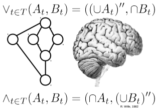
Random variables and Bayesian networks

Bayesian Statistics and Machine Learning

Dominik Endres

November 8, 2017

Philipps Universität Marburg



Outline

Random variable

Reminder:
Definition: A *probability space* is a tuple $(W, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-algebra over W and $P : \mathcal{F} \rightarrow [0, 1]$, with the properties:

P1 $P(W) = 1$
P2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$
We assume that $\mathcal{F} = 2^W$

Definition: a *random variable* X on a set of possible worlds W is a function $X : W \rightarrow Z$ from W to some range Z. If the range is the reals, i.e. $Z \subseteq \mathbb{R}$, then X is also called a *gamble*.

Notes:

- A random variable is neither random, nor is it a variable.
- But its value is unpredictable, if you don't know which $w \in W$ is the 'real world'.
- An instantiation of the value of a random variable (e.g. after you toss a coin) is called a *random variate*.

Random variable

Reminder:

Definition: A *probability space* is a tuple $(W, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-algebra over W and $P : \mathcal{F} \to [0,1]$, with the properties:

P1 $P(W) = 1$

P2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$

We assume that $\mathcal{F} = 2^W$

Definition: a *random variable* X on a set of possible worlds W is a function $X : W \to Z$ from W to some range Z. If the range is the reals, i.e. $Z \subseteq \mathbb{R}$, then X is also called a *gamble*.

Notes:

- A random variable is neither random, nor is it a variable.
- But its value is unpredictable, if you don't know which $w \in W$ is the 'real world'.
- An instantiation of the value of a random variable (e.g. after you toss a coin) is called a *random variate.*

Random variable: example 1

A coin is tossed 5 times. The outcome of a single toss is $\in \{H, T\}$. Thus, $W_5 = \{H, T\}^5$, i.e. all sequences of length 5 comprised of Hs and/or Ts.

Let N_H be the random variable $N_H : W_5 \rightarrow [0, 1, 2, 3, 4, 5]$ which represents the number of heads in a given sequence. In the world $HHHTT$, the value of N_H is 3: $N_H(HHHTT) = 3$.

Random variable: example 1

A coin is tossed 5 times. The outcome of a single toss is $\in \{H, T\}$. Thus, $W_5 = \{H, T\}^5$, i.e. all sequences of length 5 comprised of Hs and/or Ts.

Let N_H be the random variable $N_H : W_5 \rightarrow [0, 1, 2, 3, 4, 5]$ which represents the number of heads in a given sequence. In the world $HHHTT$, the value of N_H is 3: $N_H(HHHTT) = 3$.

Random variable: example 2

A die is rolled. Let $W = \{w_1, w_2, w_3, w_4, w_5, w_6\}$.

Let X be the random variable $X : W \to [1, 2, 3, 4, 5, 6]$ which represents the number on the face of the die which shows.

Question: what is the probability that X takes on a given value x?

Answer: this can be computed from the probability assigned to the elementary outcomes w. Let w_x be that w which fulfills $X(w_x) = x$, then

$$P(X = x) = P(w_x)$$

Random variable: example 2

A die is rolled. Let $W = \{w_1, w_2, w_3, w_4, w_5, w_6\}$.

Let X be the random variable $X : W \rightarrow [1, 2, 3, 4, 5, 6]$ which represents the number on the face of the die which shows.
Question: what is the probability that X takes on a given value x?
Answer: this can be computed from the probability assigned to the elementary outcomes w. Let w_x be that w which fulfills $X(w_x) = x$, then

$$P(X = x) = P(w_x)$$

Non-injective X

A die is rolled. Let $W = \{w_1, w_2, w_3, w_4, w_5, w_6\}$.

Let $X_{\frac{1}{2}}$ be the random variable $X_{\frac{1}{2}} : W \rightarrow [0, 1, 2, 3]$ which represents half of the number on the face of the die which shows, rounded to the next lower integer. Then

- $P(X_{\frac{1}{2}} = 0) = P(w_1)$
- $P(X_{\frac{1}{2}} = 1) = P(w_2) + P(w_3) = P(\{w_2, w_3\})$
- $P(X_{\frac{1}{2}} = 2) = P(w_4) + P(w_5) = P(\{w_4, w_5\})$
- $P(X_{\frac{1}{2}} = 3) = P(w_6)$
- For non-injective X, probability of $X = x$ is defined by summing over all elementary outcomes w_x for which $X(w_x) = x$. This is a consequence of P2.
- Also, $\sum_{x=0}^{3} P(X = x) = 1$ because of P1.

Probability distribution

Definition: Let Y be a random variable with range Z. A *probability distribution* is a function $P : Z \to [0,1]$ such that $\sum_{y \in Z} P(Y = y) = 1$.

Note:

- Given a probability space $(W, \mathcal{F}, Q)$, and a random variable Y, the corresponding probability distribution over Y can be obtained via $P(Y = y) = \sum_{w: w \in W, Y(w) = x} Q(w)$.
- It is customary to denote the probability distribution over Y by $P(Y)$.
- Instead of writing $P(Y = y)$ for the probability that $Y = y$ under $P(Y)$, it is customary to write $P(y)$.

Probability distribution

Definition: Let Y be a random variable with range Z. A *probability distribution* is a function $P : Z \to [0,1]$ such that $\sum_{y \in Z} P(Y = y) = 1$.

Note:

- Given a probability space $(W, \mathcal{F}, Q)$, and a random variable Y, the corresponding probability distribution over Y can be obtained via $P(Y = y) = \sum_{w: w \in W, Y(w) = x} Q(w)$.
- It is customary to denote the probability distribution over Y by $P(Y)$.
- Instead of writing $P(Y = y)$ for the probability that $Y = y$ under $P(Y)$, it is customary to write $P(y)$.

Use of random variables: computing expectations

Let Y be a *gamble*, i.e. random variable with range $Z \subseteq \mathbb{R}$ and probability distribution $P(Y)$.
The **expected value** or *expectation* of Y w.r.t. $P(Y)$ is defined as

$$\mathrm{E}_{P(Y)}(Y) = \sum_{y \in Z} yP(y)$$

Notes:

- $\mathrm{E}_{P(Y)}(Y)$ does not have to be $\in Z$.
- Let Z be the value of a fair die roll. Then $\mathrm{E}_{P(Y)}(Y) = \frac{1}{6}(1 + \ldots + 6) = 3.5$

Use of random variables: computing expectations

Let Y be a *gamble*, i.e. random variable with range $Z \subseteq \mathbb{R}$ and probability distribution $P(Y)$.

