COMP111: Artificial Intelligence

Section 9. Reasoning under Uncertainty 2

Frank Wolter

Content

- Random variables
- (Full) joint probability distribution
- Marginalization
- Probabilistic inference problem
- (Conditional) Independence and Belief Networks
- Expected Value
- Expected Value and Decision Making

Random variable

Let (S, P) be a probabality space. A random variable F is a function $F : S \to \mathbb{R}$ that assigns to every $s \in S$ a single number $F(s)$.

- Neither a variable nor random
- English translation of variabile casuale

We still assume that the sample space is finite. Thus, given a random variable F from some sample space S, the set of numbers r that are values of F is finite as well.

The event that F takes the value r, that is $\{s \mid F(s) = r\}$, is denoted $(F = r)$. The probability $(F = r)$ of the event $(F = r)$ is then

$$P(F = r) = P(\{s \mid F(s) = r\})$$

Example 1

Let

$$S = \{\text{car}, \text{train}, \text{plane}, \text{ship}\}$$

Then the function $F : S \to \mathbb{R}$ defined by

$$F(\text{car}) = 1, \quad F(\text{train}) = 1, \quad F(\text{plane}) = 2, \quad F(\text{ship}) = 2$$

is a random variable.

$(F = 1)$ denotes the event $\{s \in S \mid F(s) = 1\} = \{\text{car}, \text{train}\}$.

Define a uniform probability space (S, P) by setting

$$P(\text{car}) = P(\text{train}) = P(\text{plane}) = P(\text{ship}) = \frac{1}{4}$$

Then $P(F = 1) = P(\{s \in S \mid F(s) = 1\}) = P(\{\text{car}, \text{train}\}) = \frac{1}{2}$.

Example 2

Suppose that I roll two dice. So the sample space is

$$S = \{1, 2, 3, 4, 5, 6\}^2$$

and $P(ab) = \frac{1}{36}$ for every $ab \in S$.
Let

$$F(ab) = a + b.$$

F is a random variable. The probability that

$$F = r$$

for a number r (say, 12) is given by

$$P(F = r) = P(\{ab \mid F(ab) = r\})$$

For example, $P(F = 12) = P(\{ab \mid F(ab) = 12\}) = P(66) = \frac{1}{36}$.

Random Variable

When defining a probability distribution P for a random variable F, we often do not specify its sample space S but directly assign a probability to the event that F takes a certain value. Thus, we directly define the probability

$$P(F = r)$$

of the event that F has value r. Observe:

- $0 \leq P(F = r) \leq 1$;
- $\sum_{r \in \mathbb{R}} P(F = r) = 1$.

Thus, the events $(F = r)$ behave in the same way as outcomes of a random experiment.

Notation and Rules

- We write $\neg(F = r)$ for the event $\{s \mid F(s) \neq r\}$. For example, assume the random variable *Die* can take values $\{1, 2, 3, 4, 5, 6\}$ and

$$P(Die = n) = \frac{1}{6}$$

for all $n \in \{1, 2, 3, 4, 5, 6\}$ (thus we have a fair die). Then $\neg(Die = 1)$ denotes the event

$$(Die = 2) \text{ or } (Die = 3) \text{ or } (Die = 4) \text{ or } (Die = 5) \text{ or } (Die = 6)$$

We have the following complementation rule:

$$P(\neg(F = r)) = 1 - P(F = r)$$

- We write $(F_1 = r_1, F_2 = r_2)$ for '$(F_1 = r_1)$ and $(F_2 = r_2)$'.

Notation and Rules

- We write $(F_1 = r_1) \vee (F_2 = r_2)$ for '$(F_1 = r_1)$ or $(F_2 = r_2)$'. Then

$$\begin{aligned} P((F_1 = r_1) \vee (F_2 = r_2)) &= P(F_1 = r_1) + P(F_2 = r_2) \\ &\quad - P(F_1 = r_1, F_2 = r_2) \end{aligned}$$

- Conditional probability: if $P(F_2 = r_2) \neq 0$, then

$$P(F_1 = r_1 \mid F_2 = r_2) = \frac{P(F_1 = r_1, F_2 = r_2)}{P(F_2 = r_2)}$$

- Product rule:

$$P(F_1 = r_1, F_2 = r_2) = P(F_1 = r_1 \mid F_2 = r_2) \times P(F_2 = r_2)$$

Notation

We sometimes use symbols distinct from numbers to denote the values of random variables.
For example, for a random variable *Weather* rather than using values $1, 2, 3, 4$, we use

$$sunny, rain, cloudy, snow$$

Thus,

$$(Weather = sunny)$$

denotes the event that it is sunny.
To model a visit to a dentist, we use random variables *Toothache*, *Cavity*, and *Catch* (the dentist's steel probe catches in the tooth) that all take values 1 and 0 (for true and false).
For example, $(Toothache = 1)$ states that the person has toothache and $(Toothache = 0)$ states that the person does not have toothache.

