Bayesian Networks

1. Basic Probability Calculus

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Business Economics and Information Systems
& Institute for Computer Science
University of Hildesheim
http://www.ismll.uni-hildesheim.de

Joint probability distributions

Pain	Y				N			
Weightloss	Y		N		Y		N	
Vomiting	Y	N	Y	N	Y	N	Y	N
Adeno Y	0.220	0.220	0.025	0.025	0.095	0.095	0.010	0.010
N	0.004	0.009	0.005	0.012	0.031	0.076	0.050	0.113

Figure 1: Joint probability distribution $p(P, W, V, A)$ of four random variables P (pain), W (weightloss), V (vomiting) and A (adeno).

Joint probability distributions

Discrete JPDs are described by

- nested tables,
- multi-dimensional arrays,
- data cubes, or
- tensors

having entries in $[0, 1]$ and summing to 1.

Probability spaces

Definition 1. Let Ω be a finite set. We call Ω the **sample space** and every subset $E \subseteq \Omega$ an **event**; subsets containing exactly one element, i.e.

$$E = \{e\}, \quad e \in \Omega$$

are called **elementary events**.

A function

$$p : \mathcal{P}(\Omega) \rightarrow [0, 1]$$

with

1. p is additive, i.e. for disjunct $E, F \subseteq \Omega$:
$$p(E \cup F) = p(E) + p(F)$$
2. $p(\Omega) = 1$

is called **probability function** (axioms of probability, Kolmogorov, 1933). A pair (Ω, p) is called **probability space**.

Probability spaces

Lemma 1.

$$p(E) = \sum_{e \in E} p(\{e\}), \quad E \subseteq \Omega$$

Example 1. Throwing a dice can be described by

$$\Omega := \{1, 2, 3, 4, 5, 6\}$$

For a fair dice we have

$$p(\{1\}) = p(\{2\} = \ldots = p(\{6\}) = \frac{1}{6}$$

Then $E = \{2\}$ is the event of dicing a 2, $F = \{2, 4, 6\}$ the event of dicing an even number.

$$p(\{2, 4, 6\}) = p(\{2\}) + p(\{4\}) + p(\{6\}) = \frac{1}{2}$$

Independent events

Definition 2. Let $E, F \subseteq \Omega$ with $p(F) > 0$. Then

$$p(E|F) := p^{|F} := \frac{p(E \cap F)}{p(F)}$$

is called **conditional probability** of E given F.

Two events $E, F \subseteq \Omega$ are called **independent**, if

$$p(E \cap F) = p(E) \cdot p(F)$$

i.e., if $p(E|F) = p(E)$ or $p(E) = 0$ or $p(F) = 0$.

Independent Events / Example

Example 2. Let $F := \{2, 4, 6\}$ be the event of dicing an even number. Then the conditional probability

$$p(\{2\}|F) = \frac{1}{6} / \frac{1}{2} = \frac{1}{3}$$

describes the probability of dicing a 2 given we diced an even number.

Example 3. The events $E := \{2, 4, 6\}$ of dicing an even number and $F := \{1, 2, 3, 4\}$ of dicing a number less than 5 are independent as

$$\begin{aligned} p(E \cap F) &= p(\{2, 4\}) = \frac{1}{3} \\ &\stackrel{!}{=} p(E) \cdot p(F) = \frac{1}{2} \cdot \frac{2}{3} \end{aligned}$$

Conditional independent events

Definition 3. Let $G \subseteq \Omega$ be an event with $p(G) > 0$. Two events $E, F \subseteq \Omega$ are called **conditionally independent** given G, if

$$p(E \cap F \cap G) = p(E \cap G) \cdot p(F \cap G) / p(G)$$

i.e., if $p(E|F \cap G) = p(E|G)$ or $p(E|G) = 0$ or $p(F|G) = 0$.

Definition 4. A partition $(E_i)_{i=1,\ldots,m}$ of Ω is also called **a set of mutually exclusive and exhaustive events**, i.e.

