Subjective probability and utility

Christos Dimitrakakis

April 29, 2014

Goals of today's (?) lecture

Subjective probability

- Understand the different interpretations of probability.
- Refresh the mathematical properties of probability.
- Understand how to use probability to represent your beliefs.
- Show why probability is the right thing for this job.
- See how you can update your beliefs using probability.

Utility

- Understand the concept of preferences.
- See how utility can be used to formalize preferences.
- Show how we can combine utility and probability to deal with decision making under uncertainty.

The decision-theoretic foundations of artificial intelligence.

- Probability: how likely things are?
- Utility: which things do we want?

Interpretations of probability

- Objective: inherent randomness.
- Frequentist: long-term averages.
- Algorithmic: program complexity.
- Subjective: uncertainty.

Interpretations of utility

- Monetary.
- Psychological.
- “true” value of things?

Objective Probability

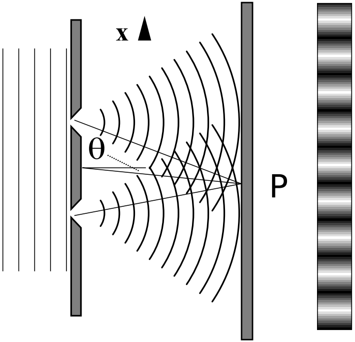


Figure : The double slit experiment

Objective Probability

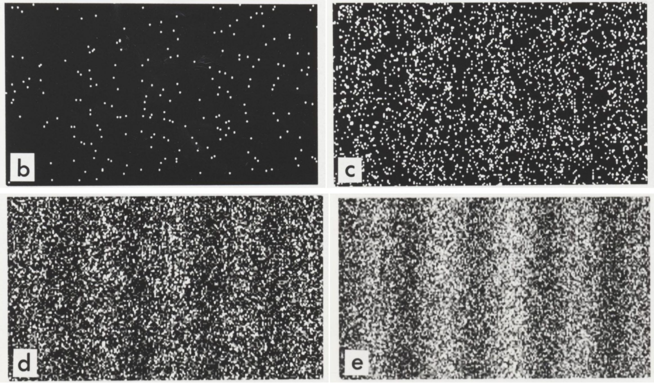


Figure : The double slit experiment

Objective Probability

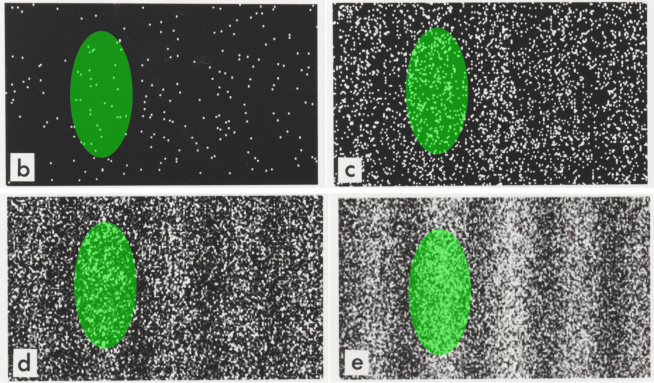


Figure : The double slit experiment

Algorithmic probability

- Consider a binary string $x = 101010001011101001010010101$.

Algorithmic probability

- Consider a binary string $x = 101010001011101001010010101$.
- Consider another string $y = 111111111111111111111111111$.

Algorithmic probability

- Consider a binary string $x = 101010001011101001010010101$.
- Consider another string $y = 111111111111111111111111111$.
- Intuitively, do you think that
  - A x is more likely than y.
  - B x is as likely as y.
  - C x is less likely than y.
  - D The question is meaningless.

  m.socrative.com – `ai-chalmers-2014`

Algorithmic probability

- Consider a binary string $x = 10101000101110100101001010 1$.
- Consider another string $y = 11111111111111111111111111111$.
- Intuitively, do you think that
  - A x is more likely than y.
  - B x is as likely as y.
  - C x is less likely than y.
  - D The question is meaningless.

  m.socrative.com – `ai-chalmers-2014`
- Intuitively, y is "simpler"… perhaps it's generated by an algorithm! But which algorithm?

Algorithmic probability

- Consider a binary string $x = 1010100010111010010100101010$.
- Consider another string $y = 11111111111111111111111111111$.
- Intuitively, do you think that
  - A x is more likely than y.
  - B x is as likely as y.
  - C x is less likely than y.
  - D The question is meaningless.

  m.socrative.com – `ai-chalmers-2014`
- Intuitively, y is "simpler"... perhaps it's generated by an algorithm! But which algorithm?

Solomonoff induction

- Occam's razor: Prefer the simplest explanation (algorithm).
- Epicurus: Do not throw away any hypothesis (algorithm).
- Weigh algorithms according to
  - Simplicity.
  - How well they fit the data.

What about everyday life?

Subjective probability

- Making decisions requires making predictions.

Subjective probability

- Making decisions requires making predictions.
- Outcomes of decisions are uncertain.

Subjective probability

- Making decisions requires making predictions.
- Outcomes of decisions are uncertain.
- How can we represent this uncertainty?

Subjective probability

- Making decisions requires making predictions.
- Outcomes of decisions are uncertain.
- How can we represent this uncertainty?

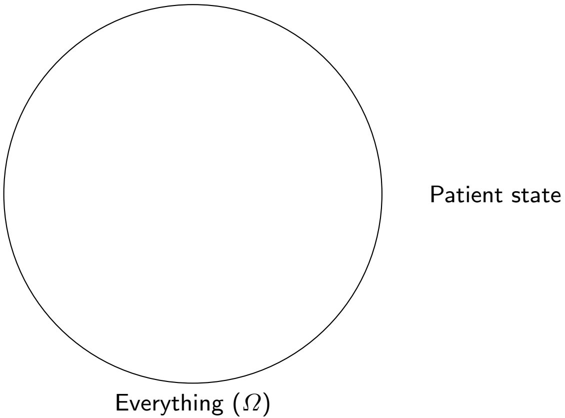
Subjective probability

- Describe which events we think are more likely.
- We quantify this with probability.

