Information, Bayes, Independence

Contents

Information and Entropy

- you run an experiment to see if sun rises in the east tomorrow

$$A := \{\text{the event the sun rises in the east}\}$$

$$\mathbb{P}(A) = 0.9999$$

- sure enough, the sun rises in the east tomorrow
- how much information did we learn about the world?

$$\log\left(\frac{1}{\mathbb{P}(A)}\right)$$

- imagine it doesn't rise in the east

$$\log\left(\frac{1}{\mathbb{P}(A^c)}\right)$$

- on average . . .

$$\text{entropy} := \mathbb{P}(A)\log\left(\frac{1}{\mathbb{P}(A)}\right) + \mathbb{P}(A^c)\log\left(\frac{1}{\mathbb{P}(A^c)}\right)$$

Back to Information and Entropy

- you run another experiment to see if two events happen:

$A := \{\text{the event the sun rises in the east}\}$

$B := \{\text{you win a million dollars in the lottery}\}$

assume A, B are independent

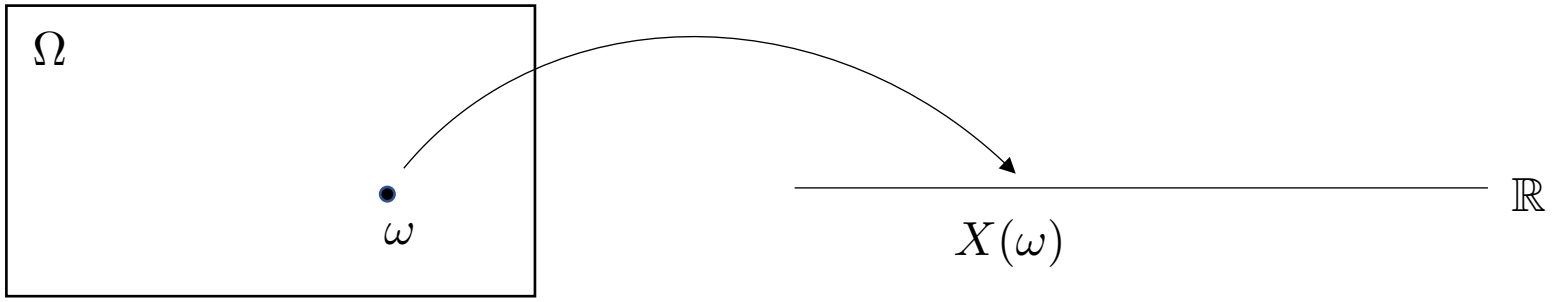
- how much information do we learn about the world if A and B happen?

$$\log\left(\frac{1}{\mathbb{P}(A,B)}\right) = \log\left(\frac{1}{\mathbb{P}(A)}\right) + \log\left(\frac{1}{\mathbb{P}(B)}\right)$$

Random variables

- probability space consists of $(\Omega, \mathcal{F}, \mathbb{P})$
  - A is a set of outcomes, i.e, a subset of Ω
    - $\mathbb{P}(A)$ is the probability that the outcome of a random experiment is in the set A
    - $\mathbb{P}(A)$ is the probability that the event A *occurs* or is *true*
- a random variable X is a real valued function $X(\omega)$ over the sample space Ω

$$X : \Omega \to \mathbb{R}$$



- use upper case letters for random variables: $X, Y, Z, \ldots$
- lower case letters for the values the random variable assumes, i.e, the event $X = x$

Random variables – shorthand notation

For $B \subset \mathbb{R}$, $\{X \in B\}$ means $\{\omega \in \Omega : X(\omega) \in B\}$.

So $\{X \in B\}$ denotes a subset of Ω.

Similarly, $\mathbb{P}(X \in B)$ means $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$.

So it is clear that the argument of $\mathbb{P}$ is ***always*** a subset of Ω.

To derive a formula expressed in shorthand, you should expand the shorthand so that you see what subsets of Ω you are dealing with.

To say that a random variable X is real valued means that for all $\omega \in \Omega, X(\omega) \in \mathbb{R}$. Hence,

$$\mathbb{P}(X \in \mathbb{R}) = \mathbb{P}(\{\omega \in \Omega: X(\omega) \in \mathbb{R}\}) = \mathbb{P}(\Omega) = 1.$$

