Entropy + Information (Selected Topics)

Entropy: a natural measure for the "richness" of a distribution

Say P is specified by $p_1, \dots, p_M$ where $\sum p_j = 1$, $p_j \geq 0$; p_j is the probability that a symbol drawn from P is "j".

$H(P)$ = # of yes-no questions, on average, required to determine which symbol is drawn.

Will show $\boxed{H(P) = -\sum p_j \log_2 p_j}$

Say: P has 2^n symbols each with $p_j = 2^{-n}$.

n yes-no questions are necessary, and sufficient

necessary: only 2^n possible sequences of answers
sufficient: dichotomy strategy

Say $n=3$:

Is it in $\{0,1,2,3\}$? $\nearrow$ yes: refine $\{0,1,2,3\}$
$\searrow$ no: refine $\{4,5,6,7\}$.

Or, more compactly:
Express j as a binary number of j digits - ask about each one.

$$-\sum_{j=1}^{2^n} 2^{-n} \log_2 (2^{-n}) = -\log_2 (2^{-n}) = n.$$

Say P has a symbols, $2^{n-1} < a < 2^n$.

$n-1$ yes-no questions will not suffice.
n questions will suffice, so $(n-1) < H < n$.

Consider symbols in pairs. (e.g., $a=7$)
$\underbrace{36}\underbrace{40}\underbrace{31}505221\ldots$

This is an alphabet of a^2 symbols.
More generally, considering symbols in k-tuples gives an alphabet of a^k symbols.

Determining a k-tuple = determining k 1-tuples.

Say n_k is s.t. $2^{n_k-1} < a^k < 2^{n_k}$.

Then $n_k - 1 < kH < n_k$.

$$2^{n_k-1} < a^k < 2^{n_k} \iff n_k - 1 < \log_2 a^k < n_k$$

$$\frac{n_k-1}{k} < \log_2 a < \frac{n_k}{k}$$

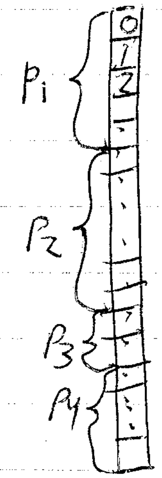
$n_k = \lceil (\log_2 a) \cdot k \rceil$ $\quad \lceil u \rceil$ = least integer $\geq u$.

$$\frac{1}{k}\left(\lceil (\log_2 (a) \cdot k) \rceil - 1\right) < H < \frac{1}{k}\lceil (\log_2 a) \cdot k \rceil, \text{ all } k.$$

$$H = \log_2 a$$

Unequal probabilities.
[sketch]

Say each $p_j = N_j / a$.



We need, on average, $\log_2 a$ bits to determine which of a "hidden" symbols is present. But, with probability p_j, we have $\log_2 N_j$ "excess" questions.

$$
\begin{aligned}
\text{So } H &= \log_2 a - \sum_j p_j \log_2 N_j \\
&= \log_2 a - \sum_j p_j \log_2 (a p_j) \\
&= \log_2 a - \sum_j p_j \log_2 a - \sum p_j \log_2 p_j \\
&= - \sum_j p_j \log_2 p_j .
\end{aligned}
$$

A few basic properties:

1. Say P & Q are independent processes. Then,
$$H(P \times Q) = H(P) + H(Q).$$

P- stream $x_1\ x_2\ x_3 \cdots\ x_k \cdots$ $\quad$ prob $(x = X_\alpha) = p_\alpha$

Q- stream $y_1\ y_2\ y_3 \cdots\ y_k$ $\quad$ prob $(y = Y_\beta) = q_\beta$

$R = P \times Q$-stream $(x_1, y_1) \cdots\ (x_k, y_k)$ $\quad$ $r_{\alpha\beta} = p_\alpha q_\beta$

prob $(x, y) = (X_\alpha, Y_\beta) = r_{\alpha\beta}$

4.

