GI12/4C59: Information Theory

Lectures 1–3

Massimiliano Pontil

About these lectures

Theme of Lectures 1–3: We provide a quick tour on basic probability which constitutes the main mathematical ingredient of Information Theory.
Other mathematical tools will be reviewed during the course when needed.

Prerequisites: familiarity with calculus (real-valued functions, limits, derivatives, Taylor series, etc..)

Lecture notes are available at
`http://www.cs.ac.uk/staff/M.Pontil/courses/`

Outline

1. Sample space and probability function
2. Conditional probability and independence
3. Random variables
4. Continuous random variables
5. Jointly random variables
6. Convergence in probability
7. Basic inequalities

Random experiment

Consider a random experiment consisting of a finite number N of mutually exclusive outcomes or *elementary events* $\omega_1, \ldots, \omega_N$. We assume for the time being that the outcomes are *equiprobable*, that is, they all have a probability of $1/N$.

Example 1: Tossing an unbiased (i.e. fair) coin. There are two outcomes, "head" and "tail". Both outcomes have probability 1/2.

Example 2: Throwing an unbiased die. Outcomes are the six possible faces of an unbiased die. Each face has probability 1/6.

The set $\Omega := \{\omega_1, \ldots, \omega_N\}$ is called the *sample space* and the elementary events the *sample points*.

Probability of an event

An event A is associated with the elementary events in Ω if for every $\omega \in \Omega$ we can always decide whether or not ω leads to the occurrence of A.

The probability $P(A)$ of the event A is defined by the formula

$$P(A) := \frac{N(A)}{N}$$

where $N(A)$ is the number of elementary events leading to A.

Example 2 (cont.) Let $A = \{$getting an even number of spots$\}$. Then, $P(A) = 3/6 = 1/2$.

More examples

Example 3: Tossing a coin twice. $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$. If $A =$"getting at least one head" then the elementary events leading to A are $(H,H),(H,T),(T,H)$ and, so, $P(A) = 3/4$.

Example 4: Throwing a pair of dice. The sample space consists of 36 elements. Let $A =$ "getting two sixes" and $B =$ "both dice show the same number of spots". We have $P(A) = \frac{1}{36}$ and $P(B) = \frac{6}{36} = \frac{1}{6}$.

Note: Apparently different events may coincide. For example, the event "the total number of spots is even" is the same as the event "both dice show an even or odd number of spots"

Combinatorial formulae

The computation of $N(A)$ requires combinatorial analysis.

Let $e_1, e_2, \ldots, e_n$ be n ordered elements of a set. Two useful formulae are

- The number of possible reordering of these elements is

$$n! := n \cdot (n-1) \cdots 2 \cdot 1$$

- The number of different unordered subgroups of size k is

$$C_k^n := \frac{n!}{(n-k)!k!}$$

Events as subsets of the sample space

Abusing notation we also let A be the set of elementary events leading to the occurrence of A.

This way, an event A is simply a subset of Ω, that is, $A \subseteq \Omega$.

In particular, the *sure event* is the set Ω and the *impossible event* is the empty set $\emptyset$.

Set operations

We define the following operations on $A, B \subseteq \Omega$:

- Union: $A \bigcup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$
  (occurrence of at least A or B).
- Intersection: $A \bigcap B \equiv AB := \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$
  (occurrence of both A and B).
  We say that A and B are *mutually exclusive* if $A \cap B = \emptyset$.
- Difference: $A - B := \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$
  (occurrence of A but not of B). Note that $A - B = A\bar{B}$.

Remark: The *complementary* of A is defined by $\bar{A} := \Omega - A$.

