Shannon's Entropy

Part I – The Problem

Rivki Gadot and Dvir Lanzberg

Shannon's entropy *
Part 1, the problem

Given a list of events and their probabilities, how much information is there in knowing which event happened?

- What is "information"?
- Why talk about events and not just one event?
- Why should we care?

Rivki Gadot and Dvir Lanzberg

<SKILLZ>
לומדים. משחקים. יוצרים.

Shannon asked the following question: Given a list of events and their probabilities, how much information is there in knowing which event happened?
But this question raises other questions. Like *
What is the exact definition of information? It is a little vague.*
Why ask "how much information is there in knowing the outcome of a coin toss" and not "how much information is there in knowing that the outcome of the toss is tails?" and *
Why bother? Why is this "information" thing so important?
We are NOT going to answer any of these questions.

Given a list of events and their probabilities, how much *information* is there in knowing which event happened?

Rivki Gadot and Dvir Lanzberg

We are going to stick to Shannon's question.
To make "information"…

Given a list of events and their probabilities, how much *information* is there in knowing which event happened?

Rivki Gadot and Dvir Lanzberg

<SKILL7>
לומדים. משחקים. יוצרים.

…a litte less vague we can think of…

Given a list of events and their probabilities, how much *surprise* is there in knowing which event happened?

- We toss a coin. How much surprise can there be in the outcome?
- How surprised can we get by knowing which permutation of an array is sorted?
- August in Tel-Aviv. It is noon. Probably it is sunny. There is only a tiny chance it is raining. What is the expected value of our surprise?

Rivki Gadot and Dvir Lanzberg

<SKILL7>
לומדים. משחקים. יוצרים.

Surprise. Which is a "psychological" term and we can intuitively grasp it *
How much surprise is there in finding out the outcome of a coin toss? *
How much surprise is there in finding out which permutation of an array is sorted? *
How much surprised can we get by knowing the weather in Tel-Aviv in August.

Shannon's Entropy

Part II – The Solution

Rivki Gadot and Dvir Lanzberg

Shannon's entropy *
Part two. The solution *

Shannon showed that, given n events (the probability of each event is p_i , $1 \leq i \leq n$ and the sum of all p_i is 1) :

$$H(p_1, \ldots, pn) = -\sum_{i=1}^{n} p_i \log p_i$$

Shannon showed that, given n events
The surprise, the information, (which is also called the entropy) H of the event's probabilities
is minus the sum of plogp over all the events *
This is a very strange formula. Where does it come from?

But first, some examples.

But first, some examples.

But first, some examples.

We toss a coin. How much surprise can there be in the outcome?

$$H(0.5,0.5) = -(0.5\log 0.5 + 0.5\log 0.5) = -\log 0.5 = \log 2$$

Rivki Gadot and Dvir Lanzberg

The entropy of a coin toss is
Minus half log half plus half log half that equals log two
The base of the log is arbitrary. We usually choose it to be two.
Pause the clip and make sure the math comes out alright.

But first, some examples.

How surprised can we get by knowing which permutation of an array is sorted?

$$H\left(\frac{1}{n!}, \ldots, \frac{1}{n!}\right) = -\left(\frac{1}{n!}\log\frac{1}{n!} + \cdots + \frac{1}{n!}\log\frac{1}{n!}\right) = -\log\frac{1}{n!} = \log(n!)$$

How surprised can we get by knowing which permutation of an array is sorted?
The entropy of n factorial events with equal probabilities is
Log of n factorial
Again, pause and make sure the math comes out alright.

But first, some examples.

August in Tel-Aviv. It is noon. Probably it is sunny. There is only a tiny chance it is raining. What is the expected value of our surprise?

The log is in base 2

$$H(0.999,0.001) = -(0.999 \log 0.999 + 0.001 \log 0.001) = 0.0085$$

August in Tel-Aviv. It is noon. Probably it is sunny. How surprised can we get by the weather?
If the probability of "sunny" is 0.999 (which makes the probability of "rainy" 0.001) and we take the log in base 2
We can see the entropy is very low: 0.0085.
Not much surprise there.
Pause the clip and check the math.