The **expected value** or *expectation* of Y w.r.t. $P(Y)$ is defined as

$$\mathrm{E}_{P(Y)}(Y) = \sum_{y \in Z} y P(y)$$

Notes:

- $\mathrm{E}_{P(Y)}(Y)$ does not have to be $\in Z$.
- Let Z be the value of a fair die roll. Then $\mathrm{E}_{P(Y)}(Y) = \frac{1}{6}(1 + \ldots + 6) = 3.5$

Joint probability distribution

Reminder:
$P(Y)$ denotes a probability distribution over random variable Y.
$P(y)$ is a shorthand for $P(Y = y)$.

Definition: Let $X_1, \ldots, X_N$ be random variables with ranges $Z_1, \ldots, Z_N$. A **joint probability distribution** $P(X_1, \ldots, X_N)$ is a function $P : \prod_{i=1}^{N} Z_i \to [0, 1]$ such that

$$\sum_{x_1 \in Z_1} \cdots \sum_{x_N \in Z_N} P(x_1, \ldots, x_N) = 1.$$

Example: joint probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die will show only the numbers 1,2,3'.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

$P(X, Y)$: joint probability distribution over the numbers shown and the fairness of the die.
$P(X = x, Y = y) = P(x, y) =$ "the probability that the die showed x and is $y \in \{\text{fair}, \text{unfair}\}$".

Example: joint probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die will show only the numbers 1,2,3'.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

$P(X, Y)$: joint probability distribution over the numbers shown and the fairness of the die.
$P(X = x, Y = y) = P(x, y) =$ "the probability that the die showed x and is $y \in \{\text{fair}, \text{unfair}\}$".

Example: joint probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die will show only the numbers 1,2,3'.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

$P(X, Y)$: joint probability distribution over the numbers shown and the fairness of the die.
$P(X = x, Y = y) = P(x, y) =$ "the probability that the die showed x and is $y \in \{\text{fair}, \text{unfair}\}$".

Use of random variables: structuring the set W

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die will show only the numbers 1,2,3'.

Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the number shown.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the fairness of the die.

Both X and Y act on W, but they extract different aspects of the possible worlds/elementary outcomes.
⇒ **Random variables** are useful for structuring and describing sets of possible worlds and/or elementary outcomes.

Marginal probability distribution

Reminder:
$P(Y)$ denotes a probability distribution over random variable Y.
$P(y)$ is a shorthand for $P(Y = y)$.
$P(X_1, \ldots, X_N)$ denotes a joint probability distribution over $X_1, \ldots, X_N$.

Definition: Let $X_1, \ldots, X_N$ be random variables with ranges $Z_1, \ldots, Z_N$, and $P(X_1, \ldots, X_N)$ be their joint probability distribution. Let $I = \{i_1, \ldots, i_K\} \subseteq \{1, \ldots, N\}$ be an index set and $J = \{1, \ldots, N\} \setminus I$ its complement.
The **marginal probability distribution** $P(X_{i_1}, \ldots, X_{i_K})$ is

$$P(x_{i_1}, \ldots, x_{i_K}) = \sum_{x_{j_1} \in Z_{j_1}} \cdots \sum_{x_{j_{N-K}} \in Z_{j_{N-K}}} P(x_1, \ldots, x_N)$$

- The marginal distribution over any subset of random variables is obtained by 'summing out' all other random variables.
- Since the joint distribution $P(X_1, \ldots, X_N)$ is normalized to 1, so are all marginals.

Marginal probability distribution

Reminder:
$P(Y)$ denotes a probability distribution over random variable Y.
$P(y)$ is a shorthand for $P(Y = y)$.
$P(X_1, \dots, X_N)$ denotes a joint probability distribution over $X_1, \dots, X_N$.

Definition: Let $X_1, \dots, X_N$ be random variables with ranges $Z_1, \dots, Z_N$, and $P(X_1, \dots, X_N)$ be their joint probability distribution. Let $I = \{i_1, \dots, i_K\} \subseteq \{1, \dots, N\}$ be an index set and $J = \{1, \dots, N\} \setminus I$ its complement.
The **marginal probability distribution** $P(X_{i_1}, \dots, X_{i_K})$ is

$$P(x_{i_1}, \dots, x_{i_K}) = \sum_{x_{j_1} \in Z_{j_1}} \cdots \sum_{x_{j_{N-K}} \in Z_{j_{N-K}}} P(x_1, \dots, x_N)$$

- The marginal distribution over any subset of random variables is obtained by 'summing out' all other random variables.
- Since the joint distribution $P(X_1, \dots, X_N)$ is normalized to 1, so are all marginals.

Example: marginal probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world. $P(X, Y)$ is joint probability distribution over the numbers shown and the fairness of the die.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

$P(X)$: probability distribution over the numbers shown by the die.
$P(X = x) = P(x) =$"the probability that the die showed x"$=\sum_y P(x, y)$, where $\sum_y = \sum_{y \in \{\text{fair},\text{loaded}\}}$

Example: marginal probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world. $P(X, Y)$ is joint probability distribution over the numbers shown and the fairness of the die.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

$P(X)$: probability distribution over the numbers shown by the die.
$P(X = x) = P(x)$ = "the probability that the die showed x" $=\sum_y P(x, y)$, where $\sum_y = \sum_{y \in \{\text{fair},\text{loaded}\}}$

Conditional probability distribution

Reminder:
$P(Y)$ denotes a probability distribution over random variable Y.
$P(y)$ is a shorthand for $P(Y = y)$.
$P(X_1, \ldots, X_N)$ denotes a joint probability distribution over $X_1, \ldots, X_N$.

Definition: Let $X_1, \ldots, X_N$ be random variables and $P(X_1, \ldots, X_N)$ be their joint probability distribution. Let $I = \{i_1, \ldots, i_K\}$ and $C = \{c_1, \ldots, c_M\}$ be two index sets such that $I \cup C = \{1, \ldots, N\}$. If $P(X_{c_1}, \ldots, X_{c_M}) > 0$, then the **conditional probability distribution** is

$$P(X_{i_1}, \ldots, X_{i_K} | X_{c_1}, \ldots, X_{c_M}) = \frac{P(X_1, \ldots, X_N)}{P(X_{c_1}, \ldots, X_{c_M})}$$

Note: $P(X_{j_1}, \ldots, X_{j_M}) > 0$ means that this marginal distribution is strictly positive for all values of $X_{j_1}, \ldots, X_{j_M}$.

Example: conditional probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

$P(X|Y)$: probability distribution over the numbers shown by the die given which die it is $= \frac{P(X,Y)}{P(Y)}$
$P(X = x|Y = y) = P(x|y) =$ "the probability that the die showed x given that it was die y" $= \frac{P(x,y)}{P(y)}$.