Examples of probabilistic models

To model a domain using probability theory, one first introduces the relevant random variables. We have seen two basic examples:

- The weather domain could be modeled using the single random variable *Weather* with values

$$(sunny, rain, cloudy, snow)$$

- The dentist domain could be modeled using the random variables *Toothache*, *Cavity*, and *Catch* with values 0 and 1 for true and false.
  We might be interested in

$$P(Cavity = 1 \mid Toothache = 1, Catch = 1)$$

Student Exam Domain

A very basic model of the performance of students in an exam could be given by the random variables

- *Grade*: takes as values the possible grades of a student in the exam;
- *Answers*: takes as values the possible answers to exam questions;
- *Background*: takes as value the school visited before going to university;
- *Works_hard*: takes as values the degree to which the student works hard.

We might be interested in

$$P(Grade = A \mid Works\_hard = 1, Background = Comprehensive)$$

Fire Alarm Domain

A basic model of a fire alarm system and reporting about it could be given by the following random variables (all take value 0 or 1):

- *Fire*: there is fire;
- *Alarm*: the alarm goes off;
- *Tampering*: there is tampering with the alarm system;
- *Smoke*: there is smoke (no smoke detector used);
- *Leaving*: people leave the building;
- *Report*: it is reported that people leave the building (reporting not always correct).

We might be interested in

$$P(\mathit{Fire} = 1 \mid \mathit{Report} = 1)$$

Probability Distribution

- The probability distribution for a random variable gives the probabilities of all the possible values of the random variable.
- For example, let *Weather* be a random variable with values

$$(sunny, rain, cloudy, snow)$$

  such that its probability distribution is given by
  - $P(Weather = sunny) = 0.7$;
  - $P(Weather = rain) = 0.2$
  - $P(Weather = cloudy) = 0.08$;
  - $P(Weather = snow) = 0.02$.
- Assume the order of the values is fixed. Then we write instead

$$\mathbf{P}(Weather) = (0.7, 0.2, 0.08, 0.02)$$

  where the bold **P** indicates that the result is a vector of numbers representing the individual values of *Weather*.

More Probability Distributions

- Assume the random variable *Die* can take the values $1, 2, 3, 4, 5, 6$ and represents a fair die. Then we can define its probability distribution as

$$\mathbf{P}(Die) = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$$

- Recall the random variable $F(ab) = a + b$ from the sample space $S = \{1, 2, 3, 4, 5, 6\}^2$ with $P(ab) = \frac{1}{36}$ for all $a, b \in \{1, 2, 3, 4, 5, 6\}$. Then $2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ are its possible values. Then

$$\mathbf{P}(F) = (\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \ldots, \frac{1}{36})$$

Joint Probability Distribution

Let $F_1, \ldots, F_k$ be random variables. A joint probability distribution for

$$F_1, \ldots, F_k$$

gives the probabilities

$$P(F_1 = r_1, \ldots, F_k = r_k)$$

for the events

$$(F_1 = r_1) \text{ and } \cdots \text{ and } (F_k = r_k)$$

that F_1 takes value r_1, F_2 takes value r_2, and so on up to k, for all possible values $r_1, \ldots, r_k$.

The joint probability distribution is denoted $\mathbf{P}(F_1, \ldots, F_k)$.

Example

A possible joint probability distribution $\mathbf{P}(Weather, Cavity)$ for the random variables *Weather* and *Cavity* is given by the following table:

$Weather =$	*sunny*	*rain*	*cloudy*	*snow*
$Cavity = 1$	0.144	0.02	0.016	0.02
$Cavity = 0$	0.576	0.08	0.064	0.08

The probabilities of the joint distribution sum to 1!

Full Joint Probability Distribution

A full joint probability distribution

$$\mathbf{P}(F_1, \ldots, F_k)$$

is a joint probability distribution for all relevant random variables $F_1, \ldots, F_k$ for a domain of interest.

Every probability question about a domain can be answered by the full joint distribution because the probability of every event is a sum of probabilities

$$P(F_1 = r_1, \ldots, F_k = r_k)$$

(The $r_1, \ldots, r_k$ are often called data points or sample points.)

Example: Full Joint Probability Distribution for Dentist Domain

Assume the random variables *Toothache*, *Cavity*, *Catch* fully describe a visit to a dentist.
Then a full joint probability distribution is given by the following table:

	$Toothache = 1$		$Toothache = 0$	
	$Catch = 1$	$Catch = 0$	$Catch = 1$	$Catch = 0$
$Cavity = 1$	0.108	0.012	0.072	0.008
$Cavity = 0$	0.016	0.064	0.144	0.576

The probabilities of the joint distribution sum to 1!

Full Joint Probability Distributions

- The full joint probability distribution for the student exam domain, denoted

$$\mathbf{P}(Grade, Answers, Background, Works\_hard)$$

  gives the probability for every possible combination of values of the random variables *Grade*, *Answers*, *Background*, and *Works_hard*.
- The full joint probability distribution for the fire alarm domain gives the probability for every possible combination of values of the random variables *Fire*, *Alarm*, *Tampering*, *Smoke*, *Leaving*, and *Report*.

Marginalization

Given a joint distribution $\mathbf{P}(F_1, \ldots, F_k)$, one can compute the marginal probabilities of the random variables F_i by summing out the remaining variables.

