Lecture 2 Entropy

September 2, 2022

Outline

Outline

- Let E be an event belonging to a given event space and having probability $p_E := Pr(E)$, where $0 \leq p_E \leq 1$.
- $\mathcal{I}(E)$, the self-information of E: the amount of information one gains when learning that E has occurred, or equivalently, the amount of uncertainty one had about E prior to learning that it has happened.

- Let E be an event belonging to a given event space and having probability $p_E := Pr(E)$, where $0 \leq p_E \leq 1$.
- $\mathcal{I}(E)$, the self-information of E: the amount of information one gains when learning that E has occurred, or equivalently, the amount of uncertainty one had about E prior to learning that it has happened.
- Question: What properties should $\mathcal{I}(E)$ have?

The properties that $\mathcal{I}(E)$ is expected to have

- $\mathcal{I}(E)$ should be a decreasing function of p_E.
  In other words, this property states that $\mathcal{I}(E) = I(p_E)$, where $I(\cdot)$ is a real-valued function defined over $[0, 1]$.
- $I(p_E)$ should be continuous in P_E.
- If E_1 and E_2 are two independent events, then $\mathcal{I}(E_1 \cap E_2) = \mathcal{I}(E_1) + \mathcal{I}(E_2)$, or equivalently, $I(p_{E_1} \times p_{E_2}) = I(p_{E_1}) + I(p_{E_2})$.

Theorem

The only function defined over $p \in [0, 1]$ and satisfying

- *$I(p)$ is monotonically decreasing in p;*
- *$I(p)$ is a continuous function of p for $0 \leq p \leq 1$;*
- *$I(p_1 \times p_2) = I(p_1) + I(p_2)$;*

is $I(p) = -c \cdot \log_b(p)$, where c is a positive constant and the base b of the logarithm is a real number larger then one.

- The constant c above is by convention normalized to $c = 1$.
- The base b of the logarithm determines the type of units used in measuring information.
- We will use the base-2 logarithm throughout unless otherwise specified.

Outline

Entropy

Definition

The entropy $H(X)$ of a discrete random variable X with probability mass distribution or probability mass function (pmf) $P_X(\cdot)$ is defined by

$$H(X) := -\sum_{x \in \mathcal{X}} P_X(x) \cdot \log_2 P_X(x) \text{ (bits)}.$$

- $H(X)$ represents the *statistical average* (mean) amount of information one gains when learning that one gains when learning that one of its $|\mathcal{X}|$ outcomes has occurred.
- $H(X) = -E[-\log_2 P_X(X)] = E[\mathcal{I}(X)]$, where $\mathcal{I}(X) := -\log_2 P_X(x)$.
- We adopt the convention $0 \cdot \log_2 0 = 0$.

Lemma

$H(X) \geq 0$.

Proof.

$0 \leq p(x) \leq 1$ implies that $\log \frac{1}{p(x)} \geq 0$. □

Lemma

$H_b(X) = (\log_b a) H_a(X)$.

Proof.

$\log_b p = \log_b a \log_a p$. □

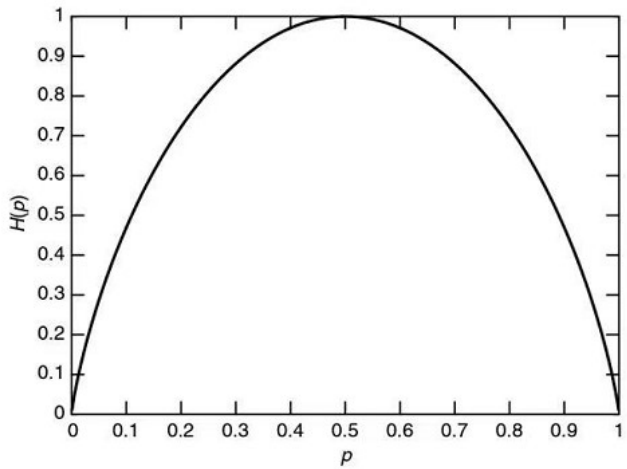
Example

Let

$$X = \begin{cases} 1 & \textit{with probability } p, \\ 0 & \textit{with probability 1-p.} \end{cases}$$

Then

$$H(X) = -p \log p - (1-p) \log(1-p) =: H(p) \ (\textit{bits}).$$



How to measure information content?