Why probability?

- Quantifies uncertainty in a "natural" way.
- A framework for drawing conclusions from data.
- Computationally convenient for decision making.

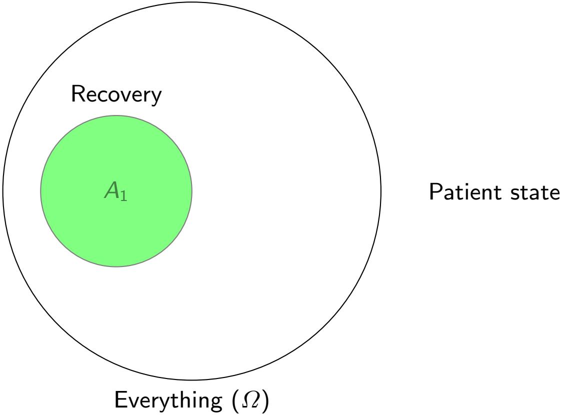
Events as sets



Example 1 (Experiment: give medication to a patient.)

- Does the patient recover?
- Does the medication have side-effects?

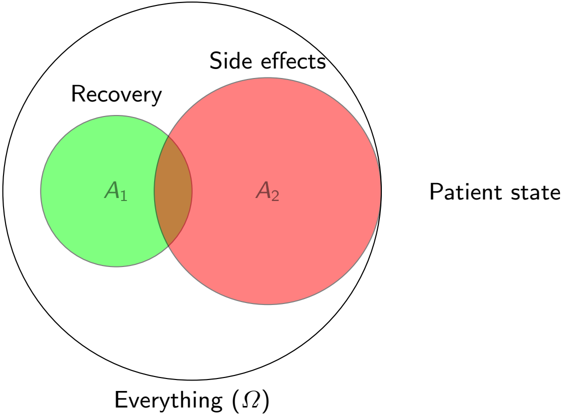
Events as sets



Example 1 (Experiment: give medication to a patient.)

- Does the patient recover?
- Does the medication have side-effects?

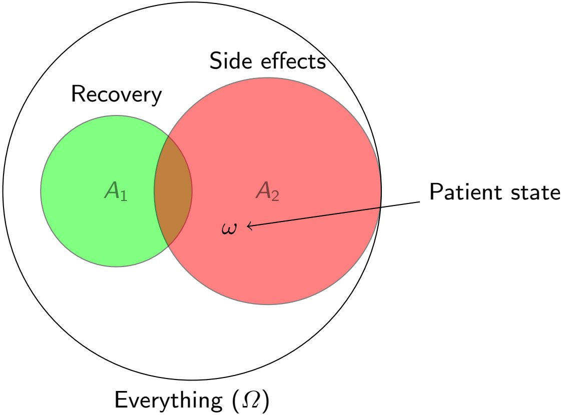
Events as sets



Example 1 (Experiment: give medication to a patient.)

- Does the patient recover?
- Does the medication have side-effects?

Events as sets



Example 1 (Experiment: give medication to a patient.)

- Does the patient recover?
- Does the medication have side-effects?

How likely are events

The relative likelihood of two events A and B

- Do you think A is more likely than B? Write $A \succ B$.
- Do you think A is less likely than B? Write $A \prec B$.
- Do you think A is as likely as B? Write $A \approx B$.

We also use $\succsim$ and $\precsim$ for at least as likely as and for no more likely than.

How likely are events

The relative likelihood of two events A and B

- Do you think A is more likely than B? Write $A \succ B$.
- Do you think A is less likely than B? Write $A \prec B$.
- Do you think A is as likely as B? Write $A \eqsim B$.

We also use $\succsim$ and $\precsim$ for at least as likely as and for no more likely than.

Functions on sets

A function P is said to agree with a relation $A \precsim B$, if it has the property that: $P(A) \leq P(B)$ if and only if $A \precsim B$.

How likely are events

The relative likelihood of two events A and B

- Do you think A is more likely than B? Write $A \succ B$.
- Do you think A is less likely than B? Write $A \prec B$.
- Do you think A is as likely as B? Write $A \approx B$.

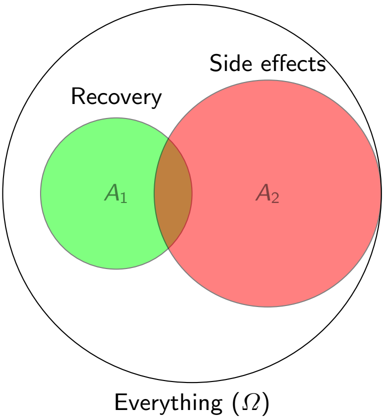
We also use $\succsim$ and $\precsim$ for at least as likely as and for no more likely than.

Functions on sets

A function P is said to agree with a relation $A \precsim B$, if it has the property that: $P(A) \leq P(B)$ if and only if $A \precsim B$.

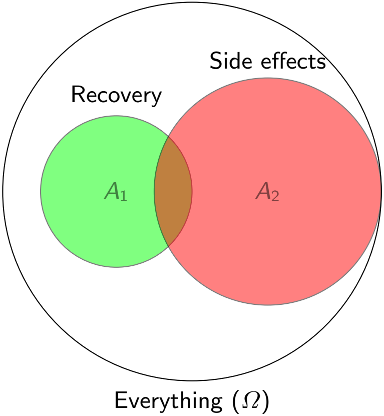
We want such a function for all events of interest.

Which events should we look at?



We wish to look at all combinations of events which are relevant. So, if we want to calculate the probability of recovery, and the probability of side effects, we must also be able to calculate the probability of recovery or side-effects, as well as the probability of no recovery. This is formally captured by the notion of a σ-field.

Which events should we look at?