For example,

$$P(Cavity = 1) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

is the sum of the entries in the first row of:

	$Toothache = 1$		$Toothache = 0$	
	$Catch = 1$	$Catch = 0$	$Catch = 1$	$Catch = 0$
$Cavity = 1$	0.108	0.012	0.072	0.008
$Cavity = 0$	0.016	0.064	0.144	0.576

Conditional Distributions

- We can also compute conditional distributions from the full joint distribution.
- We use the **P** notation for conditional distributions.
- $\mathbf{P}(F \mid G)$ gives the conditional distribution of F given G given by the probabilities $P(F = r \mid G = s)$ for all values r and s.
- Using **P** notation, the general version of the product rule is as follows:

$$\mathbf{P}(F, G) = \mathbf{P}(F \mid G)\mathbf{P}(G)$$

  stands for the list of equations:

$$\begin{aligned}
P(F = r_1, G = s_1) &= P(F = r_1 \mid G = s_1)P(G = s_1) \\
P(F = r_1, G = s_2) &= P(F = r_1 \mid G = s_2)P(G = s_2) \\
\cdots &= \cdots
\end{aligned}$$

Probabilistic Inference

Probabilistic inference can be characterized as the computation of posterior probabilities

$$\mathbf{P}(Q \mid E_1 = e_1, \ldots, E_n = e_n)$$

for query variables Q given observed evidence $e_1, \ldots, e_n$.
In principle, we can use the full joint distribution to do this:

	$Toothache = 1$		$Toothache = 0$	
	$Catch = 1$	$Catch = 0$	$Catch = 1$	$Catch = 0$
$Cavity = 1$	0.108	0.012	0.072	0.008
$Cavity = 0$	0.016	0.064	0.144	0.576

Example: $\mathbf{P}(Cavity \mid Toothache = 1)$

We want to compute the conditional probability distribution for $Cavity$ given the observation/evidence $Toothache = 1$.

Thus we want to compute:

- $P(Cavity = 1 \mid Toothache = 1)$ and
- $P(Cavity = 0 \mid Toothache = 1)$

We can easily obtain this using the table:

$$P(Cavity = 1 \mid Toothache = 1) = \frac{P(Cavity = 1, Toothache = 1)}{P(Toothache = 1)} = \frac{0.12}{0.2} = 0.6$$

$$P(Cavity = 0 \mid Toothache = 1) = \frac{P(Cavity = 0, Toothache = 1)}{P(Toothache = 1)} = \frac{0.08}{0.2} = 0.4$$

The denominator 0.2 can be viewed as a normalization constant $\frac{1}{\alpha} = 5$ for the distribution $\mathbf{P}(Cavity|Toothache = 1)$, ensuring that it adds up to 1.

Example: $\mathbf{P}(\mathit{Cavity} \mid \mathit{Toothache} = 1)$

Instead of

$$P(\mathit{Cavity} = 1 \mid \mathit{Toothache} = 1) = \frac{P(\mathit{Cavity} = 1, \mathit{Toothache} = 1)}{P(\mathit{Toothache} = 1)} = \frac{0.12}{0.2} = 0.6$$

$$P(\mathit{Cavity} = 0 \mid \mathit{Toothache} = 1) = \frac{P(\mathit{Cavity} = 0, \mathit{Toothache} = 1)}{P(\mathit{Toothache} = 1)} = \frac{0.08}{0.2} = 0.4$$

consider

$$\begin{aligned}
\mathbf{P}(\mathit{Cavity} \mid \mathit{Toothache} = 1) &= \alpha \mathbf{P}(\mathit{Cavity}, \mathit{Toothache} = 1) \\
&= \alpha(0.12, 0.08) \\
&= 5(0.12, 0.08) \\
&= (0.6, 0.4)
\end{aligned}$$

Combinatorial Explosion

This approach does not scale well: for a domain described by n random variables taking k distinct values each we face two problems:

- Writing up the full joint distribution requires $k^n - 1$ entries;
- How do we find the numbers (probabilities) for the entries?

For these reasons, the full joint distribution in tabular form is not a practical tool for building reasoning systems.

Independence to the Rescue

Random variables F and G are independent if

$$\mathbf{P}(F, G) = \mathbf{P}(F) \times \mathbf{P}(G),$$

that is, for all values r and s

$$P(F = r, G = s) = P(F = r) \times P(G = s)$$

As one's dental problems do not influence the weather, the pairs of random variables

- *Toothache*, *Weather*,
- *Catch*, *Weather*
- *Cavity*, *Weather*

are each independent.

Example: Weather and Dental Problems

The full joint probability distribution

$$\mathbf{P}(Toothache, Catch, Cavity, Weather)$$

has 32 entries. It contains four tables for the variables *Toothache*, *Catch*, *Cavity*. One for each kind of weather from

$$sunny, rain, cloudy, snow$$

Thus, we have

- 8 probabilities for $(Weather = sunny)$:

  $$P(Weather = sunny, Toothache = r_1, Catch = r_2, Cavity = r_3)$$

- 8 probabilities for $(Weather = rain)$:

  $$P(Weather = rain, Toothache = r_1, Catch = r_2, Cavity = r_3)$$

- and so on for $(Weather = cloudy)$ and $(Weather = snow)$.

