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.isml.uni-hildesheim.de

1. Events
2. Independent Events
3. Random Variables
4. Chain Rule and Bayes Formula
5. Independent Random Variables
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.

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Weightloss</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Vomiting</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Adeno</td>
<td>0.220</td>
<td>0.220</td>
<td>0.025</td>
<td>0.025</td>
<td>0.095</td>
<td>0.095</td>
</tr>
<tr>
<td></td>
<td>0.004</td>
<td>0.009</td>
<td>0.005</td>
<td>0.012</td>
<td>0.031</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.010</td>
<td>0.050</td>
<td>0.113</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Joint probability distribution \(p(P, W, V, A)\) of four random variables \(P\) (pain), \(W\) (weight-loss), \(V\) (vomiting) and \(A\) (adeno).
Probability spaces

**Definition 1.** Let $\Omega$ be a finite set. We call $\Omega$ 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

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

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 $(\Omega, p)$ is called **probability space**.

**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.

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

2. Independent Events

3. Random Variables

4. Chain Rule and Bayes Formula

5. Independent Random Variables

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

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

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

$$p(E \cap F) = p(\{2, 4\}) = \frac{1}{3}$$

$$p(E) \cdot p(F) = \frac{1}{2} \cdot \frac{2}{3}$$

---

**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 $(E_i)_{i=1,...,m}$ of $\Omega$ 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 $i \neq j$).
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

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

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

\[
p(E \cap F \cap G) = p(\{2, 4\}) = \frac{1}{3}
\]

\[
\Rightarrow p(E \cap G) \cdot p(F \cap G) / p(G) = \frac{1}{3} \cdot \frac{2}{3} / \frac{2}{3}
\]
**Definition 5.** Any function

\[ X : \Omega \rightarrow X \]

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 \( X \).

---

**Example 5.** Let \( \Omega \) 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,

\[ X : \Omega \rightarrow \mathbb{N} \\
(i,j) \mapsto i + j \]

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 \( E(X) = 7 \).
Joint probability distributions

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

$$X \times Y : \Omega \rightarrow X \times Y$$

$$e \mapsto (X(e), Y(e))$$

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

---

**Example 6.** Let $\Omega$ 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 the joint probability of the random variables $\mathcal{X} := \{X_1, \ldots, X_n\}$ and $\mathcal{Y} \subseteq \mathcal{X}$ a subset thereof. Then

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

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

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

<table>
<thead>
<tr>
<th>Vomiting</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adeno</td>
<td>0.350</td>
<td>0.350</td>
</tr>
<tr>
<td>N</td>
<td>0.090</td>
<td>0.210</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pain</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0.220</td>
<td>0.220</td>
<td>0.025</td>
<td>0.025</td>
<td>0.095</td>
<td>0.095</td>
</tr>
<tr>
<td>N</td>
<td>0.004</td>
<td>0.009</td>
<td>0.005</td>
<td>0.012</td>
<td>0.031</td>
<td>0.076</td>
</tr>
</tbody>
</table>

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

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

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

Then \( p \) is called **non-extreme**.

**Example 8.**

\[
\begin{pmatrix}
0.4 & 0.0 \\
0.3 & 0.3
\end{pmatrix}
\quad \quad \quad \quad \quad \quad
\begin{pmatrix}
0.4 & 0.1 \\
0.2 & 0.3
\end{pmatrix}
\]

Conditional probability distributions

**Definition 9.** For a JPD \( p \) and a subset \( \mathcal{Y} \subseteq \text{dom}(p) \) of its variables with \( p|_{\mathcal{Y}} > 0 \) we define

\[
p|_{\mathcal{Y}} := \frac{p}{p|_{\mathcal{Y}}}
\]

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.,

\[
(p|_{\mathcal{Y}})|_{\mathcal{Y}} \equiv 1
\]
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 $\text{dom}(R) = \text{dom}(C) = \{1, 2\}$.

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

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

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

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)

---

**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)
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].
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 $p(\mathcal{Y} = y) > 0$.

Example 11. Let $\Omega$ be the cards in an ordinary deck and

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

Cards for a single color:

2 3 4 5 6 7 8 9 10 J Q K A

| $S$ | $R$ | $T$ | $p(R, T | S)$ |
|-----|-----|-----|-------------|
| 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 |

<table>
<thead>
<tr>
<th>$R$</th>
<th>$T$</th>
<th>$p(R, T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Y</td>
<td>4/52 = 1/13</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>8/52 = 2/13</td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
<td>4/52 = 1/13</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>36/52 = 9/13</td>
</tr>
</tbody>
</table>
**Definition 11.** Let $X$, $Y$, and $Z$ be sets of variables.

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

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

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

$$p(X = x|Y = y, Z = z) = p(X = x|Z = z)$$

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

**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