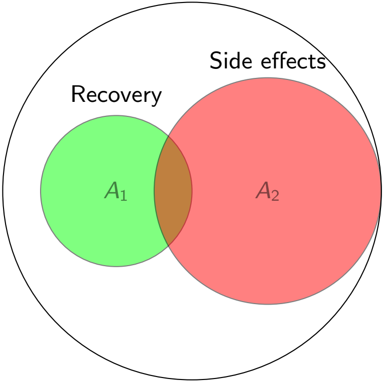


Definition 2 (σ-field on Ω)

A family $\mathcal{F}$ of sets, s.t. $\forall A \in \mathcal{F}$, $A \subset \Omega$, is called a σ-field on Ω if and only if

1. $\Omega \in \mathcal{F}$
2. if $A \in \mathcal{F}$, then $A^{\complement} \in \mathcal{F}$.
3. If $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Which events should we look at?



Definition 2 (σ-field on Ω)

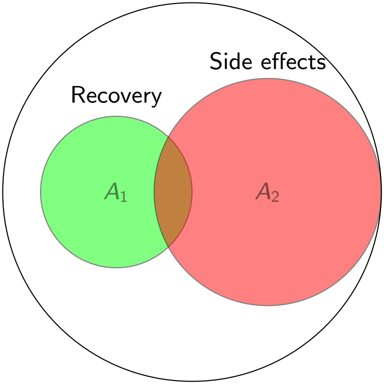
A family $\mathcal{F}$ of sets, s.t. $\forall A \in \mathcal{F}$, $A \subset \Omega$, is called a σ-field on Ω if and only if

1. $\Omega \in \mathcal{F}$
2. if $A \in \mathcal{F}$, then $A^{\complement} \in \mathcal{F}$.
3. If $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Exercise 1

Is $\mathcal{F} = \left\{ \emptyset, A_1, A_1^{\complement}, \Omega \right\}$ *a* σ*-field?*

Which events should we look at?



Definition 2 (σ-field on Ω)

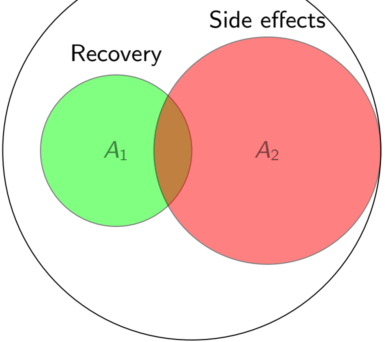
A family $\mathcal{F}$ of sets, s.t. $\forall A \in \mathcal{F}$, $A \subset \Omega$, is called a σ-field on Ω if and only if

1. $\Omega \in \mathcal{F}$
2. if $A \in \mathcal{F}$, then $A^{\complement} \in \mathcal{F}$.
3. If $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Exercise 1

Is $\mathcal{F} = \{\emptyset, A_1, A_2, \Omega\}$ a σ-field?

Which events should we look at?



Definition 2 (σ-field on Ω)

A family $\mathcal{F}$ of sets, s.t. $\forall A \in \mathcal{F}$, $A \subset \Omega$, is called a σ-field on Ω if and only if

1. $\Omega \in \mathcal{F}$
2. if $A \in \mathcal{F}$, then $A^{\complement} \in \mathcal{F}$.
3. If $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Example 3

The σ-field generated by $\{\emptyset, A_1, A_2, \Omega\}$ is:

$$\mathcal{F} = \{A_1, A_1^{\complement}, A_2, A_2^{\complement}, A_1 \cap A_2, (A_1 \cap A_2)^{\complement}, A_1 \cup A_2, (A_1 \cup A_2)^{\complement}, A_2, A_2 \setminus A_1, A_1 \setminus A_2, (A_2 \setminus A_1)^{\complement}, (A_1 \setminus A_2)^{\complement}, \emptyset, \Omega\}.$$

Subjective probability assumptions I

Our beliefs must be consistent. This can be achieved if they satisfy some assumptions:

Subjective probability assumptions I

Our beliefs must be consistent. This can be achieved if they satisfy some assumptions:

Assumption 1 (SP1)

For any events A, B, one of the following must hold: $A \succ B$, $A \prec B$, $A \approx B$.

It is always possible to say whether one event is more likely than the other.

Subjective probability assumptions I

Our beliefs must be consistent. This can be achieved if they satisfy some assumptions:

Assumption 1 (SP1)

For any events A, B*, one of the following must hold:* $A \succ B$, $A \prec B$, $A \eqsim B$.

Assumption 2 (SP2)

Let $A = A_1 \cup A_2$, $B = B_1 \cup B_2$ *with* $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$*. If* $A_i \precsim B_i$ *then* $A \precsim B$.

If we can split A, B in such a way that each part of A is less likely than its counterpart in B, then A is less likely than B.

Subjective probability assumptions I

Our beliefs must be consistent. This can be achieved if they satisfy some assumptions:

Assumption 1 (SP1)

For any events A, B, one of the following must hold: $A \succ B$, $A \prec B$, $A \eqsim B$.

Assumption 2 (SP2)

Let $A = A_1 \cup A_2$, $B = B_1 \cup B_2$ with $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$. If $A_i \precsim B_i$ then $A \precsim B$.

Assumption 3 (SP3)

For any event A, we have: $\emptyset \precsim A$ For the certain event Ω, we have: $\emptyset \prec \Omega$.

Resulting properties of relative likelihoods

Theorem 4 (Transitivity)

If A, B, D such that $A \precsim B$ and $B \precsim D$, then $A \precsim D$.

Theorem 5 (Complement)

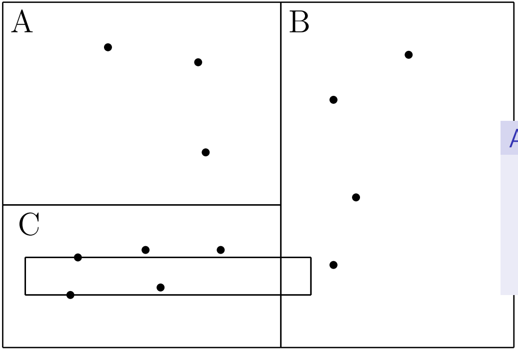
For any A, B: $A \precsim B$ iff $A^{\complement} \succsim B^{\complement}$.