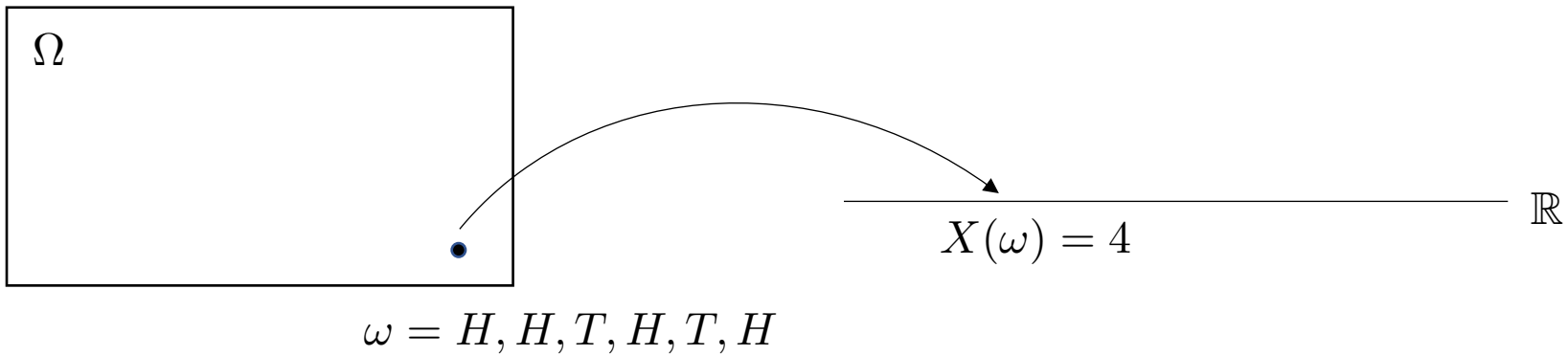
Example

- example: flipping a coin n times

$$\Omega = \{H, T\}^n$$

  - define the random variable X to be the number of heads

$$X : \Omega \to \{0, 1, \ldots, n\}$$



Example

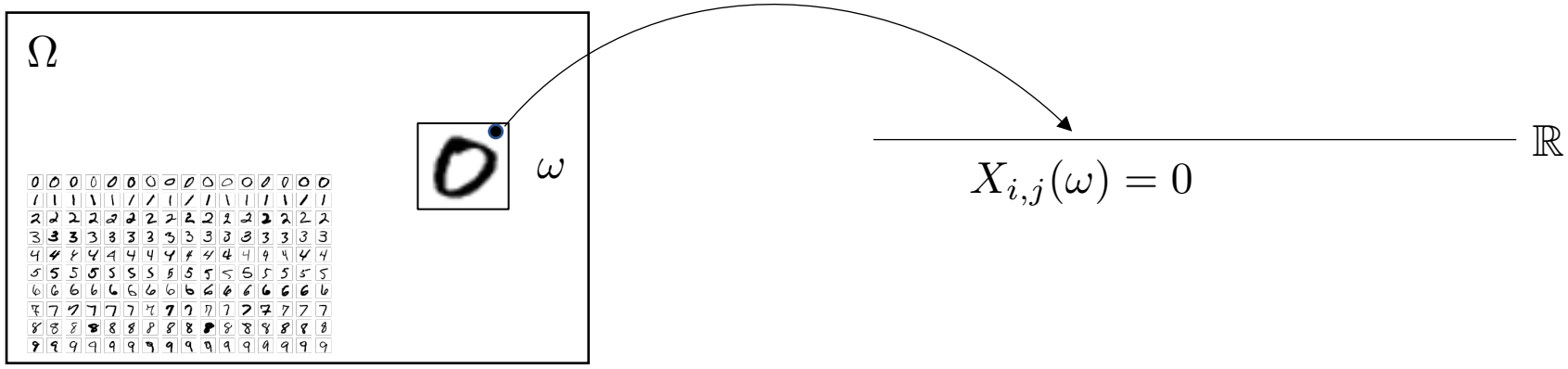
- example: ask someone to write a digit (0-9) on a piece of paper

$$\Omega = \mathcal{X} \times \mathcal{Y}$$

- $\mathcal{X}$ is the space of possible hand draw digits
- $\mathcal{Y} = \{0, 1, \ldots, 9\}$ is the writer's intention

- define $X_{i,j}$ to be the value pixel i, j after scanning in black and white

$$X_{i,j} : \Omega \to \{0, 1\}$$



Discrete Random Variables and pmfs

- definition: a random variable is a *discrete random variable* if

  $\mathbb{P}(X \in \mathcal{X}) = 1$ for some countable set $\mathcal{X} \subset \mathbb{R}$

- i.e, if you can count the possible values assumed by X, then it's a discrete R.V.

- a discrete random variable is completely specified by its probability mass function $p(x)$

$$p(x) = \mathbb{P}(X = x) \qquad \text{for all } x \in \mathcal{X}$$

properties of pmfs:

1. $p(x) \geq 0$
2. $\sum_{x \in \mathcal{X}} p(x) = 1$

- notation: $X \sim p(x)$ indicates that the random variable X has pmf $p(X)$

  note: a random variable that is not discrete *might* be *continuous* (more on this next time)

Let's derive Property 2: $\sum_x p(x) = 1.$

To do this, we expand the shorthand notation

$$p(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega\colon X(\omega) = x\}), \quad \text{for each } x \in \mathcal{X}.$$

Now put $A_x \coloneqq \{\omega \in \Omega\colon X(\omega) = x\}$ so that the property can be written as

$$\sum_{x \in \mathcal{X}} \mathbb{P}(A_x) = 1.$$

Since the A_x are pairwise disjoint, we can write

$$\sum_{x \in \mathcal{X}} \mathbb{P}(A_x) = \mathbb{P}\left(\bigcup_{x \in \mathcal{X}} A_x\right) = \mathbb{P}(X \in \mathcal{X}) = 1,$$

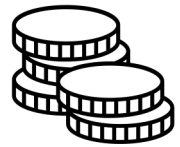
where we use the fact that

$$\bigcup_{x \in \mathcal{X}} A_x = \{\omega \in \Omega : X(\omega) \in \mathcal{X}\}.$$

This equation is understood as saying that

$$\omega \in A_x \text{ for some } x \in \mathcal{X} \Leftrightarrow X(\omega) \in \mathcal{X}.$$

Famous Discrete Random Variables



- Bernoulli: $X \sim \text{Bern}(p)$ for $0 \leq p \leq 1$

$$p(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$