Inference

Christos Dimitrakakis

March 27, 2025

Outline

Set theory

- First, consider some universal set Ω.
- A set A is a collection of points x in Ω.
- $\{x \in \Omega : f(x)\}$: the set of points in Ω with the property that $f(x)$ is true.

Unary operators

- $\neg A = \{x \in \Omega : x \notin A\}$.

Binary operators

- $A \cup B$ if $\{x \in \Omega : x \in A \vee x \in B\}$ - (c.f. $A \vee B$)
- $A \cap B$ if $\{x \in \Omega : x \in A \wedge x \in B\}$ - (c.f. $A \wedge B$)

Binary relations

- $A \subset B$ if $x \in A \Rightarrow x \in B$ - (c.f. $A \implies B$)
- $A = B$ if $x \in A \Leftrightarrow x \in B$ - (c.f. $A \Leftrightarrow B$)

The inference problem

- Given statements $A_1, \ldots, A_n$ we know to be true (i.e. a knowledge base), is another statement B true?

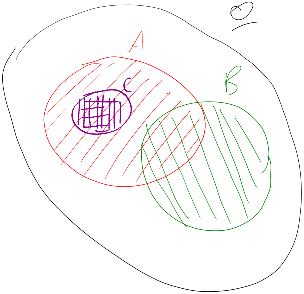
The following statements are equivalent:

- $A \implies B$ iff $(A \cap \neg B) = \emptyset$.
- $A \implies B$ iff $A \subset B$.

In addition

- If $(A \Rightarrow B) \wedge A$ then B.
- If $(A \wedge B)$ then A.

Illustration



$(A \mid C) =$

infered

known

$(B \mid C) =$

$(C \mid A) =$

$(A \wedge B \mid C)$

Events as sets

The universe and random outcomes

- The Ω contains all events that can happen.
- When something happens, we observe an element $\omega \in \Omega$.

Events in the universe

- An event is true if $\omega \in A$, and false if $\omega \notin A$.
- The negative event $\neg A = \Omega \setminus A$ is the set
- The possible events are a collection of subsets Σ of Ω so that

(i) $\Omega \in \Sigma$, (ii) $A, B \in \Sigma \Rightarrow A \cup B in \Sigma$ (iii) $A \in \Sigma \Rightarrow \neg A \in \Sigma$

Example: Traffic violation

- A car is moving with speed $\omega \in [0, \infty)$ in front of the speed camera.
- $A_0 = [0, 50]$: below the speed limit
- $A_1 = (50, 60]$: low fine
- $A_2 = (60, \infty]$: high fine
- $A_3 = (100, \infty)$: Suspension of license
- All combinations of the above events are interesting.

Probability fundamentals

Probability measure P

Probability can be seen as an area-like function assigning a likelihood to sets.

- $P : \Sigma \to [0, 1]$ gives the likelihood $P(A)$ of an event $A \in \Sigma$.
- $P(\Omega) = 1$
- For $A, B \subset \Omega$, if $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.

Marginalisation

Partition

If $A_1, \ldots, A_n$ are a partition of B then:

- $A_j \cap A_i = \emptyset$ for $i \neq j$
- $\bigcup_{i=1}^{n} A_i = B$.

Marginalisation

If $A_1, \ldots, A_n \subset \Omega$ are a partition of Ω

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i).$$

Conditional probability

Definition (Conditional probability)

The conditional probability of an event A given an event B is defined as

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

The above definition requires $P(B)$ to exist and be positive.

Conditional probabilities as a collection of probabilities

More generally, we can define conditional probabilities as simply a collection of probability distributions:

$$\{P_\theta : \theta \in \Theta\},$$

where Θ is indexing possible values of θ.