Note: writing $P(X|Y) = \frac{P(X,Y)}{P(Y)}$ means that this relationship holds point-wise, i.e. for all possible values of X and Y.

Example: conditional probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

$P(X|Y)$: probability distribution over the numbers shown by the die given which die it is $= \frac{P(X,Y)}{P(Y)}$
$P(X = x|Y = y) = P(x|y) =$ "the probability that the die showed x given that it was die y" $= \frac{P(x,y)}{P(y)}$.

Note: writing $P(X|Y) = \frac{P(X,Y)}{P(Y)}$ means that this relationship holds point-wise, i.e. for all possible values of X and Y.

Example: conditional probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

$P(X|Y)$: probability distribution over the numbers shown by the die given which die it is $= \frac{P(X,Y)}{P(Y)}$
$P(X = x|Y = y) = P(x|y) =$ "the probability that the die showed x given that it was die y" $= \frac{P(x,y)}{P(y)}$.
Note: writing $P(X|Y) = \frac{P(X,Y)}{P(Y)}$ means that this relationship holds point-wise, i.e. for all possible values of X and Y.

Product rule for probability distributions

Reminder:
Random variables $X_1, \ldots, X_N$ with joint prob. dist. $P(X_1, \ldots, X_N)$.
$I = \{i_1, \ldots, i_K\}$ and $C = \{c_1, \ldots, c_M\}$ such that $I \cup C = \{1, \ldots, N\}$.
Conditional prob. dist. :$P(X_{i_1}, \ldots, X_{i_K} | X_{c_1}, \ldots, X_{c_M}) = \frac{P(X_1, \ldots, X_N)}{P(X_{c_1}, \ldots, X_{c_M})}$

A consequence of the definition of the conditional probability distribution is the **product rule for random variables**:

$$P(X_{i_1}, \ldots, X_{i_K} | X_{c_1}, \ldots, X_{c_M}) P(X_{c_1}, \ldots, X_{c_M}) = P(X_1, \ldots, X_N)$$

Note: as before, the equality is point-wise.

Chain rule for probability distributions

Reminder:
Random variables $X_1, \ldots, X_N$ with joint prob. dist. $P(X_1, \ldots, X_N)$.
$I = \{i_1, \ldots, i_K\}$ and $C = \{c_1, \ldots, c_M\}$ such that $I \cup C = \{1, \ldots, N\}$.
Product rule for prob. dist. $P(X_{i_1}, \ldots, X_{i_K} | X_{c_1}, \ldots, X_{c_M}) P(X_{c_1}, \ldots, X_{c_M}) = P(X_1, \ldots, X_N)$

Apply product rule repeatedly:

$$\begin{aligned} P(X_1, \ldots, X_N) &= P(X_1|X_2, \ldots, X_N) P(X_2, \ldots, X_N) \\ &= P(X_1|X_2, \ldots, X_N) P(X_2|X_3, \ldots, X_N) P(X_3, \ldots, X_N) \\ &\vdots \\ &= \prod_{i=1}^{N-1} P(X_i|X_{i+1}, \ldots, X_N) P(X_N) \qquad (1) \end{aligned}$$

Holds for any ordering of the X_i !
This is the **chain rule for probability distributions**.

Independence between random variables

Reminder:
$P(X, Y)$ is joint probability distribution of X and Y.
$P(X) = \sum_y P(X, y)$ is the marginal probability distribution of X.
$P(Y) = \sum_x P(x, Y)$ is the marginal probability distribution of Y.

Definition: Two random variables X and Y are independent if and only if

$$P(X, Y) = P(X)P(Y).$$

Note: If X, Y are independent, then $P(X|Y) = \frac{P(X,Y)}{P(Y)} = P(X)$. Knowing Y does not change knowledge of X.

Motivating example: conditional independence between random variables

Example: A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Random variables:

- X_1: value of outcome of 1st roll $\in \{1; \ldots; 6\}$.
- X_2: value of outcome of 2nd roll $\in \{1; \ldots; 6\}$.
- Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.

Question: what is the joint distribution $P(X_1, X_2, Y)$? Knowing it would enable us to compute all marginals and conditionals, e.g. $P(Y|X_1, X_2)$.

Motivating example: conditional independence between random variables

Example: A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Random variables:

- X_1: value of outcome of 1st roll $\in \{1; \ldots; 6\}$.
- X_2: value of outcome of 2nd roll $\in \{1; \ldots; 6\}$.
- Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.

Question: what is the joint distribution $P(X_1, X_2, Y)$? Knowing it would enable us to compute all marginals and conditionals, e.g. $P(Y|X_1, X_2)$.

Motivating example: conditional independence between random variables

Example: A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Random variables:

- X_1: value of outcome of 1st roll $\in \{1; \ldots; 6\}$.
- X_2: value of outcome of 2nd roll $\in \{1; \ldots; 6\}$.
- Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.

Question: what is the joint distribution $P(X_1, X_2, Y)$? Knowing it would enable us to compute all marginals and conditionals, e.g. $P(Y|X_1, X_2)$.

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \dots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

Question: what is the joint distribution $P(X_1, X_2, Y)$?
Answer: use chain rule:

$$P(X_1, X_2, Y) = P(X_1|X_2, Y)P(X_2|Y)P(Y)$$

We know $P(Y)$ and $P(X_2|Y)$. What about $P(X_1|X_2, Y)$?
Once we know the die (i.e. the value of Y), the values of $P(X_i|Y)$ of each die roll should be the same, no matter how often we roll the die.

$$\Rightarrow P(X_1|X_2, Y) = P(X_1|Y).$$

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots ; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

Question: what is the joint distribution $P(X_1, X_2, Y)$?
Answer: use chain rule:

$$P(X_1, X_2, Y) = P(X_1|X_2, Y)P(X_2|Y)P(Y)$$

We know $P(Y)$ and $P(X_2|Y)$. What about $P(X_1|X_2, Y)$?
Once we know the die (i.e. the value of Y), the values of $P(X_i|Y)$ of each die roll should be the same, no matter how often we roll the die.

$$\Rightarrow P(X_1|X_2, Y) = P(X_1|Y).$$

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

Question: what is the joint distribution $P(X_1, X_2, Y)$?
Answer: use chain rule:

$$P(X_1, X_2, Y) = P(X_1|X_2, Y)P(X_2|Y)P(Y)$$

We know $P(Y)$ and $P(X_2|Y)$. What about $P(X_1|X_2, Y)$?
Once we know the die (i.e. the value of Y), the values of $P(X_i|Y)$ of each die roll should be the same, no matter how often we roll the die.

$$\Rightarrow P(X_1|X_2, Y) = P(X_1|Y).$$

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

We believe: $P(X_1|X_2, Y) = P(X_1|Y)$. Thus:

$$\begin{aligned} P(X_1, X_2, Y) &= P(X_1|Y)P(X_2|Y)P(Y) \\ &= P(X_1, X_2|Y)P(Y) \end{aligned}$$

$$\Rightarrow P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

Like the definition of independence, but everything is conditioned on Y.

