Probabilistic Reasoning

course slides to accompany syllabus

2024-2025

UU – ICS Master Programmes:
Computing Science
Artificial Intelligence

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
(covered by syllabus, hardly in lectures!)

Literature: syllabus, papers, slides & studymanual

Additional info: see Blackboard and course website:
`http://ics.uu.nl/docs/vakken/prob/`

Practical information

Course form:

- lectures, Q&A for last assignment and exam
- exercises (formative self assessment)
  (tip: discuss exercises together, e.g. in discussion forum)

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 $\boldsymbol{V} = \{V_1, \ldots, V_n\}$. 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 $\Pr(\boldsymbol{V})$, 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 $\boldsymbol{H} = \{h_1, \ldots, h_n\}, n \geq 1$, be a set of hypotheses, and let $\boldsymbol{E} = \{e_1, \ldots, e_m\}, m \geq 1$, be a set of relevant findings (evidence).

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

The approach: Compute for each $\boldsymbol{h} \subseteq \boldsymbol{H}$ the probability

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

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

Pioneering in the 1960s

Determine the diagnosis given findings $\boldsymbol{e} \subseteq \boldsymbol{E}$.

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

Then, compute for each $h_i \in \boldsymbol{H}$:

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

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 $\boldsymbol{e} = \{e_p, \ldots, e_q\}$, $1 \leq p, q \leq m$.

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

$$\Pr(h_i \mid \boldsymbol{e}) = \frac{\Pr(e_p, \ldots, e_q \mid h_i) \Pr(h_i)}{\sum_{k=1}^{n} \Pr(e_p, \ldots, e_q \mid h_k) \Pr(h_k)}$$

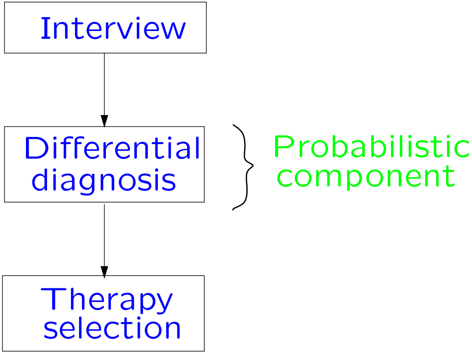
$$= \frac{\Pr(e_p \mid h_i) \cdot \ldots \cdot \Pr(e_q \mid h_i) \Pr(h_i)}{\sum_{k=1}^{n} \Pr(e_p \mid h_k) \cdot \ldots \cdot \Pr(e_q \mid h_k) \Pr(h_k)}$$

<u>**Benefit**</u>: 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:



developed with data collected from $\pm$ 1200 patients.

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 $\boldsymbol{e}$, are defined by

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

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

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

Now consider the following transformation: $10 \cdot \ln O(h \mid \boldsymbol{e}) \ldots$

The idea (cntd)

Applying the transformation $10 \cdot \ln$ to

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

results in a score s:

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

where w_i is a weight for finding e_i.

The probability $\Pr(h \mid \boldsymbol{e})$ is now computed from

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

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:

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

$$\lambda_{\mathsf{m}} = \frac{\Pr(\mathsf{m} \mid \mathsf{du})}{\Pr(\mathsf{m} \mid \mathsf{gu})} = \frac{0.68}{0.40} \sim 1.7 \implies \mathsf{w}_{\mathsf{m}} = 10 \cdot \ln \lambda_{\mathsf{m}} \sim 5$$

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

$$\lambda_{\mathsf{f}} = \frac{\Pr(\mathsf{f} \mid \mathsf{du})}{\Pr(\mathsf{f} \mid \mathsf{gu})} = \frac{0.32}{0.60} \sim 0.53 \implies \mathsf{w}_{\mathsf{f}} = 10 \cdot \ln \lambda_{\mathsf{f}} \sim -6$$

Symptoms and their weights

		duodenal ulcer ($n = 248$)	gastric ulcer ($n = 43$)	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’

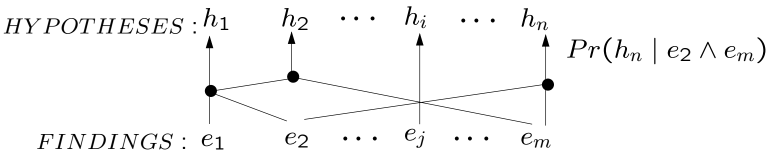
The naive Bayes approach is

- mathematically correct, and
- computationally easy.