- θ is sometimes called the model or parameter

The theorem of Bayes

Theorem (Bayes's theorem)

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

The theorem of Bayes

Theorem (Bayes's theorem)

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

The general case

If $A_1, \ldots, A_n$ are a partition of Ω, meaning that they are mutually exclusive events (i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$) such that one of them must be true (i.e. $\bigcup_{i=1}^n A_i = \Omega$), then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

and

$$P(A_j|B) = \frac{P(B|A_j)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

Bayes's theorem

As a conditional measure

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

Bayes's theorem

As a conditional measure

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

As a causal explanation

$$\mathbb{P}(\text{cause} \mid \text{effect}) = \frac{\mathbb{P}(\text{effect} \mid \text{cause})\, \mathbb{P}(\text{cause})}{\mathbb{P}(\text{effect})}$$

Bayes's theorem

As a conditional measure

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

As a causal explanation

$$\mathbb{P}(\text{cause} \mid \text{effect}) = \frac{\mathbb{P}(\text{effect} \mid \text{cause})\,\mathbb{P}(\text{cause})}{\mathbb{P}(\text{effect})}$$

As model inference

- Prior $\beta(\theta)$
- Model class $\{P_\theta(\beta) : \theta \in \Theta\}$
- Data x

$$\beta(\theta \mid x) = \frac{P_\theta(x)\beta(\theta)}{\mathbb{P}_\beta(x)} = \frac{P_\theta(x)\beta(x)}{\sum_{\theta' \in \Theta} P_{\theta'}(x)\beta(\theta')}$$

Example: COVID symptoms

Activity (with playing cards or dice)

- Pick two (x, y) from 1 to 10.
- If $(x = 1$ and $y < 9)$, or $(x$ is even and $y \geq 9)$, you have symptoms.
- Do you have COVID?

Example: COVID symptoms

Activity (with playing cards or dice)

- Pick two (x, y) from 1 to 10.
- If $(x = 1$ and $y < 9)$, or $(x$ is even and $y \geq 9)$, you have symptoms.
- Do you have COVID?

Information

- 20% of people have COVID
- 50% of people with COVID have symptoms.
- 10% of people with no COVID have symptoms.
- If you do have symptoms, what are the chances you have COVID?

Example: COVID symptoms

Activity (with playing cards or dice)

- Pick two (x, y) from 1 to 10.
- If $(x = 1$ and $y < 9)$, or $(x$ is even and $y \geq 9)$, you have symptoms.
- Do you have COVID?

Information

- 20% of people have COVID
- 50% of people with COVID have symptoms.
- 10% of people with no COVID have symptoms.
- If you do have symptoms, what are the chances you have COVID?

Formalisation

- Prior $P(C) = 0.1$:
- Likelihood: $P(S|C) = 0.5$, $P(S|\neg C) = 0.1$
- Posterior:

$$P(C|S) = \frac{P(S|C)P(C)}{P(S|C)P(C) + P(S|\neg C)P(\neg C)}$$

Random variables

A random variable $f : \Omega \to \mathbb{R}$ is a real-valued function, with $\omega \sim P$.

The distribution of f

The probability that f lies in some subset $A \subset \mathbb{R}$ is

$$P_f(A) \triangleq P(\{\omega \in \Omega : f(\omega) \in A\}),$$

and we write $f \sim P_f$.

Shorthands for RV

- For RVs $f : \Omega \to \mathbb{R}$, we write $P(f \in A)$ to mean $P_f(A)$.
- For RVs $f : \Omega \to X$, where X is a finite set e.g. $\{1, 2, \ldots, n\}$, we write $P(f = x) = P_f(\{x\})$ for any $x \in X$.

Independence of random variables

Two RVs f, g are independent in the same way that events are independent.

$$P(f \in A \wedge g \in B) = P(f \in A)P(g \in B) = P_f(A)P_g(B).$$

In that sense, $f \sim P_f$ and $g \sim P_g$.

Formal definition

More specifically, we are measuring the set of ω values for which $f(\omega) \in A$ and $g(\omega) \in B$:

$$P(\{\omega : f(\omega) \in A, g(\omega) \in B\}) = P_f(A)P_g(B).$$

Shorthand notation

Since the above is very cumbersome, we usually just write that

$$P(f, g) = P(f)P(g)$$

for any two independent random variables f, g.

