Information

Welcher Artikel hat bessere Chancen, es auf die Titelseite zu schaffen?

- Bayern München besiegt 1860 München
- 1860 München besiegt Bayern München

⇒ Je unwahrscheinlicher ein Ereignis ist, desto informativer ist es.

Informationsgehalt: $I(x) = -log_2\ p(x)$

Beispiel: Der Informationsgehalt des Ergebnisses eines Münzwurfes beträgt
$-\log 0.5 = 1\text{Bit}$

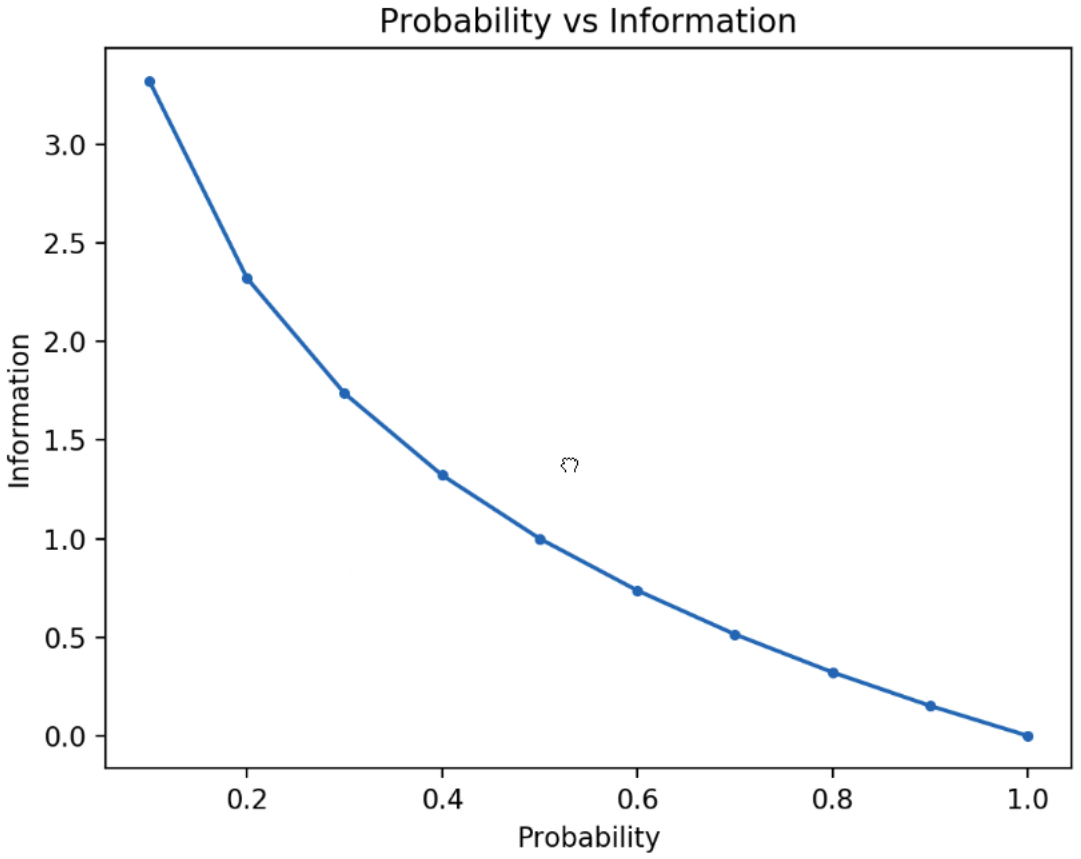
Bit ist die Maßeinheit der Information.

Wiederholung
what is an event x?
- an event is a set of results (Ergebnisse) of an observation.
- x is the set of results that are mapped by the randomvariable X to the value x

Information content of an event tells about how **surprising/more informative** the event is.
An event is surprising when it happens rare, meaning, low probablity.
“The "information content" can be viewed as how much useful information the message actually contains.”

p(head) = 0.5 , I(head) = 1 Bit (the information of the event that we get head when tossing a coin is 1 Bit
p(tail) = 0.5, I(tail) = 1 Bit

https://brilliant.org/wiki/entropy-information-theory/



Plot of Probability vs Information

Assuming we roll a die and define a random variable “Augenzahl” which can take values 1,2,3,4,5,6.

