Information Entropy

A walk-through

Richard Torkar

Department of Computer Science and Engineering
Chalmers and University of Gothenburg
Gothenburg, Sweden

torkarr@chalmers.se

December 29, 2021

CHALMERS

Properties a measure of uncertainty should possess

Properties a measure of uncertainty should possess

- Continuous.
- Increase as number of possible events increases.
- Additive.

Properties a measure of uncertainty should possess

- Continuous.
- Increase as number of possible events increases.
- Additive.

$$H(X) = -\mathrm{E}\log(X = x_i) = -\sum_{i=1}^{n} P(X = x_i)\log P(X = x_i)$$

Properties a measure of uncertainty should possess

- Continuous.
- Increase as number of possible events increases.
- Additive.

$$H(X) = -\mathrm{E}\log(X = x_i) = -\sum_{i=1}^{n} P(X = x_i)\log P(X = x_i)$$

The uncertainty contained in a probability distribution is the average log-probability of an event.

But why logarithms?

I have selected a number between 0 and 31.[1]

- Can you guess which one?

[1] @TivadarDanka provided this thought experience.

But why logarithms?

I have selected a number between 0 and 31.[1]

- Can you guess which one?
- You can ask as many questions as you want.

[1] @TivadarDanka provided this thought experience.

But why logarithms?

I have selected a number between 0 and 31.[1]

- Can you guess which one?
- You can ask as many questions as you want.
- What is the minimum number of questions you have to ask to be 100% sure?

[1] @TivadarDanka provided this thought experience.

Let's do the math

We could pick between $0, \dots, 31$, i.e., 32 numbers. Let's rewrite as an equation,

$$\log_2(32) = x$$

Proof.

If x and b are positive real numbers, and b does not equal 1, then $\log_b(x) = y$ is equivalent to $b^y = x$,

$\square$

$$2^x = 32$$

Create equivalent expressions in the equation that all have equal bases,

$$2^x = 2^5$$

Since the bases are the same, the two expressions are only equal if the exponents are also equal,

$$x = 5$$

- In other words, we need to take the base two logarithm of 32 to get the number of questions required, i.e., 5.

- In other words, we need to take the base two logarithm of 32 to get the number of questions required, i.e., 5.
- This logic applies to all numbers! If we pick a number between 0 and $n - 1$, you need $\log(2, n)$ questions to find the number.

Since we are nerds...

If we write down the answers in a row, we effectively encode the numbers in n bits.

0 00000

1 00001

2 00010

...

31 11111

Each 'code' is simply the number in base two.

But why sums?

- No matter which number we picked, 5 questions are needed to find it.
- The average number of bits needed is also 5.

But why sums?

- No matter which number we picked, 5 questions are needed to find it.
- The average number of bits needed is also 5.

However, we have an assumption here,

Corollary

Each number is picked with an equal probability.

What if this is not the case?

New thought experiment

- Let's say we're picking between 0, 1, and 2, but we're picking 0 50% of the time, while 1 and 2 only 25% of the time.

New thought experiment

- Let's say we're picking between 0, 1, and 2, but we're picking 0 50% of the time, while 1 and 2 only 25% of the time.
- How many bits do we need now?

New thought experiment

- Let's say we're picking between 0, 1, and 2, but we're picking 0 50% of the time, while 1 and 2 only 25% of the time.
- How many bits do we need now?
- We should put this into mathematical form!

New thought experiment

- Let's say we're picking between 0, 1, and 2, but we're picking 0 50% of the time, while 1 and 2 only 25% of the time.
- How many bits do we need now?
- We should put this into mathematical form!
- Let's denote the number we pick with X. This is a random variable.

New thought experiment

- Let's say we're picking between 0, 1, and 2, but we're picking 0 50% of the time, while 1 and 2 only 25% of the time.
- How many bits do we need now?
- We should put this into mathematical form!
- Let's denote the number we pick with X. This is a random variable.

$$P(X = 0) = 1/2$$
$$P(X = 1) = 1/4$$
$$P(X = 2) = 1/4$$

New thought experiment

- Let's say we're picking between 0, 1, and 2, but we're picking 0 50% of the time, while 1 and 2 only 25% of the time.
- How many bits do we need now?
- We should put this into mathematical form!
- Let's denote the number we pick with X. This is a random variable.

$$P(X = 0) = 1/2$$
$$P(X = 1) = 1/4$$
$$P(X = 2) = 1/4$$

But we can be a bit more 'careful' with our 'bits', i.e., which question to ask.

New thought experiment

- Let's say we're picking between 0, 1, and 2, but we're picking 0 50% of the time, while 1 and 2 only 25% of the time.
- How many bits do we need now?
- We should put this into mathematical form!
- Let's denote the number we pick with X. This is a random variable.

$$P(X = 0) = 1/2$$
$$P(X = 1) = 1/4$$
$$P(X = 2) = 1/4$$

But we can be a bit more 'careful' with our 'bits', i.e., which question to ask.
Consider this,

1st Q did you pick 0? If the answer is 'Yes', the 2nd question is not needed. If not, we proceed!

2nd Q did you pick 1? No matter what the answer is, we know the solution! 'Yes' implies 1, 'No' implies 2.

Let's do some math

$$
\begin{aligned}
\text{avg. num. bits} &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) \\
&= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 \\
&= \frac{1}{2} + 2\frac{1}{4} + 2\frac{1}{4} \\
&= \frac{3}{2}
\end{aligned}
$$

The general case

- Suppose we pick between $x_1, x_2, \ldots, x_n$.
- We pick x_k with probability p_k.

The general case

- Suppose we pick between $x_1, x_2, \ldots, x_n$.
- We pick x_k with probability p_k.

Then the number of questions needed to find k is the base two logarithm of $1/p_k$,

$$\begin{aligned}\text{Num. bits needed to find } x_k &= \log_2\left(\frac{1}{p_k}\right)\\ &= -\log_2 p_k\end{aligned}$$

Conclusion

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \log P(X = x_i)$$

Conclusion

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \log P(X = x_i)$$

- The entropy of a random variable is simply the average bits of information needed to guess its value successfully.
- Even though the formula might seem complicated, its meaning is simple.
- Please see [Shannon '48] for the original formulations.

References

Claude Shannon
A Mathematical Theory of Communication
1948, Bell Labs Journal