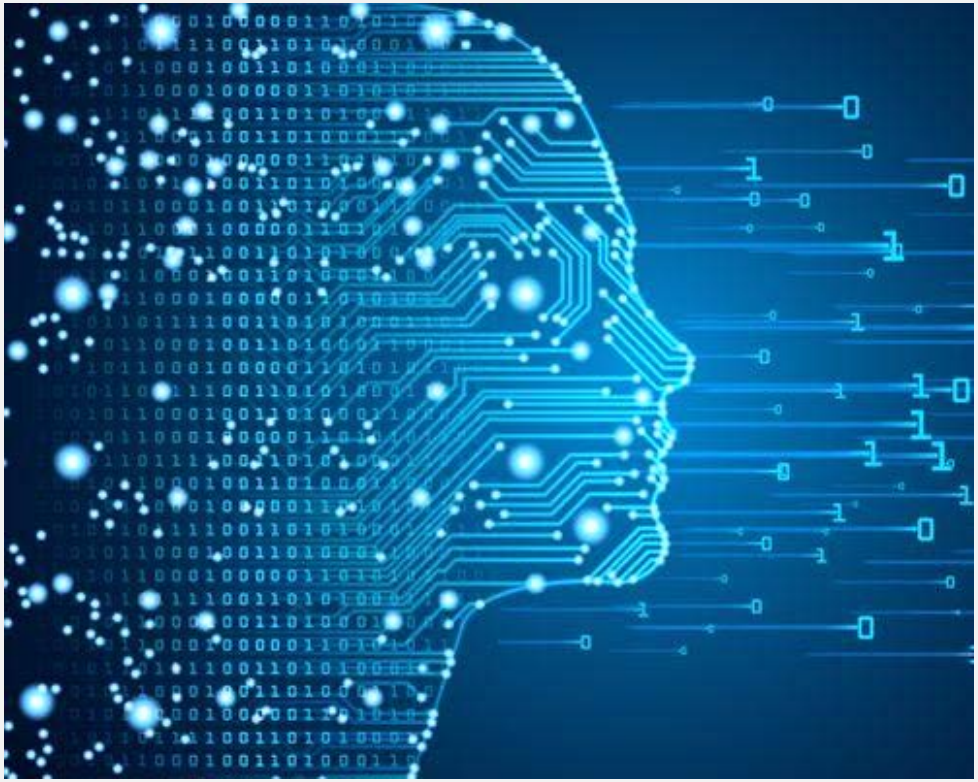
Computational Foundations of Informatics



Vasant G. Honavar
Edward Frymoyer Professor of Information Sciences and Technology
Artificial Intelligence Research Laboratory
Informatics Graduate Program
Computer Science and Engineering Graduate Program
Bioinformatics and Genomics Graduate Program
Neuroscience Graduate Program
Data Sciences Undergraduate Program
Center for Big Data Analytics and Discovery Informatics
Huck Institutes of the Life Sciences
Institute for Computational and Data Sciences
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University

Sudha Murty Distinguished Visiting Chair of Neurocomputing and Data Science
Indian Institute of Science

Elements of Information Theory

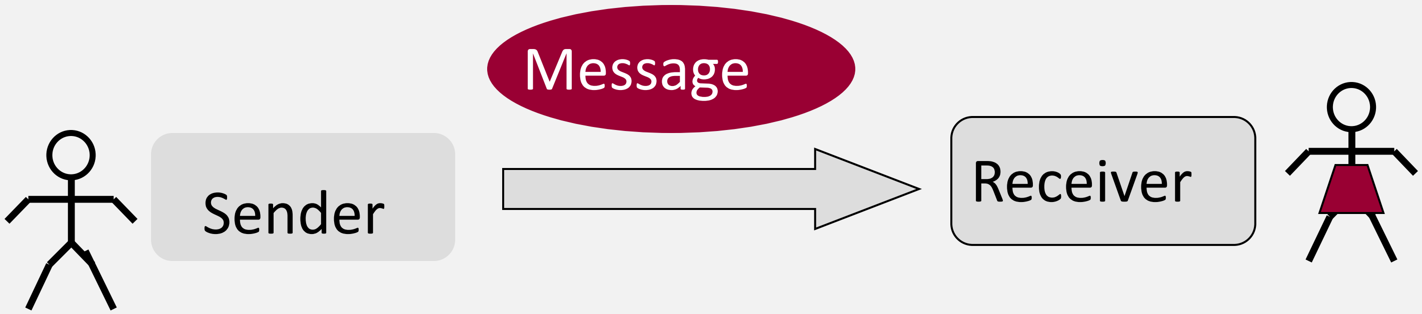
Father of Digital Communication

The roots of modern digital communication stem from the ground-breaking paper “A Mathematical Theory of Communication” by **Claude Elwood Shannon** in 1948.

What is information?

- How can we measure the information content of messages?
- Is information subjective?
- How can we transmit messages efficiently?
- How can we do so in the presence of noise?

Measuring information content of descriptions



- How much information does a message contain?
- If my message to you describes a scenario that you expect with certainty, the information content of the message for you is zero
- The more surprising the message to the receiver, the greater the amount of information conveyed by the message
- What does it mean for a message to be surprising?

Measuring information content of descriptions

- Suppose I have a coin with heads on both sides and you know that I have a coin with heads on both sides.
- I toss the coin, and without showing you the outcome, tell you that it came up heads. How much information did I give you?
- Suppose I have a fair coin and you know that I have a fair coin.
- I toss the coin, and without showing you the outcome, tell you that it came up heads. How much information did I give you?

