Artificial Intelligence 2: Introduction to Information Theory

Shan He

School for Computer Science
University of Birmingham

Outline of Topics

1. What is information theory and why it matters?

2. Shannon's information measures
   - Self Information and Entropy

What is information theory

Information theory answers:

"**How to quantify information?** "

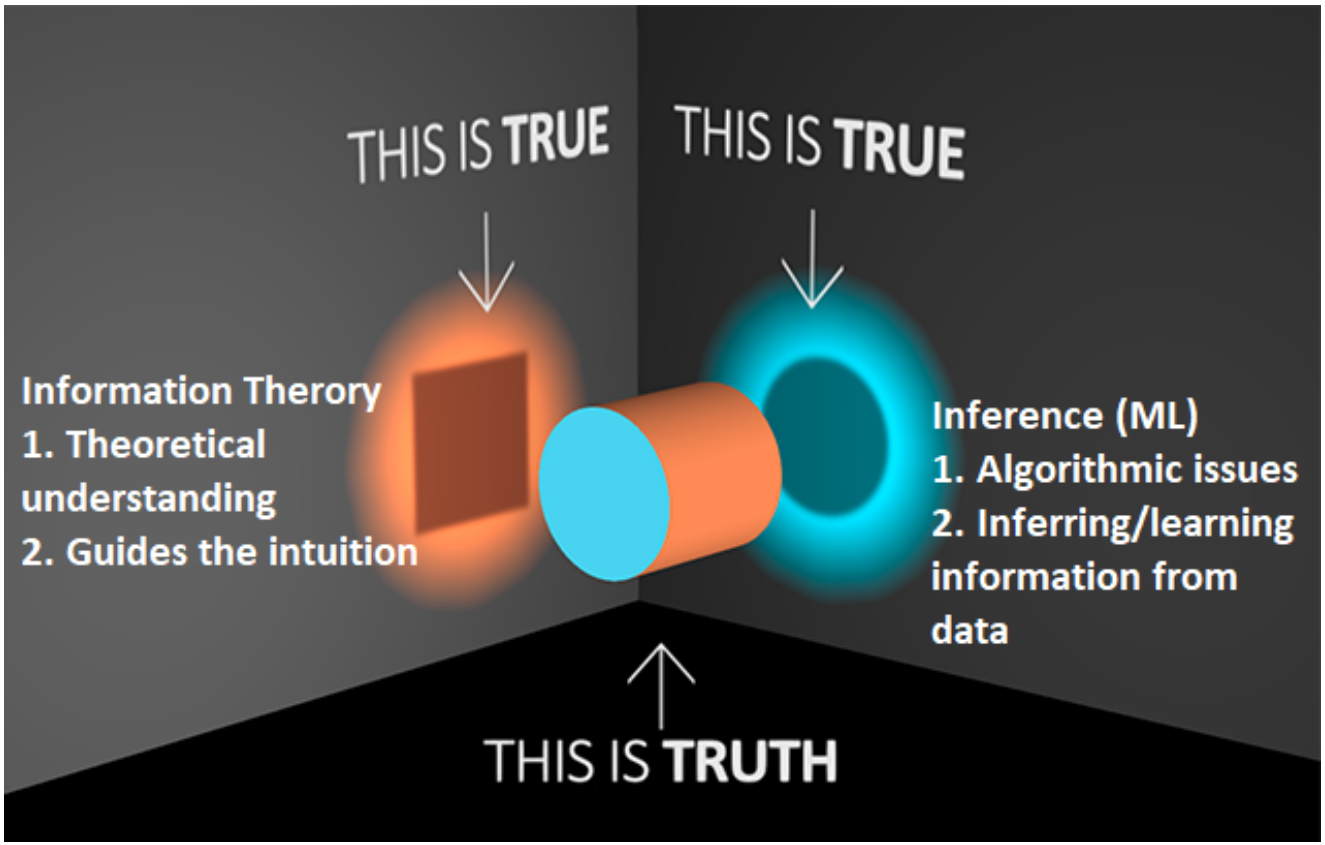
Why information theory matters?

- The foundation for practical solutions to communication problems – Shannon's original motivation
  - Source coding
  - Channel coding, i.e., error correction
  - Audio/Image/Video Compression
- Rich connections with probability and statistics, e.g., Self-information and logit, relative entropy and Kullback-Leibler divergence, etc.
- Connecting Information (AI/ML) with physics [1]
  - The maximum entropy principle
  - Maximum likelihood and maximum entropy

[1] Maximum Entropy (Most Likely) Double Helical and Double Logarithmic Spiral Trajectories in Space-Time

Why study information theory in AI2

Question: Why should we study Information Theory in AI2? Answer: Information theory and machine learning is the two side of the same coin



The book: Information Theory, Inference, and Learning Algorithms

Applications of information theory to machine learning

- Feature selection:
  - Information Theoretic Feature Selection: Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection
  - In clinical decision making, we can use information theory to choose a test that provides the most information.
- Unsupervised learning: mutual information criterion for clustering
- Supervise learning:
  - For deep learning, information bottleneck method, a method in information theory provide a plausible theoretical foundation – See Information bottleneck method
  - Decision-tree learning: information theory provides a useful criterion for choosing which property to split on – Information Gain
  - In Bayesian learning, information theory provides a basis for deciding which is the best model given some data.
  - In logistic regression and neural networks, cross entropy is used as the loss function

Information theory: motivating example

- Consider below two sentences:
  - Shan is a human.
  - Shan is a man.
- It is obvious the second sentence gives us more information.
- Questions:
  - How can we quantify the difference between two sentences?
  - Can we have a mathematical measure that tells us how much more information second sentence have as compared to the first?
- Answer from Shannon:
  - Shannon proposed that the "*semantic aspects of data are irrelevant*", i.e., the nature and meaning of data are not the information per se
  - Instead he quantified information in terms of **probability distribution and "uncertainty" (surprise)**.