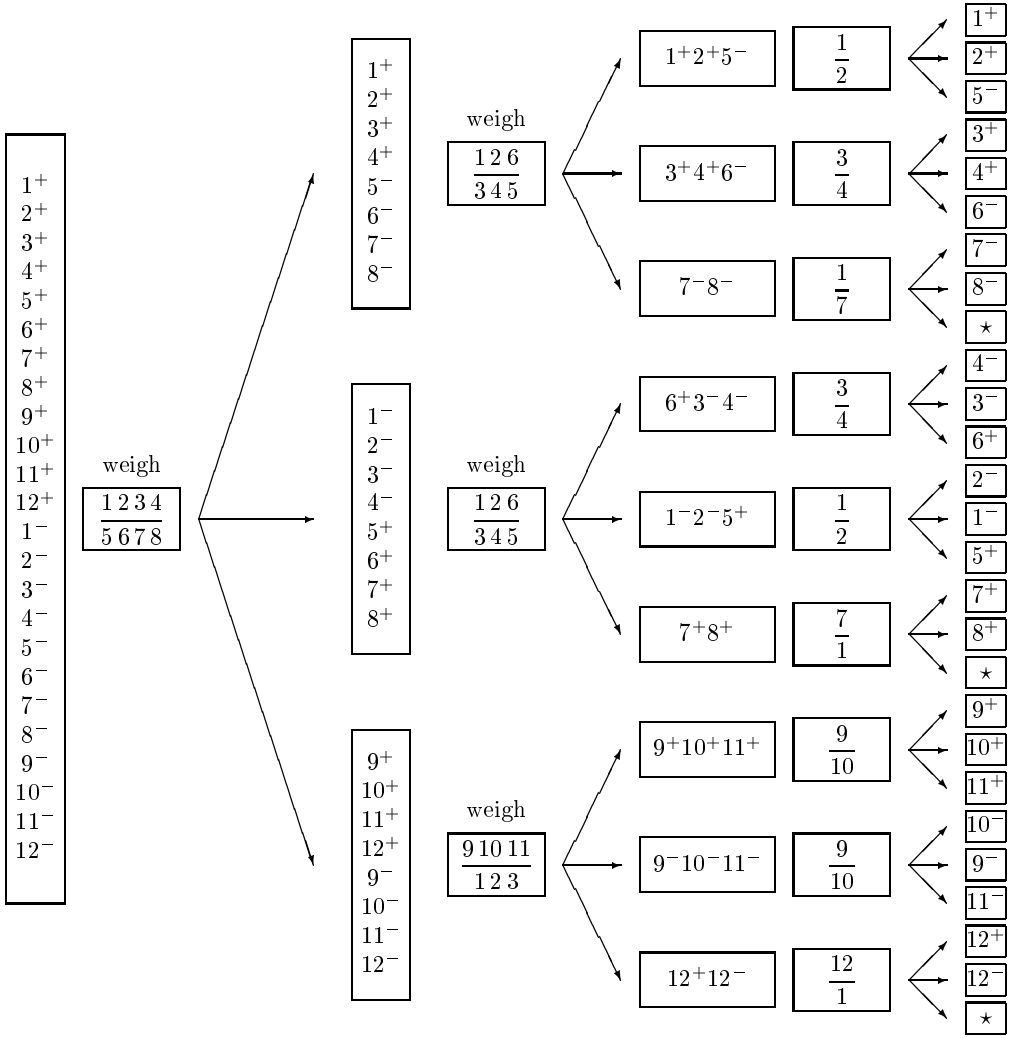
You are given 12 balls, all equal in weight expect for one that is either heavier or lighter. You are also given a two-pan balance to use. In each use of the balance you may put any number of the 12 balls on the left pan and the same number on the right pan. There are three possible outcomes: either the weights are equal, or the balls on the left are heavier, or the balls on the right are heavier.

How to measure information content?

You are given 12 balls, all equal in weight expect for one that is either heavier or lighter. You are also given a two-pan balance to use. In each use of the balance you may put any number of the 12 balls on the left pan and the same number on the right pan. There are three possible outcomes: either the weights are equal, or the balls on the left are heavier, or the balls on the right are heavier.

Your task is to design a strategy to determine which is the odd ball and whether it is heavier or lighter than the others in as few uses of the balance as possible.

