Chaos, Complexity, and Inference (36-462)

Lecture 7

Cosma Shalizi

5 February 2008

Information Theory

Entropy and Information Measuring randomness and dependence in bits

Relative Entropy The connection to statistics

Entropy and Ergodicity Long-run randomness

Single best book on information theory: [1]

Entropy

Fundamental notion in information theory
$X =$ a discrete random variable, values from $\mathcal{X}$
The **entropy of** X is

$$H[X] \equiv -\sum_{x \in \mathcal{X}} \Pr(X = x) \log_2 \Pr(X = x)$$

EXERCISE: Prove that $H[X]$ is maximal when all X are equally probable, and then $H[X] = \log_2 \#\mathcal{X}$.
EXERCISE: Prove that $H[X] \geq 0$, and $= 0$ only when $\Pr(X = x) = 1$ for some x.

Interpretations

$H[X]$ measures

- how *random* X is
- How much *variability* X has
- How *uncertain* we should be about X

  "paleface" problem

  consistent resolution leads to a completely subjective probability theory

Description Length

Another, fundamental interpretation of $H[X]$: how concise can we make a description of X?
Imagine X as text message:

```
wtf?; lol; omg; o rly?; bored now;
what u doing 4 fri pm?; no i mean rly wtf?;
in reno;
in reno send money;
in reno divorce final;
in reno send lawyers guns and money k thx bye
```

I know what X is but won't show it to you
You can guess it by asking yes/no (binary) questions

First goal: ask as few questions as possible

Making the first question "is it y?" works, if $X = y$ — but not otherwise

New goal: minimize the *mean* number of questions

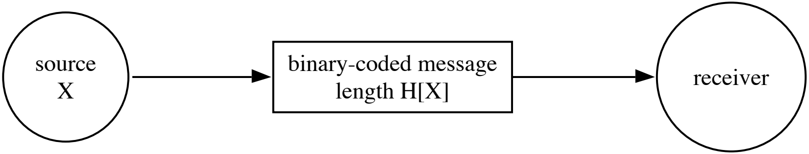
Ask about more probable messages first

Best you can do is get to x with about $-\log_2 \Pr(X = x)$ questions

Mean is then $H[X]$

$H[X]$ is the minimum mean number of binary distinctions needed to describe X

Units of $H[X]$ are **bits**



$H[f(X)] \leq H[X]$, equality if and only if f is invertible

Multiple Variables — Joint Entropy

Joint entropy of two variables X and Y:

$$H[X,Y] \equiv -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr\left(X = x, Y = y\right) \log_2 \Pr\left(X = x, Y = y\right)$$

Entropy of joint distribution
This is the minimum mean length to describe both X and Y

$$\begin{aligned} H[X,Y] &\geq H[X] \\ H[X,Y] &\geq H[Y] \\ H[X,Y] &\leq H[X] + H[Y] \\ H[f(X),X] &= H[X] \end{aligned}$$

Conditional Entropy

Entropy of conditional distribution:

$$H[X|Y=y] \equiv -\sum_{x \in \mathcal{X}} \Pr(X=x|Y=y) \log_2 \Pr(X=x|Y=y)$$

Average over y:

$$H[X|Y] \equiv \sum_{y \in \mathcal{Y}} \Pr(Y=y) H[X|Y=y]$$

On average, how many bits are needed to describe X, *after* Y is given?

$$H[X|Y] = H[X,Y] - H[Y]$$

text completion principle
Note: $H[X|Y] \neq H[Y|X]$, in general
Chain rule:

$$H[X_1^n] = H[X_1] + \sum_{t=1}^{n-1} H[X_{t+1}|X_1^t]$$

Describe one variable, then describe 2nd with 1st, 3rd with first two, etc.