Self Information

Question: How to measure the information content of an event?
Answer: Shannon proposed **self-information** based on three axioms:

- An event with probability 100% is perfectly unsurprising and yields no information.
- The less probable an event is, the more surprising it is and the more information it yields.
- If two independent events are measured separately, the total amount of information is the sum of the self-informations of the individual events.

Self Information

Self Information: Given a random variable X with probability mass function $P_X(x)$, the self-information of measuring X as outcome x is defined as

$$\mathrm{I}_X(x) = -\log_b [P_X(x)] = \log_b \frac{1}{P_X(x)}, \tag{1}$$

where different bases of the logarithm b result in different units:

- $b = 2$: bits
- $b = e$: called "natural units" or "nat",
- $b = 10$ called "dits", "bans", or "hartleys".

Self-information and log-odds

Self-information and logit (log-odds): given some event x, and $p(x)$ is the probability of x occurring, and that $p(\neg x) = 1 - p(x)$ is the probability of x not occurring. From the definition of logit:

$$\text{logit} = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$$

we have

$$\text{logit}(x) = I(\neg x) - I(x)$$

Entropy

Entropy: quantifies "*the uncertainty in a random variable X*". More formally, given a **discrete random variable** [2] X with range $R_X = \{x_1, \ldots, x_n\}$, and its probability mass function as $P_X(x)$, the entropy of X is formally defined as:

Entropy

$$H(X) \equiv E\left[I_X(x)\right] \equiv -\sum_i^n P(X = x_i) \log_b P(X = x_i)$$
$$\equiv E\left[\log_b \frac{1}{P_X(x)}\right] \equiv -E\left[\log_b P_X(x)\right] \quad (2)$$

[2] For continuous random variables we use differential entropy, see here

Example 1: Fair and biased coins

Question: Denoting the outcome of tossing a fair coin or a biased coin as X and Y, respectively. We know the biased coin comes up heads with probability of 0.7. Calculate their entropies and interpret the results
Solution:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) = -\sum_{i=1}^{2} \frac{1}{2} \log_2 \frac{1}{2} = -\sum_{i=1}^{2} \frac{1}{2} \cdot (-1) = 1$$

$$H(Y) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) = -0.7 \log_2(0.7) - 0.3 \log_2(0.3) \approx 0.8816$$

Example 2: Quantifying he information of English

Question: How to quantify the information of English?
Answer: Regarding the English language as a discrete random variable X with the range $R_X = \{1, 2, 3, \ldots, 27\}$, in which the values 1-26 represent 26 letters (a-z) occur and 27 represent the space character (to separate two words).

- Step 1: To obtain the PMF of X: we choose a book *"The Frequently Asked Questions Manual for Linux"*[3] and calculate the experimental probabilities of the 27 characters (See next page)
- Step 2: Calculate the entropy

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) = 4.11(\text{bits/letter})$$

[3] Example from Information Theory, Inference, and Learning Algorithms

x_i	character	$P_X(x_i)$	$I_X(x_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
	$\sum_i P_X(x_i) \log_2 \frac{1}{P_X(x_i)}$		4.1

The book without 'e'

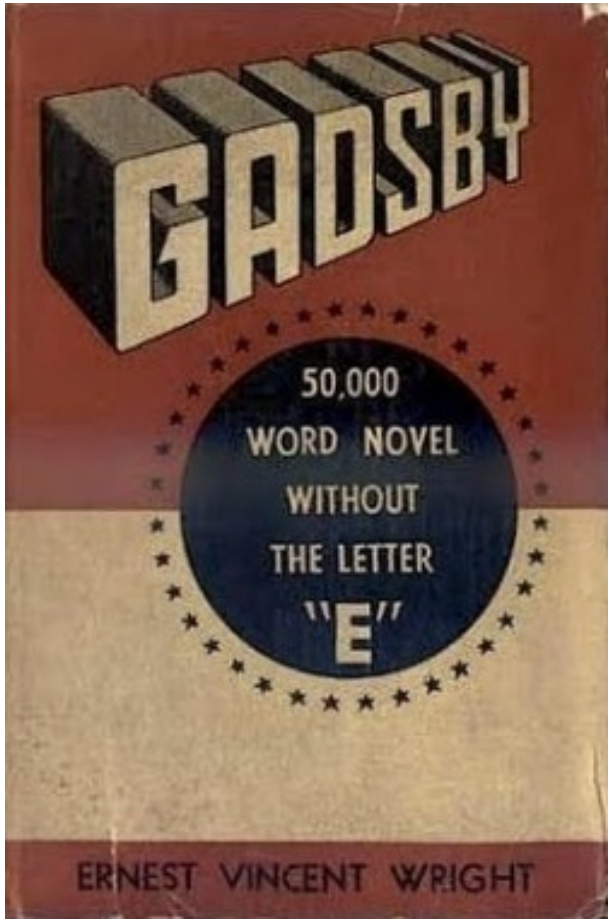


Figure: Gadsby is a 1939 novel by Ernest Vincent Wright which does not include any words that contain the letter E, the most common letter in English. From Wikipedia: Gadsby

Further readings

- Stanford Encyclopedia of Philosophy - Information Processing and Thermodynamic Entropy
- Entropy, Shannon's Measure of Information and Boltzmann's H-Theorem