Expectation

For any real-valued random variable $f : \Omega \to \mathbb{R}$, the expectation with respect to a probability measure P is

$$\mathbb{E}_P(f) = \sum_{\omega \in \Omega} f(\omega) P(\omega).$$

When Ω is continuous, we can use a density p

$$\mathbb{E}_P(f) = \int_\Omega f(\omega) p(\omega) d\omega.$$

Linearity of expectations

For any RVs x, y:

$$\mathbb{E}_P(x + y) = \mathbb{E}_P(x) + \mathbb{E}_P(y)$$

Multiple variables

The joint distribution $P(x, y)$

For two (or more) RVs $x : \Omega \to \mathbb{R}$, and $y : \Omega \to \mathbb{R}$, this is a shorthand for the distribution of $(x(\omega), y(\omega))$ when $\omega \sim P$. We can also use $P(x = i, y = j)$ for the probability that the two variables assume the values i, j respectively.

Independence

If x, y are independent RVs then $P(x, y) = P_x(x)P_y(y)$.

Correlation

If x, y are not correlated then $\mathbb{E}_P(xy) = \mathbb{E}(x)\,\mathbb{E}(y)$.

IID (Independent and Identically Distributed) random variables

A sequence x_t of r.v.s is IID if $x_t \sim P$ so that

$$(x_1, \ldots, x_t, \ldots, x_T) \sim P^T$$

i.e. a T-length sample is drawn from the product distribution $P^T = P \times P \times \cdots \times P$.

Conditional expectation

The conditional expectation of a random variable $f : \Omega \to \mathbb{R}$, with respect to a probability measure P conditioned on some event B is simply

$$\mathbb{E}_P(f|B) = \sum_{\omega \in \Omega} f(\omega) P(\omega|B).$$

Conditional expectations are similar to conditional probabilities.

Conditional probabilities of RVs

Similarly to the notation over sets,

$$P(A \cap B) = P(A \mid B)P(B),$$

when dealing with RVs, it is common to use the notation

$$P(x, y) = P(x|y)P(y)$$

This equation works for all possible values of x, y e.g.

$$P(x = 1, y = 0) = P(x = 1|y = 0)P(y = 0)$$

which then denotes the probability msas of each

Probability notation: math versus statistics

- $P(C)$: Probability of event C
- $P(A \cap B)$: Probability of A and B
- $P(A|B)$: Probability of the event A if we know B
- $P(A \cup B)$: Probability of A or B

- $P(x)$: distribution of variable x.
- $P(x, y)$: joint distribution of x, y
- $P(x|y)$: distribution of x for different values of y
- No correspondence.

Example: The k-meteorologists problem (set notation)

- R_t: The event that it rains at time t.
- A set of stations Θ, with $\theta \in \Theta$ making weather predictions:

$$P(R_{t+1} \mid R_1, \ldots, R_t, \theta),$$

- A prior probability $P(\theta)$ on the stations.
- The marginal probability

$$P(R_1 \cap \cdots \cap R_t) = \sum_{\theta \in \Theta} P(R_1 \cap \cdots \cap R_t \mid \theta) P(\theta)$$

- The posterior probability

$$\begin{aligned} P(\theta \mid R_1 \cap \cdots \cap R_t) &= \frac{P(R_1 \cap \cdots \cap R_t \mid \theta) P(\theta)}{P(R_1 \cap \cdots \cap R_t)} = \frac{\prod_{i=1}^{t} P(R_t \mid R_1 \cap \cdots \cap R_{t-1}}{P(R_1 \cap \cdots \cap R_t)} \\ &= \frac{P(R_t \mid R_1 \cap \cdots \cap R_{t-1} \mid \theta) P(\theta \mid R_1 \cap \cdots \cap R_{t-1})}{P(R_t \mid R_1 \cap \cdots \cap R_{t-1})} \end{aligned}$$

- The marginal posterior probability

$$P(R_{t+1} \mid R_1 \cap \cdots \cap R_t) = \sum_{\theta \in \Theta} P(R_{t+1} \mid R_1 \cap \cdots \cap R_t, \theta) P(\theta \mid R_1 \cap \cdots \cap R_t)$$

Example: The k-meteorologists problem (stat notation)