Mutual Information

Mutual information between X and Y

$$I[X;Y] \equiv H[X] + H[Y] - H[X,Y]$$

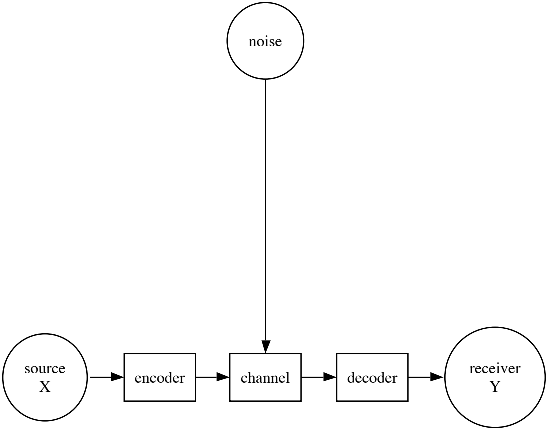
How much shorter is the *actual* joint description than the sum of the individual descriptions?
Equivalent:

$$I[X;Y] = H[X] - H[X|Y] = H[Y] - H[Y|X]$$

How much can I shorten my description of either variable by using the other?

$$0 \leq I[X;Y] \leq \min H[X], H[Y]$$

$I[X;Y] = 0$ if and only if X and Y are statistically independent



How much can we learn about what was sent from what we receive? $I[X;Y]$

Historically, this is the origin of information theory: sending coded messages efficiently [2]

channel capacity $C = \max I[X;Y]$ as we change distribution of X

Any rate of information transfer $< C$ can be achieved with arbitrarily small error rate, *no matter what the noise*

No rate $> C$ can be achieved without error

This is connected to how much money side information can make you in gambling [3]

Historical dramatization: [4]

with silly late-1990s story tacked on

This is not the only model of communication! [5, 6]

Conditional Mutual Information

$$I[X;Y|Z] = H[X|Z] + H[Y|Z] - H[X,Y|Z]$$

How much extra information do X and Y give, over and above what's in Z?

$X \perp Y|Z$ if and only if $I[X;Y|Z] = 0$

Markov property is completely equivalent to

$$I[X_{t+1}^{\infty}; X_{-\infty}^{t-1}|X_t] = 0$$

Markov property is really about information flow

Generalization to partially-observed Markov processes:

$$I[X_t^{\infty}; X_{-\infty}^{t-1}|S_t] = 0$$

Relative Entropy

P, Q = two distributions on the same space $\mathcal{X}$

$$D(P\|Q) \equiv \sum_{x\in\mathcal{X}} P(x)\log_2 \frac{P(x)}{Q(x)}$$

Or, if $\mathcal{X}$ is continuous,

$$D(P\|Q) \equiv \int_{\mathcal{X}} dx\, p(x)\log_2 \frac{p(x)}{q(x)}$$

a.k.a. **Kullback-Leibler divergence**
$D(P\|Q) \geq 0$, with equality if and only if $P = Q$
$D(P\|Q) \neq D(Q\|P)$, in general
Invariant under invertible functions

Joint and Conditional Relative Entropies

P, Q now distributions on $\mathcal{X}, \mathcal{Y}$

$$D(P\|Q) = D(P(X)\|Q(X)) + D(P(Y|X)\|Q(Y|X))$$

where

$$\begin{aligned} D(P(Y|X)\|Q(Y|X)) &= \sum_x P(x) D(P(Y|X=x)\|Q(Y|X=x)) \\ &= \sum_x P(x) \sum_y P(y|x) \log_2 \frac{P(y|x)}{Q(y|x)} \end{aligned}$$

and so on for more than two variables

Relative Entropy and Miscoding

Suppose real distribution is P but we think it's Q and we use that for coding

Our average code length (**cross-entropy**) is

$$-\sum_{x} P(x) \log_2 Q(x)$$

But the optimum code length is

$$-\sum_{x} P(x) \log_2 P(x)$$

Difference is relative entropy

Relative entropy is the extra description length from getting the distribution wrong

Relative Entropy and Hypothesis Testing

Testing P vs. Q

Optimal error rate (chance of guessing Q when really P) goes like

$$\Pr(\text{error}) \approx 2^{-nD(Q\|P)}$$

More exact statement:

$$\frac{1}{n}\log_2 \Pr(\text{error}) \rightarrow -D(Q\|P)$$

The bigger $D(Q\|P)$, the harder they are to confuse, easier to tell apart with a test

