Entropy & Information

Shannon Entropy – Recap!

Suppose we learn a random variable X

$H(X) = -\sum_x p_x \log p_x \equiv$ "uncertainty about X before learning it"

$\parallel$

"Amount of information we gain on learning X"

Note $\lim_{x \to} 0 \log 0 = 0$

$\log \equiv \log_2$

Suppose a source is producing data in the form of random variables $X_1\ X_2\ X_3 \ldots$

Suppose each random variable can take a character x_k with probability p_k.

What's the minimal physical resources required to store the data produced by the source?

Answ: $\underbrace{n \text{ symbol string can be compressed to } n H(X) \text{ symbols}}_{\text{Shannon's noiseless coding theorem}}$

eg. Suppose a source of information produces $1, 2, 3$ or 4 with probabilities $1/2,\ 1/4,\ 1/8,\ 1/8$

A naive binary encoding $1 = 00 \quad 2 = 01 \quad 3 = 10 \quad 4 = 11$

On average the length of a string with this encoding is

$$2 \times \frac{1}{2} + 2 \times \frac{1}{4} + 2 \times \frac{1}{8} + 2 \times \frac{1}{8} = 2$$

Then we can use the bias to reduce the amount of symbols required to store data from that source by using less characters to store commonly obtained symbols & more to obtain less likely ones.

eg. $1 = 0$ $2 = 10$ $3 = 110$ $4 = 111$

On average the length of a string with this encoding is

$$1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} = \frac{7}{4} \leq 2$$

New code is more efficient!

Sketch of general proof:

Binary case first - $X = 1$ with prob p
$\quad\quad = 0 \quad\quad 1-p$

Consider a data string of length n

In the limit of large n, a typical bit string will contain about $n(1-p)$ 0s and np 1s.

There are ${}^{n}C_{np}$ such typical strings

$$\log({}^{n}C_{np}) = \log\left(\frac{n!}{(np)!\, n(1-p)!}\right)$$

$$= \log n! - \log(np!) - \log(n(1-p)!)$$

Use Stirling approximation $\log(n!) \simeq n\log n - n$ (in limit of large n)

$$\log({}^{n}C_{np}) \approx n\log n - n - (np\log np - np + n(1-p)\log(n(1-p)) - n(1-p))$$

$$= -np\log p - n(1-p)\log(1-p) = n\,H(p)$$

$\Rightarrow$ no. of typical strings $\simeq 2^{nH(p)}$ binary entropy

Compression strategy – assign a positive integer to each of the possible typical bit strings.

There are $2^{nH(p)}$ such strings

so $2^{nH(p)}$ letters are required

& each letter can be encoded using $nH(p)$ bits.

Note – the completely uniform distribution cannot be compressed

ie. $H(\tfrac{1}{2}) = -\tfrac{1}{2}\log(\tfrac{1}{2}) - \tfrac{1}{2}\log(\tfrac{1}{2}) = -\log(\tfrac{1}{2}) = \log 2 = 1$

ie. n bits are 'encoded' in n bits

Generalization beyond binary case.

If letter k occurs with probability p_k in a string of length n each k will typically occur np_k times

There are $\dfrac{n!}{\prod_k (np_k)!}$ such typical strings

& $\dfrac{n!}{\prod_k (np_k)!} \simeq 2^{nH(X)} \Rightarrow nH(X)$ binary encoding possible

$\Rightarrow$ Operational interpretation of Shannon entropy!

Conditional Entropy & Mutual Information

Consider 2 random variables X & Y

- How is the information content of X related to Y?

Conditional entropy & Mutual information provide answers

But first: Joint Entropy

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

This is the total uncertainty about X & Y

Suppose we know the value of Y, so we have gained $H(Y)$ bits of information.

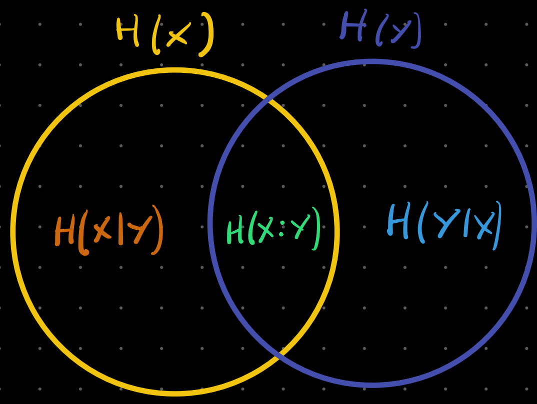
The conditional entropy of X on knowing Y is the remaining uncertainty in X on knowing Y.

$$H(X \mid Y) = H(X,Y) - H(Y)$$

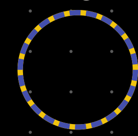
Mutual information measures the amount of information X and Y have in common - i.e. measures their correlations

$$H(X:Y) = H(X) + H(Y) - H(X,Y)$$

The following Venn diagram is a super useful tool to get a sense of their properties.



eg. if perfectly correlated



If completely uncorrelated

Can read off some properties straight from the venn:

- $0 \leq H(X|Y) \leq H(X)$
- $H(X|Y) \neq H(Y|X)$
- $0 \leq H(X:Y) \leq \min\{H(X), H(Y)\}$

Drawing Venn is helpful for providing an intuition but is not the full story – always prove inequalities also independently.

