Lecture 5

Information and Entropy

Lecture 5— Contents

Lecture 5— Contents

A measure for information

Any event from a probability space is partly unexpected (unpredictable). Thus it bears some information.
Formally,

- we want to measure how informative an event $A \in \mathcal{F}$ is, in a probability space $(\Omega, \mathcal{F}, \mathrm{P}[\cdot])$
- we do it by defining a real-valued quantity $i(A)$, named information of A,
- the information of A depends only on the probability $\mathrm{P}[A]$

$$i \ : \ \mathcal{F} \mapsto \mathbb{R} \quad , \quad i(A) = g(\mathrm{P}[A])$$

for a suitable function $g \ : \ [0, 1] \mapsto \mathbb{R}$.

▸ Go to axiomatic definition ⏩ Skip axiomatic definition

Axiomatic definition of information

We require $i(A)$ and $g(\alpha)$ to satisfy the following axioms

1. information is non negative, for all A
2. the sure event has null information, $i(\Omega) = 0$;
3. less likely events are more informative;
4. A, B independent events $\Rightarrow \; i(A \cap B) = i(A) + i(B)$

1. $g(\alpha) \geq 0$, for all $\alpha \in [0, 1]$
2. $g(1) = 0$;
3. g is a nonincreasing function in $[0, 1]$
4. $g(\alpha\beta) = g(\alpha) + g(\beta)$.

Axiomatic definition of information

The only functions that meet the above are

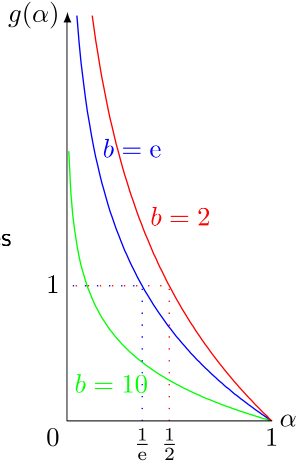
$$g(\alpha) = \log_b \frac{1}{\alpha} = \log_{1/b} \alpha$$

for $0 < \alpha \leq 1$ and the base $b > 1$

Defining information with a base or another only changes by a multiplicative constant.

It is customary to choose $b = 2$ so that an event with probability $\alpha = 1/2$ carries unit information

The unit information is called bit.



Definition of information

Definition

The information of an event A, having $\mathrm{P}[A] > 0$, is given by

$$i(A) = \log_2 \frac{1}{\mathrm{P}[A]} = \log_{1/2} \mathrm{P}[A] \quad \text{[bit]}$$

If $\mathrm{P}[A] = 0$ it is not possible to define $i(A)$.
It is also said that $i(A) = \infty$.

Show examples Skip examples

Lecture 5— Contents

Information function of a random variable

Given a discrete rv x with alphabet $\mathcal{A}_x$ and PMD $p_x(a)$, the information function of x maps any value $a \in \mathcal{A}_x$ into the information carried by x taking the value a

$$i_x \ : \ \mathcal{A}_x \mapsto \mathbb{R} \quad , \qquad i_x(a) = i\left(\{x = a\}\right) = \log_2 \frac{1}{p_x(a)}$$

Once defined i_x, we can apply it to x itself.
The rv $i_x(x)$ is the (random) information carried by x.

Definition of entropy

The mean of $i_x(x)$ represents the average information carried by x

Definition

The entropy $H(x)$ of a discrete rv x is the expectation of its information function

$$H(x) = \mathrm{E}\left[i_x(x)\right]$$

By the fundamental theorem for expectation,

$$H(x) = \sum_{a \in \mathcal{A}_x} p_x(a)\, i_x(a) = \sum_{a \in \mathcal{A}_x} p_x(a) \log_2 \frac{1}{p_x(a)}$$

$H(x)$ does not depend on the alphabet of x but only on its PMD values.

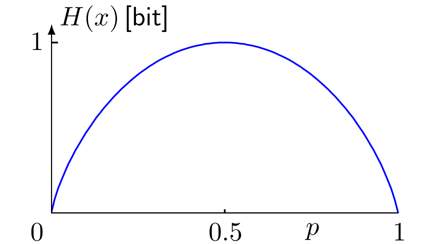
▸ Show example ▸▸ Skip example

Entropy calculation

Example (binary variable)

x binary rv with $\mathcal{A}_x = \{a_1, a_2\}$, $p_x(a_1) = p\,,\ p_x(a_2) = 1 - p$

$$\begin{aligned} H(x) &= p_x(a_1) i_x(a_1) + p_x(a_2) i_x(a_2) \\ &= p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{(1-p)} \end{aligned}$$



$p = 1/2 \quad \Rightarrow \quad H(x) = 1$ bit

any unbalancing ($p > 1/2$ or $p < 1/2$) decreases $H(x)$

for x a.s. constant ($p = 0$ or $p = 1$) $H(x) = 0$.