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \dots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

We believe: $P(X_1|X_2, Y) = P(X_1|Y)$. Thus:

$$\begin{aligned} P(X_1, X_2, Y) &= P(X_1|Y)P(X_2|Y)P(Y) \\ &= P(X_1, X_2|Y)P(Y) \end{aligned}$$

$$\Rightarrow P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

Like the definition of independence, but everything is conditioned on Y.

Conditional independence between random variables

Definition: Two random variables X_1 and X_2 are *conditionally independent* given a random variable Y if and only if

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y).$$

Alternatively, X_1 and X_2 are conditionally independent if and only if

- $P(X_1|Y) > 0$
- $P(X_2|Y) > 0$
- $P(X_1|X_2, Y) = P(X_1|Y)$
- $P(X_2|X_1, Y) = P(X_2|Y)$

Conditional independence between random variables

Definition: Two random variables X_1 and X_2 are *conditionally independent* given a random variable Y if and only if

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y).$$

Notes:

- This definition can be extended to more than 3 random variables by replacing any of X_1, X_2 or Y with a list of random variables.
- Variables that are conditionally independent are usually marginally dependent, and vice versa.

Conditional independence between random variables

Definition: Two random variables X_1 and X_2 are *conditionally independent* given a random variable Y if and only if

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y).$$

Notes:

- This definition can be extended to more than 3 random variables by replacing any of X_1, X_2 or Y with a list of random variables.
- Variables that are conditionally independent are usually marginally dependent, and vice versa.

Example: conditional independence vs. marginal dependence

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \dots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
Conditional independence: $P(X_1|X_2, Y) = P(X_1|Y)$ and $P(X_2|X_1, Y) = P(X_2|Y)$.

Die rolls are conditionally independent given Y.
Marginal probability distribution $P(X_1, X_2)$:

$$\begin{aligned} P(X_1, X_2) &= \sum_y P(X_1, X_2, y) \\ &= \sum_y P(X_1|X_2, y) P(X_2|y) P(y) \\ &= \sum_y P(X_1|y) P(X_2|y) P(y) \end{aligned}$$

Example: conditional independence vs. marginal dependence

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
Conditional independence: $P(X_1|X_2, Y) = P(X_1|Y)$ and $P(X_2|X_1, Y) = P(X_2|Y)$.
Marginal prob. dist. $P(X_1, X_2) = \sum_y P(X_1|y)P(X_2|y)P(y)$

On the other hand:

$$
\begin{aligned}
P(X_1) &= \sum_y P(X_1, y) \\
&= \sum_y P(X_1|y)P(y) \\
P(X_2) &= \sum_y P(X_2|y)P(y)
\end{aligned}
$$

Example: conditional independence vs. marginal dependence

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots ; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
Conditional independence: $P(X_1|X_2, Y) = P(X_1|Y)$ and $P(X_2|X_1, Y) = P(X_2|Y)$.
Marginal prob. dist. $P(X_1, X_2) = \sum_y P(X_1|y)P(X_2|y)P(y)$

Therefore

$$
\begin{aligned}
P(X_1)P(X_2) &= \sum_y P(X_1|y)P(y) \sum_y P(X_2|y)P(y) \\
&\neq \sum_y P(X_1|y)P(X_2|y)P(y) \\
&= P(X_1, X_2)
\end{aligned} \tag{2}
$$

$\Rightarrow$ X_1 and X_2 are marginally dependent.

$\Rightarrow$ One die roll contains information about the other if we *do not* know Y.

$\Rightarrow$ The marginal dependence goes in *both* directions.

Independent identically distributed random variables

We defined: X_1 and X_2 have the same range $Z = \{1, \ldots, 6\}$.
We believe:

- Conditional independence:
  - $P(X_1|X_2, Y) = P(X_1|Y)$
  - $P(X_2|X_1, Y) = P(X_2|Y)$
- Identical distributions:
  $\forall x \in Z: \; P(X_1 = x|Y) = P(X_2 = x|Y)$.

Random variables, which are (conditionally) independent and have the same probability distribution are called **independent identically distributed**, short **i.i.d.**.
This is an extremly common assumption in machine learning, but it is not the only possible assumption (see e.g. time series modelling).

Independent identically distributed random variables

We defined: X_1 and X_2 have the same range $Z = \{1, \dots, 6\}$.
We believe:

- Conditional independence:
    - $P(X_1|X_2, Y) = P(X_1|Y)$
    - $P(X_2|X_1, Y) = P(X_2|Y)$
- Identical distributions:
  $\forall x \in Z: \; P(X_1 = x|Y) = P(X_2 = x|Y)$.

Random variables, which are (conditionally) independent and have the same probability distribution are called **independent identically distributed**, short **i.i.d.**.

This is an extremly common assumption in machine learning, but it is not the only possible assumption (see e.g. time series modelling).

Independent identically distributed random variables

We defined: X_1 and X_2 have the same range $Z = \{1, \ldots, 6\}$.
We believe:

- Conditional independence:
  - $P(X_1|X_2, Y) = P(X_1|Y)$
  - $P(X_2|X_1, Y) = P(X_2|Y)$
- Identical distributions:
  $\forall x \in Z : P(X_1 = x|Y) = P(X_2 = x|Y)$.

Random variables, which are (conditionally) independent and have the same probability distribution are called **independent identically distributed**, short **i.i.d.**.
This is an extremly common assumption in machine learning, but it is not the only possible assumption (see e.g. time series modelling).

Summary: random variables

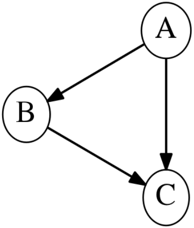
- A *random variable* X on a set of possible worlds W is a function $X : W \rightarrow Z$ from W to some range Z.
- A *probability distribution* is a function $P : Z \rightarrow [0, 1]$ such that $\sum_{x \in Z} P(X = x) = 1$.
- Chain rule: $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | X_{i+1}, \ldots, X_N)$
- Conditional independence $P(X_1, X_2 | Y) = P(X_1 | Y) P(X_2 | Y)$.
- Conditional independence $P(X_1 | X_2, Y) = P(X_1 | Y)$.
  - Expressed by omitting all variables that X_1 does not depend on after the conditioning line (here: X_2 omitted).
- Marginal probability distribution $P(X_1) = \sum_y P(X_1, y)$.
- Independent identically distibuted (i.i.d) random variables.