$$H(P+Q) = -\sum_{\alpha,\beta} r_{\alpha\beta} \log_2 r_{\alpha\beta}$$

$$= -\sum_{\alpha\beta} p_\alpha q_\beta \log_2 p_\alpha q_\beta$$

$$= -\sum_{\alpha\beta} p_\alpha q_\beta (\log_2 p_\alpha + \log_2 q_\beta)$$

$$= -\sum_{\alpha,\beta} p_\alpha q_\beta \log_2 p_\alpha - \sum_{\alpha\beta} p_\alpha q_\beta \log_2 q_\beta$$

$$= -\sum_\alpha p_\alpha \log_2 p_\alpha - \sum_\beta q_\beta \log_2 q_\beta$$

$(\sum q_\beta = 1)$ $\qquad$ $(\sum p_\alpha = 1)$

$$= H(P) + H(Q).$$

(2) Mixing. "$R_z = (1-z)P + zQ$". $\qquad \dfrac{d^2 H(R_z)}{dz^2} < 0$ if $Q \neq P$.

P & Q both distributions on same letters.

$$r_\alpha = (1-z)p_\alpha + z q_\alpha.$$

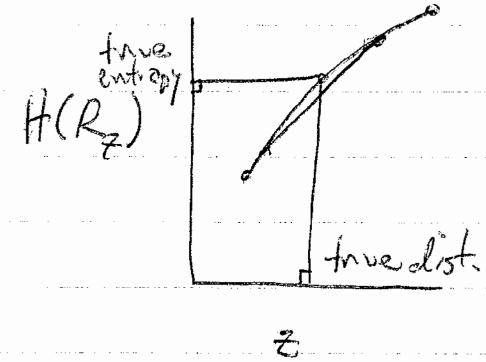
$$\frac{dH(R_z)}{dz} = \frac{d}{dz} \sum_\alpha ((1-z)p_\alpha + z q_\alpha) \log_2((1-z)p_\alpha + z q_\alpha)$$

$$= -\sum_\alpha \left[\left(\frac{-p_\alpha + q_\alpha}{\ln 2} \right) + +(-p_\alpha + q_\alpha) \log_2((1-z)p_\alpha + z q_\alpha) \right]$$

$$= -\sum_\alpha (-p_\alpha + q_\alpha) \log_2((1-z)p_\alpha + z q_\alpha)$$

$$\frac{d^2 H(R_z)}{dz^2} = -\sum_\alpha \frac{(-p_\alpha + q_\alpha)^2}{\ln 2} \left(\frac{1}{(1-z)p_\alpha + z q_\alpha} \right) < 0$$

Consequence of mixing property for estimates of entropy



"plug-in" estimate is always downward-biased.

Amount of bias, for an estimate of $p_1, \ldots, p_K$ from N samples, is

[Miller Carlton Treves Panzeri]

$$\frac{K-1}{2\ln 2}\cdot\frac{1}{N} + O\left(\frac{1}{N^2}\right) \quad \text{provided } K \ll N.$$

Caution – bias estimate is very wrong if $K \gtrsim N$.

(3) Max Ent distribution on (x_α, y_β) subj to $\sum_\beta r_{\alpha\beta} = p_\alpha$, $\sum_\alpha r_{\alpha\beta} = q_\beta$

We know (from previous maxent analyses) that

$$r_{\alpha\beta} = C\, e^{-\lambda_\alpha - \mu_\beta}$$

so, $r_{\alpha\beta}$ must be a product, so $r_{\alpha\beta} = p_\alpha q_\beta$.

Any other dist. on (x_α, y_β) with $\sum_\beta r_{\alpha\beta} = p_\alpha$, $\sum_\alpha r_{\alpha\beta} = q_\beta$ must have lower entropy (via property (2)).

So we now have a nonparametric measure of association between two variables:

Say $r_{\alpha\beta} = \text{prob}(x_\alpha, y_\beta)$; $p_\alpha = \sum_\beta r_{\alpha\beta}$, $q_\beta = \sum_\alpha r_{\alpha\beta}$.

Then $H(P) + H(Q) - H(R) \geq 0$;

this is 0 only for independence.

$H(R)$ can never be less than $H(P)$ or $H(Q)$

So this quantity can never be larger than $\min(H(P), H(Q))$.

This is the "mutual information" between x and y.

The above ideas also make sense if the sequence of symbols is <u>not</u> independent

$s_1 \, s_2 \, s_3 \, s_4 \cdots$

i.e., if $p(s_1 = x_\alpha, s_2 = x_\beta) \neq p(s_1 = x_\alpha) \cdot p(s_2 = x_\beta)$.

