

Constructive Artificial Intelligence

Information Theory

Daniel Polani

School of Computer Science
University of Hertfordshire

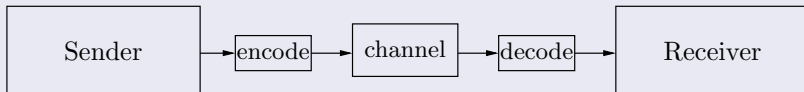
November 23, 2009

All rights reserved. Permission is granted to copy and distribute these slides in full or in part for purposes of research, education as well as private use, provided that author, affiliation and this notice is retained.

Use as part of home- and coursework is only allowed with express permission by the responsible tutor and, in this case, is to be appropriately referenced.

Task

Sending data over a (possibly noisy) channel:



Problem

Assume: channel unreliable because of noise

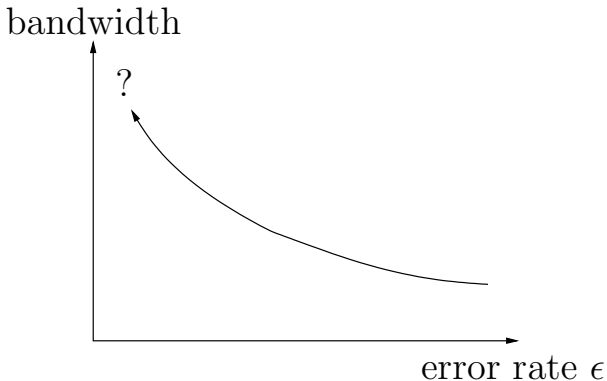
Question: how much “effort” (bandwidth) needed to ensure a **reliable transmission**?

Reliable Transmission: error rate can be made less than any ϵ

Quantification of Data Transmission

Question

- How will the required bandwidth grow for arbitrarily small error rate ϵ ?
- Would it grow to infinity?



Shannon's Theorem 1948

Shannon's Central Result

For a channel with given noise, to transmit data with an error rate $\epsilon \rightarrow 0$, only a finite bandwidth is needed.

Extensions (Jaynes 1957)

- Shannon's theory provides universal theory of uncertainty
- deep connections with physics

Shannon's Theorem 1948

Shannon's Central Result

For a channel with given noise, to transmit data with an error rate $\epsilon \rightarrow 0$, only a finite bandwidth is needed.

Extensions (Jaynes 1957)

- Shannon's theory provides universal theory of uncertainty
- deep connections with physics

Introduce now some basics, starting with probability theory

Definition (Entropy)

Let X be a random variable assuming values x in $\{x_1, x_2, \dots, x_n\}$. Then define the **entropy** of X as

$$H(X) := - \sum_x p(x) \log p(x)$$

Back to the Communication Problem

Note

If 'log' is a binary logarithm, measure $H(X)$ in *bit*.

Interpretation

Assume symbol source generating symbols x with probability $p(x)$.
Entropy $H(X)$ in bit measures

- how many symbols $\{0, 1\}$ are required *on average*
- using optimal coding
- to send signals from this source over a noiseless channel.

Alternative Interpretations

Entropy measures

- **uncertainty** about X before outcome is known
- in a stream of outcomes for X , how many yes/no questions are required *on average* to identify one occurrence of X

Mutual Information

Let X, Y be random variables with joint distribution $p(x, y)$. Then define **mutual information** $I(X; Y)$ by

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

where

- $H(Y|X) = \sum_x p(x) \cdot H(Y|X = x)$
- $H(X, Y)$ is the entropy of the joint random variable (X, Y) with probability $p(x, y)$
- $H(X)$ is the entropy of the marginal distribution $p(x)$ of (X, Y) , i.e. $p(x) = \sum_y p(x, y)$, analogously for $H(Y)$.

Properties of Mutual Information

One has the relations

$$I(X; Y) = I(Y; X) \quad (\text{symmetry})$$

$$I(X; Y) \geq 0 \quad (\text{positivity})$$

and many more (see Cover and Thomas 1991)

Notes

Call mutual information also **Shannon information** or, simply, **information**.

The Kullback-Leibler Distance

Definition (Kullback-Leibler Distance)

for a random variable with two different distributions p and q , define the Kullback-Leibler distance between p and q as

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Theorem

- 1 $D(p||q) \geq 0$ and $D(p||q) \Leftrightarrow p = q$
- 2 Let $p(x, y)$ be a the joint distribution of random variables X and Y . Write $\tilde{p}(x, y) = p(x)p(y)$ with $p(x)$ and $p(y)$ the marginal distributions of $p(x, y)$. Then

$$I(X; Y) = D(p||\tilde{p})$$

Why Information?

Interpretation

- information $I(X; Y)$ quantifies how much X tells about Y
- it serves as a measure of information transmission
- maximum possible $I(X; Y)$ called **channel capacity**

Power of Information

- universal and independent from concrete substrate
- unifies concepts from computation, physics and statistics
- can be used to quantify complexity
- encompasses many concepts, from computation to “black holes”
- minimizes arbitrariness in the description of processes

Measuring Probabilities

Definition (Information Gain)

Let a random variable X and an associated evidence E (also a random variable) be given.

Then, the knowledge of E will add the information $I(X; E)$ about the outcome of X . In particular, if we observe outcome e for the variable E , then one has

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

which is the **information gain** about X induced by observation e .

Notes

- information gain measures how good the observation “focused” the knowledge we had
- **information gain for a concrete e can be negative!** (Why?)
- information gain can help in the evaluation of the quality of an observation

References

- Cover, T. M., and Thomas, J. A., (1991). *Elements of Information Theory*. New York: Wiley.
- Jaynes, E. T., (1957). Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620–630.
- Pearl, J., (1984). *Heuristics: Intelligent Search Strategies for Computer Problem-Solving*. Addison Wesley.
- Russell, S., and Norvig, P., (2002). *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Prentice Hall. Second edition.
- Shannon, C. E., (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423.