Bayesian networks

A type of probabilistic graphical model which expresses conditional (in)dependence relationships.

Random variables	**Bayesian networks**
Random variables A, B, C	Nodes of a graph
Conditional (in)dependence	Directed edges
Chain rule decomposition	directed acyclic graph (DAG)

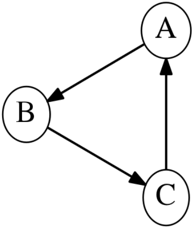
The graph represents a set of *constraints* on the joint probability distribution of the random variables.



A Bayesian network with 3 random variables A,B,C.

Example: closed loop

NOT a directed acyclic graph (DAG), thus not a Bayesian network.
Closed directed loop $A \to B \to C$.

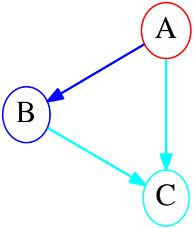


Representing constraints on joint distributions

Reminder:
Chain rule for prob. dist. $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | X_{i+1}, \ldots, X_N)$

Example: 3 random variables A, B, C. Joint distribution

$$P(A, B, C) = P(A)P(B|A)P(C|A, B)$$

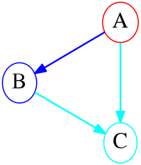


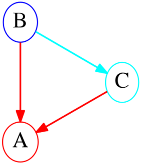
- Each node represents a random variable
- If and only if there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A, \ldots)$.

Alternative ordering of variables

Reminder:
Chain rule for prob. dist. $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|X_{i+1}, \ldots, X_N)$
Independece of random variables $P(X, Y) = P(X)P(Y)$.

Order of factorization of joint distribution can be exchanged:





$P(A, B, C) = P(A)P(B|A)P(C|A, B)$ $\quad$ $P(A, B, C) = P(B)P(C|B)P(A|B, C)$

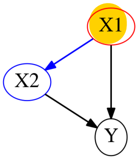
- Both graphs describe possible factorizations of $P(A, B, C)$.
- Here, both factorizations are equivalent w.r.t. the dependency structure: a given variable is conditionally dependent on all others.
  - A consequence of probabilistic (in)dependence being a mutual property.
  - Both graphs are *fully connected*.

Example: rolling a die twice. Bad ordering

Reminder:
Each node represents a random variable.
If there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A,\ldots)$.

Random variables: X_1, X_2: value of 1st and 2nd roll, Y: fairness.
Factorization of joint probability distribution:

$$P(X_1, X_2, Y) = P(X_1)P(X_2|X_1)P(Y|X_1, X_2)$$



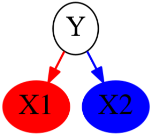
$\Rightarrow$ the factorization order $X_1 \to X_2 \to Y$ is not a good choice, because all variables are dependent on each other.

Example: rolling a die twice. Good ordering

Reminder:
Each node represents a random variable.
If there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A,...)$.

Random variables: X_1, X_2: value of 1st and 2nd roll, Y: fairness.
Factorization of joint probability distribution:

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)\underbrace{P(X_2|X_1, Y)}_{P(X_2|Y)}$$



$\Rightarrow$ the factorization order $Y \rightarrow X_1 \rightarrow X_2$ is a better choice of ordering, because conditional independence relationships are represented in the graph!

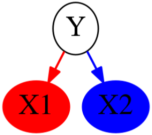
Example: rolling a die twice. Good ordering

Reminder:
Each node represents a random variable.
If there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A,...)$.

Random variables: X_1, X_2: value of 1st and 2nd roll, Y: fairness.
Factorization of joint probability distribution:

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)\underbrace{P(X_2|X_1, Y)}_{P(X_2|Y)}$$

Filled nodes:
observed variables



$\Rightarrow$ the factorization order $Y \to X_1 \to X_2$ is a better choice of ordering, because conditional independence relationships are represented in the graph!

Good vs. bad random variable ordering

Question: in what sense is the factorization

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)P(X_2|Y)$$

better than

$$P(X_1, X_2, Y) = P(X_1)P(X_2|X_1)P(Y|X_1, X_2)$$

Good vs. bad random variable ordering

Question: in what sense is the factorization

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)P(X_2|Y)$$

better than

$$P(X_1, X_2, Y) = P(X_1)P(X_2|X_1)P(Y|X_1, X_2)$$

Answer 1: consider the number of probabilities which you have to assign: if a random variable can take on N different values, then you have to guess/estimate $N-1$ probabilities to determine its probability distribution.

- Good ordering: $1 + (5 \times 2) + (5 \times 2) = 21$. Because of i.i.d. property, actually only 11.
- Bad ordering: $5 + (5 \times 6) + 1 \times (6 \times 6) = 71$.

$\Rightarrow$ far less probabilities for the good ordering.

Good vs. bad random variable ordering

Question: in what sense is the factorization

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)P(X_2|Y)$$

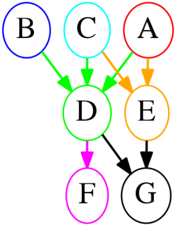
better than

$$P(X_1, X_2, Y) = P(X_1)P(X_2|X_1)P(Y|X_1, X_2)$$

Answer 2: The good ordering represents our information about the structure of the problem: die fairness determines probabilities of outcomes, not the other way round. We might say that the good ordering represents the 'causal structure' of the problem. *Caveat*: for a Bayesian network to represent causal structure, additional conditions must hold (see e.g. Pearl(2000):Causality).

Bayesian network terminology

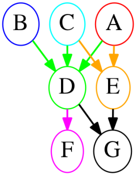
Reminder:
Each node represents a random variable.
If there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A,\ldots)$.



A,B,C are the *parents* of D. $\mathrm{pa}_D = \{A, B, C\}$.
D,E are the *children* of A.
A,B,C,D,E are the *ancestors* of G.
D,F,G are the *descendants* of B.
A,B,C are the *roots* (no parents).
F,G are the *leaves* (no children).

Conditional independence given parents

Reminder:
Each node represents a random variable.
If and only if there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A,\ldots)$.
Set of parents of node A is pa_A.