What is the relationship between these editions?

Example: Weather and Dental Problems

Clearly we can make the independence assumption that for any combination of values of the random variables *Toothache*, *Catch*, *Cavity*, the the probabilities for *Weather*. For example,

$$\begin{aligned} & P(Weather = sunny | Toothache = r_1, Catch = r_2, Cavity = r_3) \\ = \; & P(Weather = sunny) \end{aligned}$$

for all $r_1, r_2, r_3 \in \{0, 1\}$.

Thus, equivalently,

$$\begin{aligned} & P(Weather = sunny, Toothache = r_1, Catch = r_2, Cavity = r_3) \\ = \; & P(Weather = sunny) P(Toothache = r_1, Catch = r_2, Cavity = r_3) \end{aligned}$$

for all $r_1, r_2, r_3 \in \{0, 1\}$.

The same equations hold for *rain*, *cloudy*, and *snow*.

Example: Weather and Dental Problems

We have seen that the joint probability distribution

$$\mathbf{P}(\mathit{Weather}, \mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})$$

can be written as:

$$\mathbf{P}(\mathit{Weather}) \times \mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})$$

The 32-element table for four variables can be constructed from one 4-element table and one 8-element table.

Independence Analysis

- Dentist domain: the variables *Toothache*, *Catch*, and *Cavity* are all dependent on each other.
- Student exam domain with variables *Grade*, *Answers*, *Background*, *Works_hard*:
  It seems reasonable to assume that *Background* and *Works_hard* are independent. No other pair is independent.
- Fire alarm domain: *Fire*, *Alarm*, *Tampering*, *Smoke*, *Leaving*, and *Report*.
  It seems reasonable to assume that *Tampering* and *Fire*, and *Tampering* and *Smoke* are independent. No other pair is independent.

We conclude that independence is rare. We will now turn to conditional independence.

Conditional Independence

Random variables G, F are conditionally independent given $H_1, \ldots, H_n$ if

$$\mathbf{P}(G, F \mid H_1, \ldots, H_n) = \mathbf{P}(G \mid H_1, \ldots, H_n) \times \mathbf{P}(F \mid H_1, \ldots, H_n)$$

or, equivalently,

$$\mathbf{P}(G \mid F, H_1, \ldots, H_n) = \mathbf{P}(G \mid H_1, \ldots, H_n)$$

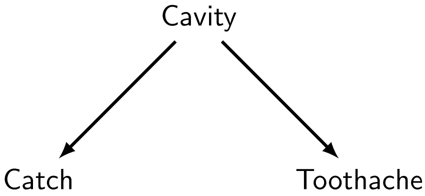
Example: Dentistry

In the dentist domain it seems reasonable to assert conditional independence of the variables *Toothache* and *Catch*, given *Cavity*:

$$\mathbf{P}(Toothache, Catch \mid Cavity) = \mathbf{P}(Toothache \mid Cavity)\mathbf{P}(Catch \mid Cavity)$$

or, equivalently,

$$\mathbf{P}(Toothache \mid Cavity) = \mathbf{P}(Toothache \mid Catch, Cavity)$$



How does this help?

Example: Dentistry

Using conditional independence of *Catch* and *Toothache* given *Cavity* we can compute the joint probability distribution

$$\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})$$

using only the probability distributions:

$$\mathbf{P}(\mathit{Toothache} \mid \mathit{Cavity}), \quad \mathbf{P}(\mathit{Catch} \mid \mathit{Cavity}), \quad \mathbf{P}(\mathit{Cavity})$$

The computation is as follows (using first multiplication rule and then conditional independence):

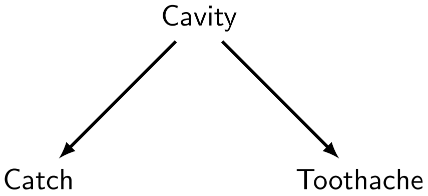
$$\begin{aligned}
& \mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity}) \\
= \; & \mathbf{P}(\mathit{Toothache}, \mathit{Catch} \mid \mathit{Cavity}) \times \mathbf{P}(\mathit{Cavity}) \\
= \; & \mathbf{P}(\mathit{Toothache} \mid \mathit{Cavity}) \times \mathbf{P}(\mathit{Catch} \mid \mathit{Cavity}) \times \mathbf{P}(\mathit{Cavity})
\end{aligned}$$

The number of probabilities needed is reduced to 5. Moreover, these probabilites can often be learned from data.

Towards Belief Networks

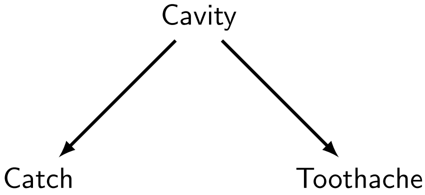
Conditional independence can be used to give concise representations of many domains.

A belief network (Bayesian network) is a graphical probabilistic model of a domain in which nodes represent random variables and arcs probabilistic dependence (often causality).