Shannon showed that, given n events (the probability of each event is p_i , $1 \leq i \leq n$ and the sum of all p_i is 1), the amount of information in knowing which event happened is:

$$H(p_1, \ldots, pn) = -\sum_{i=1}^{n} p_i \log p_i$$

How did he show that?

So, how did Shannon get to this strange formula?

Shannon's Entropy

Part III – Three Properties

Rivki Gadot and Dvir Lanzberg

Shannon's entropy *
Part three - three properties

Shannon required three properties of H:

- H should be continuous in the p_i
- $H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ should be a monotonic increasing function of n
- If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Shannon required three properties of H: *

H should be continuous. Small changes in the probabilities should not result in huge changes in the entropy *

The entropy of n event with equal probabilities, should grow with n. more events, larger entropy.*

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Shannon required three properties of H:

- H should be continuous in the p_i
- $H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ should be a monotonic increasing function of n
- If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

This demands some examples….

- If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

$P(\text{warm and dry})=1/6$ $P(\text{cold and dry})=1/6$

$P(\text{warm and rainy})=1/6$ $P(\text{cold and rainy})=1/2$

$$H\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{2}\right)$$

*

*

he entropy of the weather is the entropy of one sixth, one sixth, one sixth and a half…

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

$P(\text{warm and dry})=1/6$ $P(\text{cold and dry})=1/6$

$P(\text{warm and rainy})=1/6$ $P(\text{cold and rainy})=1/2$

$$H\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{2}\right) = H\left(\frac{1}{3}, \frac{2}{3}\right)$$

But it is also the entropy of finding whether the weather is warm or cold

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

$P(\text{warm and dry})=1/6$ $P(\text{cold and dry})=1/6$

$P(\text{warm and rainy})=1/6$ $P(\text{cold and rainy})=1/2$

$$H\left(\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{2}\right)=H\left(\frac{1}{3},\frac{2}{3}\right)+\frac{1}{3}H\left(\frac{1}{2},\frac{1}{2}\right)+\frac{2}{3}H\left(\frac{1}{4},\frac{3}{4}\right)$$

Plus the weighted entropies of finding if it is dry or rainy (given it is warm or cold)

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Three classes of 25, 30 and 35 students.
One student is chosen randomly.

$H(\frac{1}{90}, \ldots, \frac{1}{90})$

The entropy of choosing 1 student out of 90 is *
H of 90 events of equal probabilities of 1 divided by 90 *

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Three classes of 25, 30 and 35 students.
One student is chosen randomly.

$$H\left(\frac{1}{90}, \ldots, \frac{1}{90}\right) = H\left(\frac{25}{90}, \frac{30}{90}, \frac{35}{90}\right)$$

Or the entropy of choosing a class

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Three classes of 25, 30 and 35 students.
One student is chosen randomly.

$$H\left(\frac{1}{90}, \ldots, \frac{1}{90}\right) = H\left(\frac{25}{90}, \frac{30}{90}, \frac{35}{90}\right) + \frac{25}{90}H\left(\frac{1}{25}, \ldots, \frac{1}{25}\right) + \frac{30}{90}H\left(\frac{1}{30}, \ldots, \frac{1}{30}\right) + \frac{35}{90}H\left(\frac{1}{35}, \ldots, \frac{1}{35}\right)$$

Plus the weighted entropies of choosing a student from the chosen class

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Throwing m dice.

