

Information Theory

Information theory is the study of the transmission of bits across a noisy channel.

Suppose I have an alphabet $\{A, B, C, D\}$ and I want to encode it in binary. I can use:

- $A = \langle 0, 0 \rangle$
- $B = \langle 0, 1 \rangle$
- $C = \langle 1, 0 \rangle$
- $D = \langle 1, 1 \rangle$

Now, suppose someone told me that in the messages I send, A is more common than the others:

- $p(A) = \frac{1}{2}$
- $p(B) = \frac{1}{4}$
- $p(C) = \frac{1}{8}$
- $p(D) = \frac{1}{8}$

Can we do better? Sure?

- $A = \langle 0 \rangle$
- $B = \langle 1, 0 \rangle$
- $C = \langle 1, 1, 0 \rangle$
- $D = \langle 1, 1, 1 \rangle$

This is clearly unambiguous and gets us an average of 1.75 bits/character.

The minimum number of bits is the *entropy*

$$H(X) = - \sum_x p(X = x) \log_2 p(X = x)$$

Zero entropy means deterministic, high entropy means close to uniform.

$H(Y|X = x)$ is the number of bits needed to send Y , given that both the sender and recipient knew $X = x$. (Specific conditional entropy.) It's the same as entropy, but computed only over data points where $X = x$.

The full conditional entropy $H(Y|X)$ is the *expected* specific conditional entropy:

$$H(Y|X) = \sum_x p(X = x) H(Y|X = x)$$

Information gain $IG(Y|X)$ is: i must send Y — how many bits would I save if both ends knew X ?

$$IG(Y|X) = H(Y) - H(Y|X)$$

Intuitively, X has a high information gain with respect to Y if, knowing X , it takes many fewer bits to transmit Y .

A few random notes for IG with respect to decision trees:

- We usually don't need to actually compute $H(Y)$ because it's a constant for all features.
- It doesn't matter what log you use so long as you're consistent (it's just a multiplicative constant).