- How can one measure information?
- When you have identified the odd ball and whether it is heavy or light, how much information have you gained?
- Once you have designed a strategy, draw a tree showing, for each of the possible outcomes of a weighing, what weighing you perform next. At each node in the tree, how much information have the outcomes so far given you, and how much information remains to gained?
- How much information is gained on the first step of weighing problem if 6 balls are weighed against the other 6? How much is gained if 4 are weighed against 4 on the first step, leaving out 4 balls.


$1^+ 2^+ 3^+ 4^+ 5^+ 6^+ 7^+ 8^+ 9^+ 10^+ 11^+ 12^+ 1^- 2^- 3^- 4^- 5^- 6^- 7^- 8^- 9^- 10^- 11^- 12^-$
weigh
$\frac{1\,2\,3\,4}{5\,6\,7\,8}$
$1^+ 2^+ 3^+ 4^+ 5^- 6^- 7^- 8^-$
$1^- 2^- 3^- 4^- 5^+ 6^+ 7^+ 8^+$
$9^+ 10^+ 11^+ 12^+ 9^- 10^- 11^- 12^-$
weigh
$\frac{1\,2\,6}{3\,4\,5}$
weigh
$\frac{1\,2\,6}{3\,4\,5}$
weigh
$\frac{9\,10\,11}{1\,2\,3}$
$1^+2^+5^-$
$3^+4^+6^-$
7^-8^-
$6^+3^-4^-$
$1^-2^-5^+$
7^+8^+
$9^+10^+11^+$
$9^-10^-11^-$
12^+12^-
$\frac{1}{2}$
$\frac{3}{4}$
$\frac{1}{7}$
$\frac{3}{4}$
$\frac{1}{2}$
$\frac{7}{1}$
$\frac{9}{10}$
$\frac{9}{10}$
$\frac{12}{1}$
$1^+ 2^+ 5^- 3^+ 4^+ 6^- 7^- 8^- \star$
$4^- 3^- 6^+ 2^- 1^- 5^+ 7^+ 8^+ \star$
$9^+ 10^+ 11^+ 10^- 9^- 11^- 12^+ 12^- \star$

Outline

Joint entropy

Definition

The *joint entropy* $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y),$$

which can also be expressed as

$$H(X,Y) = -E\log p(X,Y).$$

Conditional entropy

Definition

If $(X, Y) \sim p(x, y)$, the conditional entropy $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -E \log p(Y|X). \end{aligned}$$

Chain rule

Theorem (Chain rule)

$$H(X,Y) = H(X) + H(Y|X). \tag{3.1}$$

Proof.

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)p(y|x)) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= H(X) + H(Y|X). \quad \square
\end{aligned}
$$

Equivalently, we can write

$$\log p(X, Y) = \log p(X) + \log p(Y|X)$$

and take the expectation of both sides of the equation to obtain the theorem.

Corollary

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z).$$

Example

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

Let (X, Y) have the above joint distribution. The marginal distributions of X and Y are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ and $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ respectively, and hence $H(X) = \frac{7}{4}$ bits, $H(Y) = 2$ bits. Also

$$H(X|Y) = \sum_{i=1}^{4} p(Y = i) H(X|Y = i) = \frac{11}{8} \text{ bits.}$$

Similarly $H(Y|X) = \frac{13}{8}$ bits, and $H(X, Y) = \frac{27}{8}$ bits.

Outline

Relative entropy

Definition

The *relative entropy or Kullback-Leibler distance* between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(x)}{q(x)}.$$

Mutual information

Definition

Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The *mutual information* $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y) \parallel p(x)p(y)) \\ &= E_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}. \end{aligned}$$

