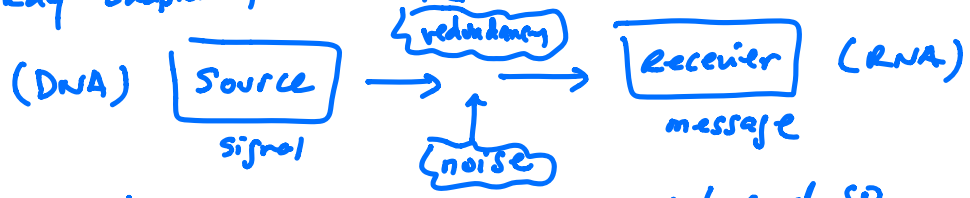
Information theory

Claude Shannon (1912-2001)

"A mathematical theory of communication" (1948)

Mackay chapter 2, lectures 1, 2



how much information do you need to send so that message is reproduced faithfully ?

need to quantify information

1. I had breakfast this morning
2. Today is not my birthday
3. Today it's my birthday
4. I run the 1st Cambridge ½ marathon
5. I run the 1st + 2nd " "
6. I run th 1st + 2nd + 3rd " "
7. I have been to Antarctica

Which of these messages carries more info?

→ BREAKROOMS 5'

Information content (I) $\propto$ (proportional to) inverse of how frequent an event is

p - probability of an event

$$I(e) \propto \frac{1}{p}$$

- the rarer an event is, the lower it's prob. and the more info you would obtain by running the experiment
- the more ignorant about an outcome (many possibilities) each has lower prob., and the more info you get by listening at the end of the channel

Information content should be additive

e_1, e_2 independent events

$$I(e_1, e_2) = I(e_1) + I(e_2)$$

Putting both together

$$I(e) \propto 1/P(e)$$

$$I(e_1 e_2) = I(e_1) + I(e_2)$$

→ Shannon proposed

$$I(e) = \log \frac{1}{P(e)}$$

in base 2 $I(e) = \log_2 \frac{1}{P}$ bits

In shannon own words:

- "useful"
- "intuitive"
- "suitable"

} subjective!

$P(e_1, e_2) = P(e_1) \cdot P(e_2)$ if $e_1 \perp e_2$ (independent)

$$I(e_1 e_2) = \log \frac{1}{P(e_1, e_2)} = \log \frac{1}{P(e_1) P(e_2)}$$

$\log(a \cdot b) = \log a + \log b$

$$= \log \frac{1}{P(e_1)} + \log \frac{1}{P(e_2)}$$

$$= I(e_1) + I(e_2)$$

$$\log(ab) = \log(a) + \log(b)$$

$$\log(a/b) = \log(a) - \log(b)$$

log base

(natural) $\log a = b \longrightarrow a = e^b$ (nats)

$\log_2 a = b_2 \longrightarrow a = 2^{b_2}$ (bits)

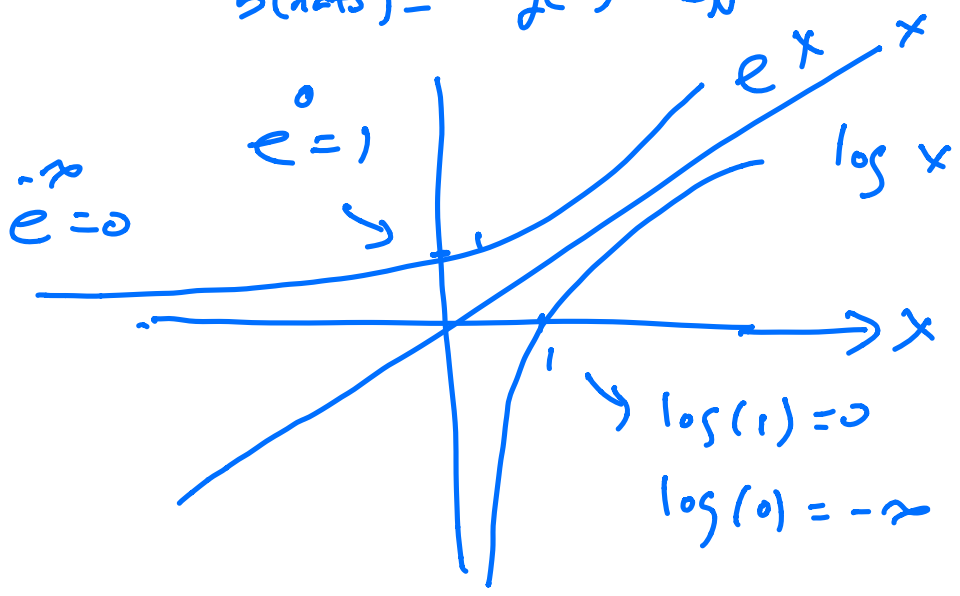
$\log_{10} a = b_{10} \longrightarrow a = 10^{b_{10}}$

