Probability theory

Mark Andrews
Psychology Department, Nottingham Trent University

@xmjandrews
mark.andrews@ntu.ac.uk

Probability space

- Probability theory begins with a *random experiment*. This is anything that can have a number of different possible outcomes.
- The set of possible outcomes is the *sample space* that we can denote as follows.

$$\Omega = \{\omega_1, \omega_2 \ldots \omega_n \ldots\},$$

  where each ω_i is a possible outcome.
- We then define the set of possible *events* of Ω. An *event* is any subset of Ω.
- The set of all possible events in a finite (or countably infinite) Ω is its power set. In an uncountably infinite Ω, we must define a σ-algebra. We denote the set of events by $\mathcal{F}$.
- We then define a function P on $\mathcal{F}$

$$P : \mathcal{F} \mapsto [0, 1].$$

  which is known as a *probability measure*, that obeys three axioms.
- The triplet $(\Omega, \mathcal{F}, P)$ is a *probability space*.

Probability axioms

- A probability measure obeys three axioms, usually known as the *Kolmolgorov axioms*.
  - For any set $A \in \mathcal{F}$, $P(A) \leqslant 0$.
  - $P(\Omega) = 1$.
  - For any disjoint set $A_1, A_2 \ldots A_3 \ldots$, where $A_i \in \mathcal{F}$, then

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

- From this we can derive various theorems. For example,
  - $0 \leqslant P(A) \leqslant 1$.
  - $P(\emptyset) = 0$.
  - $P(A^c) = 1 - P(A)$, where A^c is the complement of A.
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Conditional probability

- For any two events A and B

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

- This effectively defines a new probability space with B as the sample space.
- Given that

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

and

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

we therefore have

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

Random variables

- A random variable X is a deterministic mapping from Ω to a measurable space. This space is usually $\mathbb{R}$ and so

$$X\colon \Omega \mapsto \mathbb{R}$$

- For any $s \subseteq \mathbb{R}$, the probability that X takes a value in s is

$$P(X \in s) = P(\{\omega\colon X(\omega) \in s\}).$$

- We call $P(X)$ is probability distribution of the random variable X whose *support* is $\mathbb{R}$.

Joint random variables

- Given a Ω, we may define multiple random variable, e.g. $X, Y \ldots$.
- For any $s \subseteq \mathbb{R}$, $r \subseteq \mathbb{R}$, the probability that X takes a value in s *and* Y takes a value in r is

$$P(X \in s, Y \in r) = P(\{\omega \colon X(\omega) \in s \wedge Y(\omega) \in r\}).$$

- The conditional probability distribution of X given Y is

$$P(X \in s | Y \in r) = \frac{P(X \in s, Y \in r)}{P(Y \in r)}.$$

- The marginal probability that $X \in s$ is

$$P(X \in s) = \sum_{r \in \mathbb{R}} P(X \in s, X \in r).$$

Independent random variables

- Two variables X and Y are *independent* if and only if

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

- Two variables X and Y are *conditionally independent* given a third variable Z is

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

Chain rule

- For random variables X, Y

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

.

- For X, Y, Z,

$$\begin{aligned} P(X, Y, Z) &= P(X|Y, Z)P(Y, Z), \\ &= P(Y|X, Z)P(X, Z), \\ &= P(Z|X, Y)P(X, Y) \end{aligned}$$

Reasoning with probabilistic models

- If we have, e.g., a set of random variables $y_1, y_2 \ldots y_n$ that are independently and identically distributed as $N(\mu, \sigma^2)$, where μ and σ also defined by probability distributions, our probabilistic model is a probability distribution over $n + 2$ random variables

$$P(y_1, y_2, y_2 \ldots y_n, \mu, \sigma^2).$$

- Because of conditional independence, this decomposes as

$$\prod_{i=1}^{n} P(y_i|\mu, \sigma^2)P(\mu, \sigma^2).$$

- The marginal probability of $y_i \ldots y_n$ is

$$P(y_1 \ldots y_n) = \int P(y_1 \ldots y_n, \mu, \sigma) d\mu d\sigma^2$$

Bayes theorem

- Given a joint probability distribution over $y_1 \dots y_n$ and μ, σ^2, we have

$$P(\mu, \sigma^2 | y_1 \dots y_n) = \frac{P(y_1 \dots y_n, \mu, \sigma^2)}{P(y_1 \dots y_n)}$$

and

$$P(\mu, \sigma^2 | y_1 \dots y_n) = \frac{\prod_{i=1}^n P(y_i | \mu, \sigma^2) P(\mu, \sigma^2).}{P(y_i \dots y_n)}$$

and

$$P(\mu, \sigma^2 | y_1 \dots y_n) = \frac{\prod_{i=1}^n P(y_i | \mu, \sigma^2) P(\mu, \sigma^2).}{\int \prod_{i=1}^n P(y_i | \mu, \sigma^2) P(\mu, \sigma^2) d\mu d\sigma^2.}$$