Rivki Gadot and Dvir Lanzberg

Throwing m dice

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Throwing m dice.

$$H\left(\frac{1}{6^m}, \frac{1}{6^m}, \ldots, \frac{1}{6^m}\right) =$$

There are 6 to the power of m possible outcomes and the entropy is
H of 6 to the power of m events with probability 1 divided by 6 to the power of m

If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Throwing m dice.

$$H\left(\frac{1}{6^m}, \frac{1}{6^m}, \ldots, \frac{1}{6^m}\right) =$$

$$H\left(\frac{1}{6}, \ldots, \frac{1}{6}\right) + \ldots + H\left(\frac{1}{6}, \ldots, \frac{1}{6}\right) = m \cdot H\left(\frac{1}{6}, \ldots, \frac{1}{6}\right)$$

But it is also m times the entropy of throwing one dice

Shannon's Entropy

Part IV – Equal Probabilities

Rivki Gadot and Dvir Lanzberg

Shannon's entropy *
Part four – equal probabilities

First, we discuss n events with equal probabilities.

$$H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = A(n)$$

Shannon started by considering n events with equal probabilities.
(then he moved to discuss rational probabilities, and after that he considered the general case of real probabilities)
the discussion of equal probabilities is mathematically complicated
So embrace yourselves.
We will call the entropy of n events with equal probabilities A of n

First, we discuss n events with equal probabilities.

$$H\left(\frac{1}{n},\ldots,\frac{1}{n}\right)=A(n)$$

If $n=s^m$ then by (3):

$$A(s^m)=mA(s)$$

If n is equal s to the power of m then A of s to the power of m
Is equal m times A of s (its like the example of throwing m dice we discussed before)

First, we discuss n events with equal probabilities.

$$H\left(\frac{1}{n},\ldots,\frac{1}{n}\right)=A(n)$$

If $n=s^m$ then by (3):

$$A(s^m)=mA(s)$$

We will call this X

We will call this X

Given s and t, we can choose n (as big as we want) and choose m accordingly such that:

$$s^m \leq t^n \leq s^{m+1}$$

Given s and t, we can choose n (as big as we want) and choose m accordingly such that:
T to the power of n is “sandwiched” between s to the power of m and s to the power of m plus 1.

Given s and t, we can choose n (as big as we want) and choose m accordingly such that:

$$s^m \leq t^n \leq s^{m+1}$$

Take log and divide by $n \log s$:

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n}$$

If we take log and divide by n log s we get this…
Pause and make sure the math comes out alright.

Given s and t, we can choose n (as big as we want) and choose m accordingly such that:

$$s^m \leq t^n \leq s^{m+1}$$

Take log and divide by $n \log s$:

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n}$$

For positive ε, as small as we want, we can choose a big enough n, so that:

$$\left| \frac{m}{n} - \frac{\log t}{\log s} \right| < \varepsilon$$

Because n is as big as we want, m divided by n can be as close
As we want to log of t divided by log of s
We remember that and start all over again…

Given s and t, we can choose n (as big as we want) and choose m accordingly such that:

$$s^m \leq t^n \leq s^{m+1}$$

This is where we started *

Monotonicity:

$$s^m \leq t^n \leq s^{m+1}$$

Using the monotonicity of H
If the left term is no bigger than the middle term
and the middle term is no bigger than the right term

Monotonicity:

$$A(s^m) \leq A(t^n) \leq A(s^{m+1})$$

Then we can show that.
And go on like the previous slides..

Monotonicity:

$$A(s^m) \leq A(t^n) \leq A(s^{m+1})$$

Use X and divide by $n\,\mathrm{A}(s)$:

$$\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n}$$

For ε, as small as we want, we can choose a big enough *n*, so that:

$$\left|\frac{m}{n} - \frac{A(t)}{A(s)}\right| < \varepsilon$$

And conclude that m divided by n can be as close as we want to a of t divided by a of s
Pause and make sure the math comes out all right.
Let's look at the two results….