Towards Belief Networks

Informally, if there is an arc from a random variable F to another random variable G then G depends on F. F is called a parent of G. It is assumed that there are no cycles and that any random variable G is conditionally independ of any non-parent variable G' given the parents of G if G' cannot be reached by a sequence of arcs from G. For example:



The full joint probability distribution is then given as

$$\prod_{F \text{ in network}} \mathbf{P}(F \mid \text{parents}(F))$$

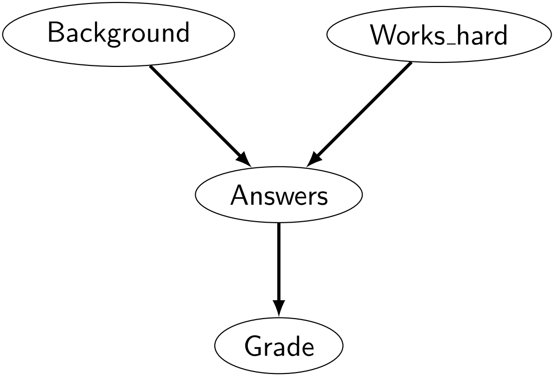
In the example $\mathbf{P}(\mathit{Toothache} \mid \mathit{Cavity}) \times \mathbf{P}(\mathit{Catch} \mid \mathit{Cavity}) \times \mathbf{P}(\mathit{Cavity})$.

Student Exam Domain

Variables: *Grade*, *Answers*, *Background*, *Works_hard*.
Then it seems reasonable to assume that

- *Works_hard* and *Background* are independent;
- *Grade* and *Works_hard* are independent given *Answers* and *Grade* and *Background* are independent given *Answers*.

We represent this modeling of the domain using the Belief Network:



Fire Alarm Domain

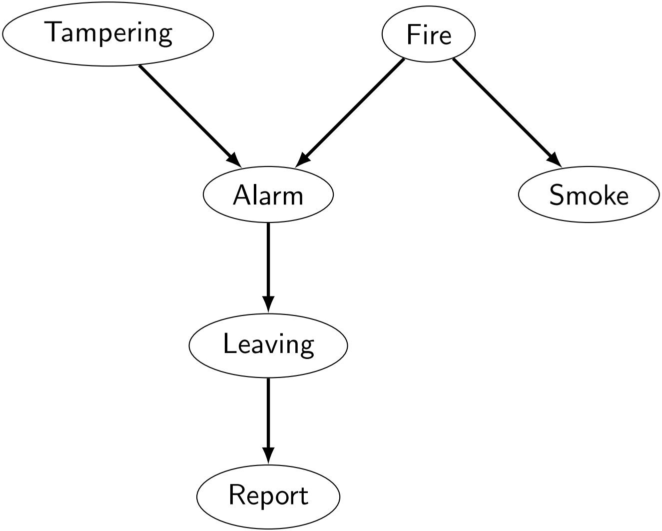
Recall the fire alarm system and reporting domain with random variables (all take value 0 or 1):

- *Fire*: there is fire;
- *Alarm*: the alarm goes off;
- *Tampering*: there is tampering with the alarm system;
- *Smoke*: there is smoke (no smoke detector used);
- *Leaving*: people leave the building;
- *Report*: it is reported that people leave the building (reporting not always correct).

Assume

- Fire is conditionally independent of Tampering;
- Alarm depends on Fire and Tampering;
- Smoke depends only on Fire and is conditionally independent of Tampering and Alarm given Fire;
- Leaving is conditionally independent of the other variables above given Alarm;
- Report only directly depends on Leaving.

Fire Alarm Domain



Joint Probability Distribution from $\mathbf{P}(F \mid \text{parents}(F))$

Given a Belief Network, we can always assume an ordering

$$F_1, \ldots, F_n$$

of its random variables such that for all i, j:

$$F_i \rightarrow F_j \quad \text{implies} \quad i < j$$

In our example, we can order the random variables as follows:

$$\mathit{Background}, \mathit{Works\_hard}, \mathit{Answers}, \mathit{Grade}$$

$$\mathit{Tampering}, \mathit{Fire}, \mathit{Alarm}, \mathit{Smoke}, \mathit{Leaving}, \mathit{Report}$$

Joint Probability Distribution from $\mathbf{P}(F \mid \text{parents}(F))$

According to the Chain Rule (proof on Exercise 8), given $F_1, \dots, F_n$, we have for all $r_1, \dots, r_n$:

$$\begin{aligned} P(F_1 = r_1, \dots, F_n = r_n) \quad = \quad & P(F_1 = r_1) \times \\ & P(F_2 = r_2 \mid F_1 = r_1) \times \\ & P(F_3 = r_3 \mid F_1 = r_1, F_2 = r_2) \times \\ & \dots \\ & P(F_n = r_n \mid F_1 = r_1, \dots, F_{n-1} = r_{n-1}) \end{aligned}$$

Using bold **P** notation this means:

$$\begin{aligned} \mathbf{P}(F_1, \dots, F_n) \quad = \quad & \mathbf{P}(F_1) \times \\ & \mathbf{P}(F_2 \mid F_1) \times \\ & \mathbf{P}(F_3 \mid F_1, F_2) \times \\ & \dots \\ & \mathbf{P}(F_n \mid F_1, \dots, F_{n-1}) \end{aligned}$$

