(true), F(false) $\in \mathcal{V}$;
- for all variables $V_i \in \mathcal{V}$ we have that $v_i \in \mathcal{V}$;
- for all $x \in \mathcal{V}$ we have that $\neg x \in \mathcal{V}$;
- for all $x, y \in \mathcal{V}$ we have that $x \wedge y \in \mathcal{V}$ and $x \vee y \in \mathcal{V}$.

The elements of \mathcal{V} obey the usual rules of propositional logic.

The joint probability distribution

Definition:

Let \mathcal{V} be the Boolean Algebra of propositions spanned by a set of random variables V . Let $\text{Pr} : \mathcal{V} \rightarrow [0, 1]$ be a function such that

- Pr is **positive**: for each $x \in \mathcal{V}$ we have that $\text{Pr}(x) \geq 0$ and, more specifically, $\text{Pr}(F) = 0$;
- Pr is **normed**: $\text{Pr}(T) = 1$;
- Pr is **additive**: we have, for each $x, y \in \mathcal{V}$ with $x \wedge y \equiv F$, that $\text{Pr}(x \vee y) = \text{Pr}(x) + \text{Pr}(y)$.

The function Pr is a **joint probability distribution** on V ; the function value $\text{Pr}(x)$ is the **probability** of x .

Independence of propositions

Definition: Let \mathcal{V} be the Boolean Algebra of propositions spanned by a set of random variables V . Let Pr be a joint probability distribution on V .

Propositions $x, y \in \mathcal{V}$ are called **independent** in Pr if

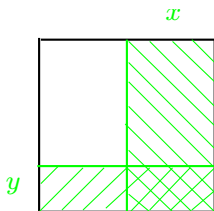
$$\text{Pr}(x \wedge y) = \text{Pr}(x) \cdot \text{Pr}(y)$$

Propositions $x, y \in \mathcal{V}$ are called **conditionally independent** given the proposition $z \in \mathcal{V}$ if we have that

$$\text{Pr}(x \wedge y \mid z) = \text{Pr}(x \mid z) \cdot \text{Pr}(y \mid z)$$

The two notions of independence (1)

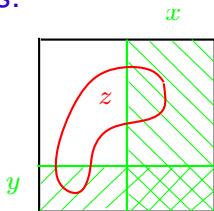
- Consider two propositions $x, y \in \mathcal{V}$ such that x and y are independent¹:



$$\begin{aligned}\Pr(x) &= \frac{1}{2}; \Pr(y) = \frac{1}{4}; \\ \Pr(x \wedge y) &= \frac{1}{8} \\ &= \Pr(x) \cdot \Pr(y)\end{aligned}$$

Can $z \in \mathcal{V}$ exist such that x and y are dependent given z ?

- Yes:

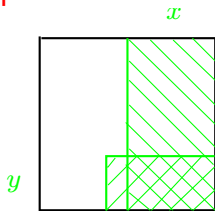


$$\begin{aligned}\Pr(x | z) &> 0; \Pr(y | z) > 0; \\ \Pr(x \wedge y | z) &= 0 \\ &\neq \Pr(x | z) \cdot \Pr(y | z)\end{aligned}$$

¹The square has area 1, representing the total probability mass.

The two notions of independence (2)

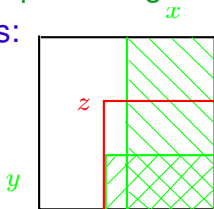
- Consider two propositions $x, y \in \mathcal{V}$ such that x and y are **dependent**:



$$\begin{aligned}\Pr(x) &= \frac{1}{2}; \Pr(y) = \frac{1}{5}; \\ \Pr(x \wedge y) &= \frac{1}{7} \\ &> \Pr(x) \cdot \Pr(y)\end{aligned}$$

Can $z \in \mathcal{V}$ exist such that x and y are **conditionally independent given z** ?

- Yes:



$$\begin{aligned}\Pr(x | z) &= \frac{4}{5}; \Pr(y | z) = \frac{1}{2}; \\ \Pr(x \wedge y | z) &= \frac{4}{10} \\ &= \Pr(x | z) \cdot \Pr(y | z)\end{aligned}$$

Configurations

Let \mathcal{V} be spanned by random variables V and let $W \subseteq V$.

- proposition $w \in \mathcal{V}$ is called a **configuration** of W iff it is a **conjunction** of value assignments to the variables from W ;
- c_W is used to denote an **arbitrary configuration** of W ;
- W also indicates **all possible configurations** to the set W (notation abuse!): W is then considered to be a **template** for **all possible configurations** c_W ;
- if $W = \emptyset$, then by convention $c_W = c_\emptyset = \top$.

Example: Let $W = \{V_1, V_3, V_7\}$, with $W = V_1 \wedge V_3 \wedge V_7$ the associated configuration template. Some configurations c_W captured by this template are:

$$\begin{array}{l} V_1 = true \quad \wedge \quad V_3 = true \quad \wedge \quad V_7 = false \\ v_1 \quad \wedge \quad \neg v_3 \quad \wedge \quad v_7 \\ \neg v_1 \quad \wedge \quad v_3 \quad \wedge \quad \neg v_7 \end{array}$$



Conventions and notation

	set (bold faced)	singleton
variables/templates (capital)	\mathbf{V}	V
values/configurations	$c_{\mathbf{V}}, \mathbf{v}$	c_V, v

- conjunctions are often left implicit: e.g. $v_1 v_2$ denotes $v_1 \wedge v_2$;
- note the following differences (!)

probabilities: $\Pr(c_{\mathbf{V}}), \Pr(c_V), \Pr(\mathbf{v}), \Pr(v), \Pr(v \mid c_{\mathbf{E}})$

distributions: $\Pr(\mathbf{V}), \Pr(V), \Pr(\mathbf{V} \mid \mathbf{e})$

distribution sets: $\Pr(\mathbf{V} \mid \mathbf{E}), \Pr(V \mid \mathbf{E})$

Independence of variables

Definition: Let V be a set of random variables and let $X, Y, Z \subseteq V$. Let \Pr be a joint distribution on V .

X is called **conditionally independent** of Y given Z in \Pr , if

$$\Pr(X | Y \wedge Z) = \Pr(X | Z)$$

If this holds for $Z = \emptyset$ then X is (marginally) **independent** of Y .

Remarks:

- Note the template notation: equation should hold for all c_X , c_Y and c_Z !
- $\Pr(X | Y \wedge Z) = \Pr(X | Z) \Rightarrow$
 $\Pr(X \wedge Y | Z) = \Pr(X | Z) \cdot \Pr(Y | Z)$ ■

(See syllabus exercise 2.6)