▸ Show more examples   ▸▸ Skip other examples

Bounds on entropy values

Proposition (lower bound)

1. *If x is a.s. constant, then $H(x) = 0$.*
2. *Otherwise, $H(x) > 0$.*

Proof.

Skip

1. $\mathcal{A}_x = \{a_1\}$ and $i_x(a_1) = 0$.
   There is only one term in the sum $H(x) = \sum_{a \in \mathcal{A}_x} p_x(a) i_x(a)$ and it is null.
2. For all $a \in \mathcal{A}_x$ we have $0 < p_x(a) < 1$ and $i_x(a) > 0$.
   All terms in $H(x) = \sum_{a \in \mathcal{A}_x} p_x(a) i_x(a)$ are strictly positive. $\square$

Bounds on entropy values

Proposition (upper bound)

Let x be a rv with a finite alphabet of M values

1. *If all $a \in \mathcal{A}_x$ are equally likely with probability $p_x(a) = 1/M$, then $H(x) = \log_2 M$.*
2. *Otherwise, $H(x) < \log_2 M$.*

Proof.

Skip

1. $$i_x(a) = \log_2 M \quad , \qquad \text{for all } a \in \mathcal{A}_x$$

   $i_x(x)$ is a.s. constant, its expectation $H(x) = \log_2 M$ □

Bounds on entropy values

Proof.

Skip

2. use Jensen's inequality, with the function $h(z) = \log_2 z$ strictly concave and the rv $z = 1/p_x(x)$ not a.s. constant

$$H(x) = \mathrm{E}\left[\log_2 \frac{1}{p_x(x)}\right] < \log_2 \mathrm{E}\left[\frac{1}{p_x(x)}\right]$$

$$\log_2 \mathrm{E}\left[\frac{1}{p_x(x)}\right] = \log_2 \left(\sum_{a \in \mathcal{A}_x} p_x(a) \frac{1}{p_x(a)}\right) = \log_2 M$$

□

Nominal information, efficency and redundancy

If x has an infinite alphabet, there is no upper bound. ▸ See example
For a finite rv x, the upper bound value $\log_2 M$ is called the nominal information.
The ratio of entropy to nominal information

$$\eta_x = \frac{H(x)}{\log_2 M} \quad , \quad 0 \leq \eta_x \leq 1$$

is called efficiency of x.
Its complement $1 - \eta_x$ is called redundancy.

Lecture 5— Contents

Entropy of a random vector

For a discrete random vector $\boldsymbol{x} = [x_1, \dots, x_n]$ we define information function

$$i_{\boldsymbol{x}} \; : \; \mathcal{A}_{\boldsymbol{x}} \mapsto \mathbb{R} \quad , \qquad i_{\boldsymbol{x}}(\boldsymbol{a}) = i\left(\{\boldsymbol{x} = \boldsymbol{a}\}\right) = \log_2 \frac{1}{p_{\boldsymbol{x}}(\boldsymbol{a})}$$

and entropy

$$H(\boldsymbol{x}) = \mathrm{E}\left[i_{\boldsymbol{x}}(\boldsymbol{x})\right] \; .$$

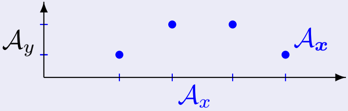
$H(\boldsymbol{x}) = H(x_1, \dots, x_n)$ is also called the joint entropy of the rvs $x_1, \dots, x_n$

Joint and single entropies

We want to relate $H(\boldsymbol{x}) = H(x_1, \ldots, x_n)$ and $H(x_1), \ldots, H(x_n)$. Start with two variables, $\boldsymbol{x} = [x, y]$.

Proposition (lower bound)

1. *If y is a.s. a function of x then $H(x, y) = H(x)$.*
2. *Otherwise, $H(x, y) > H(x)$.*



Proof.