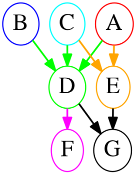
$$\mathrm{pa}_A = \mathrm{pa}_B = \mathrm{pa}_C = \emptyset$$
$$\mathrm{pa}_D = \{A, B, C\}$$
$$\mathrm{pa}_E = \{A, C\}$$
$$\mathrm{pa}_F = \{D\}$$
$$\mathrm{pa}_G = \{D, E\}$$

Factorization of joint distribution: choose an ordering such that pa_X always precede X in the factorization chain. Always possible because graph is a DAG. Let $P(X|\emptyset) = P(X)$.

$$\begin{aligned} P(A,B,C,D,E,F,G) &= P(A)\,P(B)\,P(C) \\ &\times P(D|A,B,C)\,P(E|A,C) \\ &\times P(G|D,E)\,P(F|D) \end{aligned}$$

Conditional independence given parents

Reminder:
Each node represents a random variable.
If and only if there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A,\ldots)$.
Set of parents of node A is pa_A.



$$\mathrm{pa}_A = \mathrm{pa}_B = \mathrm{pa}_C = \emptyset$$
$$\mathrm{pa}_D = \{A, B, C\}$$
$$\mathrm{pa}_E = \{A, C\}$$
$$\mathrm{pa}_F = \{D\}$$
$$\mathrm{pa}_G = \{D, E\}$$

Alternatively, we can write this as

$$P(A,B,C,D,E,F,G) =$$

$$\begin{array}{llll} = & P(A)\,P(B)\,P(C) & = & P(A|\mathrm{pa}_A)\,P(B|\mathrm{pa}_B)\,P(C|\mathrm{pa}_C) \\ \times & P(D|A,B,C)\,P(E|A,C) & \times & P(D|\mathrm{pa}_D)\,P(E|\mathrm{pa}_E) \\ \times & P(F|D)\,P(G|D,E) & \times & P(F|\mathrm{pa}_F)\,P(G|\mathrm{pa}_G) \end{array}$$

Translating a graph structure into a factorization

The expression for the joint distribution

$$
\begin{aligned}
P(A, B, C, D, E, F, G) &= P(A|\mathrm{pa}_A)\, P(B|\mathrm{pa}_B)\, P(C|\mathrm{pa}_C) \\
&\times P(D|\mathrm{pa}_D)\, P(E|\mathrm{pa}_E) \\
&\times P(F|\mathrm{pa}_F)\, P(G|\mathrm{pa}_G)
\end{aligned}
$$

no longer depends on the chosen factorization order, only on the parent-child relationships expressed in the graph!
(because multiplication is commutative).
Algorithm for translating a Bayesian network into a factorization of a joint distribution:

- Given: random variables $X_1, \ldots, X_N$ and a DAG G with nodes labeled $X_1, \ldots, X_N$.
- For all X_i, identify pa_{X_i} from G.
- Output $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\mathrm{pa}_{X_i})$

Translating a graph structure into a factorization

The expression for the joint distribution

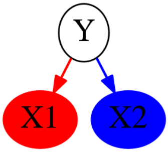
$$\begin{aligned} P(A, B, C, D, E, F, G) &= P(A|\mathrm{pa}_A)\, P(B|\mathrm{pa}_B)\, P(C|\mathrm{pa}_C) \\ &\times P(D|\mathrm{pa}_D)\, P(E|\mathrm{pa}_E) \\ &\times P(F|\mathrm{pa}_F)\, P(G|\mathrm{pa}_G) \end{aligned}$$

no longer depends on the chosen factorization order, only on the parent-child relationships expressed in the graph!
(because multiplication is commutative).
Algorithm for translating a Bayesian network into a factorization of a joint distribution:

- Given: random variables $X_1, \ldots, X_N$ and a DAG G with nodes labeled $X_1, \ldots, X_N$.
- For all X_i, identify pa_{X_i} from G.
- Output $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\mathrm{pa}_{X_i})$

Example: from graph to factorization

Random variables: X_1, X_2: value of 1st and 2nd roll, Y: fairness.



- $\text{pa}_Y = \emptyset$
- $\text{pa}_{X_1} = \{Y\}$
- $\text{pa}_{X_2} = \{Y\}$

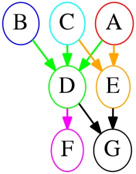
$\Rightarrow P(X_1, X_2, Y) = P(Y)P(X_1|Y)P(X_2|Y)$

Given a factorization:

$$
\begin{aligned}
P(A,B,C,D,E,F,G) &= P(A)\,P(B)\,P(C) \\
&\times P(D|A,B,C)\,P(E|A,C) \\
&\times P(G|D,E)\,P(F|D)
\end{aligned}
$$

building the graph is straightforward:

1. Identify and draw the roots: A, B, C
2. Find all children of the roots: D, E
3. Draw arrows for each cond. dependence
4. Iterate 2. and 3. until leaves are reached



Summary: Bayesian networks

A type of probabilistic graphical model which expresses conditional (in)dependence relationships.

Random variables	**Bayesian networks**
Random variables A, B, C	Nodes of a graph
Conditional (in)dependence	Directed edges
Chain rule decomposition	directed acyclic graph (DAG)

- Good decompositions keep the number of probabilities to estimate small.
- Good decompositions represent our knowledge/assumptions about probabilistic (in)dependence relationships between the random variables involved.
- A given chain-rule factorization can translated into a DAG.
- A given DAG can be translated into a chain-rule factorization.

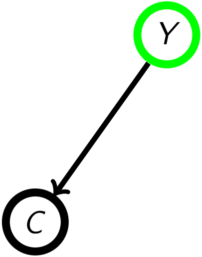
Causality example: coffee network

Random variables

- **C**: coffee in the pot
- **Y**: yesterday's coffee is still there
- **R**: coffee machine was recently run
- **T**: time of day

C=1 happens $\approx$ 3 times a day: in the morning, after lunch and at 4pm.

Causality example: coffee network

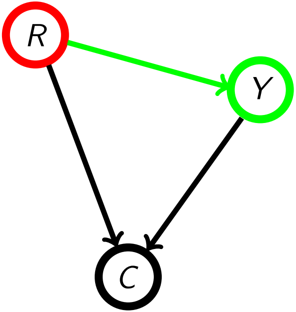


Random variables

- **C**: coffee in the pot
- **Y**: yesterday's coffee is still there
- **R**: coffee machine was recently run
- **T**: time of day

Y=1: typically in the morning, if at all.

Causality example: coffee network

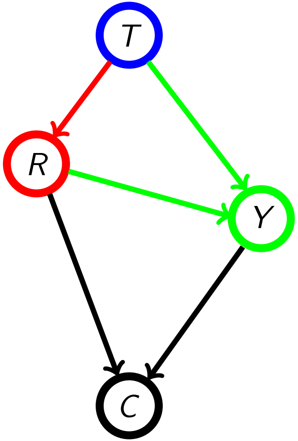


Random variables

- **C**: coffee in the pot
- **Y**: yesterday's coffee is still there
- **R**: coffee machine was recently run
- **T**: time of day

R=1 happens $\approx$ 3 times a day: in the morning, after lunch and at 4pm.