Set properties

For every $A, B \subseteq \Omega$ we have that

i) if $C = A \cup B$ then $\bar{C} = \bar{A} \cap \bar{B}$

ii) if $C = A \cap B$ then $\bar{C} = \bar{A} \cup \bar{B}$

iii) if $A \subset B$ then $\bar{B} \subset \bar{A}$ (follows from (i))

Note: Properties (i) and (ii) are known as the De Morgan laws. More generally, for every sequence of events $\{A_n : A_n \subset \Omega : n \in \mathbb{N}\}$ we have that

$$\overline{\left(\bigcup_{n\in\mathbb{N}} A_n\right)} = \bigcap_{n\in\mathbb{N}} \bar{A}_n, \quad \overline{\left(\bigcap_{n\in\mathbb{N}} A_n\right)} = \bigcup_{n\in\mathbb{N}} \bar{A}_n$$

Axioms of probability

If Ω is a (finite, countably infinite or uncountable) set, a function P defined on the "measurable" subsets of Ω is called a probability function if

1. For every $A \subseteq \Omega$, $P(A) \in [0,1]$.
2. $P(\Omega) = 1$.
3. For every sequence of mutually exclusive events $A_1, A_2, \ldots$, $P\left(\bigcup_{k\in\mathbb{N}} A_k\right) = \sum_{k\in\mathbb{N}} P(A_k)$.

Note: The above case where $\Omega = \{\omega_1, \ldots, \omega_N\}$ and $P(\omega_n) = 1/N$ satisfies these axioms.

Properties of P

For every $A, B \subseteq \Omega$ we have that

(a) $P(A - B) = P(A) - P(A \cap B)$

(b) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(c) $P(B) \leq P(A)$ if $B \subset A$.

Proof: (a): note that A is the union of the mutually exclusive events $(A - B)$ and $(A \cap B)$ and apply axiom 3.
(b): note that $A \cup B$ is the union of the mutually exclusive events $A - B$, $B - A$ and $A \cap B$. Then apply twice (a) and axiom 3 to get the result.
(c): We have that $A \cap B = B$ and, so, by (a) the result follows.

Note: A particular case of (a) is $A = \Omega$ in which case $A - B = \bar{B}$ and, so, (a) says that $P(\bar{B}) = 1 - P(B)$.

Dependent events and conditional probability

If $A, B \subseteq \Omega$ and $P(B) > 0$ we define the conditional probability of A given B by

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

which implies that $P(AB) = P(B)P(A|B)$.

It can be shown that, for every $B \subseteq \Omega$, $P(\cdot|B)$ verifies the axioms of probability.

In general we have the *multiplication rule*

$$P(A_1 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2A_1) \cdots P(A_n|A_1 \cdots A_{n-1})$$

Dependent events and conditional probability

Example 4 (cont.) Throwing a pair of dice. Let $A =$ "getting two sixes" and $B =$ "both dice show the same number of spots". Then $P(A) = \frac{1}{36}$, $P(B) = \frac{1}{6}$ and since $A \subset B$ we have that $AB = B$. So, we conclude that

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

and

$$P(B|A) = \frac{P(AB)}{P(A)} = 1.$$

An important decomposition

Note that for every $A, B \subseteq \Omega$, $A = AB \cup A\bar{B}$ and $AB \cap A\bar{B} = \emptyset$ (AB and $A\bar{B}$ are mutually exclusive). Thus, by axiom (3) we have the useful formula

$$\begin{aligned} P(A) &= P(AB) + P(A\bar{B}) \qquad (1) \\ &= P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \\ &= P(A|B)P(B) + P(A|\bar{B})(1 - P(B)) \end{aligned}$$

Bayes formula

If B_k, $k = 1, \ldots, n$ are mutually exclusive and $\bigcup_{k=1}^{n} B_k = \Omega$ then, for every $A \subseteq \Omega$ we have that

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}, \quad k = 1, \ldots, n$$

Proof: Note that $A = \bigcup_{k=1}^{n} AB_k$ and the sets AB_k are mutually exclusive. Consequently, by axiom (3)

$$P(A) = \sum_{k=1}^{n} P(AB_k) = \sum_{k=1}^{n} P(A|B_k)P(B_k)$$

from which the result follows.