Theorem 6 (Fundamental property of relative likelihoods)

If $A \subset B$ then $A \precsim B$. Furthermore, $\emptyset \precsim A \precsim S$ for any event A.

What functions can agree with a relative likelihood?

- For any events $P(A) > P(B)$, $P(A) < P(B)$ or $P(A) = P(B)$.
- If A_i, B_i are disjoint sets, $\forall i : P(A_i) \leq P(B_i) \Rightarrow P(A) \leq P(B)$.
- For any A, $P(\emptyset) \leq P(A)$ and $P(\emptyset) < P(\Omega)$.



Area

- A: $4 \times 5 = 20m^2$.
- B: $6 \times 4 = 24m^2$.
- C: $2 \times 5 = 10m^2$.

Figure : A fashionable apartment

Measure the sets: $\mathcal{F} = \{\emptyset, A, B, C, A \cup B, A \cup C, B \cup C, A \cup B \cup C\}$.
Note that all those measures have an additive property.

Measure theory primer

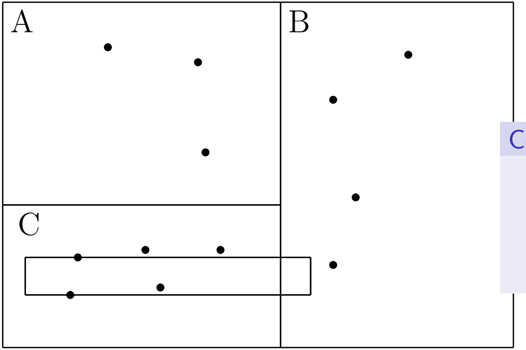


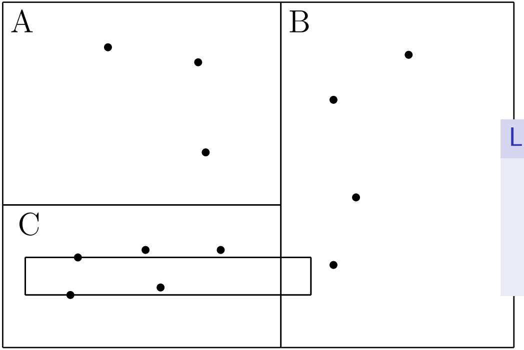
Figure : A fashionable apartment

Coins on the floor

- A: 3.
- B: 4
- C: 5.

Measure the sets: $\mathcal{F} = \{\emptyset, A, B, C, A \cup B, A \cup C, B \cup C, A \cup B \cup C\}$.
Note that all those measures have an additive property.

Measure theory primer



Length of red carpet

- A: $0m$
- B: $0.5m$
- C: $4.5m$.

Figure : A fashionable apartment

Measure the sets: $\mathcal{F} = \{\emptyset, A, B, C, A \cup B, A \cup C, B \cup C, A \cup B \cup C\}$.
Note that all those measures have an additive property.

Definition 7 (Measure)

A measure λ on $(\Omega, \mathcal{F})$ is a function $\lambda : \mathcal{F} \to \mathbb{R}^+$ such that

1. $\lambda(\emptyset) = 0$.
2. $\lambda(A) \geq 0$ for any $A \in \mathcal{F}$.
3. For any collection of subsets $A_1, A_2, \ldots$ with $A_i \in \mathcal{F}$ and $A_i \cap A_j = \emptyset$.

$$\lambda\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \lambda(A_i) \tag{2.1}$$

Measure and probability

Definition 7 (Probability measure)

A probability measure P on $(\Omega, \mathcal{F})$ is a function $P : \mathcal{F} \to [0, 1]$ such that:

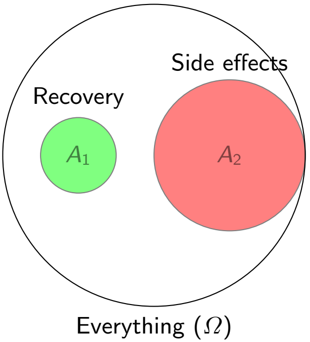
1. $P(\Omega) = 1$
2. $P(\emptyset) = 0$
3. $P(A) \geq 0$ for any $A \in \mathcal{F}$.
4. If $A_1, A_2, \ldots$ are disjoint then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \qquad \text{(union)}$$

$(S, \mathcal{F}, P)$ is called a *probability space*.

So, probability is just a special type of measure.

Logical interpretation: Mutually exclusive and independent events

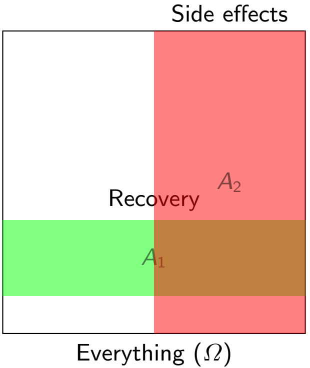


Definition 8 (Mutually exclusive events)

If A, B are disjoint (i.e. $A \cap B = \emptyset$) then they are *mutually exclusive*. Since P is a measure,

$$P(A \cup B) = P(A) + P(B).$$

Logical interpretation: Mutually exclusive and independent events



Definition 8 (Independent events)

Events A, B are independent iff

$$P(A \cap B) = P(A)P(B). \tag{2.1}$$

Thus, the probability of either A occuring does not depend on whether B occurs.

Logical interpretation: Mutually exclusive and independent events

Definition 8 (Mutually exclusive events)

If A, B are disjoint (i.e. $A \cap B = \emptyset$) then they are *mutually exclusive*. Since P is a measure,

$$P(A \cup B) = P(A) + P(B).$$

Definition 9 (Independent events)

Events A, B are independent iff

$$P(A \cap B) = P(A)P(B). \tag{2.1}$$

Thus, the probability of either A occuring does not depend on whether B occurs.