$\log_N a = b_N \longrightarrow a = N^{b_N}$

$e^b = 2^{b_2} \rightarrow \log(e^b) = \log(2^{b_2})$

$b = b_2 \log 2$

$b(\text{nats}) = \log(N) \cdot b_N$



	P	I	bits
1. I had breakfast today	1	$\log 1$	0
2. Today not my b-day	$\frac{364}{365}$	$\log \frac{365}{364}$	2.6
3. " is "	$\frac{1}{365}$	$-\log 365$	8.5
4. I am 1st Cam ½	$\frac{4500}{4.7 \cdot 10^3}$		10.0
5. I " 1st + 2nd Cam ½	$\frac{4500}{4.7 \cdot 10^3} \cdot \frac{6500}{4.7 \cdot 10^3}$		13.5

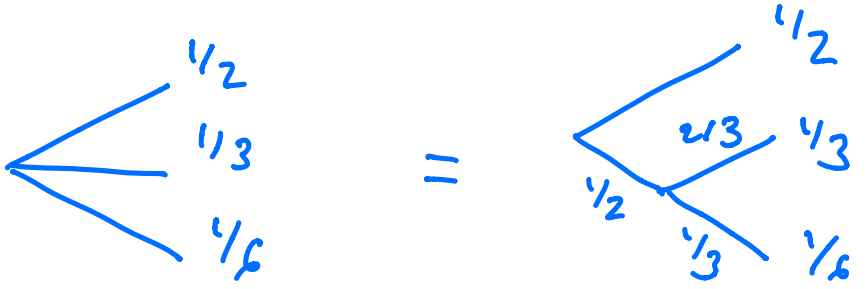
Entropy average information in a probability distribution

X $P(x)$ probability density (pdf)

$$H(X) = \langle \log P(x) \rangle = \int_X \log P(x) \cdot P(x)\, dx$$

$$H(x) \geq 0 \qquad H(x) = 0 \iff x = 0$$

A desirable property described in Shannon's paper is



$$H(1/2, 1/3, 1/6) = H(1/2, 1/2) + \frac{1}{2} H(2/3, 1/3)$$

(exercise to proof this)