- $x_t \in \{0,1\}$: A random variable, telling us whether it rains at time t.
- A set of stations Θ, with $\theta \in \Theta$ making weather predictions:

$$P_\theta(x_{t+1} \mid x_1, \ldots, x_t)$$

- A prior probability $\beta(\theta)$ on the stations.
- The marginal probability

$$\mathbb{P}_\beta(x_1, \ldots, x_t) = \sum_{\theta \in \Theta} P_\theta(x_1, \ldots, x_t)\beta(\theta)$$

- The posterior probability

$$\beta(\theta \mid x_1, \ldots, x_t) = \frac{P_\theta(x_1, \ldots, x_t)\beta(\theta)}{\mathbb{P}_\beta(x_1, \ldots, x_t)} = \frac{\prod_{i=1}^t P_\theta(x_t \mid x_1, \ldots, x_{t-1})\beta(\theta)}{\mathbb{P}_\beta(x_1, \ldots, x_t)}$$
$$= \frac{P_\theta(x_t \mid x_1, \ldots, x_{t-1})\beta(\theta \mid x_1, \ldots, x_{t-1})}{\mathbb{P}_\beta(x_t \mid x_1, \ldots, x_{t-1})}$$

- The marginal posterior probability

$$\mathbb{P}_\beta(x_{t+1} \mid x_1, \ldots, x_t) = \sum_{\theta \in \Theta} P_\theta(x_{t+1} \mid x_1, \ldots, x_t)\beta(\theta \mid x_1, \ldots, x_t)$$

Graphical models

- A graphical model or Bayesian network is used to model dependencies between random variables.
- Each node in the graph is a variable.
- The arcs show what variable is an input to which variable.

Independence

Independent events $A \perp\!\!\!\perp B$

- A, B are independent iff $P(A \cap B) = P(A)P(B)$.
- Knowing if A happened, does not tell us anything about whether B happened

Conditional independence $A \perp\!\!\!\perp B \mid C$

- A, B are conditionally independent given C iff $P(A \cap B|C) = P(A|C)P(B|C)$.
- Knowing if C happened tells us all we need to know about A and B.

For random variables

- Independence: $P(x, y) = P(x)P(y)$.
- Conditional independence: $P(x, y|z) = P(x|z)P(y|z)$.

Model specification: Independent

$$
\begin{aligned}
f &= \mathrm{Bernoulli}(1/2) \\
g &= \mathrm{Bernoulli}(0.8) \\
x_1 &\sim f \\
x_2 &\sim g
\end{aligned}
$$

```
def f():
  return np.random.choice(2)
def g:
  return np.random.choice(2, [0.2, 0.8])
x1 = f()
x2 = g()
```

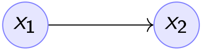
Model specification: Gaussian Dependent variables

$$x_1 \longrightarrow x_2$$

$$\begin{aligned} f &= \mathrm{Normal}(0,1) \\ g(a) &= \mathrm{Normal}(a,1) \\ x_1 &\sim f \\ x_2 | x_1 = a &\sim g(a) \end{aligned}$$

```
def f():
  return np.random.normal(0,1)
def g(a):
  return np.random.normal(a)
x1 = f()
x2 = g(x1)
```

Model specification: Bernoulli Dependent variables



$$
\begin{aligned}
f &= \text{Bernoulli}(1/2) \\
g(a) &= \text{Bernoulli}(\theta_a) \\
x_1 &\sim f \\
x_2 | x_1 = a &\sim g(a) \\
\theta &= (0.6, 0.5)
\end{aligned}
$$

```
def f():
  return np.random.choice(2)
def g(a):
  theta = [0.6, 0.5]
  return np.random.choice(2,
 [1 - theta[a], theta[a]])
x1 = f()
x2 = g(x1)
```

Model specification: Chain

$$x_1 \longrightarrow x_2 \longrightarrow x_3$$

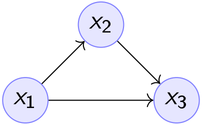
$$x_1 \sim f \tag{1}$$

$$x_2 \mid x_1 = a \sim g(a) \tag{2}$$

$$x_3 \mid x_2 = b \sim h(b), \tag{3}$$

```
def f():
  return np.random.uniform()
def g(a):
  return np.random.uniform() + a
def h(b)
  return np.random.uniform * b
x1 = f()
x2 = g(x1)
x3 = h(x2)
```

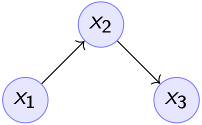
Graphical models



- Variables: x_1, x_2, x_3
- Arrows denote dependencies between variables.
- In this example the value of x_3 is a function of x_1, x_2, as well as a random input.