But we can still talk about the entropy per symbol,

$$H = \lim_{k \to \infty} \frac{1}{k} \{\text{entropy of } k\text{-tuples}\}.$$

Example:

Sequence of 0's & 1's, equally probable, but

$$p(0\,0) = c$$

$$p(0,1) = \tfrac{1}{2} - c \qquad \text{since } p(0,0) + p(0,1) = \tfrac{1}{2}$$

$$p(1,0) = \tfrac{1}{2} - c \qquad \text{"} \quad p(0,0) + p(1,0) = \tfrac{1}{2}$$

$$p(1,1) = c \qquad \text{"} \quad p(1,1) + p(0,1) = \tfrac{1}{2}$$

Recalling orig. def. of entropy (# of yes-no questions required to specify):

$$0\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 0\ \ldots$$

can be represented by

$$0\ \Delta = = \Delta\ \Delta\ \Delta\ \Delta = \Delta\ \Delta\ \Delta \ldots$$

but $=$ and Δ are independent, so,

$$p(=) = 2c$$
$$p(\Delta) = 1 - 2c$$

k-symbol entropy =

$$\underbrace{1}_{\text{initial symbol}} + (k-1)\left[-2c\log_2 2c - (1-2c)\log_2(1-2c)\right]$$

So $H = -2c\log_2 2c - (1-2c)\log_2(1-2c)$.

(maximum at $c = \frac{1}{4}$; $H = 1$).

Extends to arbitrary-order Markov processes.

8.

Transmitted information [Mutual Information]

P: a symbol sequence (for simplicity, independent) $\bar{c}$ probabilities $p_1, \cdots, p_M$

Q: a symbol sequence, possibly dependent on P, but no other serial dependence $q_1, \cdots, q_N$.

Idea: Information that Q has about P =

$\boxed{H_1}$ # of bits, on average, required to determine a sample of P [without looking at Q]

$\boxed{H_2}$ − # of bits, on average, required to determine a sample of P [after observing]

Use $r_{\alpha\beta}$ to describe coupling of P & Q. $\left(p_\alpha = \sum_\beta r_{\alpha\beta} \quad q_\beta = \sum_\alpha r_{\alpha\beta} \right)$

$$H_1 = -\sum_\alpha p_\alpha \log_2 p_\alpha.$$

$$H_2 = \sum_\beta q_\beta \left\{ -\sum_\alpha \text{prob}(\alpha|\beta) \log_2 \text{prob}(\alpha|\beta) \right\}$$

where $\text{prob}(\alpha|\beta) = r_{\alpha\beta}/q_\beta$. So

$$H_2 = -\sum_{\alpha,\beta} r_{\alpha\beta} \left(\log_2 r_{\alpha\beta} - \log_2 q_\beta \right)$$

$$= -\sum_{\alpha,\beta} r_{\alpha\beta} \log_2 r_{\alpha\beta} + \sum_\beta q_\beta \log_2 q_\beta$$

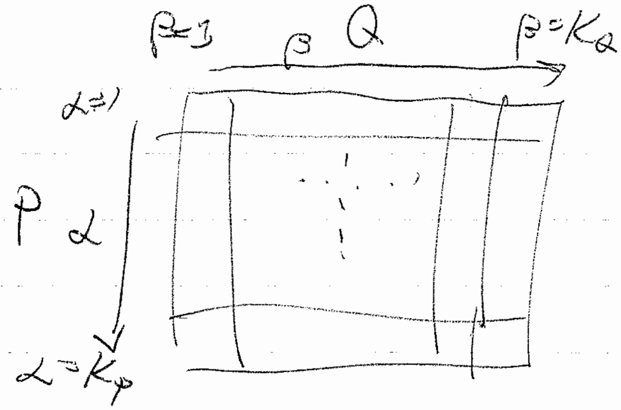
Info of Q about P = $H_1 - H_2 = -\sum_\alpha p_\alpha \log_2 p_\alpha - \sum_\beta q_\beta \log_2 q_\beta + \sum_{\alpha\beta} r_{\alpha\beta} \log_2 r_{\alpha\beta}$

$= H(P) + H(Q) - H(R)$ [Mutual Info of P & Q]

9.