For dependent data, substitute sum of conditional relative entropies for nD

Relative entropy can be the basic concept

$$H[X] = \log_2 m - D(U\|P)$$

where $m = \#\mathcal{X}$, $U =$ uniform dist on $\mathcal{X}$, $P =$ dist of X

$$I[X;Y] = D(J\|P \times Q)$$

where $P =$ dist of X, $Q =$ dist of Y, $J =$ joint dist

Maximum likelihood and relative entropy

Data $= X$
True distribution of $= P$
Model distributions $= Q_\theta$, $\theta =$ parameter
Look for the Q_θ which will best describe new data

Best-fitting distribution

$$\begin{aligned}
\theta^* &= \underset{\theta}{\operatorname{argmin}}\, D(P \| Q_\theta) \\
&= \underset{\theta}{\operatorname{argmin}} \sum_x P(x) \log_2 \frac{P(x)}{Q_\theta(x)} \\
&= \underset{\theta}{\operatorname{argmin}} \sum_x P(x) \log_2 P(x) - P(x) \log_2 Q_\theta(x) \\
&= \underset{\theta}{\operatorname{argmin}} -H_P[X] - \sum_x P(x) \log_2 Q_\theta(x) \\
&= \underset{\theta}{\operatorname{argmin}} -\sum_x P(x) \log_2 Q_\theta(x) \\
&= \underset{\theta}{\operatorname{argmax}} \sum_x P(x) \log_2 Q_\theta(x)
\end{aligned}$$

This is the *expected log-likelihood*

We don't know P but we do have a sample, the **empirical distribution** $\hat{P}_n$
For IID case

$$\begin{aligned}\hat{\theta} &= \operatorname*{argmax}_{\theta} \sum_{t=1}^{n} \log Q_{\theta}(x_t) \\ &= \operatorname*{argmax}_{\theta} \frac{1}{n} \sum_{t=1}^{n} \log_2 Q_{\theta}(x_t) \\ &= \operatorname*{argmax}_{\theta} \sum_{x} \hat{P}_n(x) \log_2 Q_{\theta}(x)\end{aligned}$$

So $\hat{\theta}$ comes from approximating P by $\hat{P}_n$
$\hat{\theta} \rightarrow \theta^*$ because $\hat{P}_n \rightarrow P$

Non-IID case (e.g. Markov) goes similarly, more notation

This is related to the general problem of **large deviations**, and the theory showing that large deviations are exponentially rare [7]

In general:

$-H[X] - D(P\|Q) =$ expected log-likelihood of Q

$-H[X] =$ optimal expected log-likelihood

Relative Entropy and Fisher Information

$$
\begin{aligned}
I_{uv}(\theta_0) &\equiv -\mathbf{E}_{\theta_0}\left[\left.\frac{\partial^2 \log Q_{\theta_0}(X)}{\partial\theta_u \partial\theta_v}\right|_{\theta=\theta_0} \right] \\
&= \left.\frac{\partial^2}{\partial\theta_u \partial\theta_v} D(Q_{\theta_0} \| Q_\theta)\right|_{\theta=\theta_0}
\end{aligned}
$$

Fisher information is how quickly the relative entropy grows with small changes in parameters

$$D(\theta_0 \| \theta_0 + \epsilon) \approx \epsilon^T I \epsilon + O(\|\epsilon\|^2)$$

Intuition: “easy to estimate” is the same as “easy to reject sub-optimal values”

