Basics of Information Theory

Samuel Brody

February 2009

Entropy

- $p(x)$ is a probability mass function of variable X , over a discrete set of symbols (or alphabet) $X = \{x_1, x_2, ..., x_n\}$:

$$p(x_i) = Prob(X = x_i), \quad x_i \in X$$

- The entropy (or self-information) is the average uncertainty of a single random variable:

$$H(p) = H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

Entropy - cont.

Definition:

- $H(X) = -\sum_{x \in X} p(x) log_2 p(x)$
- (weighted) average uncertainty
- the number of *yes / no* questions
- complete certainty*= 0, complete uncertainty = $log_2 |X|$

*by definition, $0 log 0 = 0$

Some Intuition

The amount of uncertainty depends on:

- the number of choices

- the distribution of probabilities

Joint & Conditional Entropy

The joint entropy of a pair of discrete random variables $X, Y \sim p(x, y)$ is the amount of information needed on average to specify both their values.

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

The conditional entropy of a discrete random variable Y given another variable X is the remaining uncertainty of Y given that the other party knows the value of X :

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

Chain Rule

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

$$H(X_1, ..., X_n) = H(X_1) + H(X_2|X_1) + ... + H(X_n|X_1, ..., X_{n-1})$$

Mutual Information

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

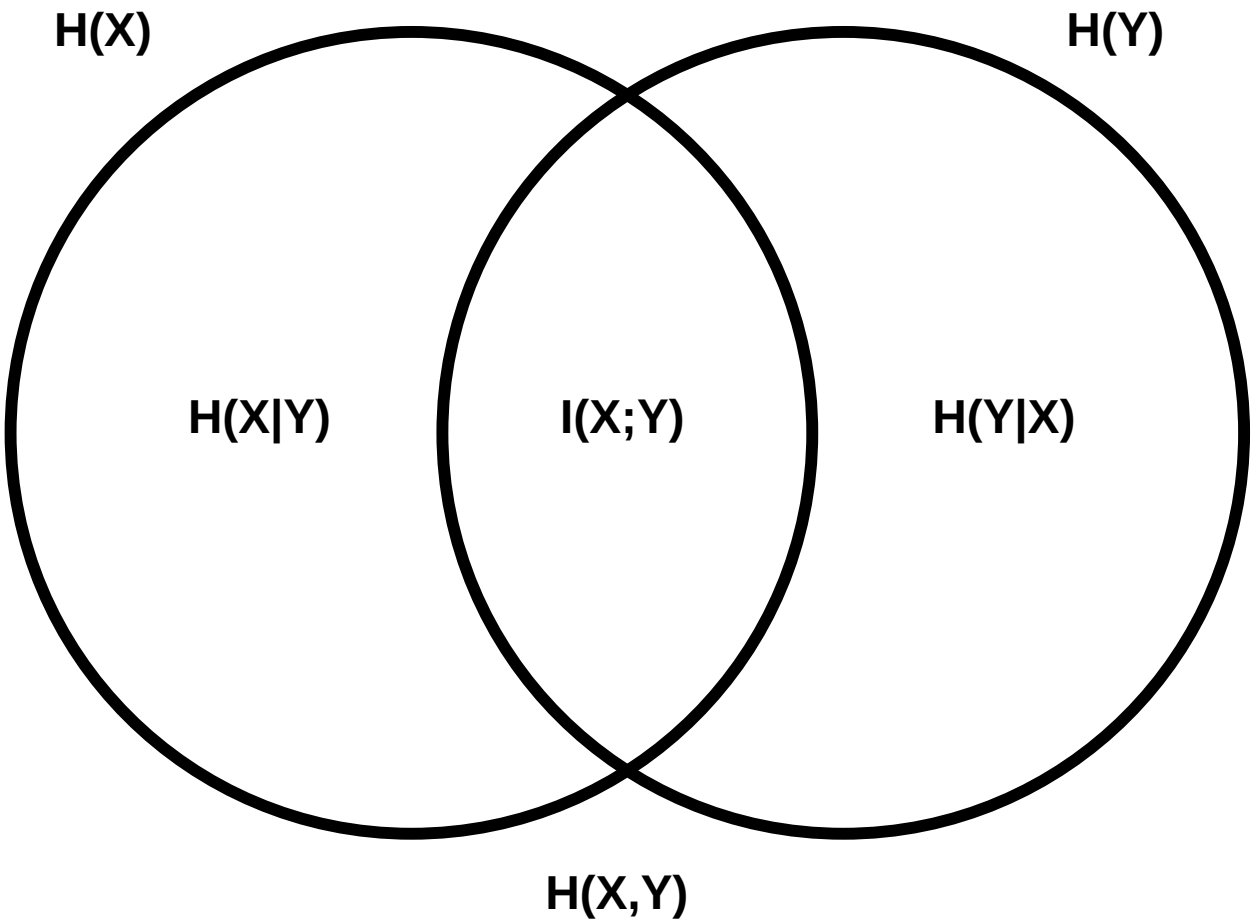
$$I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

- symmetric
- non-negative
- zero if variables are independent

- Also, since $H(X|X) = 0$ it follows that:

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

Mutual Information



Summary

Entropy

- expresses the amount of uncertainty about the state of a variable
- a function of a probability distribution over the possible states

Mutual Information

- the amount of information shared by two different variables

or

- how much one variable tells us about the other

Measuring the Difference

- Often, it is useful to be able to compare two distributions.
- Do two variables behave similarly?
- How far away is one variable from another?
  (e.g., the dice we are using vs. a fair pair)