1. $E_i \neq \emptyset$,
2. $\bigcup_{i=1}^{m} E_i = \Omega$, and
3. E_i are pairwise disjunct (i.e., $E_i \cap E_j = \emptyset$ for $i \neq j$).

Conditional independent events / Example

Example 4. The events

- $E := \{2, 4, 6\}$ of dicing an even number and
- $F := \{1, 2, 3, 4, 5\}$ of dicing anything but 6

are dependent as

$$p(E \cap F) = p(\{2, 4\}) = \frac{1}{3} \overset{!}{\neq} p(E) \cdot p(F) = \frac{1}{2} \cdot \frac{5}{6}$$

But given the event

- $G := \{1, 2, 3, 4\}$ of dicing a number less than 5,

E and F are conditionally independent given G as

$$p(E \cap F \cap G) = p(\{2, 4\}) = \frac{1}{3}$$
$$\overset{!}{=} p(E \cap G) \cdot p(F \cap G)/p(G) = \frac{1}{3} \cdot \frac{2}{3} / \frac{2}{3}$$

Random variables and probability distributions

Definition 5. Any function

$$X : \Omega \to X$$

is called a **random variable** (by abuse of notation we label both, the map and the target space with X).

We assign each value $x \in X$ a probability via

$$p(X = x) := p(X^{-1}(x))$$

p is called the **probability distribution of** X.

If X is numeric, e.g., $X = \mathbb{R}$, we call

$$E(X) := \sum_{x \in X} x \cdot p(x)$$

the **expected value** of X.

Random variables and probability distributions

Example 5. Let Ω contain the outcomes of a throw of two (distinguishable) dice, i.e.

$$\Omega := \{(1,1),(1,2),\ldots,(1,6),\\ (2,1),(2,2),\ldots,(6,5),(6,6)\}$$

Then the sum of the two dice,

$$\begin{aligned} X: \quad \Omega \quad &\to \mathbb{N} \\ (i,j) \quad &\mapsto i+j \end{aligned}$$

is a random variable.

The value $X = 3$ then represents the event $X^{-1}(3) = \{(1,2),(2,1)\}$ and thus $p(X = 3) = \frac{2}{36}$.

The expected value of X is $E(X) = 7$.

X	2	3	4	5	6	7	8	9	10	11	12
$p(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Joint probability distributions

Definition 6. Let X and Y be two random variables. Then their cartesian product

$$\begin{aligned} X \times Y : \Omega &\rightarrow X \times Y \\ e &\mapsto (X(e), Y(e)) \end{aligned}$$

is again a random variable; its distribution is called **joint probability distribution** of X and Y.

Bayesian Networks / 3. Random Variables

Joint probability distributions

Example 6. Let Ω be the outcomes of a throw of two dices and X the sum of their numbers as before. Let Y be

$$Y(i,j) := \begin{cases} \text{odd}, & \text{if } i \text{ and } j \text{ is odd} \\ \text{even}, & \text{if } i \text{ or } j \text{ is even} \end{cases}$$

Then the probability of

$$p(X = 4, Y = \text{odd}) = p(\{(1,3),(3,1)\}) = \frac{2}{36}$$

In general,

$$p(X = x, Y = y) \neq p(X = x) \cdot p(Y = y)$$

as can be seen here:

$$p(X = 4) = p(\{(1,3),(3,1),(2,2)\}) = \frac{3}{36}$$

$$p(Y = \text{odd}) = \frac{9}{36}$$

Marginal probability distributions

Definition 7. Let p be a the joint probability of the random variables $\mathcal{X} := \{X_1, \ldots, X_n\}$ and $\mathcal{Y} \subseteq \mathcal{X}$ a subset thereof. Then

$$p(\mathcal{Y} = y) := p^{\downarrow \mathcal{Y}}(y) := \sum_{x \in \operatorname{dom} \mathcal{X} \setminus \mathcal{Y}} p(\mathcal{X} \setminus \mathcal{Y} = x, \mathcal{Y} = y)$$

is a probability distribution of $\mathcal{Y}$ called **marginal probability distribution**.