Joint Probability Distribution from $\mathbf{P}(F \mid \text{parents}(F))$

$$
\begin{aligned}
\mathbf{P}(F_1, \ldots, F_n) \;\; = \;\; & \mathbf{P}(F_1) \times \\
& \mathbf{P}(F_2 \mid F_1) \times \\
& \mathbf{P}(F_3 \mid F_1, F_2) \times \\
& \ldots \\
& \mathbf{P}(F_n \mid F_1, \ldots, F_{n-1})
\end{aligned}
$$

As $\text{parents}(F_i) \subseteq \{F_1, \ldots, F_{i-1}\}$ conditional independence implies:

$$
\mathbf{P}(F_i \mid F_1, \ldots, F_{i-1}) = \mathbf{P}(F_i \mid \text{parents}(F_i))
$$

Thus:

$$
\begin{aligned}
\mathbf{P}(F_1, \ldots, F_n) \;\; = \;\; & \mathbf{P}(F_1) \times \\
& \mathbf{P}(F_2 \mid \text{parents}(F_2)) \times \\
& \mathbf{P}(F_3 \mid \text{parents}(F_3)) \times \\
& \ldots \\
& \mathbf{P}(F_n \mid \text{parents}(F_n))
\end{aligned}
$$

Example: Student Exam Domain

The full joint probability distribution

$$\mathbf{P}(Background, Works\_hard, Answers, Grade)$$

can then be computed as

$$\begin{aligned}
&\mathbf{P}(Background) \times \\
&\mathbf{P}(Works_hard) \times \\
&\mathbf{P}(Answers \mid Background, Works_hard) \times \\
&\mathbf{P}(Grade \mid Answers)
\end{aligned}$$

Example: Fire Alarm

The full joint probability distribution

$$\mathbf{P}(\mathit{Tampering}, \mathit{Fire}, \mathit{Alarm}, \mathit{Smoke}, \mathit{Leaving}, \mathit{Report})$$

can then be computed as

$$\begin{aligned}
&\mathbf{P}(\mathit{Tampering}) \times \\
&\mathbf{P}(\mathit{Fire}) \times \\
&\mathbf{P}(\mathit{Alarm} \mid \mathit{Tampering}, \mathit{Fire}) \times \\
&\mathbf{P}(\mathit{Smoke} \mid \mathit{Fire}) \times \\
&\mathbf{P}(\mathit{Leaving} \mid \mathit{Alarm}) \times \\
&\mathbf{P}(\mathit{Report} \mid \mathit{Leaving}) \times
\end{aligned}$$

The full joint probability table requires $2^6 - 1$ entries. Only 12 entries are needed for the conditional probabilities.

Fire Alarm Domain

Assume the following (conditional) probabilities (where we write $P(A \mid B)$ for $P(A = 1 \mid B = 1)$, $P(\neg A \mid B)$ for $P(A = 0 \mid B = 1)$ and so on):

- $P(\mathit{Tampering}) = 0.02$
- $P(\mathit{Fire}) = 0.01$
- $P(\mathit{Smoke} \mid \mathit{Fire}) = 0.9$
- $P(\mathit{Smoke} \mid \neg \mathit{Fire}) = 0.01$
- $P(\mathit{Alarm} \mid \mathit{Fire} \wedge \mathit{Tampering}) = 0.5$
- $P(\mathit{Alarm} \mid \mathit{Fire} \wedge \neg \mathit{Tampering}) = 0.99$
- $P(\mathit{Alarm} \mid \neg \mathit{Fire} \wedge \mathit{Tampering}) = 0.85$
- $P(\mathit{Alarm} \mid \neg \mathit{Fire} \wedge \neg \mathit{Tampering}) = 0.0001$
- $P(\mathit{Leaving} \mid \mathit{Alarm}) = 0.88$
- $P(\mathit{Leaving} \mid \neg \mathit{Alarm}) = 0.001$
- $P(\mathit{Report} \mid \mathit{Leaving}) = 0.75$
- $P(\mathit{Report} \mid \neg \mathit{Leaving}) = 0.01$

Querying

If Report is observed, then the probability of Fire and Tampering go up:

- $P(Fire) = 0.01$ and $P(Fire \mid Report) = 0.2305$
- $P(Tampering) = 0.02$ and $P(Tampering \mid Report) = 0.399$

If, in addition, Smoke is observed, then probability of Fire goes up further but Tampering goes down:

- $P(Fire \mid Report \wedge Smoke) = 0.964$
- $P(Tampering \mid Report \wedge Smoke) = 0.0284$

If, however, $\neg$Smoke is observed, then the probability of Fire goes down:

- $P(Fire \mid Report \wedge \neg Smoke) = 0.0294$
- $P(Tampering \mid Report \wedge \neg Smoke) = 0.501$

Summary Belief Networks

- Belief networks are a representation of conditional independence in probabilistic models.
- Querying can often be done using exact inference (with algorithmic tricks).
- Sometimes exact inference too hard. There are also approximate algorithms.
- Lots of research on learning belief networks from data. Either learning the conditional probabilities or even the structure of a belief network.