Exercise 1

Can mutually exclusive events be independent?

You can think of $A \cap B$ as $A \wedge B$, i.e. "A and B".
You can think of $A \cup B$ as $A \vee B$, i.e. "A or B".

A probability measure can satisfy our assumptions

Exercise 2

(i) *For any events* $P(A) > P(B)$, $P(A) < P(B)$ *or* $P(A) = P(B)$.

(ii) *If* A_i, B_i *are partitions of* A, B, $\forall i P(A_i) \leq P(B_i) \Rightarrow P(A) \leq P(B)$.

(iii) *For any* A, $P(\emptyset) \leq P(A)$ *and* $P(\emptyset) < P(\Omega)$

From events to variables

Let $\omega \sim P$ denote that ω is selected according to P.

Events as indicator functions

Until now we were just considering simple events: where $\omega \in A$. Each event A can be seen as a function $\mathbb{1}_A : \Omega \to \{0, 1\}$

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \text{otherwise} \end{cases}$$

Then the probability that $\omega \in A$ is simply $P(A)$.

Definition 10 (Random variable)

However, we can also define some arbitrary other function $x : \Omega \to \mathbb{R}$. This function is called a random variable, because it is a variable whose value depends on the random outcome ω.

Example 11 (Functions of the patient state)

Temperature, blood pressure, heart rate, ...

Probabilities and expectations of random variables

Given a random variable $x : \Omega \to \mathbb{R}$, we can naturally ask things such as what value x takes on average:

Definition 12 (Expectation of a random variable)

If $\omega \sim P$, then:

$$\mathbb{E}_P(x) \triangleq \sum_{\omega \in \Omega} x(\omega) P(\omega) \qquad \text{(discrete case)}$$

(general case)

(For the discrete case, it is usual to write $P(\omega)$ to mean $P(\{\omega\})$).

Probabilities and expectations of random variables

Given a random variable $x : \Omega \to \mathbb{R}$, we can naturally ask things such as what value x takes on average:

Definition 12 (Expectation of a random variable)

If $\omega \sim P$, then:

$$\mathbb{E}_P(x) \triangleq \sum_{\omega \in \Omega} x(\omega) P(\omega) \qquad \text{(discrete case)}$$

$$\mathbb{E}_P(x) \triangleq \int_{\Omega} x(\omega)\, \mathrm{d}P(\omega) \qquad \text{(general case)}$$

(For the discrete case, it is usual to write $P(\omega)$ to mean $P(\{\omega\})$).

Probabilities and expectations of random variables

Given a random variable $x : \Omega \to \mathbb{R}$, we can naturally ask things such as what value x takes on average:

Definition 12 (Expectation of a random variable)

If $\omega \sim P$, then:

$$\mathbb{E}_P(x) \triangleq \sum_{\omega \in \Omega} x(\omega) P(\omega) \qquad \text{(discrete case)}$$

$$\qquad \text{(general case)}$$

(For the discrete case, it is usual to write $P(\omega)$ to mean $P(\{\omega\})$).

Definition 13 (Distribution of a random variable)

If $\omega \sim P$, then $x \sim P_x$ with:

$$P_x(A) \triangleq \sum_{\omega \in \Omega} \mathbb{1}_A(x(\omega)) P(\omega) \qquad \text{(discrete case)}$$

Recap of fundamental probability

- Subjective probability can be used to represent uncertainty.
- Events can be represented as sets in a space of outcomes Ω.
- The set of all possible event combinations $\mathcal{F}$ is a σ-field in Ω.
- The relative likelihood between events $A, B \subset \Omega$ is our subjective belief of which one is more likely.
- If we think A is more likely than B, we write $A \succ B$.
- The likelihood relation can be captured via probabilities:

$$P(A) > P(B) \Leftrightarrow A \succ B.$$

- Probabilities are measures, e.g. similar to *area, length, mass*, etc.
- Mutually exclusive events are disjoint.
- Independent events have product joint probability.
- Random variables are simply functions on outcomes.
- The expectation of a r.v. is the sum of its values for each outcome, weighed by the outcome's probability.

Conditional likelihood

- A likelihood relation encodes our prior opinions.
- Sometimes we need to take into account evidence.
- For example, ordinarily we may think that $A \precsim B$.
- However, we may have additional information D ...

Example 14

Conditional likelihood

- A likelihood relation encodes our prior opinions.
- Sometimes we need to take into account evidence.
- For example, ordinarily we may think that $A \precsim B$.
- However, we may have additional information D ...

Example 14

- Say that A is the event that it rains in Gothenburg tomorrow.

Conditional likelihood

- A likelihood relation encodes our prior opinions.
- Sometimes we need to take into account evidence.
- For example, ordinarily we may think that $A \precsim B$.
- However, we may have additional information D ...

Example 14

- Say that A is the event that it rains in Gothenburg tomorrow.
- Clearly, $A \succsim A^{\complement}$.

Conditional likelihood

- A likelihood relation encodes our prior opinions.
- Sometimes we need to take into account evidence.
- For example, ordinarily we may think that $A \precsim B$.
- However, we may have additional information $D \ldots$

Example 14

- Say that A is the event that it rains in Gothenburg tomorrow.
- Clearly, $A \succsim A^{\complement}$.
- Let D denote a good forecast!

Conditional likelihood

- A likelihood relation encodes our prior opinions.
- Sometimes we need to take into account evidence.
- For example, ordinarily we may think that $A \precsim B$.
- However, we may have additional information D ...

Example 14

- Say that A is the event that it rains in Gothenburg tomorrow.
- Clearly, $A \succsim A^{\complement}$.
- Let D denote a good forecast!
- I personally believe that $(A \mid D) \precsim (A^{\complement} \mid D)$.