Bias (large N limit)

$$= \frac{(K_P-1)(K_Q-1)-(K_R-1)}{2\ln 2 \cdot N}$$



[$K_R \ll N$]

If all cells are occupiable,

$K_R = K_P K_Q$, → bias < 0 (plug-in estimate is an overestimate)

But not all cells need be occupiable. Bias can be + or −.

Data Processing Thm.

$$X \longrightarrow Y \longrightarrow Z$$

$$p_\alpha \qquad q_\beta \qquad s_\gamma$$

Say $r_{\alpha\beta}$ indicates relationship of p_α, q_β - indep. of γ

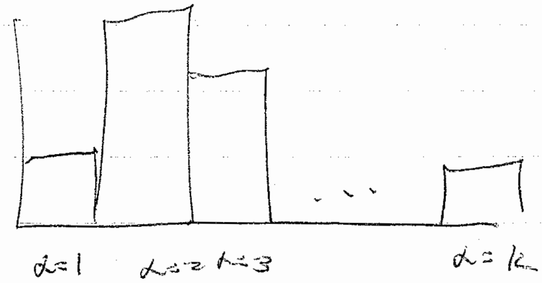
$t_{\beta\gamma}$ " " " q_β, s_γ - indep. of α

Information of Z about X ≤ Information of Y about X.

Rel. of s_γ, p_α given by $u_{\alpha\gamma} = \sum_\beta r_{\alpha\beta} t_{\beta\gamma}$.

Continuum case.

Replace P_α



by $p(x)$ where $\int p(x)\,dx = 1$.

(x may be a vector).

What is $\lim_{\Delta x \to 0} -\sum_i \left(p(x_i)\Delta x_i\right) \log_2 \left(p(x_i)\Delta x_i\right)$?

This is a natural notion for entropy of P, but, note that this =

$$\lim_{\Delta x \to 0} -\sum_i p(x_i)\Delta x \left(\log_2 p(x_i) + \log_2 \Delta x\right)$$

$$= \underbrace{-\int p(x)\log_2 p(x)\,dx}_{\text{"Differential Entropy" of } P} \quad \underbrace{-\log_2 \Delta x}_{\text{limit does not exist}}$$

Can still compare entropies, and, can still calculate mutual informations.

11.

Differential entropy of a Gaussian, covariance matrix V

$$p(\vec{x}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{\sqrt{\det V}} e^{-x^T V^{-1} x/2}$$

$$-\ln p(\vec{x}) = \frac{d}{2} \ln 2\pi + \frac{1}{2} \ln \det V + \frac{1}{2} x^T V^{-1} x .$$

$$\text{Differential entropy} = \frac{1}{\ln 2} \int \left(\frac{d}{2} \ln 2\pi + \frac{1}{2} \ln \det V + \frac{1}{2} x^T V^{-1} x \right) p(\vec{x}) d\vec{x}$$

$$\int x^T V^{-1} x \, p(\vec{x}) d\vec{x} = \int (y^T y) \cdot \frac{1}{(2\pi)^{d/2}} e^{-y^T y/2} d\vec{y} = d$$

$$y \text{ s.t. } y^T y = x^T V^{-1} x \, ; \; d\vec{y} = \frac{1}{\sqrt{\det V}} d\vec{x}$$

$$\text{So, diff. entropy} = \frac{1}{\ln 2} \left[\frac{d}{2} (1 + \ln 2\pi) + \frac{1}{2} \ln \det V \right]$$

<u>Mutual Info in the continuous setting:</u>

No new issues. And, since $\log(\Delta x \Delta y) = \log \Delta x + \log \Delta y$, the annoying term drops out.

Consequence of Data Processing Inequality.

$$X \longrightarrow Y \underset{\text{invertible}}{\longleftrightarrow} Y' \Rightarrow MI(X, Y) = MI(X, Y')$$

since $X \to Y \to Y' \to Y \quad MI(X,Y) \le MI(X, Y')$

and $X \to Y \to Y' \quad \Rightarrow \quad MI(X, Y') \le MI(X, Y)$

Important example!

Gaussian signal, additive Gaussian noise

s: possibly multivariate signal with covariance $\langle s s^T \rangle = V_S$

r: response. $r = As + x$, x = Gaussian noise; noise indep. of s; $\bar{c}$ covariance $\langle x x^T \rangle = V_X$.