Example 7. Marginal $p(V, A)$:

Vomiting	Y	N
Adeno Y	0.350	0.350
N	0.090	0.210

Pain	Y				N			
Weightloss	Y		N		Y		N	
Vomiting	Y	N	Y	N	Y	N	Y	N
Adeno Y	0.220	0.220	0.025	0.025	0.095	0.095	0.010	0.010
N	0.004	0.009	0.005	0.012	0.031	0.076	0.050	0.113

Figure 2: Joint probability distribution $p(P, W, V, A)$ of four random variables P (pain), W (weightloss), V (vomiting) and A (adeno).

Marginal probability distributions / example

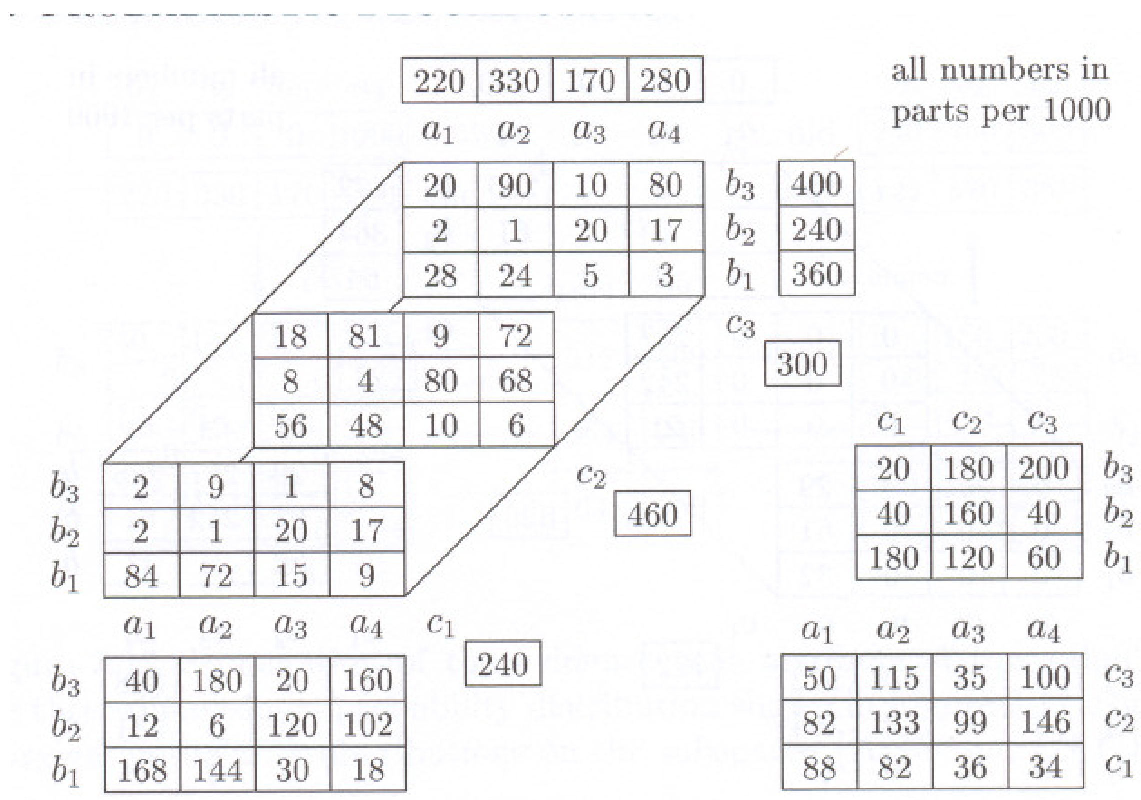


Figure 3: Joint probability distribution and all of its marginals [BK02, p. 75].

Extreme and non-extreme probability distributions

Definition 8. By $p > 0$ we mean

$$p(x) > 0, \quad \text{for all } x \in \prod \operatorname{dom}(p)$$

Then p is called **non-extreme**.