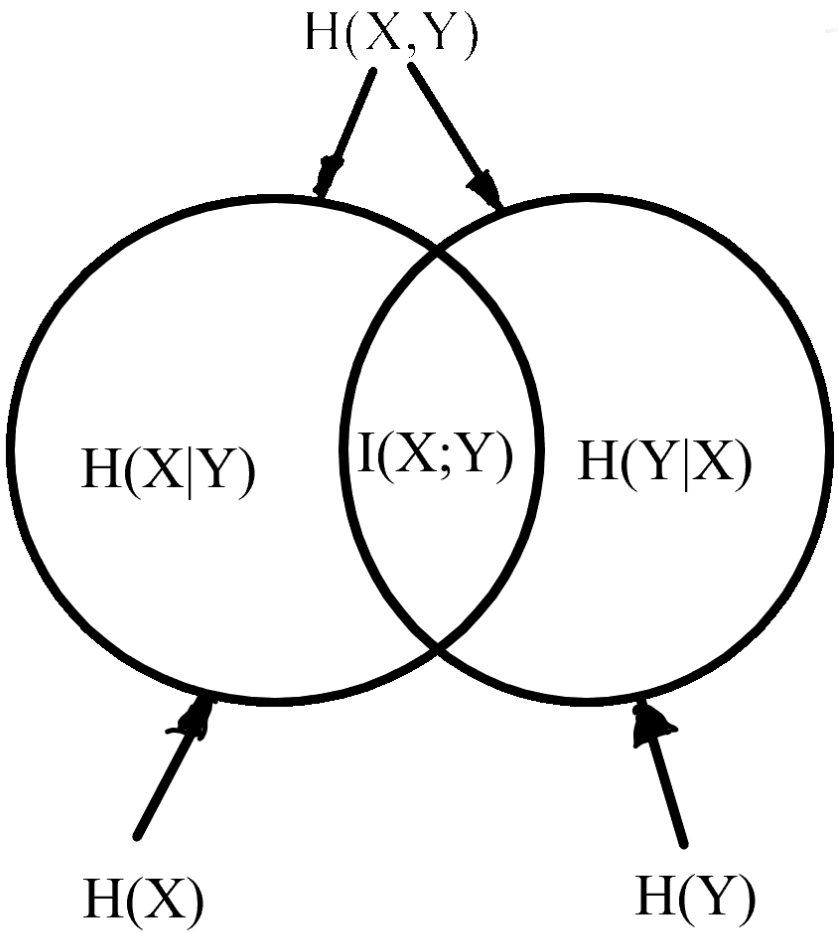
The definition of mutual information can be rewritten as

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\ &= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\ &= -\sum_{x} p(x) \log p(x) - (-\sum_{x,y} p(x,y) \log p(x|y)) \\ &= H(X) - H(X|Y). \end{aligned}$$

Similarly,

$$I(X;Y) = H(Y) - H(Y|X).$$

Relationship between entropy and mutual information



Proposition

The mutual information between a random variable X and itself is equal to the entropy of X, i.e., $I(X;X) = H(X)$.

Example

Let $\mathcal{X} = \{0, 1\}$, *and consider two distributions* p q *on* $\mathcal{X}$. *Let* $p(0) = 1 - r,\ p(1) = r$ *and* $q(0) = 1 - s,\ q(1) = s$. *Then*

$$D(p\|q) = (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s}$$

and

$$D(q\|p) = (1-s)\log\frac{1-s}{1-r} + r\log\frac{s}{r}.$$

If $r = s$ *then* $D(p\|q) = D(q\|p) = 0$. *Note that in general* $D(p\|q) \neq D(q\|p)$. *For example, if* $r = \frac{1}{2},\ s = \frac{1}{4}$, *then*

$$D(p\|q) = 0.2075 \text{ bits},\ D(q\|p) = 0.1887 \text{ bits}.$$

Outline

Theorem

Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1).$$

conditional mutual information

Definition

The *conditional mutual information* of random variables X and Y given Z is defined by

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}. \end{aligned}$$

Chain rule for mutual information

Theorem

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \cdots, X_1).$$

Proof.

$$
\begin{aligned}
& I(X_1, X_2, \cdots, X_n; Y) \\
= \; & H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n | Y) \\
= \; & \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1) - \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1, Y) \\
= \; & \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \cdots, X_1). \quad \square
\end{aligned}
$$

conditional relative entropy

Definition

For joint probability mass functions $p(x, y)$ and $q(x, y)$, the *conditional relative entropy* $D(p(y|x)\|q(y|x))$ is the average of the relative entropies between the conditional probability mass function $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$. More precisely,

$$\begin{aligned} D(p(y|x)\|q(y|x)) &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}. \end{aligned}$$

Chain rule for relative entropy

Theorem (Chain rule for relative entropy)

$$D(p(x,y)\|q(x,y)) = D(p(x)\|q(x)) + D(p(y|x)\|q(y|x)).$$

Proof.

$$
\begin{aligned}
& D(p(x,y)\|q(x,y)) \\
= & \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
= & \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\
= & \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)} \\
= & D(p(x)\|q(x)) + D(p(y|x)\|q(y|x)). \quad \square
\end{aligned}
$$