Entropy

The information content of an event σ is defined to be:

$$\log_2 \frac{1}{P(\sigma)}$$

and it is usually denoted by $I(\sigma)$ or $h(\sigma)$.
This definition is chosen because it meets several conditions:
1) If there is no randomness in an event (i.e. $P(\sigma)=1$), then its information content is 0
2) The less probable an event is, the more information it yields
3.) Information content of two independent events is the sum of the information contents of the events.

It can be shown that up to a scaling factor, $\log_2 \frac{1}{P(\sigma)}$ is the unique function with the above properties.

P1 We are rolling two fair dice. What is the information content of the event that the sum is 7. What about the event that the sum is 2.

Sol: Our sample space is $\Omega = \{(i,j) : 1 \leq i,j \leq 6\}$, and let X be the random variable defined as the of the dice, i.e. $X(i,j) = i+j$. Then

$$\{X=7\} = \{(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)\},$$ hence

$$P(X=7) = \frac{|\{X=7\}|}{|\Omega|} = \frac{6}{36} = \frac{1}{6},$$ and the information content is $h(X=7) = \log_2 \frac{1}{\frac{1}{6}} = \log_2 6 = 2.585$

For the event $\{X=2\} = \{(1,1)\}$, we have

$P(X=2) = \frac{1}{36}$, so $h(X=2) = \log_2 \frac{1}{\frac{1}{36}} = \log_2 36 = 5.1699$

Given a discrete random variable X, with possible outcomes $a_1, \dots, a_n$; then the **entropy** of X is defined as:

$$\boxed{H(X) = \sum_{i=1}^{n} P(X=a_i) \cdot \log_2 \frac{1}{P(X=a_i)}}$$

(i.e. the entropy of X is the expected value of the random variable $h(X) = \frac{1}{P(X)}$, which is the information content of X.)

P2 What is the entropy of the random variable X used in P1 (i.e. the sum of the values of two fair dice)?

Sol: We first need to compute the probability mass function of X. Possible values of X are: $2, 3, \dots, 12$.

If $2 \leq s \leq 7$, then $|\{X=s\}| = s-1$, since for every $i \in \{1, 2, \dots s-1\}$ the number $s-i$ is the unique number s.t. $X(i, s-i) = i + s - i = s$ and $s-i \in \{1, \dots 6\}$.

Similarly, for $7 \leq s \leq 12$: $|\{X=s\}| = 6-(s-7) = 13-s$, since for every $i \in \{s-7+1, \dots, 6\}$ the number $s-i$ is the unique number s.t. $s-i \in \{1, \dots, 6\}$ and $X(i, s-i) = s$.

So probability mass function of X is: $p(i) = \frac{i-1}{36}$, if $i \in \{2, 3, \dots, 7\}$ and $p(i) = \frac{13-i}{36}$ if $i \in \{8, \dots, 12\}$.

From this we can compute the entropy of X as:

$$H(X) = \sum_{s=2}^{12} P(X=s) \cdot \log_2 \frac{1}{P(X=s)} =$$

$$= \frac{1}{36}\log_2 36 + \frac{2}{36}\log_2 \frac{36}{2} + \dots + \frac{6}{36}\log_2 \frac{36}{6} + \frac{5}{36}\log_2 \frac{36}{5} + \dots + \frac{1}{36}\log_2 36$$

$$= 3.2744$$

P3 Consider the discrete probability space $\Omega = \{-4, -3, \dots, 3, 4\}$ and P is the uniform distribution on Ω. Define random variables X and Y as:

$$X(k) = k^2 \;;$$

$$Y(k) = \begin{cases} -1 & \text{if } k < 0 \\ 1 & \text{if } k \geq 0 \end{cases},$$

a) Compute the probability mass functions of X and Y.
b) Compute the entropies of X and Y
c) Compute the entropy of the joint distribution $P_{X,Y}$ of X and Y.

Sol: a) $p_X(0) = \frac{1}{9}$ and $p_X(k^2) = \frac{2}{9}$ if $k \in \{1, \dots, 4\}$

$$P_Y(-1) = \frac{4}{9} \quad \text{and} \quad P_Y(1) = \frac{5}{9}$$

b)

$$H(X) = \frac{1}{9}\log 9 + 4 \cdot \frac{2}{9}\log \frac{9}{2} = 2.281$$

$$H(Y) = \frac{4}{9}\log \frac{9}{4} + \frac{5}{9}\log \frac{9}{5} = 0.991$$

(sum is 3.272)

c) If we know k^2 and the sign of k, then we know k, so $P_{X,Y}$ is uniformly distributed over 9 points, hence

$$H(X,Y) = 9 \cdot \frac{1}{9} \log 9 = \log 9 = 3.17$$

P4 The input source to a noisy communication channel is a random variable X over the four symbols a, b, c, d. The output from the channel is a r.v. Y over the same symb. The joint distribution of these two random variables is:

	$x=a$	$x=b$	$x=c$	$x=d$
$y=a$	$1/8$	$1/16$	$1/16$	$1/4$
$y=b$	$1/16$	$1/8$	$1/16$	0
$y=c$	$1/32$	$1/32$	$1/16$	0
$y=d$	$1/32$	$1/32$	$1/16$	0

a) Write the distribution of X and compute $H(X)$
b) Do the same for Y.
c) Find $H(X,Y)$!