A simple example

We have two coins. The first coin is fair while the second one is biased towards tail with $P(T) = \frac{3}{4}$. Consider the experiment where we randomly choose one of the two coins and flip it twice. What is the probability that the biased coin was flipped if we obtain two heads?

Solution: Let $A =$ "getting two heads" and $B =$ "the biased coin was flipped". We have $P(B) = P(\bar{B}) = \frac{1}{2}$, $P(A|B) = \frac{1}{16}$ and $P(A|\bar{B}) = \frac{1}{4}$. Consequently

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) = \frac{1}{16} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} = \frac{5}{32}$$

and we conclude that

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{\frac{1}{32}}{\frac{5}{32}} = \frac{1}{5}$$

Independent events

In general $P(A|B) \neq P(A)$. If the equality holds we say that A is independent of B. This implies that $P(AB) = P(A)P(B)$, and, so, that B is independent of A. Thus we say that A and B are independent if $P(AB) = P(A)P(B)$.

Note that if A and B are independent so are A and $\bar{B}$. In fact, we saw before that $A = AB \cup A\bar{B}$ and $AB \cap A\bar{B} = \emptyset$ and, so,...

Example 4 (cont.) Let $A =$"the sum of the dice is 7", B="the first die is 4", and $C =$"the second die is 3". Note that each pair consists of independent events. However, for example, A and BC are not independent!

Independent events (cont.)

The events A, B, C are said independent if every pair of them is independent and $P(ABC) = P(A)P(B)P(C)$.

Likewise, we say that the the events in the set $\{A_j : j = 1, \ldots, n\}, n \geq 2$ are independent if, for every $r \leq n$ and $1 \leq j_1 < j_2 < \cdots < j_r \leq n$, we have that

$$P(A_{j_1} \cdots A_{j_r}) = \prod_{k=1}^{r} P(A_{j_k})$$

Random variables

A random variable (r.v.) is a function $X : \Omega \to \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}$.

A is called a *discrete* r.v. if its range is either finite or countably infinite. The *Probability mass* function $p : \mathcal{X} \to [0, 1]$ of a discrete r.v. X is defined by

$$p(x) := P(\{X = x\}) = P(\{\omega \in \Omega : X(\omega) = x\}), \quad x \in \mathcal{X}$$

Example 4: Let X be the r.v. "number of heads appearing after tossing one coin three times". In this case $\mathcal{X} = \{0, 1, 2, 3\}$ and $P(X = 0) = P(X = 3) = \frac{1}{8}$, $P(X = 1) = P(X = 2) = \frac{3}{8}$.

Note: when it is clear in the text we write $P(\{X \in B\})$ as $P(X \in B)$, $B \subseteq \mathbb{R}$.

Some important quantities

- The *Expectation* of X is defined by $E[X] := \sum_{n \in \mathcal{X}} xp(x)$
- The *Variance* of X is defined by $Var(X) := E[(X - E[X])^2]$

The variance can be written as $Var(X) = E[X^2] - (E[X])^2$.

The expectation of X^n is called the $n-$order momentum of X.

Binomial random variable

Let $\mathcal{X} = \{0, 1, \ldots, n\}$. A binomial random variable has probability mass function:

$$p(k) = C_k^n p^k (1-p)^{n-k}, \quad p \in (0,1)$$

X can be interpreted as the number of successes occurring over n independent trials, each having probability p of success.

We have $E[X] = np$, $Var(X) = np(1-p)$.

Cumulative distribution function

The function $F(x) := P(\{X \leq x\})$ is called the *cumulative distribution* function (cdf) of X.

If X is discrete then $F(x) = \sum_{t \leq x} p(t)$

Example 5 (cont.): Let $X =$ "number of heads appearing after tossing one coin three times". In this case $\mathcal{X} = \{0, 1, 2, 3\}$ and $P(X = 0) = P(X = 3) = \frac{1}{8}$, $P(X = 2) = P(X = 4) = \frac{3}{8}$. Compute the cdf of X.