Causality example: coffee network

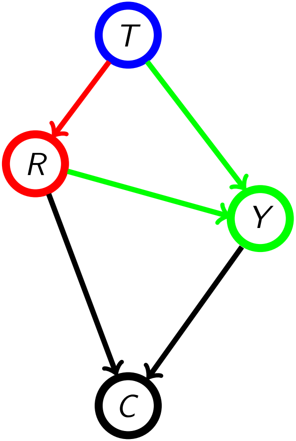


Random variables

- **C**: coffee in the pot
- **Y**: yesterday's coffee is still there
- **R**: coffee machine was recently run
- **T**: time of day

T allows prediction of **R** and **Y**, which mediate influence of **T** on **C**

Causality example: coffee network



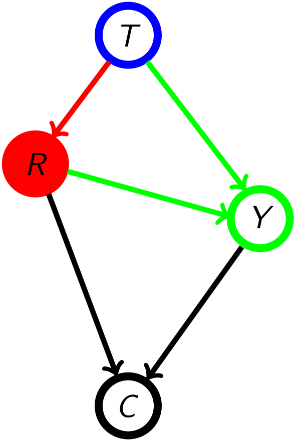
Random variables

- **C**: coffee in the pot
- **Y**: yesterday's coffee is still there
- **R**: coffee machine was recently run
- **T**: time of day

$$P(C, Y, R, T) = P(C|Y, R)P(Y|R, T)P(R|T)P(T)$$

Observing R: inference

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day.



Assume: Observe $R = 1$.

$\Rightarrow$ Since $P(C = 1|R = 1) \approx 1$, we expect **C**='true'.

Marginalize **C** and **Y**. Using Bayes' rule:

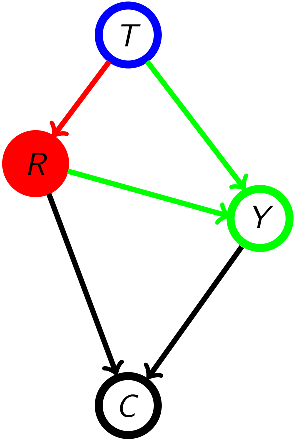
$$P(T|R) = \frac{P(R|T)P(T)}{P(R)}$$

with $P(R) = \sum_T P(R|T)P(T)$.

$\Rightarrow$ $P(T = 4\text{pm}|R = 1) > P(T = 4\text{pm})$. Likewise for T='after lunch' or T='morning'.

Observing R: inference

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day.



Assume: Observe $R = 1$.
$\Rightarrow$ Since $P(C = 1|R = 1) \approx 1$, we expect **C**='true'.

Marginalize **C** and **Y**. Using Bayes' rule:

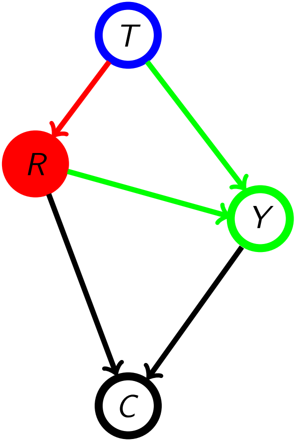
$$P(T|R) = \frac{P(R|T)P(T)}{P(R)}$$

with $P(R) = \sum_T P(R|T)P(T)$.

$\Rightarrow P(T = 4\text{pm}|R = 1) > P(T = 4\text{pm})$. Likewise for T='after lunch' or T='morning'.

Observing R: inference

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day.



Assume: Observe $R = 1$.
$\Rightarrow$ Since $P(C = 1|R = 1) \approx 1$, we expect **C**='true'.

Marginalize **C** and **Y**. Using Bayes' rule:

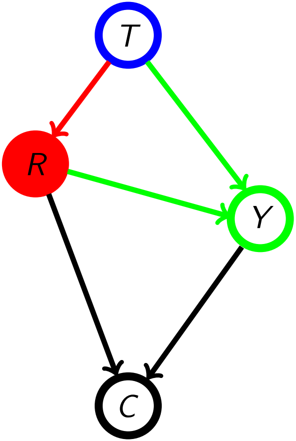
$$P(T|R) = \frac{P(R|T)P(T)}{P(R)}$$

with $P(R) = \sum_T P(R|T)P(T)$.

$\Rightarrow$ $P(T = 4\text{pm}|R = 1) > P(T = 4\text{pm})$.
Likewise for T='after lunch' or T='morning'.

Observing R: inference

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day.



Assume: Observe $R = 1$.
$\Rightarrow$ Since $P(C = 1|R = 1) \approx 1$, we expect **C**='true'.

Marginalize **C** and **Y**. Using Bayes' rule:

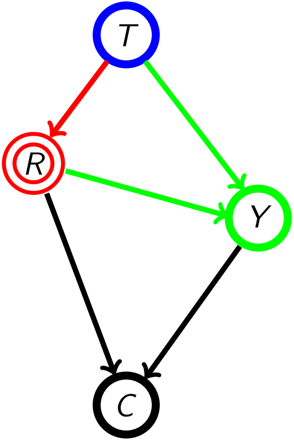
$$P(T|R) = \frac{P(R|T)P(T)}{P(R)}$$

with $P(R) = \sum_T P(R|T)P(T)$.

$\Rightarrow P(T = 4\text{pm}|R = 1) > P(T = 4\text{pm})$. Likewise for T='after lunch' or T='morning'.

Acting on R: causal inference

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day.



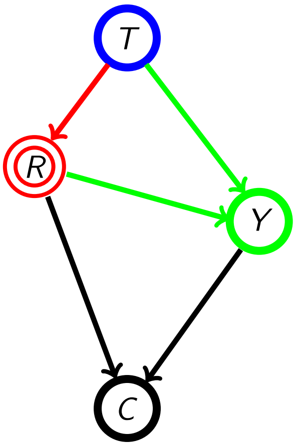
Assume: I set $R = 1$, i.e. I run the coffee machine
Expressed graphically by double circle, in formula by 'do($R = 1$)'
I still think $P(C = 1|\text{do}(R = 1)) \approx 1$, thus I expect **C**=1.

But what about
$P(T = 4\text{pm}|\text{do}(R = 1)) > P(T = 4\text{pm})$?
Does not make sense. Running the coffee machine tells me **nothing** about the time of the day.

$\Rightarrow$ this graph structure is wrong if I **act** on the variable R.

Acting on R: causal inference

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day.