Conditional likelihoods

Assumption 4 (CP)

For any events A, B, D,

$$(A \mid D) \precsim (B \mid D) \quad \textit{iff} \quad A \cap D \precsim B \cap D.$$

Theorem 15

If a relation $\precsim$ *satisfies assumptions SP1 to SP5 and CP, then* P *is the unique probability distribution such that:*
For any A, B, D *such that* $P(D) > 0$,

$$(A \mid D) \precsim (B \mid D) \quad \textit{iff} \quad P(A \mid D) \leq P(B \mid D)$$

Definition 16 (Conditional probability)

$$P(A \mid D) \triangleq \frac{P(A \cap D)}{P(D)} \tag{2.2}$$

A simple exercise in updating beliefs

Forecaster	Saturday	Sunday	Monday	Tuesday
A	Rain	Rain	Rain	Rain
B	Sun	Rain	Rain	Sun
C	Clouds	Clouds	Rain	Storms
D	Sun	Clouds	Rain	Clouds
E	Clouds	Rain	Clouds	Sun
Outcome				

Table : Five weather forecasters

A simple exercise in updating beliefs

Forecaster	Saturday	Sunday	Monday	Tuesday
A	Rain	Rain	Rain	Rain
B	Sun	Rain	Rain	Sun
C	Clouds	Clouds	Rain	Storms
D	Sun	Clouds	Rain	Clouds
E	Clouds	Rain	Clouds	Sun
Outcome	Clouds			

Table : Five weather forecasters

A simple exercise in updating beliefs

Forecaster	Saturday	Sunday	Monday	Tuesday
A	Rain	Rain	Rain	Rain
B	Sun	Rain	Rain	Sun
C	Clouds	Clouds	Rain	Storms
D	Sun	Clouds	Rain	Clouds
E	Clouds	Rain	Clouds	Sun
Outcome	Clouds	Rain		

Table : Five weather forecasters

A simple exercise in updating beliefs

Forecaster	Saturday	Sunday	Monday	Tuesday
A	Rain	Rain	Rain	Rain
B	Sun	Rain	Rain	Sun
C	Clouds	Clouds	Rain	Storms
D	Sun	Clouds	Rain	Clouds
E	Clouds	Rain	Clouds	Sun
Outcome	Clouds	Rain	Rain	

Table : Five weather forecasters

A simple exercise in updating beliefs

Forecaster	Saturday	Sunday	Monday	Tuesday
A	Rain	Rain	Rain	Rain
B	Sun	Rain	Rain	Sun
C	Clouds	Clouds	Rain	Storms
D	Sun	Clouds	Rain	Clouds
E	Clouds	Rain	Clouds	Sun
Outcome	Clouds	Rain	Rain	Sun

Table : Five weather forecasters

Updating beliefs

Theorem 17 (Bayes' theorem)

Let $A_1, A_2, \ldots$ be a (possibly infinite) sequence of disjoint events such that $\bigcup_{i=1}^{n} A_i = \Omega$ and $P(A_i) > 0$ for all i. Let B be another event with $P(B) > 0$. Then

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)} \tag{2.3}$$

Proof.

By definition, $P(A_i \mid B) = P(A_i \cap B)/P(B)$, and $P(A_i \cap B) = P(B \mid A_i)P(A_i)$, so:

□

Updating beliefs

Theorem 17 (Bayes' theorem)

Let $A_1, A_2, \ldots$ be a (possibly infinite) sequence of disjoint events such that $\bigcup_{i=1}^n A_i = \Omega$ and $P(A_i) > 0$ for all i. Let B be another event with $P(B) > 0$. Then

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^n P(B \mid A_j)P(A_j)} \tag{2.3}$$

Proof.

By definition, $P(A_i \mid B) = P(A_i \cap B)/P(B)$, and $P(A_i \cap B) = P(B \mid A_i)P(A_i)$, so:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)}, \tag{2.4}$$

□

Updating beliefs

Theorem 17 (Bayes' theorem)

Let $A_1, A_2, \ldots$ *be a (possibly infinite) sequence of disjoint events such that* $\bigcup_{i=1}^{n} A_i = \Omega$ *and* $P(A_i) > 0$ *for all* i*. Let* B *be another event with* $P(B) > 0$*. Then*

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)} \tag{2.3}$$

Proof.

By definition, $P(A_i \mid B) = P(A_i \cap B)/P(B)$, and $P(A_i \cap B) = P(B \mid A_i)P(A_i)$, so:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)}, \tag{2.4}$$

As $\bigcup_{i=1}^{n} A_i = \Omega$, we have $B = \bigcup_{j=1}^{n}(B \cap A_j)$.

□

Updating beliefs

Theorem 17 (Bayes' theorem)

Let $A_1, A_2, \dots$ be a (possibly infinite) sequence of disjoint events such that $\bigcup_{i=1}^{n} A_i = \Omega$ and $P(A_i) > 0$ for all i. Let B be another event with $P(B) > 0$. Then

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)} \tag{2.3}$$

Proof.

By definition, $P(A_i \mid B) = P(A_i \cap B)/P(B)$, and $P(A_i \cap B) = P(B \mid A_i)P(A_i)$, so:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)}, \tag{2.4}$$

As $\bigcup_{i=1}^{n} A_i = \Omega$, we have $B = \bigcup_{j=1}^{n}(B \cap A_j)$. Since A_i are disjoint, so are $B \cap A_i$.

□

Updating beliefs

Theorem 17 (Bayes' theorem)

Let $A_1, A_2, \ldots$ be a (possibly infinite) sequence of disjoint events such that $\bigcup_{i=1}^{n} A_i = \Omega$ and $P(A_i) > 0$ for all i. Let B be another event with $P(B) > 0$. Then

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)} \tag{2.3}$$

Proof.