Cumulative distribution function

The following properties follows from the axioms of probability.

- F is a nondecreasing function: if $x < t$, then $F(x) \leq F(t)$.
- $\lim_{x \to \infty} F(x) = 1$
- $\lim_{x \to -\infty} F(x) = 0$
- F is right continuous: for every non decreasing sequence $\{x_n, x \geq 1\}$ converging to x, $\lim_{x \to x_0} = F(x_0)$
- $P(s < X \leq t) = F(t) - F(s)$ for every $s < t$.

Continuous random variables

X is called a *continuous* r.v. if there is a *probability density* function $f : \mathbb{R} \to [0, \infty)$ such that for every "measurable" set $B \subset \mathbb{R}$,

$$P(X \in B) = \int_B f(x)dx$$

Examples: X represents the time a train arrives at a specified station.

Note: Measurable sets include all sets of "practical interests", e.g. $(a, b), a, b \in \mathbb{R} \cup \{-\infty, \infty\}$. A rigorous treatment or continuous r.v. requires Measure Theory.

Continuous random variables (cont.)

The probability density f satisfies the following properties:

(i) $\int_{-\infty}^{\infty} f(x) = 1$

(ii) $P(a \leq X \leq b) = \int_a^b f(x)dx$

(iii) $P(X = a) = 0$.

(iv) $f(x) = \frac{dF}{dx}$

(v) $P(a \leq X \leq b) = F(b) - F(a)$

Continuous random variables (cont.)

The expectation of a continuous r.v X is given by

$$E[X] := \int_{-\infty}^{\infty} x f(x) dx$$

If X is a nonnegative r.v. it can be shown that

$$E[X] = \int_{0}^{\infty} P(X > x) dx$$

The Gaussian distribution

The most important continuous r.v. is a the Gaussian (or normal) r.v. whose distribution is defined, for $\mu \in \mathbb{R}$ and $\sigma > 0$, by

$$f(x) = N(\mu, \sigma)(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

We have that $E[X] = \mu$ and $Var(X) = \sigma^2$.

It can be shows that for n large the binomial distribution is approximated by $N(np, np(1-p))$.

Jointly random variable

The above ideas can be extended to a set of r.v. (vector-valued r.v.). In particular, if X and Y are two discrete r.v. with set values $\mathcal{X}$ and $\mathcal{Y}$, we define their jointly mass function by

$$p(x,y) := P(\{X = x, Y = y)\}, \quad x \in \mathcal{X}, y \in \mathcal{Y}$$

The (marginal) mass function of X is obtained by the formula

$$p_X(x) := P(\{X = x\}) = \sum_{y \in \mathcal{Y}} p(x,y).$$

The *conditional mass function* of X given that $Y = y$ is defined by.

$$P_{X|Y}(x,y) := P(\{X = x\}|\{Y = y\}) = \frac{p(x,y)}{p_Y(y)}$$

Jointly random variable (cont.)

X and Y are said to be jointly continuous r.v. if there exists a *probability density* function $f : \mathbb{R}^2 \to [0, \infty)$ such that, for every "measurable" set $D \in \mathbb{R}^2$

$$P(\{X \in A, Y \in B\}) = \int_A \int_B f(x,y) dx dy$$

If X and Y are jointly continuous then they are also individually continuous and their (marginal) density functions are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y) dy \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y) dx$$

The conditional density function of X given that $Y = y$ is given by

$$f_{X|Y}(x|y) := \frac{f(x,y)}{f_Y(y)}$$

Jointly independent r.v.

X and Y are said *independent* if, for every set $A, B \subseteq \mathbb{R}$

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Two discrete r.v. X and Y are independent if and only if their probability mass function is given by $p(x, y) = p_X(x)p_Y(y)$.