Conditional independence

Example



Graphical model for the factorisation $\mathbb{P}(x_3 \mid x_2)\,\mathbb{P}(x_2 \mid x_1)\,\mathbb{P}(x_1)$.

Definition

- Consider variables $x_1, \ldots, x_n$.
- Let B, D be subsets of $[n]$.

We say x_i is conditionally independent of $\boldsymbol{x}_B$ given $\boldsymbol{x}_D$ and write

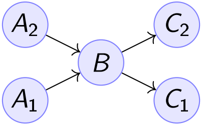
$$x_i \perp\!\!\!\perp \boldsymbol{x}_B \mid \boldsymbol{x}_D$$

if and only if:

$$\mathbb{P}(x_i, \boldsymbol{x}_B \mid \boldsymbol{x}_D) = \mathbb{P}(x_i \mid \boldsymbol{x}_D)\,\mathbb{P}(\boldsymbol{x}_B \mid \boldsymbol{x}_D).$$

Conditional independence

For any set of random variables $x_1, x_2, x_3, \ldots$, we can write their joint as $\prod_i P(x_i \mid x_1, \ldots, x_{i-1})$. However, we can use a Bayesian network to define conditional independence structures.



If A is a parent of B and C is a child of B, and there are no other paths from A to C then the following conditional independence holds:

$$P(C \mid B, A) = P(C \mid B)$$

i.e. C is conditionally independent of A given B.

Conditional probability tables

We can now write the distribution of the above example as

$$P(B, C_1, C_2) = P(A_1)P(A_2)P(B|A_1, A_2)P(C_1|B)P(C_2|B).$$

Smoking and lung cancer

$$S \longrightarrow C \longleftarrow A$$

Smoking and lung cancer graphical model, where S: Smoking, C: cancer, A: asbestos exposure.

- Here, S, A are independent
- However, they become dependent if we know C.

Smoking and lung cancer

$$S \longrightarrow C \longleftarrow A$$

Smoking and lung cancer graphical model, where S: Smoking, C: cancer, A: asbestos exposure.

- Here, S, A are independent
- However, they become dependent if we know C.

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

Smoking and lung cancer

$$S \longrightarrow C \longleftarrow A$$

Smoking and lung cancer graphical model, where S: Smoking, C: cancer, A: asbestos exposure.

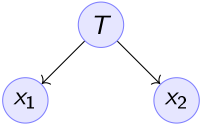
- Here, S, A are independent
- However, they become dependent if we know C.

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

$$P(A, S|C) = P(A|S, C)P(S|C) = \frac{P(C|A, S)P(A|S)}{P(C|S)} \frac{P(C|S)P(S)}{P(C)}$$

$$= \frac{P(C|A, S)P(A|S)}{P(S|C)P(C)/P(S)} \frac{P(C|S)P(S)}{P(C)}$$

Time of arrival at work

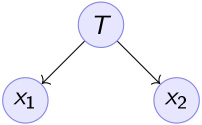


Time of arrival at work graphical model where T is a traffic jam and x_1 is the time John arrives at the office and x_2 is the time Jane arrives at the office.

*Conditional independence:

- Even though x_1, x_2 are not independent, they become independent once you know T.

Time of arrival at work



Time of arrival at work graphical model where T is a traffic jam and x_1 is the time John arrives at the office and x_2 is the time Jane arrives at the office.

*Conditional independence:

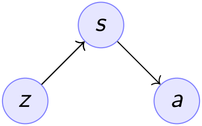
- Even though x_1, x_2 are not independent, they become independent once you know T.