By definition, $P(A_i \mid B) = P(A_i \cap B)/P(B)$, and $P(A_i \cap B) = P(B \mid A_i)P(A_i)$, so:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)}, \tag{2.4}$$

As $\bigcup_{i=1}^{n} A_i = \Omega$, we have $B = \bigcup_{j=1}^{n}(B \cap A_j)$. Since A_i are disjoint, so are $B \cap A_i$. As P is a probability, the union property and an application of 2.4 gives

$$P(B) = P\left(\bigcup_{j=1}^{n}(B \cap A_j)\right) = \sum_{j=1}^{n} P(B \cap A_j) = \sum_{j=1}^{n} P(B \mid A_j)P(A_j).$$

□

Updating beliefs: addendum

Interpreting Bayes's theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- $P(A)$: our prior belief that hypothesis A is true (use Occam's razor!)
- $P(B \mid A)$: how much does hypothesis A agree with the evidence B?
- $P(B)$: marginal probability of the evidence B according to all hypotheses (Epicurean principle)
- $P(A \mid B)$: our posterior belief that hypothesis A is true given evidence B.

Exercise 3

Recall that

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}$$

is only a definition. Give plausible alternatives.

Updating beliefs

Consider the forecasters actually giving probabilities for rain.

Forecaster	Saturday	Sunday	Monday	Tuesday
A_1	60%	70%	80%	90%
A_2	10%	50%	60%	20%
A_3	20%	25%	40%	100%
A_4	10%	15%	30%	25%
A_5	30%	40%	35%	10%
Outcome				

Table : Five weather forecasters

Let $P(A_i) = 1/5$ be our prior belief that A_i is correct. Then:

A_1	A_2	A_3	A_4	A_5

Updating beliefs

Consider the forecasters actually giving probabilities for rain.

Forecaster	Saturday	Sunday	Monday	Tuesday
A_1	60%	70%	80%	90%
A_2	10%	50%	60%	20%
A_3	20%	25%	40%	100%
A_4	10%	15%	30%	25%
A_5	30%	40%	35%	10%
Outcome	Clouds			

Table : Five weather forecasters

Let $P(A_i) = 1/5$ be our prior belief that A_i is correct. Then:

A_1	A_2	A_3	A_4	A_5
0.11	0.25	0.22	0.25	0.19

Updating beliefs

Consider the forecasters actually giving probabilities for rain.

Forecaster	Saturday	Sunday	Monday	Tuesday
A_1	60%	70%	80%	90%
A_2	10%	50%	60%	20%
A_3	20%	25%	40%	100%
A_4	10%	15%	30%	25%
A_5	30%	40%	35%	10%
Outcome	Clouds	Rain		

Table : Five weather forecasters

Let $P(A_i) = 1/5$ be our prior belief that A_i is correct. Then:

A_1	A_2	A_3	A_4	A_5
0.35	0.25	0.13	0.08	0.2

Updating beliefs

Consider the forecasters actually giving probabilities for rain.

Forecaster	Saturday	Sunday	Monday	Tuesday
A_1	60%	70%	80%	90%
A_2	10%	50%	60%	20%
A_3	20%	25%	40%	100%
A_4	10%	15%	30%	25%
A_5	30%	40%	35%	10%
Outcome	Clouds	Rain	Rain	

Table : Five weather forecasters

Let $P(A_i) = 1/5$ be our prior belief that A_i is correct. Then:

A_1	A_2	A_3	A_4	A_5
0.33	0.25	0.17	0.13	0.15

Updating beliefs

Consider the forecasters actually giving probabilities for rain.

Forecaster	Saturday	Sunday	Monday	Tuesday
A_1	60%	70%	80%	90%
A_2	10%	50%	60%	20%
A_3	20%	25%	40%	100%
A_4	10%	15%	30%	25%
A_5	30%	40%	35%	10%
Outcome	Clouds	Rain	Rain	Sun

Table : Five weather forecasters

Let $P(A_i) = 1/5$ be our prior belief that A_i is correct. Then:

A_1	A_2	A_3	A_4	A_5
0.04	0.32	0	0.30	0.36

Simplified notation and capturing dependencies

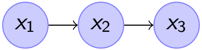
Consider random variables $x_i : \Omega \to S_i$, $i = 1, \ldots, n$. As a shorthand, especially in computer science, we may write their joint distribution as

$$P(x_1, \ldots, x_n),$$

instead of

$$P_{x_1,\ldots,x_n}(\cdot),$$

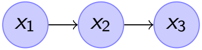
as is usually done in statistics.
Graphs can be used to capture independence between these variables. For example:



Means that $P(x_3, x_2, x_1) = P(x_3 \mid x_2)P(x_2 \mid x_1)P(x_1)$

Marginalisation (variable elimination)

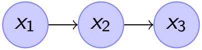
Consider the example network $P(x_3, x_2, x_1) = P(x_3 \mid x_2)P(x_2 \mid x_1)P(x_1)$.



This means that to express the joint distribution of the variables $x_i(\omega)$ we only need to model the conditional distributions $P(x_i \mid x_j)$.

Marginalisation (variable elimination)

Consider the example network $P(x_3, x_2, x_1) = P(x_3 \mid x_2)P(x_2 \mid x_1)P(x_1)$.



This means that to express the joint distribution of the variables $x_i(\omega)$ we only need to model the conditional distributions $P(x_i \mid x_j)$.

Inference via marginalisation

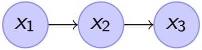
What is the distribution of x_3, ignoring the other variables?

$$P(x_3) = \sum_{x_1 \in S_1} \sum_{x_2 \in S_2} P(x_1, x_2, x_3). = \sum_{x_1 \in S_1} \sum_{x_2 \in S_2} P(x_3 \mid x_2)P(x_2 \mid x_1)P(x_1). \quad (2.5)$$

This follows from the disjoint property of measures, as illustrated in the proof of Bayes' theorem.

Marginalisation (variable elimination)

Consider the example network $P(x_3, x_2, x_1) = P(x_3 \mid x_2)P(x_2 \mid x_1)P(x_1)$.



This means that to express the joint distribution of the variables $x_i(\omega)$ we only need to model the conditional distributions $P(x_i \mid x_j)$.

