

A Gentle Tutorial on Information Theory and Learning

Roni Rosenfeld

Carnegie Mellon University



Outline

- First part based very loosely on [Abramson 63].
- Information theory usually formulated in terms of information channels and coding — will not discuss those here.

1. Information

2. Entropy

3. Mutual Information

4. Cross Entropy and Learning

Information

- information \neq knowledge
Concerned with abstract possibilities, not their meaning
- information: reduction in uncertainty

Imagine:

#1 you're about to observe the outcome of a coin flip

#2 you're about to observe the outcome of a die roll

There is more uncertainty in #2

Next:

1. You observed outcome of #1 \rightarrow uncertainty reduced to zero.
2. You observed outcome of #2 \rightarrow uncertainty reduced to zero.

\implies more information was provided by the outcome in #2

Definition of Information

(After [Abramson 63])

Let E be some event which occurs with probability $P(E)$. If we are told that E has occurred, then we say that we have received

$$I(E) = \log_2 \frac{1}{P(E)}$$

bits of information.

- Base of log is unimportant — will only change the units
We'll stick with bits, and always assume base 2
- Can also think of information as amount of "surprise" in E
(e.g. $P(E) = 1, P(E) = 0$)
- Example: result of a fair coin flip ($\log_2 2 = 1$ bit)
- Example: result of a fair die roll ($\log_2 6 \approx 2.585$ bits)

Information is Additive

- $I(k \text{ fair coin tosses}) = \log \frac{1}{1/2^k} = k \text{ bits}$
- So:
 - random word from a 100,000 word vocabulary:
 $I(\text{word}) = \log 100,000 = 16.61 \text{ bits}$
 - A 1000 word document from same source:
 $I(\text{document}) = 16,610 \text{ bits}$
 - A 480x640 pixel, 16-greyscale video picture:
 $I(\text{picture}) = 307,200 \cdot \log 16 = 1,228,800 \text{ bits}$
- \implies A (VGA) picture is worth (a lot more than) a 1000 words!
- (In reality, both are gross overestimates.)

Entropy

A *Zero-memory information source* S is a source that emits symbols from an alphabet $\{s_1, s_2, \dots, s_k\}$ with probabilities $\{p_1, p_2, \dots, p_k\}$, respectively, where the symbols emitted are statistically independent.

What is the average amount of information in observing the output of the source S ?

Call this **Entropy**:

$$H(S) = \sum_i p_i \cdot I(s_i) = \sum_i p_i \cdot \log \frac{1}{p_i} = E_P \left[\log \frac{1}{p(s)} \right]$$

Alternative Explanations of Entropy

$$H(S) = \sum_i p_i \cdot \log \frac{1}{p_i}$$

1. avg amt of info provided per symbol
2. avg amount of surprise when observing a symbol
3. uncertainty an observer has before seeing the symbol
4. avg # of bits needed to communicate each symbol
(Shannon: there are codes that will communicate these symbols with efficiency arbitrarily close to $H(S)$ bits/symbol; there are no codes that will do it with efficiency $< H(S)$ bits/symbol)

Entropy as a Function of a Probability Distribution

Since the source S is fully characterized by $P = \{p_1, \dots, p_k\}$ (we don't care what the symbols s_i actually are, or what they stand for), entropy can also be thought of as a property of a probability distribution function P : the avg uncertainty in the distribution. So we may also write:

$$H(S) = H(P) = H(p_1, p_2, \dots, p_k) = \sum_i p_i \log \frac{1}{p_i}$$

(Can be generalized to continuous distributions.)

Properties of Entropy

$$H(P) = \sum_i p_i \cdot \log \frac{1}{p_i}$$

1. Non-negative: $H(P) \geq 0$

2. Invariant wrt permutation of its inputs:

$$H(p_1, p_2, \dots, p_k) = H(p_{\tau(1)}, p_{\tau(2)}, \dots, p_{\tau(k)})$$

3. For any *other* probability distribution $\{q_1, q_2, \dots, q_k\}$:

$$H(P) = \sum_i p_i \cdot \log \frac{1}{p_i} < \sum_i p_i \cdot \log \frac{1}{q_i}$$

4. $H(P) \leq \log k$, with equality iff $p_i = 1/k \ \forall i$

5. The further P is from uniform, the lower the entropy.

Special Case: $k = 2$

Flipping a coin with $P(\text{“head”})=p$, $P(\text{“tail”})=1-p$

$$H(p) = p \cdot \log \frac{1}{p} + (1 - p) \cdot \log \frac{1}{1 - p}$$

Notice:

- zero uncertainty/information/surprise at edges
- maximum info at 0.5 (1 bit)
- drops off quickly

Special Case: $k = 2$ (cont.)