Skip

1. in this case $i_{\boldsymbol{x}}(a, b) = i_x(a)$, for all $[a, b] \in \mathcal{A}_{\boldsymbol{x}}$ and we get

$$H(x, y) = \mathrm{E}\left[i_{\boldsymbol{x}}(x, y)\right] = \mathrm{E}\left[i_x(x)\right] = H(x)$$

□

Joint and single entropies

Proof.

Skip

2. If y is not a function of x, there exist points $[a, b] \in \mathcal{A}_{\boldsymbol{x}}$ with $p_{\boldsymbol{x}}(a, b) < p_x(a)$. For such points $i_{\boldsymbol{x}}(a, b) > i_x(a)$.
   For all other points in $\mathcal{A}_{\boldsymbol{x}}$, $i_{\boldsymbol{x}}(a, b) \geq i_x(a)$.

$$
\begin{aligned}
H(x, y) &= \sum_{[a,b] \in \mathcal{A}_{\boldsymbol{x}}} p_{\boldsymbol{x}}(a, b) i_{\boldsymbol{x}}(a, b) \\
&> \sum_{[a,b] \in \mathcal{A}_{\boldsymbol{x}}} p_{\boldsymbol{x}}(a, b) i_x(a) = \mathrm{E}\left[i_x(x)\right] = H(x)
\end{aligned}
$$

by interpreting $i_x(x)$ as a function of the rve $[x, y]$. $\square$

Joint and single entropies

Proposition (upper bound)

1. *If x and y are statistically independent, then $H(x, y) = H(x) + H(y)$.*
2. *Otherwise, $H(x, y) < H(x) + H(y)$.*

Proof.

Skip

1. $\{x = a\}$ and $\{y = b\}$ are statistically independent

$$i_{\boldsymbol{x}}(a, b) = i_x(a) + i_y(b) \quad , \quad \text{for all } [a, b] \in \mathcal{A}_{\boldsymbol{x}}$$

and by the linearity of expectation

$$H(x, y) = \mathrm{E}\left[i_{\boldsymbol{x}}(x, y)\right] = \mathrm{E}\left[i_x(x)\right] + \mathrm{E}\left[i_y(y)\right] = H(x) + H(y)$$

$\square$

Joint and single entropies

Proof.

Skip

2. We will show that $H(x, y) - H(x) - H(y) < 0$. Write it as

$$
\begin{aligned}
H(x, y) - H(x) - H(y) &= \mathrm{E}\left[i_{\boldsymbol{x}}(x, y) - i_x(x) - i_y(y)\right] \\
&= \mathrm{E}\left[\log_2 \frac{p_x(x)p_y(y)}{p_{\boldsymbol{x}}(x, y)}\right]
\end{aligned}
$$

Since x, y are not independent, the rv $z = \frac{p_x(x)p_y(y)}{p_{\boldsymbol{x}}(x,y)}$ is not a.s. constant. Apply Jensens' inequality ($\log_2$ concave)

$$
\begin{aligned}
\mathrm{E}\left[\log_2 \frac{p_x(x)p_y(y)}{p_{\boldsymbol{x}}(x, y)}\right] &< \log_2 \mathrm{E}\left[\frac{p_x(x)p_y(y)}{p_{\boldsymbol{x}}(x, y)}\right] \\
&= \log_2 \sum_{[a,b]\in\mathcal{A}_{\boldsymbol{x}}} p_x(a)p_y(b)
\end{aligned}
$$

Joint and single entropies

Proof (continued).

Since $\mathcal{A}_{\boldsymbol{x}} \subset \mathcal{A}_x \times \mathcal{A}_y$

$$\begin{aligned}
\log_2 \sum_{[a,b]\in\mathcal{A}_{\boldsymbol{x}}} p_x(a)p_y(b) &\leq \log_2 \sum_{[a,b]\in\mathcal{A}_x\times\mathcal{A}_y} p_x(a)p_y(b) \\
&= \log_2 \left(\sum_{a\in\mathcal{A}_x} p_x(a)\right)\left(\sum_{b\in\mathcal{A}_y} p_y(b)\right) \\
&= \log_2 1 = 0
\end{aligned}$$

□

Summary and generalization to n variables

We obtained the bounds for the joint entropy

$$\underbrace{\max\{H(x), H(y)\}}_{\substack{\uparrow \\ \text{one a function} \\ \text{of the other}}} \leq H(x,y) \leq \underbrace{H(x)+H(y)}_{\substack{\uparrow \\ \text{statistically} \\ \text{independent}}}$$