Let's summarize:

For ε, as small as we want, we can choose a big enough n, so that:

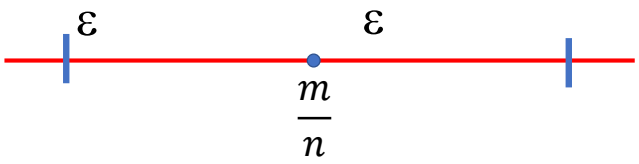
$$\left|\frac{m}{n}-\frac{A(t)}{A(s)}\right|<\varepsilon \qquad \left|\frac{m}{n}-\frac{\log t}{\log s}\right|<\varepsilon$$

Both A of t divided by A of s, and log t divided by log of s are as close to m divided by n as we want

Let's summarize:

For ε, as small as we want, we can choose a big enough n, so that:

$$\left|\frac{m}{n}-\frac{A(t)}{A(s)}\right|<\varepsilon \qquad \left|\frac{m}{n}-\frac{\log t}{\log s}\right|<\varepsilon$$

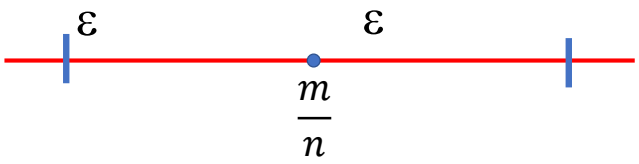


That means they are both in the epsilon neighborhood of n divided by m
And the distance between them

Let's summarize:

For ε, as small as we want, we can choose a big enough n, so that:

$$\left|\frac{m}{n}-\frac{A(t)}{A(s)}\right|<\varepsilon \qquad \left|\frac{m}{n}-\frac{\log t}{\log s}\right|<\varepsilon$$



$$\left|\frac{A(t)}{A(s)}-\frac{\log t}{\log s}\right|<2\varepsilon$$

Rivki Gadot and Dvir Lanzberg

Is as small as we want…

$$\left|\frac{A(t)}{A(s)} - \frac{\log t}{\log s}\right| < 2\varepsilon$$

$$A(t) = K \log t$$

$$H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = K \log n$$

K must be positive – monotonicity.

And because all these functions are continuous we can conclude that A of t must be klogt
Which is what we wanted to show.
Usually we take k to be 1 and the base of the log 2

Shannon's Entropy

Part V – Rational Probabilities

Rivki Gadot and Dvir Lanzberg

Shannon's entropy *
Part 5 – rational probabilities

We look now at the (almost) general case:

The probabilities are rational numbers.

We look now at rational probabilities. Real probabilities will be treated soon.
This is much simpler than equal probabilities.

We look now at the (almost) general case:

The probabilities are rational numbers.

Represent the probabilities as fractions with the same denominator:
(numerators and denominator are natural numbers)

$$p_i = \frac{q_i}{Q} \quad (1 \le i \le n)$$

Calculate:

$$H(p_1, p_2, \ldots, p_n)$$

We want to know what is the entropy if all probabilities are q sub i divided by Q
Where q sub I and Q are positive integers.
But let's look at something else…

Q events with the same probability,

Divided into n groups of q_i events $(1 \leq i \leq n)$.

Q events with the same probabilities divided into n groups – just like the example of choosing a student.

Q events with the same probability,
divided into n groups of q_i events $(1 \le i \le n)$.

Choose, randomly, one of the groups.
Choose, randomly, an event from the chosen group.
The expected amount of surprise is:

$$H\left(\frac{q_1}{Q}, \ldots, \frac{q_n}{Q}\right) + \sum_{i=1}^{n} \frac{q_i}{Q} A(q_i)$$

One way to find the entropy of these Q events is
To choose, randomly, one of the groups.
And then to choose, randomly, an event from the chosen group.

Q events with the same probability,
divided into n groups of q_i events ($1 \le i \le n$).