Expectation: motivation

- Consider a random variable F with values from the real numbers $x_1, \ldots, x_n$.
- Assume $P(F = x_i) = 1/n$ for all x_i.
- Then the expected value $E[F]$ of F is the average over the possible values $x_1, \ldots, x_n$ of F:

$$\frac{x_1 + \cdots + x_n}{n} = x_1 \frac{1}{n} + \cdots + x_n \frac{1}{n} = \sum_{x_i} x_i P(F = x_i)$$

- If the probabilities $P(F = x_i)$ are not all equal, we take the probability-weighted average.

Expectation

Assume that $x_1, \ldots, x_n$ are the values a random variable F can take. Then the expected value of F is defined as follows:

$$E[F] = x_1 P(F = x_1) + \cdots + x_n P(F = x_n) = \sum_x x P(F = x)$$

In other words, $E[F]$ is the probability-weighted average of all possible values of F.

$E[F]$ is sometimes called the expectation of F or the mean of F.

Example

Suppose you roll a fair die.

To model this take a random variable F with values in

$$\{1, 2, 3, 4, 5, 6\}$$

and set $P(F = x) = 1/6$ for all $x \in \{1, 2, 3, 4, 5, 6\}$.

The expected value of the random variable F is

$$\begin{aligned} E[F] &= \sum_x xP(F = x) \\ &= 1P(F = 1) + 2P(F = 2) + 3P(F = 3) \\ &\quad +4P(F = 4) + 5P(F = 5) + 6P(F = 6) \\ &= \tfrac{1}{6} + \tfrac{2}{6} + \tfrac{3}{6} + \tfrac{4}{6} + \tfrac{5}{6} + \tfrac{6}{6} \\ &= \tfrac{7}{2} \end{aligned}$$

Example

Suppose I pay, in p, the face value of a fair die if it comes up odd and earn the face value if it comes up even. What are my expectations?

- Consider the random variable F that takes values $\{-1, -3, -5, 2, 4, 6\}$ and $P(F = x) = 1/6$ for all $x \in \{-1, -3, -5, 2, 4, 6\}$.
- For example, $(F = -1)$ is the event that the face value is 1 and, thus, I pay $1p$.

The expected value of the random variable F is

$$\begin{aligned} E[F] &= \sum_x xP(F = x) \\ &= -P(F = -1) + 2P(F = 2) - 3P(F = -3) \\ &\quad +4P(F = 4) - 5P(F = -5) + 6P(F = 6) \\ &= -\tfrac{1}{6} + \tfrac{2}{6} - \tfrac{3}{6} + \tfrac{4}{6} - \tfrac{5}{6} + \tfrac{6}{6} \\ &= \tfrac{1}{2} \end{aligned}$$

Another way to compute $E[F]$

Let F be a random variable from the probability space (S, P). Then

$$E[F] = \sum_{s \in S} F(s)P(s)$$

Proof. Assume F takes the values $\{x_1, \ldots, x_n\}$. Then

$$\begin{aligned}
\sum_{s \in S} F(s)P(s) &= \sum_{F(s)=x_1} F(s)P(s) + \cdots + \sum_{F(s)=x_n} F(s)P(s) \\
&= x_1 \sum_{F(s)=x_1} P(s) + \cdots + x_n \sum_{F(s)=x_n} P(s) \\
&= x_1 P(F = x_1) + \cdots + x_n P(F = x_n) \\
&= E[F]
\end{aligned}$$

Linearity of Expectations

Let F and G be random variables from the same probability space (S, P) and let λ be a real number. Define new random variables $F + G$ and λF by setting:

$$(F + G)(s) = F(s) + G(s), \quad (\lambda F)(s) = \lambda F(s)$$

Then

$$E[F + G] = E[F] + E[G], \quad E[\lambda F] = \lambda E[F]$$

Proof. We prove the first claim using the reformulation of the expected value from the previous slide.

$$\begin{aligned}
E[F + G] &= \sum_{s \in S} (F + G)(s) P(s) \\
&= \sum_{s \in S} (F(s)P(s) + G(s)P(s)) \\
&= \sum_{s \in S} F(s)P(s) + \sum_{s \in S} G(s)P(s) \\
&= E[F] + E[G]
\end{aligned}$$

Example: Rolling two dice again

What is the expected value of the random variable F with $F(ab) = a + b$ from

$$S = \{1, 2, 3, 4, 5, 6\}^2$$

where $P(ab) = \frac{1}{36}$ for every $ab \in S$?

Note that $F(ab) = F_1(ab) + F_2(ab)$, where F_1 is the random variable from S with

$$F_1(ab) = a$$

and F_2 is the random variable from S with

$$F_2(ab) = b$$

We have

$$E(F) = E(F_1 + F_2) = E(F_1) + E(F_2) = \frac{7}{2} + \frac{7}{2} = 7$$

Expectation does not distribute over multiplication

If F and G are two random variables then, $E[F \times G]$ does not always equal $E[F] \times E[G]$.

Proof by counterexample. Consider the sample space (S, P) with

$$S = \{H, T\}, \quad P(H) = P(T) = \frac{1}{2}$$

Define random variables F_h and F_t by setting

$$F_h(H) = 1, \quad F_h(T) = 0, \quad F_t(H) = 0, \quad F_t(T) = 1$$

and define the random variable F by setting $F(s) = F_h(s) \times F_t(s)$.