Example 8.

$$\begin{pmatrix} 0.4 & 0.0 \\ 0.3 & 0.3 \end{pmatrix} \qquad \Bigg| \qquad \begin{pmatrix} 0.4 & 0.1 \\ 0.2 & 0.3 \end{pmatrix}$$

Conditional probability distributions

Definition 9. For a JPD p and a subset $\mathcal{Y} \subseteq \operatorname{dom}(p)$ of its variables with $p^{\downarrow \mathcal{Y}} > 0$ we define

$$p^{|\mathcal{Y}} := \frac{p}{p^{\downarrow \mathcal{Y}}}$$

as **conditional probability distribution of p w.r.t. $\mathcal{Y}$**.

A conditional probability distribution w.r.t. $\mathcal{Y}$ sums to 1 for all fixed values of $\mathcal{Y}$, i.e.,

$$(p^{|\mathcal{Y}})^{\downarrow \mathcal{Y}} \equiv 1$$

Conditional probability distributions / example

Example 9. Let p be the JPD

$$p := \begin{pmatrix} 0.4 & 0.1 \\ 0.2 & 0.3 \end{pmatrix}$$

on two variables R (rows) and C (columns) with the domains $\mathrm{dom}(R) = \mathrm{dom}(C) = \{1, 2\}$.

The conditional probability distribution w.r.t. C is

$$p^{|C} := \begin{pmatrix} 2/3 & 1/4 \\ 1/3 & 3/4 \end{pmatrix}$$

Chain rule

Lemma 2 (Chain rule)**.** *Let p be a JPD on variables $X_1, X_2, \ldots, X_n$ with $p(X_1, \ldots, X_{n-1}) > 0$. Then*

$$p(X_1, X_2, \ldots, X_n) = p(X_n | X_1, \ldots, X_{n-1}) \cdots p(X_2 | X_1) \cdot p(X_1)$$

The chain rule provides a **factorization** of the JPD in some of its conditional marginals.

The factorizations stemming from the chain rule are trivial as they have as many parameters as the original JPD:

$$\#\text{parameters} = 2^{n-1} + 2^{n-2} + \cdots + 2^1 + 2^0 = 2^n - 1$$

(example computation for all binary variables)

Bayes formula

Lemma 3 (Bayes Formula)**.** *Let p be a JPD and $\mathcal{X}, \mathcal{Y}$ be two disjoint sets of its variables. Let $p(\mathcal{Y}) > 0$. Then*

$$p(\mathcal{X} \,|\, \mathcal{Y}) = \frac{p(\mathcal{Y} \,|\, \mathcal{X}) \cdot p(\mathcal{X})}{p(\mathcal{Y})}$$

Thomas Bayes (1701/2–1761)

Bayes formula / Example

Example 10. Assign each object in fig. 4 an equal probability $\frac{1}{13}$.
Let X be the label of the outcome (1 or 2) and
Y be the color of the outcome (black or white).

Then

$$p(X = 1|Y = \text{black}) \\ = \frac{p(Y = \text{black}|X = 1)\, p(X = 1)}{p(Y = \text{black}|X = 1)\, p(X = 1) + p(Y = \text{black}|X = 2)\, p(X = 2)} \\ = \frac{\frac{3}{5} \cdot \frac{5}{13}}{\frac{3}{5} \cdot \frac{5}{13} + \frac{6}{8} \cdot \frac{8}{13}} = \frac{1}{3}$$

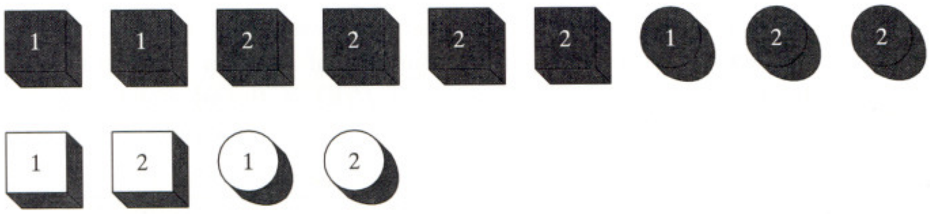


Figure 4: 13 objects with different shape, color, and label [Nea03, p. 8].