Measuring information content of descriptions

- Without loss of generality, assume that messages are binary – made of 0s and 1s.
- Conveying the outcome of a fair coin toss requires 1 bit of information – need to identify one out of two equally likely outcomes
- Conveying the outcome one of an experiment with 8 equally likely outcomes requires 3 bits
- Conveying an outcome of that is certain takes 0 bits
- In general, if an outcome has a probability *p*, the information content of the corresponding message is

$$I(p) = -\log_2 p \qquad I(0) = 0$$

Measuring information content of descriptions

- Suppose there are 3 agents – David, Sam, Aria, in a world where a dice has been tossed. David observes the outcome is a “6” and whispers to Sam that the outcome is “even” but Aria knows nothing about the outcome.

- Probability assigned by Sam to the event “6” is a subjective measure of Sam’s belief about the state of the world.
- Information gained by David by looking at the outcome of the dice $=\log_2 6$ bits.
- Information conveyed by David to Sam = $\log_2 6 - \log_2 3$ bits
- Information conveyed by David to Aria = 0 bits

Uncertainty and Probability

Suppose that you have a distribution

$$p_1 = \frac{1}{n}, \ldots, p_n = \frac{1}{n}$$

This is clearly very uncertain.

The other end

Consider a probability distribution like:

$$p_1 = 1, p_2 = 0, \ldots, p_n = 0.$$

We have a lot more “information.”

Conveying Information

Suppose that we want to convey the results of an election. There are 5 politicians running: Barak, Hillary, John, Mike and Maggie. It would normally take 3 bits to convey the result.

Suppose that the probabilities of winning are:

$$B : \frac{1}{2}, H : \frac{1}{4}, J : \frac{1}{8}, Mi, Ma : \frac{1}{16}$$

We can encode the results as:

$$B : 0, H : 01, J : 001, M_i : 0001, M_a : 0000$$

Which uses only $1\frac{7}{8}$ bits on the average.

What do we want?

We want a definition that satisfies the following conditions:

For a point distribution the uncertainty is 0
For a uniform distribution the uncertainty is maximized.

When we combine systems the uncertainty is **additive**

As we vary the probabilities the uncertainty changes

continuously

Entropy

$$H(p_1, \ldots, p_n) = -\sum_i p_i \log_2 p_i$$

- $H(0, 0, \ldots, 1, 0, \ldots, 0) = 0$
- $H(\frac{1}{n}, \ldots, \frac{1}{n}) = \log_2 n.$
- Clearly continuous.

Are there other candidates?

Entropy is the unique continuous function that is:

- maximized by the uniform distribution
- minimized by the point distribution
- additive when you combine systems
- and ....

Information and Shannon Entropy

- Suppose we have a message that conveys the result of a random experiment with *m* possible discrete outcomes, with probabilities

$$p_1, p_2, \cdots p_m$$

The expected information content of such a message is called the entropy of the probability distribution

$$H(p_1, p_2, ..p_m) = \sum_{i=1}^{m} p_i I(p_i)$$

$$I(p_i) = -\log_2 p_i \quad \text{provided} \quad p_i \neq 0$$

$$I(p_i) = 0 \quad \text{otherwise}$$

Random variables

A discrete probability space is a finite set Ω equipped with a probability distribution

$$Pr : \Omega \to [0, 1]$$

satisfying

$$\sum_{\omega \in \Omega} Pr(\omega) = 1.$$

A *random variable* X is a function from Ω to a finite set S.

A random variable induces a distribution on S via:

$$Pr_X(s \in S) = Pr(\{\omega : X(\omega) = s\})$$

$$Pr(X = s) = Pr(\{\omega : X(\omega) = s\})$$

Entropy of a Random Variable

$$H(X) = -\sum_{s \in S} Pr(X = s) \log_2 Pr(X = s)$$

We will just write $p(s)$ for $Pr(X = s)$ if the context is clear.

Joint Entropy

Consider a pair of random variables X, Y taking values in sets $\mathcal{X}, \mathcal{Y}$ with a joint distribution $p(x, y)$.

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

Nothing new, yet!

Conditional Entropy

$H(Y|X=x)$ is the entropy of the random variable Y *given* that you know that X is x.

The conditional entropy is just the weighted sum:

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X=x)$$

$$H(Y|X) \leq H(Y)$$

The Chain Rule

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

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

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

But

$$H(X|Y) \neq H(Y|X)$$

Mutual Information

The reduction in the uncertainty of one RV given another.

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

Recall, from the last slide, this means:

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

Hence, $\qquad I(X;Y) = I(Y;X)$

How far apart are distributions ?

We want a "distance" between distributions.

$$KL(p \mapsto q) = n \sum_{s \in S} p(s)[\log_2 p(s) - \log_2 q(s).$$

Recall that it takes $H(p)$ bits to describe a set distributed according to p. What if we used q instead?

It would require $H(p) + KL(p \mapsto q)$ bits.

Relative Entropy

The Kullback-Leibler distance is often called *relative entropy.*

Suppose $S = \{a, b\}$ and $p(a) = \frac{1}{2} = p(b)$ while $q(a) = \frac{1}{4}$ and $q(b) = \frac{3}{4}$.

$$KL(p \mapsto q) = 0.2075 \text{ and } KL(q \mapsto p) = 0.1887.$$

Relative Entropy and Mutual Information

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log\{\frac{p(x,y)}{p(x)p(y)}\}$$

which is equal to:

$$KL(p(x,y) \mapsto p(x)q(x))$$

A measure of how far you are from independence!

Some basic properties

There are chain rules for mutual information and relative entropy.

$$KL(p \mapsto q) \geq 0.$$

hence

$$I(X;Y) \geq 0.$$

And Information Theory has Applied to

- All kinds of Communications,
- Stock Market, Economics
- Game Theory and Gambling,
- Quantum Physics,
- Cryptography,
- Biology and Genetics,
- and many more…