However

- underlying assumptions usually unacceptable;
- and, *at the time*, for larger applications
  - number of hypotheses often large $\rightarrow$ undoable to compute each $\Pr(h_i \mid e)$;
  - 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 $\Rightarrow$ allows generic algorithms for
  - inference (computing probabilities)
  - learning
  - …

The Probabilistic Graphical Model framework

Probabilistic Graphical Model: a compact representation of a joint probability distribution $\Pr$ over a set of random variables, comprised of:

- qualitative knowledge of $\Pr$: a graph representation of the independences between the variables involved;
- quantitative knowledge of $\Pr$: functions that capture part of $\Pr$ '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 explainable 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.

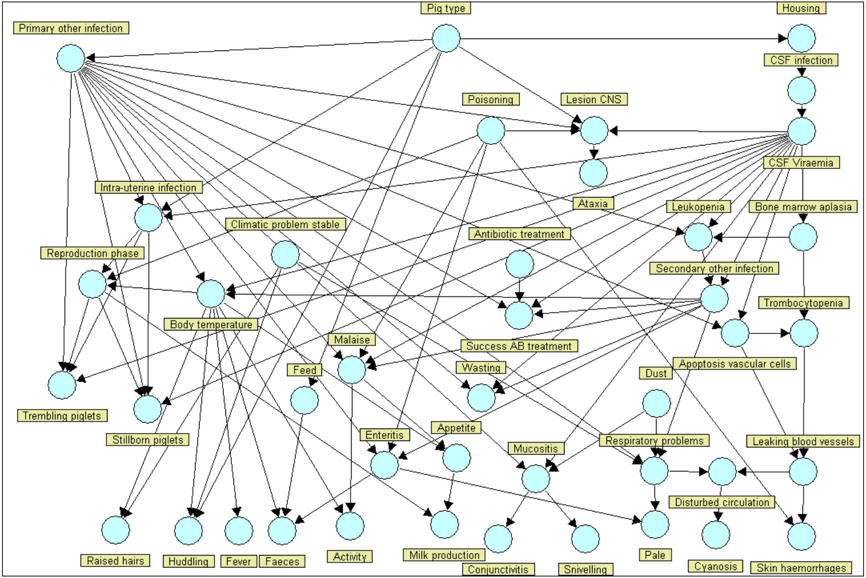
Also see "Inference in Bayesian Networks: a Historical Perspective", by Adnan Darwiche

An example: Classical Swine Fever (CSF)

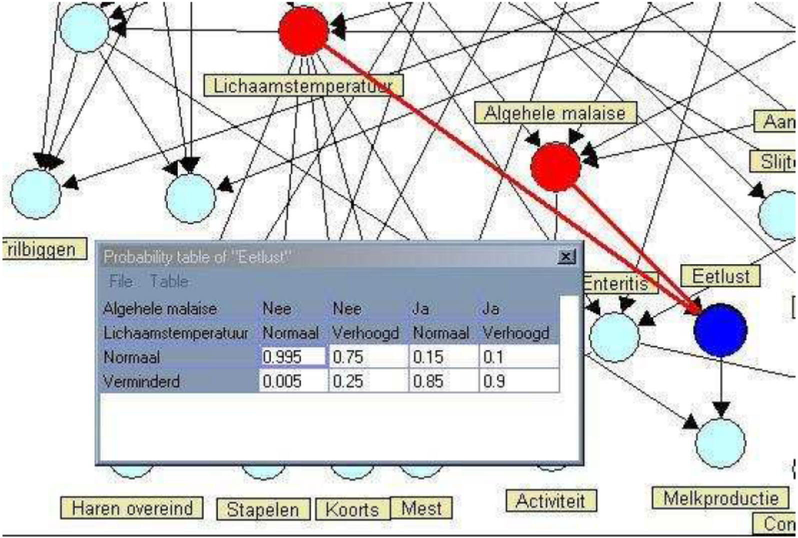
The classical swine fever network is a decision-support system for the early detection of classical swine fever (varkenspest).