Each event has p = 1/6

What is I(1) ?

what is I(2) ?

Informationsgehalt: $I(x) = -log_2\ p(x)$

Calculate the Information for an Event

Quantifying information is the foundation of the field of information theory.

The intuition behind quantifying information is the idea of measuring how much surprise there is in an event. Those events that are rare (low probability) are more surprising and therefore have more information than those events that are common (high probability).

- **Low Probability Event**: High Information (*surprising*).
- **High Probability Event**: Low Information (*unsurprising*).

> *The basic intuition behind information theory is that learning that an unlikely event has occurred is more informative than learning that a likely event has occurred.*

— Page 73, Deep Learning, 2016.

Rare events are more uncertain or more surprising and require more information to represent them than common events.

https://machinelearningmastery.com/what-is-information-entropy/

We can calculate the amount of information there is in an event using the probability of the event. This is called "*Shannon information*," "*self-information*," or simply the "*information*," and can be calculated for a discrete event *x* as follows:

- information(x) = -log(p(x))

Where *log()* is the base-2 logarithm and *p(x)* is the probability of the event *x*.

The negative sign ensures that the result is always positive or zero.

Information will be zero when the probability of an event is 1.0 or a certainty, e.g. there is no surprise.

Running the example prints the probability of the event as 50% and the information content for the event as 1 bit.

```
p(x)=0.500, information: 1.000 bits
```

If the same coin was flipped n times, then the information for this sequence of flips would be n bits.

If the coin was not fair and the probability of a head was instead 10% (0.1), then the event would be more rare and would require more than 3 bits of information.

```
p(x)=0.100, information: 3.322 bits
```

Entropy:

A way quantify how much information there is in a random variable.

$$H(X) = H(p) = -\sum_{x \in X} p(x) \log_2 p(x)$$

The "entropy" can be viewed as how much **useful information** a message is **expected** to contain.

It also provides a lower bound for the "size" of an encoding scheme, in the sense that the expected number of bits to be transmitted under some encoding scheme is **at least** the entropy of the message.

An encoding scheme that manages to achieve this lower bound is called lossless.

A Short Introduction to Entropy, Cross-Entropy and KL-Divergence (https://www.youtube.com/watch?v=ErfnhcEV1O8)
Shannon Entropy and Information Gain (https://www.youtube.com/watch?v=9r7FIXEAGvs)

Entropy

AAAAAAAA Entropy = 0

BAADABAC Entropy = 1.75

DBCADACB Entropy = 2

Sequence 1

P(A) = 1

AAAAAAAA

No Question is needed, since the letter is an “A”

Entropy = 0

Entropy: Average number of yes/no questions we need to ask to guess what letter we picked

Entropie

Die **Entropie** misst, wieviel Information ein Zufallsereignis im Mittel enthält.

Entropie einer Zufallsvariablen X mit der Verteilungsfunktion $p(x)$:

$$H(X) = H(p) = -\sum_{x \in X} p(x) \log_2 p(x) = E(log_2 \frac{1}{p(x)})$$

Beispiel: Die Entropie beim Wurf eines Würfels beträgt
$-6 \cdot 1/6 \cdot \log_2 1/6 = \log_2 6 = 2,58 Bit$

Bezug zur Kodierungstheorie: Die Entropie ist eine Untergrenze für die Zahl der Bits, die im Mittel benötigt werden, um die Ergebnisse einer Folge von unabhängigen, identisch verteilten Zufallsvariablen zu kodieren.

What is the formula for Entropy?
What information do we need to know in order to calculate an Entropy of a random variable?