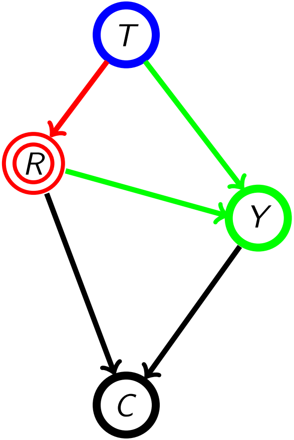
Assume: I set $R = 1$, i.e. I run the coffee machine
Expressed graphically by double circle, in formula by 'do($R = 1$)'
I still think $P(C = 1|\text{do}(R = 1)) \approx 1$, thus I expect **C**=1.

But what about
$P(T = 4\text{pm}|\text{do}(R = 1)) > P(T = 4\text{pm})$?
Does not make sense. Running the coffee machine tells me **nothing** about the time of the day.

$\Rightarrow$ this graph structure is wrong if I **act** on the variable R.

Acting on R: causal inference

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day.



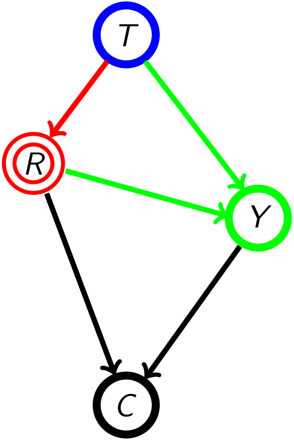
Assume: I set $R = 1$, i.e. I run the coffee machine
Expressed graphically by double circle, in formula by 'do($R = 1$)'
I still think $P(C = 1|\text{do}(R = 1)) \approx 1$, thus I expect **C**=1.

But what about
$P(T = 4\text{pm}|\text{do}(R = 1)) > P(T = 4\text{pm})$?
Does not make sense. Running the coffee machine tells me **nothing** about the time of the day.

$\Rightarrow$ this graph structure is wrong if I **act** on the variable R.

Acting on R: causal inference

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day.



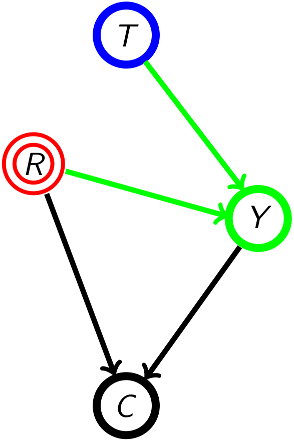
Assume: I set $R = 1$, i.e. I run the coffee machine
Expressed graphically by double circle, in formula by 'do($R = 1$)'
I still think $P(C = 1|\text{do}(R = 1)) \approx 1$, thus I expect **C**=1.

But what about
$P(T = 4\text{pm}|\text{do}(R = 1)) > P(T = 4\text{pm})$?
Does not make sense. Running the coffee machine tells me **nothing** about the time of the day.

$\Rightarrow$ this graph structure is wrong if **I act** on the variable R.

Acting on R: cutting the edge from T to R.

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day. Acting on R: do($R = 1$).



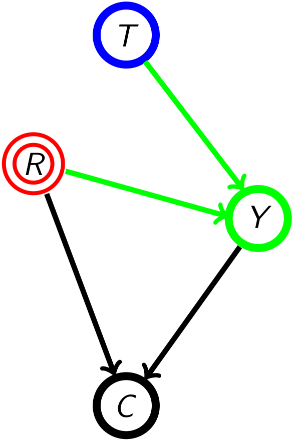
Assume: do($R = 1$)
Perhaps I should remove the edge from T to R ?
I still think $P(C = 1|\text{do}(R = 1)) \approx 1$, thus I expect **C**=1.

Now T and R are independent, if Y and C are unobserved.
$\Rightarrow P(T = 4\text{pm}|\text{do}(R = 1)) = P(T = 4\text{pm})$
This seems more sensible!

$$P(C, Y, R, T) = P(C|Y, R)P(Y|R, T)P(R)P(T)$$

Acting on R: cutting the edge from T to R.

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day. Acting on R: do$(R = 1)$.



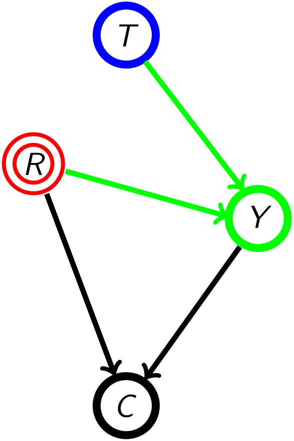
Assume: do$(R = 1)$
Perhaps I should remove the edge from T to R ?
I still think $P(C = 1|\text{do}(R = 1)) \approx 1$, thus I expect **C**=1.

Now T and R are independent, if Y and C are unobserved.
$\Rightarrow P(T = 4\text{pm}|\text{do}(R = 1)) = P(T = 4\text{pm})$
This seems more sensible!

$$P(C, Y, R, T) = P(C|Y, R)P(Y|R, T)P(R)P(T)$$

Acting on R: cutting the edge from T to R.

Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day. Acting on R: do($R = 1$).



Assume: do($R = 1$)
Perhaps I should remove the edge from T to R ?
I still think $P(C = 1|\text{do}(R = 1)) \approx 1$, thus I expect **C**=1.

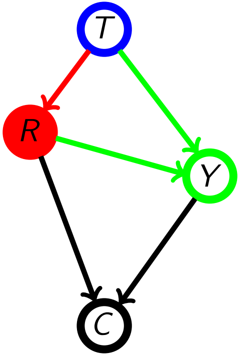
Now T and R are independent, if Y and C are unobserved.
$\Rightarrow P(T = 4\text{pm}|\text{do}(R = 1)) = P(T = 4\text{pm})$
This seems more sensible!

$$P(C, Y, R, T) = P(C|Y, R)P(Y|R, T)P(R)P(T)$$

Seeing versus doing

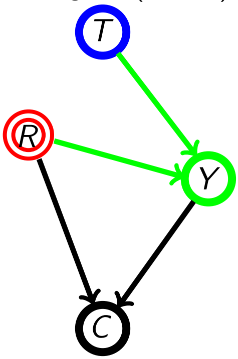
Reminder:
C: coffee in the pot, **Y**: yesterday's coffee is still there, **R**: coffee machine was recently run
T: time of day. Acting on R: $\mathrm{do}(R = 1)$.

Observing $R = 1$.



$P(C, Y, R, T) =$
$P(C|Y, R)P(Y|R, T)P(R|T)P(T)$

Acting: $\mathrm{do}(R = 1)$



$P(C, Y, R, T) =$
$P(C|Y, R)P(Y|R, T)P(R)P(T)$