Choose, randomly, one of the groups.
Choose, randomly, an event from the chosen group.
The expected amount of surprise is:

$$H\left(\frac{q_1}{Q}, \ldots, \frac{q_n}{Q}\right) + \sum_{i=1}^{n} \frac{q_i}{Q} A(q_i)$$

Choose, randomly, one of the Q events.
The expected amount of surprise is:

$$A(Q)$$

A second way is to choose one of the Q events
(we saw all that in the past).
We can write this equation…

$$H\left(\frac{q_1}{Q}, \ldots, \frac{q_n}{Q}\right) + \sum_{i=1}^{n} \frac{q_i}{Q} A(q_i) = A(Q)$$

And from it find the entropy of probabilities q sub 1 divded by Q, q sub 2 divided by Q … and so on.

$$H\left(\frac{q_1}{Q},\dots,\frac{q_n}{Q}\right)+\sum_{i=1}^{n}\frac{q_i}{Q}A(q_i)=A(Q)$$

$$H(p_1,\dots,p_n)=A(Q)-\sum_{i=1}^{n}\left(p_iA(q_i)\right)=K\log Q-K\sum_{i=1}^{n}p_i\log q_i$$

$$=K\log Q\sum_{i=1}^{n}p_i-K\sum_{i=1}^{n}p_i\log q_i=K\sum_{i=1}^{n}p_i\log Q-K\sum_{i=1}^{n}p_i\log q_i$$

$$=-K\sum_{i=1}^{n}p_i\log\frac{q_i}{Q}=-K\sum_{i=1}^{n}p_i\log p_i$$

It takes some algebra but it is quite simple if we remember
The three equations in red…

$$H\left(\frac{q_1}{Q},\dots,\frac{q_n}{Q}\right)+\sum_{i=1}^{n}\frac{q_i}{Q}A(q_i)=A(Q)$$

$$H(p_1,\dots,p_n)=A(Q)-\sum_{i=1}^{n}\left(p_iA(q_i)\right)=K\log Q-K\sum_{i=1}^{n}p_i\log q_i$$

$$=K\log Q\sum_{i=1}^{n}p_i-K\sum_{i=1}^{n}p_i\log q_i=K\sum_{i=1}^{n}p_i\log Q-K\sum_{i=1}^{n}p_i\log q_i$$

$$=-K\sum_{i=1}^{n}p_i\log\frac{q_i}{Q}=-K\sum_{i=1}^{n}p_i\log p_i$$

$$\boldsymbol{p_i=\frac{q_i}{Q}\quad(1\le i\le n)}$$

$$\boldsymbol{A(t)=K\log t}$$

$$\boldsymbol{\sum_{i=1}^{n}p_i=1}$$

Rivki Gadot and Dvir Lanzberg

Pause the clip and make sure the math is correct.

Shannon's Entropy

Part VI – Real Probabilities

Shannon's entropy *
Part 6 – real probabilities

Every real number can be "sandwiched" between two rational numbers, so the distance between them is as small as we want.

The entropy is continuous (one of the properties Shannon demanded)

Shannon's formula for entropy holds for rational probabilities

So, Shannon's formula for entropy will hold, also, for real probabilities

Every real number can be "sandwiched" between two rational numbers, so the distance between them is as small as we want. *
The entropy is continuous (one of the properties Shannon demanded) *
Shannon's formula for entropy holds for rational probabilities *
So, Shannon's formula for entropy will hold, also, for real probabilities
That was simple so we have time to recap

Summary

We started with a question *

Summary

Given a list of events and their probabilities, how much information is there in knowing which event happened?

Given a list of events and their probabilities, how much information is there in knowing which event happened?

Summary

Given a list of events and their probabilities, how much information is there in knowing which event happened?

1. H should be continuous in the p_i
2. $H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ should be a monotonic increasing function of n
3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Then we defined three properties the function of information (or entropy) must have

Summary

Given a list of events and their probabilities, how much information is there in knowing which event happened?

1. H should be continuous in the p_i
2. $H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ should be a monotonic increasing function of n
3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H

For equal probabilities, rational probabilities and real probabilities:

$$H(p_1, \ldots, pn) = -K \sum_{i=1}^{n} p_i \log p_i$$

And then we showed that for equal probabilities, rational probabilities and real probabilities
The only function that has those three properties is the minus sum of p log p function (with any positive K and any base for the log)