Sol: a) X is uniformly distr., hence $H(X) = \log_2 4 = 2$

b) $$P_Y = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$$

$$H(Y) = \frac{1}{2} + \frac{1}{2} + \frac{3}{4} = \frac{7}{4} = 1.75$$

c) $$H(X,Y) = \frac{1}{4} \cdot 2 + 2 \cdot \frac{1}{8} \cdot 3 + \cancel{6}^{3} \cdot \frac{1}{\cancel{16}_{8}} \cdot 4 + \cancel{4}^{1} \cdot \frac{1}{\cancel{32}_{8}} \cdot 5 = \frac{27}{8} = 3.375$$

P2 $\frac{1}{2}$ Calculate the probability that if somebody is „tall", that person must be a male! Suppose that 20% males are tall, and that 10% females are tall.
If you know that somebody is male, how much info you gain by learning that he is aslo tall? How much do you gain by learning that a female is tall? Finally, how much do you gain by learning that a tall person is a female?

Sol:

$$P(M \cap T) = P(T|M) \cdot P(M) = 0.2 \cdot \frac{1}{2} = 0.1$$

$$P(T) = P(T|M) \cdot P(M) + P(T|F) \cdot P(F) =$$
$$= 0.1 + 0.05 = 0.15$$

$$P(M|T) = \frac{\frac{1}{10}}{\frac{15}{100}} = \frac{\cancel{100}^{2}}{\cancel{10}\,\cancel{15}_{3}} = \frac{2}{3}$$

$$P(F|T) = \frac{\frac{5}{\cancel{100}}}{\frac{15}{\cancel{100}}} = \frac{5}{15} = \frac{1}{3}$$

$$h(P(M|T)) = \log_2 \frac{1}{\frac{2}{3}} = \log_2 \frac{3}{2} = \underbrace{\log_2 3}_{1.585} - 1 = 0.585$$

$$h(P(F|T)) = \log_2 \frac{1}{\frac{1}{3}} = \log_2 3 = 1.585$$

$$h(P(T|M)) = \log_2 5 = 2.322$$

$$h(P(T|F)) = \log_2 10 = 3.322$$

Jensen's inequality:

strictly-conv.: if = holds only for $x_1=x_2$

Let f be a convex function $\left(f((1-t)x_1 + t x_2) \leq (1-t)f(x_1) + t f(x_2) \right)$

Then if $t_1, t_2, \dots, t_n$ are positive real numbers such that $\sum t_i = 1$ and $x_1 \dots x_n$ are any points in the domain of f, then:

$$f\left(\sum_{i=1}^{n} t_i x_i\right) \leq \sum_{i=1}^{n} t_i f(x_i) \qquad (*)$$

Th: If f has a second derivative that is non-negative (positive) over an interval, then f is convex (strictly convex)

We say that f is concave if $-f$ is convex. A version of Jensen's inequality is still true, just with $\geq$ instead of $\leq$ in $(*)$, when f is concave.

Some examples:

$\log x$ is concave on $(0, +\infty)$ since:

$$(\log x)'' = \left(\frac{1}{x}\right)' = -\frac{1}{x^2} < 0 \quad \forall x \in (0, +\infty).$$

Then $-\log x = \log \frac{1}{x}$ is convex since $\log x$ is concave.

$x \log x$ is convex on the interval $(0, +\infty)$ since:

$$(x \log x)'' = \left(\log x + \cancel{x} \cdot \frac{1}{\cancel{x}}\right)' = \frac{1}{x} > 0 \text{ for } x \in (0, +\infty).$$

P6 Let X be a random variable that takes values in an alphabet $\mathcal{A}_X = \{a_1, \dots, a_n\}$ with probabilities $p_1, \dots, p_n$. Prove that:

$$H(X) \leq \log |\mathcal{A}_X| \left(= \log n\right)$$

Sol:

$$H(X) = \sum p_i \log \frac{1}{p_i}$$, now, $\log x$ is a concave function, so setting $x_i := \frac{1}{p_i}$, and $t_i := p_i$ and $f(x) = \log(x)$ in Jensen's inequality we get:

$$H(X) = \sum t_i \log x_i \leq \log\left(\sum t_i x_i\right) = \log \sum p_i \frac{1}{p_i} = \log |\mathcal{A}_X|.$$

Hence $H(X) \leq \log |\mathcal{A}_X|$.

P7 Let $P(x)$ and $Q(x)$ be two probability distr. defined over the same alphabet $\mathcal{A}_X$. Then the relative entropy or Kullback-Leibler divergence between $P(x)$ and $Q(x)$ is:

$$D_{KL}(P \| Q) = \sum_{x \in \mathcal{A}_X} P(x) \cdot \log \frac{P(x)}{Q(x)}.$$

(Gibbs' inequality)

Prove that the relative entropy satisfies $D_{KL}(P \| Q) \geq 0$.

Sol: We will prove that $-D_{KL}(P \| Q) \leq 0$. We have

$$-D_{KL}(P \| Q) = \sum_{x \in \mathcal{A}_X} P(x) \log \frac{Q(x)}{P(x)} \underset{\text{Jensen's inequality}}{\leq} \log \sum P(x) \frac{Q(x)}{P(x)} = \log 1 = 0.$$

P8 Using Gibbs' inequality solve P7.

Sol: Let $Q(x)$ be the uniform distribution over $\mathcal{A}$.
i.e. $Q(x) = \frac{1}{|\mathcal{A}|}$ $\forall x \in \mathcal{A}$. Then

$$D_{KL}(P\|Q) \geq 0$$

$$\sum P(x) \log \frac{P(x)}{Q(x)} = \sum P(x) \cdot \log |\mathcal{A}| \cdot P(x) = \log |\mathcal{A}| + \underbrace{\sum P(x) \log P(x)}_{= -H(X)}$$

$\implies$ $\log |\mathcal{A}| \geq H(X)$.