- 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



$$\Pr(Appetite \mid BodyTemp \wedge Malaise)$$

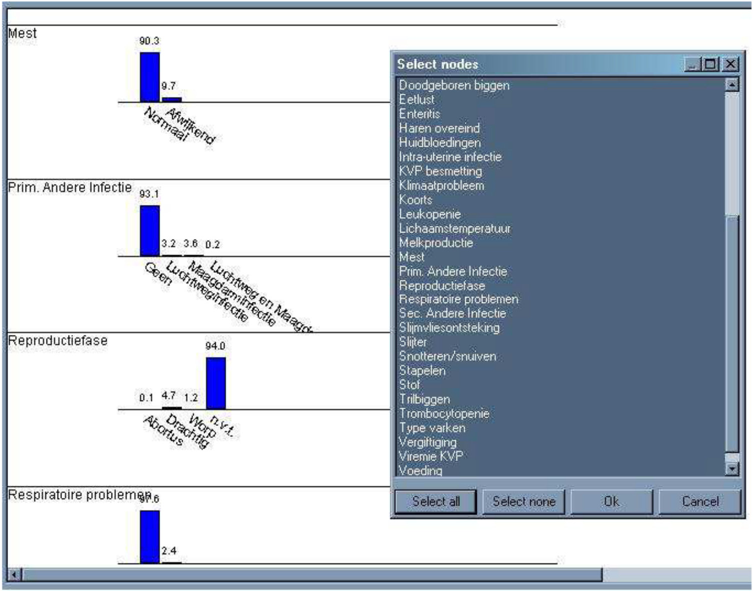
Classical swine fever: prior probabilities

Faeces

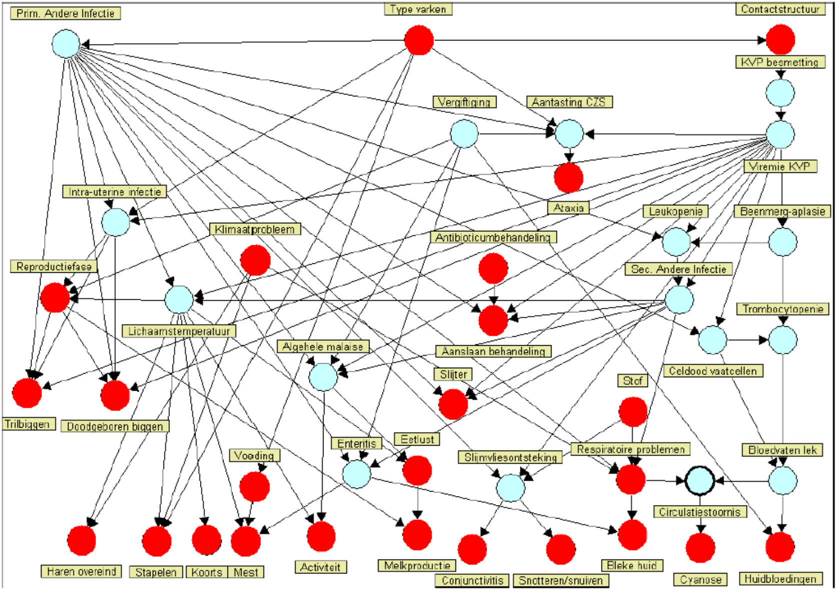
Prim. Other Infection

Reproduction phase

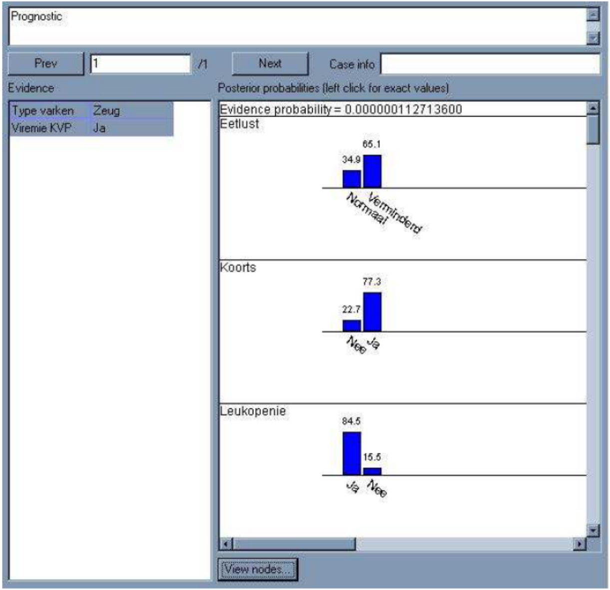
Respiratory problems



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* $\boldsymbol{V}$;
- $\Gamma = \{\gamma_{V_i} \mid V_i \in \boldsymbol{V}\}$ is a set of *assessment functions*.