Note $\langle r s^T \rangle = \langle (As + x) s^T \rangle = A V_S \quad (\langle x s^T \rangle = 0)$

$\langle s r^T \rangle = \langle s (As + x)^T \rangle = V_S A^T$

$\langle r r^T \rangle = \langle (As+x)(As+x)^T \rangle = A V_S A^T + V_X = V_R$

So $\begin{pmatrix} s \\ r \end{pmatrix} (s^T \; r^T) = \begin{pmatrix} V_S & V_S A^T \\ A V_S & A V_S A^T + V_X \end{pmatrix} = V_{SR}$

Mutual information is

$$\frac{1}{\ln 2} \Big[\frac{d_S}{2}(1 + \ln 2\pi) + \frac{1}{2} \ln \det V_S$$

$$+ \frac{d_r}{2}(1 + \ln 2\pi) + \frac{1}{2} \ln \det V_R$$

$$- \frac{d_{Sr}}{2}(+ \ln 2\pi) + \frac{1}{2} \ln \det V_{SR} \Big]$$

$d_{SR} = d_S + d_R$

$$= \frac{1}{\ln 2} \cdot \frac{1}{2} \Big[\ln (\det V_S)(\det V_R) / \det V_{SR} \Big]$$

13.

Elem row op. on V_{SR}

$$\det\begin{pmatrix} V_S & V_S A^T \\ AV_S & AV_S A^T + V_X \end{pmatrix} = \det\begin{pmatrix} V_S & AV_S A^T \\ 0 & V_X \end{pmatrix}$$

(subtract $A(V_S \quad V_S A^T)$ from 2nd row

So $\det V_{SR} = \det V_S \det V_X$.

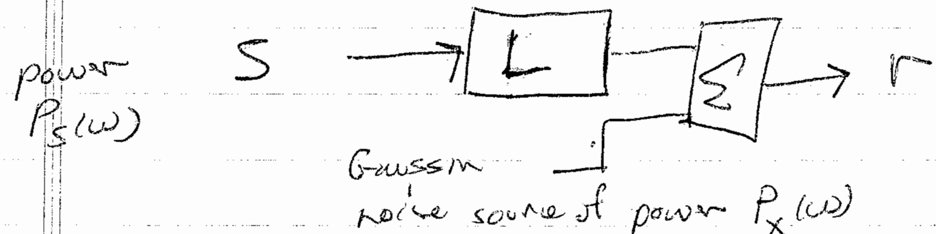
Mutual info =

$$\frac{1}{\ln 2} \cdot \frac{1}{2} \ln \frac{\det V_R}{\det V_X} = \frac{1}{2} \log_2 \frac{\det(AV_S A^T + V_X)}{\det V_X}$$

$$= \log_2 \sqrt{\det(1 + \underbrace{AV_S A^T}_{\text{signal}} / \underbrace{V_X}_{\text{noise}})}$$

Works in frequency domain too – each frequency can be considered separately (if linear, stationary)

Information rate for a Gaussian channel



$$\boxed{P_R(\omega) = |L(\omega)|^2 P_S(\omega) + P_X(\omega)}$$

Over a time T, and sampling at Δt, the relevant frequencies are $\frac{2\pi k}{T}$, $k = 1, \ldots, \frac{1}{2}\left(\frac{T}{\Delta t}\right)$.

At each frequency, the response from time 0 to T has a ~~Fourier~~ component $\tilde{r}(\omega)$ whose real & imaginary parts are each independently Gaussian distributed, with variance $\frac{1}{2} T P_R(\omega)$

Transmitted info =

$$\underbrace{2}_{\text{cos \& sine}} \sum_{k=1}^{\frac{1}{2}\frac{T}{\Delta t}} \log_2 \sqrt{\left(\frac{\frac{1}{2} T P_R(\omega)}{\frac{1}{2} T P_X(\omega)} \right)} \qquad \omega_k = \frac{2\pi k}{T}$$

$$= \sum_{k=1}^{\frac{1}{2}\frac{T}{\Delta t}} \log_2 \left(\frac{P_X(\omega_k) + |L(\omega_k)|^2 P_S(\omega_k)}{P_X(\omega_k)} \right) \qquad \Delta\omega = \frac{2\pi}{\Delta t}$$

$$\rightarrow \int_0^\infty \log_2 \left(1 + |L(\omega)|^2 \frac{P_S(\omega)}{P_X(\omega)} \right) d\omega$$