Inference via marginalisation

What is the distribution of x_3, ignoring the other variables?

$$P(x_3) = \sum_{x_1 \in S_1} \sum_{x_2 \in S_2} P(x_1, x_2, x_3). = \sum_{x_1 \in S_1} \sum_{x_2 \in S_2} P(x_3 \mid x_2)P(x_2 \mid x_1)P(x_1). \quad (2.5)$$

This follows from the disjoint property of measures, as illustrated in the proof of Bayes' theorem. What is the distribution of x_3, given x_1?

$$P(x_3 \mid x_1) = \sum_{x_2 \in S_2} P(x_2, x_3 \mid x_1) = \sum_{x_2 \in S_2} P(x_3 \mid x_2)P(x_2 \mid x_1) \qquad (2.6)$$

Application to Bayesian inference

Consider now that you have a set of models $\{\omega_i \mid i = 1, \ldots\}$, each making a different prediction for tomorrow's weather x_{t+1}, given the weather in the past $x_1, \ldots, x_t$.

$$P(x_{t+1} \mid x_1, \ldots, x_t, \omega_i)$$

Let $P(\omega_i)$ be your prior probability on each model. Then the marginal probability is going to be

$$P(x_{t+1}) = \sum_i P(x_{t+1} \mid \omega_i) P(\omega_i).$$

Given some weather observations, you can now estimate a posterior distribution

$$P(\omega_i \mid x_1, \ldots, x_t) = \frac{P(x_1, \ldots, x_t \mid \omega_i) P(\omega_i)}{\sum_j P(x_1, \ldots, x_t \mid \omega_j) P(\omega_j)}$$

You can now calculate a new marginal probability for the weather,

$$P(x_{t+1} \mid x_1, \ldots, x_t) = \sum_i P(x_{t+1} \mid x_1, \ldots, x_t, \omega_i) P(\omega_i \mid x_1, \ldots, x_t).$$

Exercise

Abdul Alhazred claims that he is psychic and can always predict a coin toss. You use a fair coin, such that the probability of it coming heads is $1/2$. You throw the coin 4 times, and AA guesses correctly all four times.
If $P(A) = 2^{-16}$ is your prior belief that AA is a psychic, then what is your posterior belief (approximately), given that AA has guessed correctly?

Posterior distributions for multiple observations

Assume that we observe a value $x^n \triangleq x_1, \ldots, x_n$ drawn from some distribution $P(x^n \mid \omega)$, with $\omega \in \Omega$. We have a prior P on Ω. For the observations, we write:

Observation probability given history x^{n-1} and parameter ω

$$P(x_n \mid x^{n-1}, \omega) = \frac{P(x^n \mid \omega)}{P(x^{n-1} \mid \omega)}$$

Posterior recursion

$$P(\omega \mid x^n) = \frac{P(x^n \mid \omega)P(\omega)}{P(x^n)} = \frac{P(x_n \mid x^{n-1}, \omega)P(\omega \mid x^{n-1})}{P(x_n \mid x^{n-1})}. \tag{2.7}$$

The posterior can be used as a new prior distribution.

Recap

- Conditional likelihood represents the likelihood of an event given another event.
- If A is a hypothesis, and B is a predicted event, $(A \mid B)$ is the likelihood of the event under hypothesis A.
- Conditional probabilities $P(A \mid B)$ can be defined analogously to normal probabilities.
- This gives us a numerical procedure for updating our beliefs about which hypotheses are true.
- This is easy to perform for finite numbers of events and hypotheses.
- Finally, the conditional structure of a problem can be captured via a graph.

Things to remember

- Probability is a measure with the property that $P(\Omega) = 1$. So it also satisfies:
  1. $P(\emptyset) = 0$
  2. $P(A) \geq 0$.
  3. If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.
- Consequently if $A \subset B$ (i.e. A implies B, or B follows logically form A), then $P(A) \leq P(B)$.
- In addition A, B are called independent if $P(A \cap B) = P(A)P(B)$.
- The conditional probability of A given B is defined as $P(A \mid B) \triangleq P(A \cap B)/P(B)$.
- The marginalisation property allows us to eliminate variables:

$$P(B) = P(B \cap A) + P(B \cap A^{\complement})$$

- Bayes' theorem states that

$$P(A \mid B) = P(B \mid A)P(A)/P(B)$$

Symbol index

- The symbol $\triangleq$ indicates a definition.
- If an element x belongs to a set A, we write $x \in A$. If it does not, we write $x \notin A$.
- We say that A is a subset of B or that B contains A, and write $A \subset B$, iff, $x \in B$ for any $x \in A$.
- Events are sets. The sample space Ω is the certain event. Any other event A is a subset of Ω.
- $B \setminus A \triangleq \{x \mid x \in B, x \notin A\}$ is the set difference.
- The negation of an event $A \subset \Omega$ is the complement $A^{\complement} \triangleq \Omega \setminus A$.
- The union of n sets: $A_1, \ldots, A_n$ is $\bigcup_{i=1}^{n} A_i = A_1 \cup \cdots \cup A_n$. This can be interpreted as logical OR ($\vee$) of events.
- The intersection of n sets $A_1, \ldots, A_n$ is $\bigcap_{i=1}^{n} A_i = A_1 \cap \cdots \cap A_n$. This can be interpreted as logical AND ($\wedge$) of events.
- The empty set is $\emptyset = \Omega^{\complement}$ and contains no elements.
- A and B are disjoint if $A \cap B = \emptyset$. Then they are mutually exclusive events.
- $A \triangle B \triangleq (B \setminus A) \cup (A \setminus B)$ is the symmetric set difference.

[1] Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, 1970.

[2] Milton Friedman and Leonard J. Savage. The expected-utility hypothesis and the measurability of utility. *The Journal of Political Economy*, 60(6):463, 1952.

[3] Joseph Y. Halpern. *Reasoning about uncertainty*. MIT Press, 2003.

[4] Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, 1972.