Relative entropy the Kullback–Liebler divergence

To compare 2 pdf's $p(x)$, $g(x)$

$$D_{KL}(p\|g) = \int_x p(x) \log \frac{p(x)}{g(x)}$$

i) $D_{KL}(p|g) \neq D_{KL}(g|p)$

ii) $D_{KL}(p|g) \geq 0$ and $=0 \Leftrightarrow p=g$

Mutual information MI X, Y $p(x,y) - p(x)p(y) > g$

$$D_{KL}(p_{xy}, p_x p_y) =$$

$$MI = \int_x \int_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

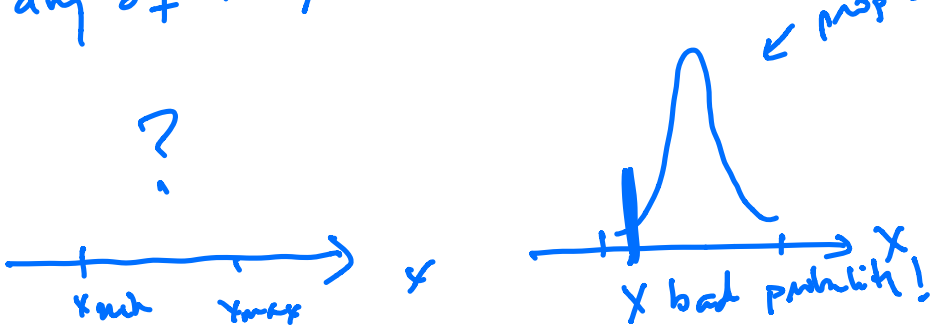
$X \perp Y$: $p(xy) = p(x)p(y) \Rightarrow MI = 0$

the principle of Maximum Entropy

$X \rightarrow$ which is the P_X that has max entropy?

it is the distribution that favors the least

any of the possible outcomes



Optimization question.

given X which is P_X such that

$H(P_X)$ is maximal

$$H(P_X) = \int_X P_X(x) \log P_X(x)\, dx$$

$$L = \int_X P_X(x) \log \frac{1}{P_X(x)}\, dx - \lambda \left(\int_X P_X(x)\, dx - 1 \right)$$

$$\frac{\delta L}{\delta P_X(y)} = \log P_X(y) - \frac{P_X(y)}{P_X(y)} - \lambda = 0$$

$$\log P_X^*(y) = -\lambda - 1$$

$$P_X^*(y) = e^{-\lambda - 1} \rightarrow \text{constant!}$$

$$\int_y P_X^*(y)\, dy = e^{-\lambda-1} \cdot (b-a) = 1 \qquad e^{-\lambda-1} = \frac{1}{b-a}$$

the uniform distribution

$$\boxed{P_X^*(x) = \frac{1}{b-a}}$$

If I know the means

$$L = -\int_y P_X(x) \log P_X(x)\, dy - \lambda \left[\int_y P_X(x) dy - 1 \right] - \lambda_\mu \left[\int_y y P_X(y) dy - \mu \right]$$

$$\frac{\delta L}{\delta P_X(x)} = -\log P_X(x) - 1 - \lambda - \lambda_\mu x = 0$$

$$\log P_X^*(x) = -(1+\lambda) - \lambda_\mu x$$

$$P_x^*(x) = e^{-(\lambda+1)} e^{-\lambda_\mu x}$$

$$\int_a^b P_X^*(x) dx = e^{-(\lambda+1)} \int_a^b e^{-\lambda_\mu x} dx = 1$$

$$P_X^*(x) = \frac{e^{-\lambda_\mu x}}{\int_y e^{-\lambda_\mu y} dy}$$

an exponential distribution

makes sense if $\lambda > 0$

$y \in [0, +\infty)$

When you know the variance σ^2

$$\sigma^2 = \int (x-\mu)^2 P_X(x)\, dx$$

$$L = -\int_y P_X(y) \log P_X(y)\, dy - \lambda \left[\int_y P_X(y)\, dy - 1 \right]$$

$$- \lambda_\sigma \left[\int_y (y-\mu)^2 P_X(y)\, dy - \sigma^2 \right]$$

$$\frac{\delta L}{\delta P_X(x)} = -\log P_X(x) - 1 - \lambda - \lambda_\sigma (y-\mu)^2 = 0$$

$$\log P_X^*(x) = -(1+\lambda) - \lambda_\sigma (x-\mu)^2$$

$$P_X^*(x) = e^{-(1+\lambda)}\, e^{-\lambda_\sigma (x-\mu)^2}$$

$$P_X^*(x) = \frac{e^{-\lambda_\sigma (x-\mu)^2}}{\int_y e^{-\lambda_\sigma (y-\mu)^2}} \rightarrow \text{Normal (Gaussian) dist}$$

$$\sigma^2 = \frac{1}{\lambda_\sigma}$$

makes sense if $\lambda_\sigma > 0$

$x \in (-\infty, +\infty)$