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

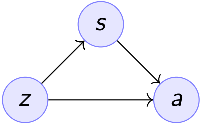
School admission

School	Male	Female
A	62	82
B	63	68
C	37	34
D	33	35
E	28	24
F	6	7

- z: gender
- s: school applied to
- a: admission

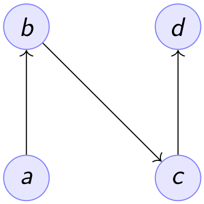


Is admission independent of gender?



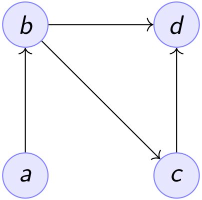
How about here?

What is the model for this graph?



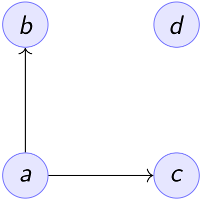
$$P(a, b, c, d) = \cdots$$

What is the model for this graph?



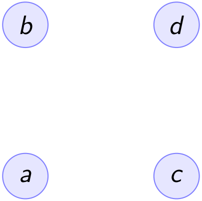
$$P(a, b, c, d) =$$

What is the model for this graph?



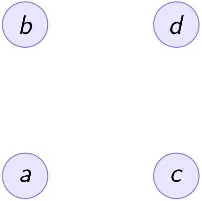
$$P(a, b, c, d) =$$

Draw the graph for this model



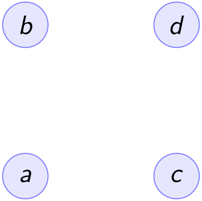
$$P(a, b, c, d) = P(a)P(b|a)P(c|b)P(d|b)$$

Draw the graph for this model



$$P(a,b,c,d) = P(a)P(b|a)P(d|c)P(c)$$

Draw the graph for this model



$$P(a, b, c, d) = P(a)P(b|a)P(c|a)P(d|b, c)$$

Example: COVID test

Information

- 10% of people have COVID
- 50% of people with COVID have a positive test
- 50% of people with COVID have symptoms
- 10% of people without COVID have a positive test
- 20% of people without COVID have symptoms

Example: COVID test

Information

- 10% of people have COVID
- 50% of people with COVID have a positive test
- 50% of people with COVID have symptoms
- 10% of people without COVID have a positive test
- 20% of people without COVID have symptoms

Formalisation

- Prior: $P(C = 1) = 0.1$
- Likelihood: $P(T, S|C) = P(T|C)P(S|C)$, $P(T, S|\neg C)$ for all va43lues of T, S, C.
- Posterior:

$$P(C|T, S) = \frac{P(S|C)P(T|C)P(C)}{\sum_{i=0}^{1} P(S|C = i)P(T|C = i)P(C = i)}$$

Example: Naive Bayes models

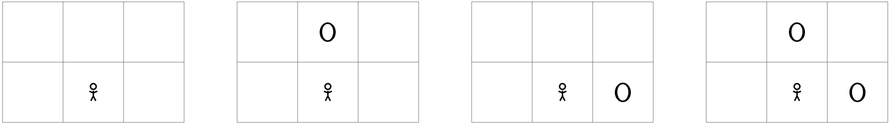
Sometimes we observe multiple effects that have a common cause, but which are otherwise independent:

$$\mathbb{P}(\text{effect}_1, \dots \text{effect}_n \mid \text{cause}) = \prod_{i=1}^{n} \mathbb{P}(\text{effect}_i \mid \text{cause})$$

Naive Bayes model

- Observations $(\boldsymbol{x}_t, y_t)_{t=1}^{T}$ with $\boldsymbol{x}_t = (x_{t,1}, \dots, x_{t,n})$.
- Probability models $P_\theta(y \mid \boldsymbol{x}) = \prod_{i=1}^{n} P_\theta(y \mid x_i)$.

Example: Wumpus world



Details

- Probability of each world A_i being true: 1/4
- Probability of each hole generating a breeze: $P(B_1|A_2 \cup A_4) = P(B_2|A_3 \cup A_4)$ with B_1, B_2 conditionally independent given A.

Questions

- What is the probability of feeling a breeze $B = B_1 \cup B_2$ in each world?
- What is the probability of a hole above if you feel a breeze?
- What is the probability of a hole above f you don't feel a breeze?