The generalization to n variables gives

$$\max_i \{H(x_i)\} \leq H(x_1, \ldots, x_n) \leq \sum_{i=1}^{n} H(x_i)$$

Conditional information and entropy

Starting from a discrete rve $\boldsymbol{x} = [x, y]$ and the conditional statistical description of x given y, we can give the following

Definition

Conditional information of x given y

$$i_{x|y} \ : \ \mathcal{A}_{\boldsymbol{x}} \mapsto \mathbb{R} \quad , \quad i_{x|y}(a|b) = \log_2 \frac{1}{p_{x|y}(a|b)}$$

Definition

Conditional entropy of x given y

$$H(x|y) = \mathrm{E}\left[i_{x|y}(x|y)\right] \ .$$

Conditional information and entropy

Since $i_{x|y}(x|y)$ is a function of both rvs x and y, the expectation must be taken with respect to their joint statistical description

$$H(x|y) = \sum_{[a,b]\in\mathcal{A}_{\boldsymbol{x}}} p_{\boldsymbol{x}}(a,b) i_{x|y}(a|b) = \sum_{[a,b]\in\mathcal{A}_{\boldsymbol{x}}} p_{\boldsymbol{x}}(a,b) \log_2 \frac{1}{p_{x|y}(a|b)} \ .$$

Observe that the conditional PMD is used in the logarithm and the joint PMD in the expectation.

Conditional, joint & single entropies

Proposition

Given a discrete rve $\boldsymbol{x} = [x, y]$, *we have*

$$i_{x|y}(a|b) = i_{\boldsymbol{x}}(a, b) - i_y(b) \quad , \quad H(x|y) = H(x, y) - H(x) \ .$$

Proof.

Skip

$$i_{x|y}(a|b) = \log_2 \frac{1}{p_{x|y}(a|b)} = \log_2 \frac{p_y(b)}{p_{xy}(a, b)}$$
$$= \log_2 \frac{1}{p_{xy}(a, b)} - \log_2 \frac{1}{p_y(b)} = i_{\boldsymbol{x}}(a, b) - i_y(b)$$

Substitute a, b with rvs x, y and take expectations

$$\mathrm{E}\left[i_{x|y}(x|y)\right] = \mathrm{E}\left[i_{\boldsymbol{x}}(x, y)\right] - \mathrm{E}\left[i_y(y)\right]$$

$\square$

Bounds for the conditional entropy

From the bounds for the joint entropy we get

$$\underset{\substack{\uparrow \\ x \text{ a function of } y}}{0} \leq H(x|y) \leq \underset{\substack{\uparrow \\ \text{statistically} \\ \text{independent}}}{H(x)}$$

We can therefore think of $H(x|y)$ as a measure of the average information (uncertainty) carried by x once we know y.

Mutual information

Definition

The mutual information between two rvs x and y is

$$I(x,y) = H(x) + H(y) - H(x,y) \ .$$

Properties

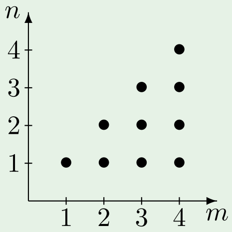
1. $I(x,y) = H(x) - H(x|y)$ is the difference between the *a priori* uncertainty on x, and the uncertainty on x once we know y.
2. $I(x,y) = I(y,x)$ is symmetrical

Bounds

$$\underset{\substack{\uparrow \\ \text{statistically} \\ \text{independent}}}{0} \leq I(x,y) \leq \underset{\substack{\uparrow \\ \text{one a function} \\ \text{of the other}}}{\min\{H(x), H(y)\}}$$

Calculation Example

Example



$\mathcal{A}_{[x,y]} = \left\{[m,n] \in \mathbb{Z}^2 \,:\, 1 \le n \le m \le 4\right\}$

$p_{x,y}(m,n) = \frac{1}{4m} \quad \Rightarrow \quad i_{x,y}(m,n) = \log_2 m + 2$

marginal x:

$p_x(m) = \sum_n p_{x,y}(m,n) = \frac{1}{4}\sum_{n=1}^{m}\frac{1}{m} = \frac{1}{4}\;,\;\forall m$

$\Rightarrow \quad i_x(m) = 2 \quad , \qquad H(x) = 2\,\text{bit}$

conditional, y given x:

$i_{y|x}(n|m) = i_{x,y}(m,n) - i_x(m) = \log_2 m$

$H(y|x) = \mathrm{E}\left[\log_2 x\right] = \sum_{m=1}^{4}\frac{1}{4}\log_2 m \simeq 1.15\,\text{bit}$

joint entropy: $H(x,y) = H(x) + H(y|x) \simeq 3.15\,\text{bit}$

marginal y: $p_y(1) = \frac{25}{48}\;,\; p_y(2) = \frac{13}{48}\;,\; p_y(3) = \frac{7}{48}\;,\; p_y(4) = \frac{1}{16}$

$H(y) = 4 + \frac{15}{16}\log_2 3 - \frac{25}{24}\log_2 5 - \frac{7}{48}\log_2 7 - \frac{13}{48}\log_2 13 \simeq 1.66\,\text{bit}$

conditional, x given y: $H(x|y) = H(x,y) - H(y) \simeq 1.49\,\text{bit}$

mutual information: $I(x,y) = H(x) - H(x|y) \simeq 0.51\,\text{bit}$

Appendix to Lecture 5

Backup slides

Definition of information: examples

Example

We randomly pick a card out of a regular 52-card pack. The event

$$A = \{\text{The suit of the extracted card is hearts}\}$$

has probability $\mathrm{P}[A] = 1/4$. Hence its information is $i(A) = 2$ bit. The event

$$B = \{\text{The value of the extracted card is 7}\}$$

has probability $\mathrm{P}[B] = 1/13$. Hence its information is $i(B) = \log_2 13 \simeq 3.7$ bit.

Example

In a message with iid binary symbols having equally likely values 0 and 1, a 3-symbol string is observed. The event

$$A = \{\text{The observed string is '010'}\}$$

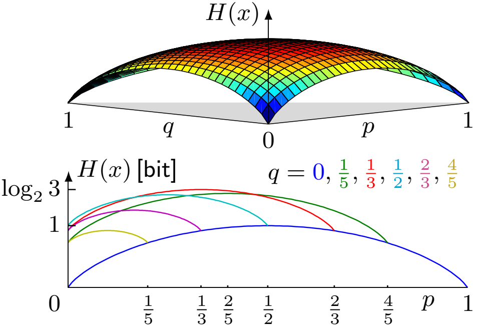
has probability $\mathrm{P}[A] = (1/2)^3 = 1/8$. Hence, $i(A) = 3$ bit. If, on the contrary, $\mathrm{P}[0] = 3/4$ and $\mathrm{P}[1] = 1/4$ (not equally likely), we have $\mathrm{P}[A] = 9/64$ and $i(A) = 6 - 2\log_2 3 \simeq 2.83$ bit.

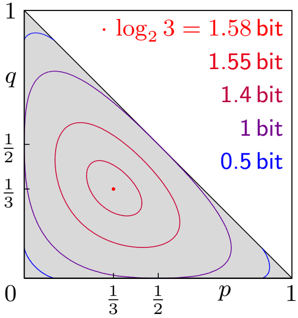
Entropy calculation

Example (ternary variable)

$$\mathcal{A}_x = \{a_1, a_2, a_3\}\ ,\ p_x(a_1) = p\ ,\ \ p_x(a_2) = q\ ,\ \ p_x(a_3) = 1 - p - q$$

$$H(x) = p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q} + (1-p-q) \log_2 \frac{1}{(1-p-q)}$$





Show next example | Return to entropy bounds

Entropy calculation

Example (geometric rv)

$\mathcal{A}_x = \{0, 1, 2, \ldots\} \quad , \quad p_x(k) = (1-p)p^k$

information of a value $k \in \mathcal{A}_x$

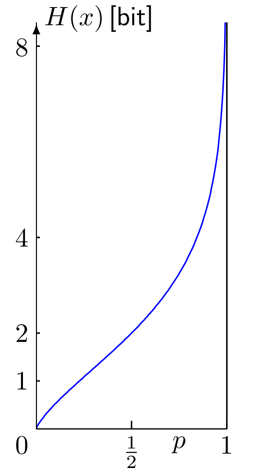
$$i_x(k) = \log_{1/2}(1-p) + k \log_{1/2} p$$

(random) information of x

$$i_x(x) = \log_{1/2}(1-p) + x \log_{1/2} p$$

take expectation with $\mathrm{E}[x] = p/(1-p)$

$$H(x) = \log_{1/2}(1-p) + \frac{p}{1-p} \log_{1/2} p$$



Return to entropy bounds   Return to nominal info