Tow continuous r.v. X and Y are independent if and only if their probability density function is given by $f(x, y) = f_X(x)f_Y(y)$.

Jointly cumulative distribution functions

The *jointly cumulative distribution* function of X and Y is define by

$$F(x, y) = P(X \leq x, Y \leq y), \quad x, y \in \mathbb{R}$$

The individual cdf are obtained as

$$F_X(x) = \lim_{y \to \infty} F(x, y), \quad F_Y(y) = \lim_{x \to \infty} F(x, y).$$

Expectation

If $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we have that

$$E[g(X,Y)] = \begin{cases} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} g(x,y)p(x,y) & \text{if } X, Y, \text{ are discrete} \\ \int_{x\in\mathbb{R}} \int_{y\in\mathbb{R}} g(x,y)f(x,y) & \text{if } X, Y, \text{ are continuous} \end{cases}$$

In particular, this implies that

$$E[X+Y] = E[X] + E[Y]$$

Conditional Expectation

The conditional expectation of X given that $Y = y$ is defined by

$$E[X|Y=y] = \begin{cases} \sum_{x\in\mathcal{X}} x p_{X|Y}(x|y) & \text{if } X, Y, \text{ are discrete} \\ \int_{x\in\mathbb{R}} x f_{X|Y}(x|y) & \text{if } X, Y, \text{ are continuous} \end{cases}$$

It can be shown that if $E[X|Y] : \Omega \to \mathbb{R}$ is defined by the formula $E[X|Y](y) = E[X|Y=y],\ y \in \mathcal{Y}$ then we have that

$$E[X] = E[E[X|Y]] = \begin{cases} \sum_{y\in\mathcal{Y}} E[X|Y=y]p_Y(y) & \text{if } X, Y, \text{ are discrete} \\ \int_{y\in\mathbb{R}} E[X|Y=y]f_Y(y) & \text{if } X, Y, \text{ are continuous} \end{cases}$$

Convergence in probability

We say that the sequence of r.v. $\{X_n : n \in \mathbb{N}\}$ converges *in probability* to X if, for every $t > 0$, we have that

$$\lim_{n\to\infty} P(\{|X_n - X| \geq t\}) = 0$$

Convergence with probability 1

We say that the sequence of r.v. $\{X_n : n \in \mathbb{N}\}$ converges *with probability* 1 to X if the following holds

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

Remark: Convergence with probability 1 is also called *almost sure* convergence. When the above equation holds true we say that the sequence $\{X_n : n \in \mathbb{N}\}$ converges to X almost surely.

If can be shown that if $X_n \to X$ almost surely then $X_n \to X$ in probability. The converse, however, is not true.

Some useful inequalities

Markov's inequality: If X is nonnegative r.v. then we have, for every $a > 0$, that

$$P(\{X \geq a\}) = \frac{E[X]}{a}$$

Proof: For $a > 0$ we define the binary r.v. Z as $Z = 1$ if $X \geq a$ and zero otherwise. Then, $Z \leq \frac{X}{a}$ and, so, $E[Z] \leq \frac{E[X]}{a}$. The result follows by observing that $P(\{X \geq a\}) = E[Z]$.

Some useful inequalities (cont.)

Chebyshev's inequality: If X has mean μ and variance σ then we have, for every $t > 0$, that

$$P(\{|X - \mu| \geq t\sigma\}) \leq \frac{1}{t^2}$$

Proof: Apply Markov's inequality to the r.v. $Y := (X - \mu)^2$.

Some topics we have not covered

- Moment generating functions
- Law of large numbers
- Central limit theorem
- Markov chains

Bibliography

Main reference:
S.Ross, A First Course in Probability, Prentice-Hall, 2002.

A concise introduction to probability is:
Y.A. Rozanov, Introductory Probability Theory, Prentice-Hall, 1969

A more advanced book is: P. Billingsley, Probability and Measure, Wiley series in probability and mathematical statistics, 1998.

In the Web:
`http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html`