Independent variables

Definition 10. Two sets $\mathcal{X}, \mathcal{Y}$ of variables are called **independent**, when

$$p(\mathcal{X} = x, \mathcal{Y} = y) = p(\mathcal{X} = x) \cdot p(\mathcal{Y} = y)$$

for all x and y or equivalently

$$p(\mathcal{X} = x | \mathcal{Y} = y) = p(\mathcal{X} = x)$$

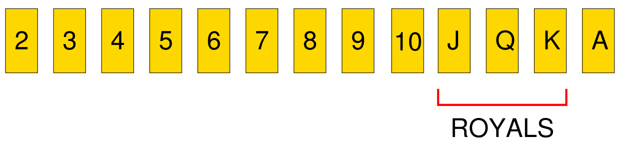
for y with $p(\mathcal{Y} = y) > 0$.

Independent variables / example

Example 11. Let Ω be the cards in an ordinary deck and

- $R =$ true, if a card is royal,
- $T =$ true, if a card is a ten or a jack,
- $S =$ true, if a card is spade.

Cards for a single color:



S	R	T	$p(R, T \| S)$
Y	Y	Y	1/13
		N	2/13
	N	Y	1/13
		N	9/13
N	Y	Y	3/39 = 1/13
		N	6/39 = 2/13
	N	Y	3/39 = 1/13
		N	27/39 = 9/13

R	T	$p(R, T)$
Y	Y	4/52 = 1/13
	N	8/52 = 2/13
N	Y	4/52 = 1/13
	N	36/52 = 9/13

Conditionally independent variables

Definition 11. Let $\mathcal{X}, \mathcal{Y}$, and $\mathcal{Z}$ be sets of variables.

$\mathcal{X}, \mathcal{Y}$ are called **conditionally independent given** $\mathcal{Z}$, when for all events $\mathcal{Z} = z$ with $p(\mathcal{Z} = z) > 0$ all pairs of events $\mathcal{X} = x$ and $\mathcal{Y} = y$ are conditionally independend given $\mathcal{Z} = z$, i.e.

$$p(\mathcal{X} = x, \mathcal{Y} = y, \mathcal{Z} = z) = \frac{p(\mathcal{X} = x, \mathcal{Z} = z) \cdot p(\mathcal{Y} = y, \mathcal{Z} = z)}{p(\mathcal{Z} = z)}$$

for all x, y and z (with $p(\mathcal{Z} = z) > 0$), or equivalently

$$p(\mathcal{X} = x | \mathcal{Y} = y, \mathcal{Z} = z) = p(\mathcal{X} = x | \mathcal{Z} = z)$$

We write $I_p(\mathcal{X}, \mathcal{Y} | \mathcal{Z})$ for the statement, that $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given $\mathcal{Z}$.

Conditionally independent variables / Example

Example 12. Assume S (shape), C (color), and L (label) be three random variables that are distributed as shown in figure 5.

We show $I_p(\{L\}, \{S\} | \{C\})$, i.e., that label and shape are conditionally independent given the color.

C	S	L	$p(L\|C,S)$
black	square	1	2/6 = 1/3
		2	4/6 = 2/3
	round	1	1/3
		2	2/3
white	square	1	1/2
		2	1/2
	round	1	1/2
		2	1/2

C	L	$p(L\|C)$
black	1	3/9 = 1/3
	2	6/9 = 2/3
white	1	2/4 = 1/2
	2	2/4 = 1/2

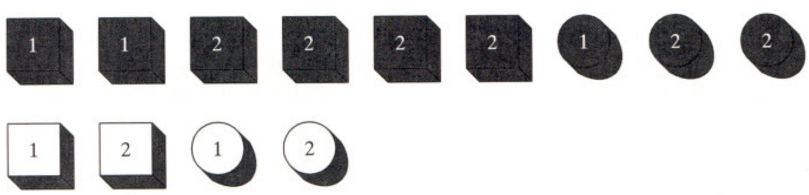


Figure 5: 13 objects with different shape, color, and label [Nea03, p. 8].

References

[BK02] Christian Borgelt and Rudolf Kruse. *Graphical Models*. Wiley, New York, 2002.

[Nea03] Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2003.