Property:

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

defines a *joint probability distribution* $\Pr$ on $\boldsymbol{V}$ 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 $\boldsymbol{V} = \{V_1, \ldots, V_n\}$, $n \geq 1$, be a set of random variables.
Each variable $V_i \in \boldsymbol{V}$ can take on one of $m \geq 2$ values.

For ease of exposition we mostly consider 'binary' variables:

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

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

- T(rue), F(alse) $\in \mathcal{V}$;
- for all variables $V_i \in \boldsymbol{V}$ 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 $\boldsymbol{V}$. Let $\Pr : \mathcal{V} \rightarrow [0, 1]$ be a function such that

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

The function $\Pr$ is a joint probability distribution on $\boldsymbol{V}$; the function value $\Pr(x)$ 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 $\boldsymbol{V}$. Let $\Pr$ be a joint probability distribution on V.

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

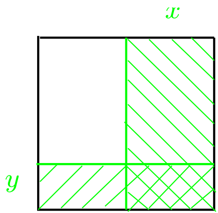
$$\Pr(x \wedge y) = \Pr(x) \cdot \Pr(y)$$

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

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

The two notions of independence (1)

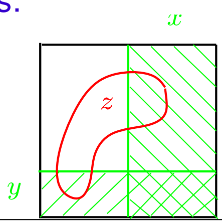
- Consider two propositions $x, y \in \mathcal{V}$ such that x and y are independent [1]:



$$\Pr(x) = \tfrac{1}{2}; \Pr(y) = \tfrac{1}{4};$$
$$\Pr(x \wedge y) = \tfrac{1}{8} = \Pr(x) \cdot \Pr(y)$$

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

- Yes:

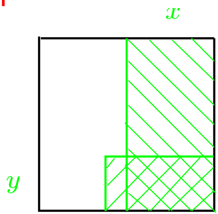


$$\Pr(x \mid z) > 0; \Pr(y \mid z) > 0;$$
$$\Pr(x \wedge y \mid z) = 0 \neq \Pr(x \mid z) \cdot \Pr(y \mid z)$$

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

The two notions of independence (2)

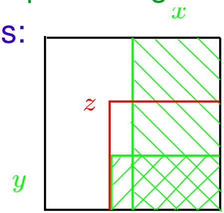
- Consider two propositions $x, y \in \mathcal{V}$ such that x and y are dependent:



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

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

- Yes:



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

Configurations

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

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

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

$$\begin{array}{rcrcr} V_1 = \mathit{true} & \wedge & V_3 = \mathit{true} & \wedge & V_7 = \mathit{false} \\ v_1 & \wedge & \neg v_3 & \wedge & v_7 \\ \neg v_1 & \wedge & v_3 & \wedge & \neg v_7 \end{array}$$

■

Conventions and notation

	set (bold faced)	singleton
variables/templates (capital)	$\boldsymbol{V}$	V
values/configurations	$c_{\boldsymbol{V}}$, $\boldsymbol{v}$	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: $\Pr(c_{\boldsymbol{V}})$, $\Pr(c_V)$, $\Pr(\boldsymbol{v})$, $\Pr(v)$, $\Pr(v \mid c_{\boldsymbol{E}})$
distributions: $\Pr(\boldsymbol{V})$, $\Pr(V)$, $\Pr(\boldsymbol{V} \mid \boldsymbol{e})$
distribution *sets*: $\Pr(\boldsymbol{V} \mid \boldsymbol{E})$, $\Pr(V \mid \boldsymbol{E})$

Independence of variables

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

$\boldsymbol{X}$ is called conditionally independent of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ in $\Pr$, if

$$\Pr(\boldsymbol{X} \mid \boldsymbol{Y} \wedge \boldsymbol{Z}) = \Pr(\boldsymbol{X} \mid \boldsymbol{Z})$$

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

Remarks:

- Note the template notation: equation should hold for all $c_{\boldsymbol{X}}$, $c_{\boldsymbol{Y}}$ and $c_{\boldsymbol{Z}}$!
- $\Pr(\boldsymbol{X} \mid \boldsymbol{Y} \wedge \boldsymbol{Z}) = \Pr(\boldsymbol{X} \mid \boldsymbol{Z}) \Rightarrow$
  $\Pr(\boldsymbol{X} \wedge \boldsymbol{Y} \mid \boldsymbol{Z}) = \Pr(\boldsymbol{X} \mid \boldsymbol{Z}) \cdot \Pr(\boldsymbol{Y} \mid \boldsymbol{Z})$ ■

(See syllabus exercise 2.6)