Kullback Leibler (KL) divergence

- The *Kullback Leibler (KL) divergence* between two probability distributions $p(x)$ and $q(x)$ is defined as:

$$D_{KL}[p||q] = \sum_x p(x) log \frac{p(x)}{q(x)}$$

where $0 log \frac{0}{q} = 0$ and $p log \frac{p}{0} = \infty$

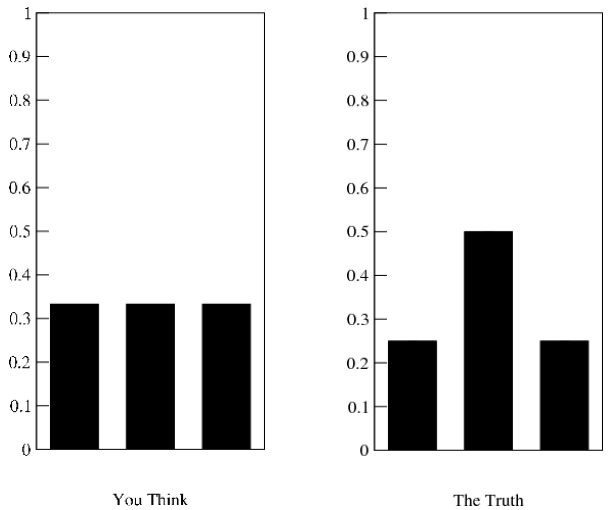
- $D_{KL}[p; q] \geq 0$
- $D_{KL}[p; q] = 0 \quad \Leftrightarrow \quad p(x) = q(x) \quad \forall x \in X$

Kullback Leibler (KL) divergence

$$D_{KL}[p||q] = E_p\left(log\frac{p(x)}{q(x)}\right)$$

- The divergence $D_{KL}[p||q]$ is the average amount of information lost by thinking you're dealing with q when actually it's p.

Example:
Your friend picks a number from zero to two, and you try to guess.
You think it's completely random (equal probability). But actually, he's tossing a coin twice and counting the heads.



KL divergence & MI

Mutual information is actually just a measure of how far a joint distribution is from independence:

$$I(X;Y) = D_{KL}[p(x,y)||p(x)p(y)]$$

Kullback Leibler (KL) divergence

- The *Kullback Leibler (KL) divergence* is also called
  - *relative entropy*
  - and sometimes *KL distance*
- However, it does not satisfy the conditions of a *distance metric*
  - it is not symmetric in p and q, i.e., it does not hold that

$$D_{KL}[p||q] = D_{KL}[q||p]$$

  - it does not satisfy triangle inequality, i.e., it does not hold that

$$D_{KL}[p||r] \leq D_{KL}[p||q] + D_{KL}[q||r]$$

- Also, if $q(x) = 0$ for some $x \in X$, we can get infinite divergence.

Jensen-Shannon (JS) Divergence

Jensen-Shannon (JS) divergence solves some of the problems of KL divergence.

$$JS_{\Pi}[p||q] = \pi_1 \cdot D_{KL}[p||r] + \pi_2 \cdot D_{KL}[q||r]$$

where $\Pi = \{\pi_1, \pi_2\}$ is a distribution, and we define:

$$r(x) = \pi_1 \cdot p(x) + \pi_2 \cdot q(x) \quad \forall x \in X$$

- if $\Pi = \{\frac{1}{2}, \frac{1}{2}\}$ is symmetric, then:
  - $JS[p||q]$ is symmetric in p and q
  - $JS[p||q] \neq \infty$
- JS is still not a metric, since is does not satisfy triangle inequality.

Cross Entropy and Perplexity

Motivation:

- In many fields of research, models are proposed to explain observed data.
- Given two models/explanations, how do we know which is better?
- Intuitively, predictive power is a big factor.
- How do we quantify this?
- How can information theory help us?

Cross Entropy

The cross entropy between a random variable X with true probability distribution p(x) and another probability distribution q (normally a model of p) is given by:

$$H(X,q) = H(X) + D_{KL}(p||q) = -\sum_{x \in X} p(x) log_2 q(x)$$

- i.e., the uncertainty inherent in X plus the uncertainty added by incorrectly modeling X.
- $p(x)$, and therefore $H(X)$ is usually unknown.
- We can get a approximation of it, through a test set:

$$H(T,q) \approx -\frac{1}{|T|} \sum_{t_i \in T} log_2 q(t_i)$$

Cross Entropy - Examples

1. entropy of the English language. In other words, how uncertain are we about the next letter (+ space) we read?

Model	C.E.	
zeroth order	4.76	uniform model, so log 27
first order	4.03	approx. distribution of letters
second order	2.8	approx. distribution based on prev. letter
Shannon's experiment	1.3	human subjects

2. entropy of protein sequences.

   The Shannon information entropy of protein sequences,
   B.J. Straita and T.G. Deweya

   Biophysical Journal, Volume 71, Issue 1, Pages 148-155

Cross Entropy

$$H(X,q) = H(X) + D(p||q) = -\sum_{x \in X} p(x) log_2 q(x)$$

- The better the model, the lower the cross entropy of the test set.
- For any interesting distribution, $H(X) > 0$ and unknown.
- Therefore, the cross entropy will be > 0 and meaningful only for comparing different models.

Perplexity

In many fields, *perplexity* is commonly used to measure how well the model fits the data.

The relationship between perplexity and cross entropy is simple:

$$perplexity(T, q) = 2^{H(T,q)} = \prod_{t_i \in T} q(t_i)^{-\frac{1}{|T|}}$$