Entropy Rate

Entropy rate, a.k.a. **Shannon entropy rate**, a.k.a. **metric entropy rate**

$$h_1 \equiv \lim_{n \to \infty} H[X_n | X_1^{n-1}]$$

Limit exists for any stationary process (and some others)
(Strictly, Strongly) Stationary: for any $k > 0$, $T > 0$, for all $w \in \mathcal{X}^k$

$$\Pr\left(X_1^k = w\right) = \Pr\left(X_{1+T}^{k+T} = w\right)$$

Or: Probability distribution is invariant under the shift

Examples of entropy rates

IID $H[X_n|X_1^{n-1}] = H[X_1] = h_1$

Markov $H[X_n|X_1^{n-1}] = H[X_n|X_{n-1}] = H[X_2|X1] = h_1$

k^{th}-order Markov $h_1 = H[X_{k+1}|X_1^k]$

SFA $H[X_n|X_1^n] \rightarrow H[X_n|S_n] = H[X_1|S_1] = h_1$

Metric vs. Topological Entropy Rate

Using chain rule, can re-write h_1 as

$$h_1 = \lim_{n\to\infty} \frac{1}{n} H[X_1^n]$$

Remember topological entropy rate:

$$h_0 = \lim_{n\to\infty} \frac{1}{n} \log_2 W_n$$

where $W_n = \#$ allowed words of length n
$H[X_1^n] = \log_2 W_n$ if and only if each word is equally probable
Otherwise $H[X_1^n] < \log_2 W_n$

$h_0 =$ growth rate in number of allowed words, counting all equally
$h_1 =$ growth rate, counting more probable words more heavily — *effective* number of words
So:

$$h_0 \geq h_1$$

2^{h_1} is the *effective* number of choices of how to continue a long symbol sequence

Entropy Rate Measures Randomness

h_1 = growth rate of mean description length of *trajectories*
Chaos needs $h_1 > 0$
For symbolic dynamics, each partition $\mathcal{B}$ has its own $h_1(\mathcal{B})$
Kolmogorov-Sinai (KS) entropy rate:

$$h_{KS} = \sup_{\mathcal{B}} h_1(\mathcal{B})$$

THEOREM If $\mathcal{G}$ is a generating partition, then $h_{KS} = h_1(\mathcal{G})$
h_{KS} is the *asymptotic randomness* of the dynamical system
or, the rate at which the symbol sequence provides *new information* about the initial condition

Entropy Rate and Lyapunov Exponents

In general (Ruelle's inequality),

$$h_{KS} \leq \sum_{i=1}^{d} \lambda_i \mathbf{1}_{x>0}(\lambda_i)$$

If the invariant measure is smooth, this is equality (Pesin's identity)

Asymptotic Equipartition Property

When n is large, for any word x_1^n, either

$$\Pr\left(X_1^n = x_1^n\right) \approx 2^{-nh_1}$$

or

$$\Pr\left(X_1^n = x_1^n\right) \approx 0$$

More exactly, it's almost certain that

$$-\frac{1}{n}\log\Pr\left(X_1^n\right) \rightarrow h_1$$

This is the **entropy ergodic theorem** or **Shannon-MacMillan-Breiman theorem**

Relative entropy version:

$$-\frac{1}{n}\log Q_\theta(X_1^n) \rightarrow h_1 + d(P\|Q_\theta)$$

where

$$d(P\|Q_\theta) = \lim_{n\rightarrow\infty} \frac{1}{n} D(P(X_1^n)\|Q_\theta(X_1^n))$$

Relative entropy AEP is less general than entropy AEP

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[2] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423, 1948. URL `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html`. Reprinted in [8].

[3] William Poundstone. *Fortune's Formula: The Untold Story of the Scientific Betting Systems That Beat the Casinos and Wall Street*. Hill and Wang, New York, 2005.

[4] Neal Stephenson. *Cryptonomicon*. Avon Books, New York, 1999.

[5] Dan Sperber and Deirdre Wilson. *Relevance: Cognition and Communication*. Basil Blackwell, Oxford, second edition, 1995.

[6] Dan Sperber and Deirdre Wilson. Rhetoric and relevance. In David Wellbery and John Bender, editors, *The Ends of*

Rhetoric: History, Theory, Practice, pages 140–155, Stanford, 1990. Stanford University Press. URL `http://dan.sperber.com/rhetoric.htm`.

[7] James A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation*. Wiley-Interscience, New York, 1990.

[8] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, 1963.