Relates to: "20 questions" game strategy (halving the space).

So a sequence of (independent) 0's-and-1's can provide up to 1 bit of information per digit, provided the 0's and 1's are equally likely at any point. If they are not equally likely, the sequence provides less information *and can be compressed*.

The Entropy of English

27 characters (A-Z, space).

100,000 words (avg 5.5 characters each)

- Assuming independence between successive characters:
 - uniform character distribution: $\log 27 = 4.75$ bits/character
 - true character distribution: 4.03 bits/character
- Assuming independence between successive *words*:
 - uniform word distribution: $\log 100,000/6.5 \approx 2.55$ bits/character
 - true word distribution: $9.45/6.5 \approx 1.45$ bits/character
- True Entropy of English is much lower!

Two Sources

Temperature T : a random variable taking on values t

$$P(T=\text{hot})=0.3$$

$$P(T=\text{mild})=0.5$$

$$P(T=\text{cold})=0.2$$

$$\implies H(T)=H(0.3, 0.5, 0.2) = 1.48548$$

humidity M : a random variable, taking on values m

$$P(M=\text{low})=0.6$$

$$P(M=\text{high})=0.4$$

$$\implies H(M) = H(0.6, 0.4) = 0.970951$$

T, M not independent: $P(T = t, M = m) \neq P(T = t) \cdot P(M = m)$

Joint Probability, Joint Entropy

	cold	mild	hot	
low	0.1	0.4	0.1	0.6
high	0.2	0.1	0.1	0.4
	0.3	0.5	0.2	1.0

- $H(T) = H(0.3, 0.5, 0.2) = 1.48548$
- $H(M) = H(0.6, 0.4) = 0.970951$
- $H(T) + H(M) = 2.456431$
- **Joint Entropy:** consider the space of (t, m) events $H(T, M) = \sum_{t,m} P(T = t, M = m) \cdot \log \frac{1}{P(T=t, M=m)}$
 $H(0.1, 0.4, 0.1, 0.2, 0.1, 0.1) = 2.32193$

Notice that $H(T, M) < H(T) + H(M)$!!!

Conditional Probability, Conditional Entropy

$$P(T = t|M = m)$$

	cold	mild	hot	
low	1/6	4/6	1/6	1.0
high	2/4	1/4	1/4	1.0

Conditional Entropy:

- $H(T|M = low) = H(1/6, 4/6, 1/6) = 1.25163$
- $H(T|M = high) = H(2/4, 1/4, 1/4) = 1.5$
- **Average Conditional Entropy** (aka equivocation):
 $H(T/M) = \sum_m P(M = m) \cdot H(T|M = m) =$
 $0.6 \cdot H(T|M = low) + 0.4 \cdot H(T|M = high) = 1.350978$

How much is M telling us on average about T ?

$$H(T) - H(T|M) = 1.48548 - 1.350978 \approx 0.1345 \text{ bits}$$

Conditional Probability, Conditional Entropy

$$P(M = m|T = t)$$

	cold	mild	hot
low	1/3	4/5	1/2
high	2/3	1/5	1/2
	1.0	1.0	1.0

Conditional Entropy:

- $H(M|T = \text{cold}) = H(1/3, 2/3) = 0.918296$
- $H(M|T = \text{mild}) = H(4/5, 1/5) = 0.721928$
- $H(M|T = \text{hot}) = H(1/2, 1/2) = 1.0$
- Average Conditional Entropy (aka Equivocation):
 $H(M/T) = \sum_t P(T = t) \cdot H(M|T = t) =$
 $0.3 \cdot H(M|T = \text{cold}) + 0.5 \cdot H(M|T = \text{mild}) + 0.2 \cdot H(M|T = \text{hot}) = 0.8364528$

How much is T telling us on average about M ?

$$H(M) - H(M|T) = 0.970951 - 0.8364528 \approx 0.1345 \text{ bits}$$

Average Mutual Information

$$\begin{aligned} I(X; Y) &= H(X) - H(X/Y) \\ &= \sum_x P(x) \cdot \log \frac{1}{P(x)} - \sum_{x,y} P(x, y) \cdot \log \frac{1}{P(x|y)} \\ &= \sum_{x,y} P(x, y) \cdot \log \frac{P(x|y)}{P(x)} \\ &= \sum_{x,y} P(x, y) \cdot \log \frac{P(x, y)}{P(x)P(y)} \end{aligned}$$

Properties of Average Mutual Information:

- Symmetric (but $H(X) \neq H(Y)$ and $H(X/Y) \neq H(Y/X)$)
- Non-negative (but $H(X) - H(X/y)$ may be negative!)
- Zero iff X, Y independent
- Additive (see next slide)

Mutual Information Visualized

$$H(X, Y) = H(X) + H(Y) - I(X; Y)$$

Three Sources

From Blachman:

(" / " means "given". " ; " means "between". " , " means "and".)