What is the formula for Entropy?
What information do we need to know in order to calculate an Entropy of a random variable? #die Wahrschienlichkeitverteilung von X

$$H(X) = H(p) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Übung:
1. Die Entropie beim Wurf eines Würfels beträgt ?
x ? 1,2,3,4,5.6
p(1) = 1

2. Die Entropie beim Wurf einer Münze beträgt

Gegeben eine Zufallsvariable X mit x= a, e, i, o (x soll eigentlich reellen Zahlen sein, die a, e,i,o representiert), und die WK-Verteilung p(x), und -log_2 p(x).
“X= a” bedeutet das Ereignis, das “a” vorkommt
Berechne die Entropy von dieser Zufallsvariable H(X)

x	p(x)	$-\log_2 p(x)$
a	0.5	1
e	0.25	2
i	0.125	3
o	0.125	3

x	p(x)	$-\log_2 p(x)$
a	0.5	1
e	0.25	2
i	0.125	3
o	0.125	3

$$H(p) = -\sum_x p(x) logp(x) = 0.5*1+0.25*2+ = .125*3+0.125*3 = 1.75$$

Die Entropy kann die minimale Bits für die Codierung der Ereignisse sagen.
Anwendung:
Wenn wir eine Codierung haben, wie können die Erwartungswert der Codelänge mit der Entropy vergleichen. Wenn unsere Codierung braucht viel mehr Bit als die Entropy, dann ist sie nicht optimal.

Gemeinsame Entropie

Die **gemeinsame Entropie** zweier Zufallsvariablen ist wie folgt definiert

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y)\ log_2\ p(x,y)$$

Die Entropie H(W 1 , W 2) beim Wurf von 2 Würfeln beträgt ?

W_1 = Augenzahl des 1. Würfels
W_2 = Augenzahl des 2. Würfels

Gemeinsame Entropie

Die **gemeinsame Entropie** zweier Zufallsvariablen ist wie folgt definiert

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y)\ log_2\ p(x,y)$$

Die Entropie H(W 1 , W 2) beim Wurf von 2 Würfeln beträgt ?

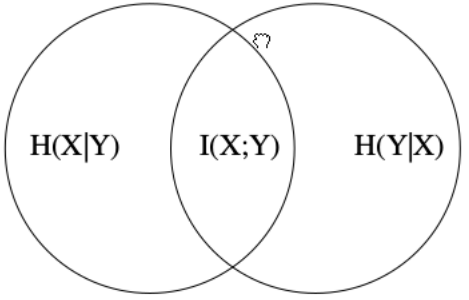
W_1 = Augenzahl des 1. Würfels
W_2 = Augenzahl des 2. Würfels

$$-36 \cdot 1/36 \cdot \log_2 1/36 = \log_2 36 = 5,17 \text{ Bit}$$

Mutual Information

Wegen $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$

gilt auch $H(X) - H(X|Y) = H(Y) - H(Y|X) =: I(X;Y)$



Die Mutual Information $I(X;Y)$ ist die "Schnittmenge" der Informationgehalte der beiden Zufallsvariablen X unf Y.

Relative Entropie

Die **Relative Entropie** (Kullback-Leibler-Abstand) zwischen zwei Verteilungsfunktionen $p(x)$ und $q(x)$ ist wie folgt definiert:

$$D(p||q) = \sum_{x \in X} p(x) \, log_2 \, \frac{p(x)}{q(x)}$$

Dieses Abstandsmaß zwischen Verteilungen gibt an, wieviele Bits im Mittel verschenkt werden, wenn Ereignisse mit einer Verteilung p mit einem Code kodiert werden, der optimal für die Verteilung q ist.

Die relative Entropie ist nie negativ und genau dann 0, wenn $p = q$.

Bezug zur Mutual Information: MI misst, wie weit die gemeinsame Verteilung $p(x, y)$ zweier Zufallsvariablen entfernt ist von einer statistisch unabhängigen Verteilung $p(x)p(y)$:

$$I(X; Y) = D(p(x, y) \, || \, p(x)p(y))$$

Crossentropie

Die **Crossentropie** zwischen zwei Verteilungen p und q

$$\begin{aligned} H(p,q) &= -\sum_x p(x)\ log_2\ q(x) \\ &= E_p(\log_2 \frac{1}{q(x)}) \\ &= H(p) + D(p||q) \end{aligned}$$

Die **Crossentropie** eines Korpus $x_1^n \overset{\circlearrowleft}{=} x_1 x_2 ... x_n$ wird folgendermaßen definiert:

$$H(x_1^n, p) = -\frac{1}{n}\ log_2\ p(x_1^n)$$

Die **Perplexität** ist eng mit der Crossentropie verwandt:

$$perp(x_1^n, p) = 2^{H(x_1^n, p)} = p(x_1^n)^{-\frac{1}{n}}$$