(Problem sheet for this week will provide many)

Relative Entropy – a measure of the closeness of 2 distributions. Useful for proving stuff.

$$H(p(x) \,\|\, q(x)) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) \longrightarrow = 0 \text{ if } p(x) = q(x) \; \forall x$$

$$= -H(X) - \Sigma_x \, p(x) \log(q(x))$$

- $H(p(x) \,\|\, q(x)) \geq 0$ (to prove this use $-\log(x) \geq \frac{1-x}{\ln 2}$

$$\Rightarrow H(p(x) \,\|\, q(x)) \geq \frac{1}{\ln 2} \underbrace{\sum_x p(x)\left(1 - \frac{q(x)}{p(x)}\right)}_{1-1=0})$$

- $H(p(x) \,\|\, {}^{1}/_{d} \,\|) = -H(X) - \Sigma_x \, p(x) \log(1/d)$

$$= \log(d) - H(X)$$

Shannon entropy $\equiv$ relative entropy to max uncertain distribution

Von Neumann Entropy

eigs of ρ

$$S(\rho) = -\mathrm{Tr}(\rho \log \rho) = -\sum_i \lambda_i \log \lambda_i$$
$$= H(\{\lambda_i\})$$

Similarly to classical case $S(\rho)$ quantifies the compressibility of quantum information

$\rho^{\otimes n}$ can be compressed to σ that lives on a Hilbert space H_c

with $\dim(H_c) = 2^{n S(\rho)}$

Intuition roughly the same $\Rightarrow$ can look at subspace corresponding to typical sequences of eigenvalues.

See Preskill's notes for a proof.

Important Properties

1) Pure states have zero entropy

$\rho = |\psi\rangle\langle\psi| \qquad \lambda = 1 \qquad S(\rho) = \log(1) = 0$

2) Invariance : $S(U\rho U^\dagger) = S(\rho)$ (eigenvalues are left unchanged)

3) Maximum : $\mathrm{Max}\, S(\rho) = S(I/d) = \log(d)$

4) Entropy of measurement:

Say you measure $M = \sum_j m_j |m_j\rangle\langle m_j|$

$$p(m_j) = \langle m_j|\rho|m_j\rangle$$

$$Y = \{m_j, p(m_j)\}$$

$$\Rightarrow \quad H(Y) \geq S(\rho)$$

Equivalent to the statement that replacing ρ in any basis with its decohered variant increases entropy.

ie. Killing off coherences increases entropy.

5) Additivity: $S(\rho_A \otimes \rho_B) = S(\rho_A) + S(\rho_B)$

"eigenvalues multiply – take log – entropies add"

6) Triangle Inequality

$$|S(\rho_A) - S(\rho_B)| \leq S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B)$$

Use Klein's Inequality: $S(\rho) \leq -\mathrm{Tr}(\rho \log \sigma)$

Let $\rho = \rho_{AB}$ & $\sigma = \rho_A \otimes \rho_B$ $\rightsquigarrow$ $S(\rho) \leq -\mathrm{Tr}(\rho^{AB}(\log(\rho^A) + \log(\rho^B)))$

$= -\mathrm{Tr}(\rho^A \log \rho^A) - \mathrm{Tr}(\rho^B \log \rho^B)$

$= S(\rho_A) + S(\rho_B)$

7) Concavity: $S(\sum_i p_i \rho_i) \geq \sum_i p_i S(\rho_i)$

"extra randomness only increases uncertainty"

$\rho_i = \sum_j \lambda_j^i |\lambda_j^i\rangle\langle\lambda_j^i|$

Let $\rho_{AB} = \sum_i p_i \rho_i \otimes |i\rangle\langle i|$

$\rho_A = \sum_i p_i \rho_i \qquad \rho_B = \sum_i p_i |i\rangle\langle i|$

$S(\sum_{ij} p_i \lambda_j^i |\lambda_j^i i\rangle\langle\lambda_j^i i|)$

$= -\sum_{ij} p_i \lambda_j^i \log p_i \lambda_j^i$

$- \sum_{ij} p_i \lambda_j^i \log p_i + \sum_{ij} p_i \lambda_j^i \log \lambda_j^i$

$H(\{p\}) + \sum_i p_i S(\rho_i)$

$S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B)$

$S(\rho_A)$ is $S(\sum_i p_i \rho_i)$, $S(\rho_B) \to H(\{p\})$

$\Rightarrow \cancel{H(\{p\})} + \sum_i p_i S(\rho_i) \leq S(\sum_i p_i \rho_i) + \cancel{H(\{p\})}$ ✓

Analogously to the classical case we can define:

Joint entropy $\quad S(\rho_{AB}) = -\mathrm{Tr}(\rho_{AB}\log(\rho_{AB}))$

Conditional entropy $\quad S(\rho_A|\rho_B) = S(\rho_{AB}) - S(\rho_B)$

Mutual information $\quad S(\rho_A:\rho_B) = S(\rho_A) + S(\rho_B) - S(\rho_{AB})$

Note that the Venn diagram breaks down in this case

eg. Conditional entropy can be negative

Say $\rho_{AB} = |\phi^+\rangle\langle\phi^+|$ $\quad \rho_A = \rho_B = \frac{I}{2}$

$S(\rho_{AB}) = 0$ $\quad S(\rho_A) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log(2) = 1$

$$S(\rho_A|\rho_B) = -1\,!$$

"Uncertainty in joint state is less than the reduced states"

Relative Entropy

- Not a distance measure (3)
- But can be used to measure the similarity between two quantum states (1) (2)

$$S(\rho\|\sigma) := \mathrm{Tr}\Big(\overset{\sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|}{\rho}\log\rho - \rho\log\overset{\sum_j \mu_j |\mu_j\rangle\langle\mu_j|}{\sigma}\Big)$$

$$= \sum_i \lambda_i\Big(\log\lambda_i - \sum_j |\langle\mu_j|\lambda_i\rangle|^2 \log(\mu_j)\Big)$$

Reduces to the classical relative entropy if diagonal in same basis but more generally depends on the overlap between their eigenbasis

Properties :

1) Positivity : $S(\rho\|\sigma) \geq 0$
2) Faithful : $S(\rho\|\sigma) = 0$ iff $\rho = \sigma$

clear from

3) Asymmetric : $S(\rho\|\sigma) \neq S(\sigma\|\rho)$
4) Unitarily invariant : $S(U\rho U^\dagger \| U\sigma U^\dagger) = S(\rho\|\sigma)$

Data processing Inequality

= holds for Unitary evolutions

$$S(\mathcal{E}(\rho) \| \mathcal{E}(\sigma)) \leq S(\rho \| \sigma) \qquad \forall\, \mathcal{E}$$

"There is no channel you can apply that will make ρ & σ more distinguishable"

Same also holds for 1 norm

$$\| \mathcal{E}(\rho) - \mathcal{E}(\sigma) \|_1 \leq \| \rho - \sigma \|_1 \qquad \forall\, \mathcal{E}$$

(But it doesn't hold for 2-norm)

Mixed State Fidelity

$$F(\rho,\sigma) = \mathrm{Tr}\left(\sqrt{\rho^{1/2}\sigma\rho^{1/2}}\right)$$

Case 1: ρ & σ commute $\rightarrow$ $\rho = \sum_i r_i |i\rangle\langle i|$ $\quad \sigma = \sum_i s_i |i\rangle\langle i|$

$$\begin{aligned} F(\rho,\sigma) &= \mathrm{Tr}\left(\sqrt{\textstyle\sum_i r_i s_i |i\rangle\langle i|}\right) \\ &= \mathrm{Tr}\left(\textstyle\sum_i \sqrt{r_i s_i}\, |i\rangle\langle i|\right) \\ &= \textstyle\sum_i \sqrt{r_i s_i} \\ &= F(\underline{r}, \underline{s}) \end{aligned}$$

Classical fidelity

Case 2: σ, $\rho = |\psi\rangle\langle\psi|$ $\qquad (|\psi\rangle\langle\psi|)^2 = \rho \Rightarrow \rho^{1/2} = |\psi\rangle\langle\psi|$

$$\begin{aligned} F(\rho,\sigma) &= \mathrm{Tr}\left(\sqrt{|\psi\rangle\langle\psi|\sigma|\psi\rangle\langle\psi|}\right) \\ &= \mathrm{Tr}\left(\sqrt{\langle\psi|\sigma|\psi\rangle}\ \underbrace{\sqrt{|\psi\rangle\langle\psi|}}_{|\psi\rangle\langle\psi|}\right) \\ &= \sqrt{\langle\psi|\sigma|\psi\rangle} \end{aligned}$$

Fidelity between pure and mixed state is equal to the overlap

Case 2b. $\sigma = |\phi\rangle\langle\phi|$ $\qquad F(\rho,\sigma) = |\langle\psi|\phi\rangle|$

Note the lack of mod square here → this is a matter of convention I'm following N&C here.

General case? Operational interpretation provided by Uhlmann's Theorem.

Uhlmann's Theorem:

max over all possible purifications of ψ & φ

$$F(\rho,\sigma) = \max_{|\psi\rangle,|\varphi\rangle} |\langle\psi|\varphi\rangle|$$

where $\rho_S = Tr_R(|\psi\rangle\langle\psi|_{RS})$ & $\sigma_S = Tr_R(|\psi\rangle\langle\psi|_{RS})$

proof - exercise sheet this week.

Data processing inequality also holds here

$$F(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \geq F(\rho,\sigma)$$