- $H(X, Y/Z) = H(\{X, Y\} / Z)$
- $H(X/Y, Z) = H(X / \{Y, Z\})$
- $I(X; Y/Z) = H(X/Z) - H(X/Y, Z)$
-

$$\begin{aligned} I(X; Y; Z) &= I(X; Y) - I(X; Y/Z) \\ &= H(X, Y, Z) - H(X, Y) - H(X, Z) - H(Y, Z) + H(X) + H(Y) \end{aligned}$$

\implies Can be negative!

- $I(X; Y, Z) = I(X; Y) + I(X; Z/Y)$ (additivity)
- But: $I(X; Y) = 0, I(X; Z) = 0$ doesn't mean $I(X; Y, Z) = 0!!!$

A Markov Source

Order- k Markov Source: A source that "remembers" the last k symbols emitted.

Ie, the probability of emitting any symbol depends on the last k emitted symbols: $P(s_{T=t} | s_{T=t-1}, s_{T=t-2}, \dots, s_{T=t-k})$

So the last k emitted symbols define a *state*, and there are q^k states.

First-order markov source: defined by $q \times q$ matrix: $P(s_i | s_j)$

Example: $S_{T=t}$ is position after t random steps

Approximating with a Markov Source

A non-Markovian source can still be approximated by one.

Examples: English characters: $C = \{c_1, c_2, \dots\}$

1. Uniform: $H(C) = \log 27 = 4.75$ bits/char
2. Assuming 0 memory: $H(C) = H(0.186, 0.064, 0.0127, \dots) = 4.03$ bits/char
3. Assuming 1st order: $H(C) = H(c_i/c_{i-1}) = 3.32$ bits/char
4. Assuming 2nd order: $H(C) = H(c_i/c_{i-1}, c_{i-2}) = 3.1$ bits/char
5. Assuming large order: Shannon got down to ≈ 1 bit/char

Modeling an Arbitrary Source

Source $\mathcal{D}(Y)$ with unknown distribution $P_{\mathcal{D}}(Y)$

(recall $H(P_{\mathcal{D}}) = E_{P_{\mathcal{D}}}[\log \frac{1}{P_{\mathcal{D}}(Y)}]$)

Goal: Model (approximate) with learned distribution $P_M(Y)$

What's a good model $P_M(Y)$?

1. *RMS error* over \mathcal{D} 's parameters \Rightarrow but \mathcal{D} is unknown!
2. *Predictive Probability*: Maximize the expected log-likelihood the model assigns to future data from \mathcal{D}

Cross Entropy

$$\begin{aligned} M^* &= \arg \max_M E_{\mathbf{D}}[\log P_M(Y)] \\ &= \arg \min_M E_{\mathbf{D}}\left[\log \frac{1}{P_M(Y)}\right] \\ &= CH(P_{\mathbf{D}}; P_M) \leftarrow \text{Cross Entropy} \end{aligned}$$

The following are equivalent:

1. Maximize Predictive Probability of P_M
2. Minimize Cross Entropy $CH(P_{\mathbf{D}}; P_M)$
3. Minimize the difference between $P_{\mathbf{D}}$ and P_M (in what sense?)

A Distance Measure Between Distributions

Kullback-Liebler distance:

$$\begin{aligned} KL(P_D; P_M) &= CH(P_D; P_M) - H(P_D) \\ &= E_{P_D} \left[\log \frac{P_D(Y)}{P_M(Y)} \right] \end{aligned}$$

Properties of KL distance:

1. Non-negative. $KL(P_D; P_M) = 0 \iff P_D = P_M$
2. Generally non-symmetric

The following are equivalent:

1. Maximize Predictive Probability of P_M for distribution D
2. Minimize Cross Entropy $CH(P_D; P_M)$
3. Minimize the distance $KL(P_D; P_M)$