Then $F(s) = F_h(s) \times F_t(s) = 0$ for all $s \in S$ and so $E[F] = 0$ but $E(F_h) \times E(F_t) = \frac{1}{4}$ since

$$E[F_h] = 1 \times \frac{1}{2} + 0 \times \frac{1}{2} = \frac{1}{2}, \quad E[F_t] = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{1}{2}$$

Expectation and Decision Making

Consider an agent who has to deliver a mail. There is a short way and a long way to deliver it. On the short way, it is more likely that the agent will have an accident. Suppose the agent can get protectors that will not change the probability of an accident but will make one less severe. The protectors are expensive, however. Going the long way reduces the probability of an accident, but takes much longer.

Problem: the agent has to decide whether to wear protectors and which way to go.

We model this situation using

- Two decision variables, *Which_way* and *Protector*. The agent gets to choose the value for each decision variable.
- A random variable, *Accident*, which represents whether an accident happens.
- A utility function that gives the utility of every possible outcome.

Conditional Probability Distribution

Recall that the probability of an accident only depends on whether the long or short way is chosen but not on whether the agent is wearing protectors. Thus, all we need is the conditional probability distribution

$$\mathbf{P}(\mathit{Accident} \mid \mathit{Which\_way})$$

Assume this is given by

- $P(\mathit{Accident} = 1 \mid \mathit{Which\_way} = \mathit{short}) = 0.2$
- $P(\mathit{Accident} = 0 \mid \mathit{Which\_way} = \mathit{short}) = 0.8$
- $P(\mathit{Accident} = 1 \mid \mathit{Which\_way} = \mathit{long}) = 0.01$
- $P(\mathit{Accident} = 0 \mid \mathit{Which\_way} = \mathit{long}) = 0.99$

Utility

The utility of the outcome depends on whether a short or long way is chosen, whether the agent wears protectors, and whether the agent has an accident. Assume the utility values are as follows:

Protector	*Which_way*	*Accident*	*Utility*
true	*short*	*true*	35
true	*short*	*false*	95
true	*long*	*true*	30
true	*long*	*false*	75
false	*short*	*true*	3
false	*short*	*false*	100
false	*long*	*true*	0
false	*long*	*false*	80

What to do? Maximize Expected Utility!

Choose *Protector* and *Which_way* in such a way that the expected utility is maximal: choose values r, s such that

$$E[Utility \mid Protector = r, Which\_way = s]$$

is maximal.

We compute $E[Utility \mid Protector = 1, Which\_way = short]$ by taking the sum of

- the value 35 of the utility of $(Protector \wedge short \wedge Accident)$ multipied by $P(Accident = 1 \mid Which\_way = short)$;
- the value 95 of the utility of $(Protector \wedge short \wedge \neg Accident)$ multiplied by $P(Accident = 0 \mid Which\_way = short)$.

We obtain:

$$\begin{aligned} 83 \quad = \quad & 35 \times P(Accident = 1 \mid Which\_way = short) + \\ & 95 \times P(Accident = 0 \mid Which\_way = short) \end{aligned}$$

Maximize Expected Utility

$E[Utility \mid Protector = 1, Which\_way = long]$ equals

$$\begin{aligned} 74.55 &= 30 \times P(Accident = 1 \mid Which\_way = long) + \\ & \quad 75 \times P(Accident = 0 \mid Which\_way = long) \end{aligned}$$

$E[Utility \mid Protector = 0, Which\_way = short]$ equals

$$\begin{aligned} 80.6 &= 3 \times P(Accident = 1 \mid Which\_way = short) + \\ & \quad 100 \times P(Accident = 0 \mid Which\_way = short) \end{aligned}$$

$E[Utility \mid Protector = 0, Which\_way = long]$ equals

$$\begin{aligned} 79.2 &= 0 \times P(Accident = 1 \mid Which\_way = long) + \\ & \quad 80 \times P(Accident = 0 \mid Which\_way = long) \end{aligned}$$

As $E[Utility \mid Protector = 1, Which\_way = short]$ is maximal, a rational agent should wear protectors and take the short way.

Journey to Manchester Airport

Recall the following example from the introduction: When going to the airport by car, how early should I start? 45 minutes should be enough from Liverpool to Manchester Airport, but only under the assumption that there are no accidents, no lane closures, that my car does not break down, and so on. This uncertainty is hard to eliminate, but still an agent has to make a decision.

Within the framework of "maximizing expected utility" this problem can be modeled as follows.

Journey to Manchester Airport

Consider a random variable Arrival taking values:

- *miss_plane*,
- *wait_0_minutes*, *wait_5_minutes*, *wait_10_minutes*, and so on

Consider the decision variable Start taking values:

- *45minutes_early*, *50minutes_early*, and so on.

Assume from experience/data we have a probability distribution

$$\mathbf{P}[Arrival \mid Start]$$

Finally, consider utility values associated with the values of the random variables Arrival: *miss_plane* has very low utility, *wait_0_minutes* high utilility, and so on.

Then the agent should choose the value r for Start such that the expected utility

$$E[Utility \mid Start = r]$$

is maximal.