Math 5251 Probability (Some of §1.4, 1.5)

We want to talk about the average length of code words when we imagine the source words

$$W = \{w_1, w_2, \dots, w_m\}$$

being emitted randomly with certain probabilities $\{p_1, p_2, \dots, p_m\}$ from a memoryless source, meaning the previous words don't affect the probability p_i that the next word is w_i.

Not a reasonable model for most messages locally, but not so unreasonable for long messages from a source with known word frequencies.

EXAMPLE For one of our earlier encodings

$W = \{A, B, C, D, E\}$ with $\Sigma = \{0, 1, 2\}$

$$\begin{array}{c} W = \{A, B, C, D, E\} \\ f \downarrow \\ \Sigma^* \quad \{0, 1, 20, 21, 22\} = C \end{array}$$

if we assume source words appear with these probabilities

$$P(\overset{w_1}{A}) = \overset{p_1}{\tfrac{1}{2}}$$

$$P(\underset{w_2}{B}) = P(\underset{w_3}{C}) = P(\underset{w_4}{D}) = P(\underset{w_5}{E}) = \underset{p_2, p_3, p_4, p_5}{\tfrac{1}{8}}$$

then what is the average length of a codeword ?

DEF'N: average codeword length

$$:= p_1 \ell(w_1) + \dots + p_m \ell(w_m)$$

$$= \frac{1}{2} \cdot \underbrace{1}_{\ell(0)} + \frac{1}{8} \cdot \underbrace{1}_{\ell(1)} + \frac{1}{8} \cdot \underbrace{2}_{\ell(20)} + \frac{1}{8} \cdot \underbrace{2}_{\ell(21)} + \frac{1}{8} \cdot \underbrace{2}_{\ell(22)}$$

$$= \frac{1}{2} + \frac{1}{8} + \frac{6}{8} = \frac{11}{8} = 1.375$$

This is an example of the expected value of a random variable on a probability space...

DEF'N: A finite probability space is a finite set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$ (like $W = \{w_1, \ldots, w_m\}$) with probabilities $\mathbb{P}(\omega_i) = p_i$ assigned to each ω_i

"the probability that sampling from Ω produces ω_i is p_i"

such that
$$\begin{cases} p_i \in [0,1] \\ p_1 + p_2 + \ldots + p_m = 1. \end{cases}$$

DEF'N: A random variable X on Ω is a function
$$X : \Omega \longrightarrow \mathbb{R}$$
$$\omega_i \longmapsto X(\omega_i)$$

and its expected value
$$\mathbb{E}X := \sum_{i=1}^{m} p_i X(\omega_i)$$

EXAMPLES

(1) $\Omega = W = \{\overset{\omega_1}{A}, \overset{\omega_2}{B}, \overset{\omega_3}{C}, \overset{\omega_4}{D}, \overset{\omega_5}{E}\}$

with $(p_1, p_2, p_3, p_4, p_5)$

$$= \left(\tfrac{1}{2}, \tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{8}\right)$$

is a probability space, and we have

a random variable $X : \Omega \longrightarrow \mathbb{R}$

$\overset{\shortparallel}{\text{codeword length}}$ $\qquad \omega_i \longmapsto \ell(f(\omega_i))$

where f:

$$\begin{aligned} A &\longmapsto 0 \\ B &\longmapsto 1 \\ C &\longmapsto 20 \\ D &\longmapsto 21 \\ E &\longmapsto 22 \end{aligned}$$

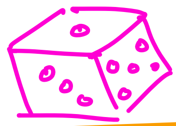
whose expected value

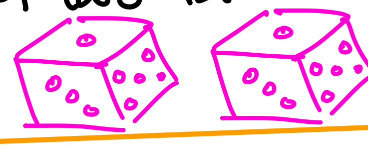
$$\mathbb{E}X = \text{average length} = \sum_{i=1}^{5} p_i\, \ell(f(\omega_i))$$

$$= \frac{1}{8}\cdot 1 + \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 2$$

$$= \frac{11}{8} = 1.375 \text{ from before}$$

(2) What is the expected value for the roll of one fair 6-sided die? Total of two fair dice?





$$\Omega_1 = \{\overset{\omega_1}{1}, \overset{\omega_2}{2}, \overset{\omega_3}{3}, \overset{\omega_4}{4}, \overset{\omega_5}{5}, \overset{\omega_6}{6}\} \text{ = outcomes for one die}$$

$$P(\omega_i) = \frac{1}{6} \ \forall i$$

called the uniform probability space on Ω when $P(\omega_i) = \dfrac{1}{|\Omega|} \ \forall i$

$$\Omega_1 \xrightarrow{X} \mathbb{R} \qquad X(i) = i$$

$$\mathbb{E}X = \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 2 + \dots + \frac{1}{6}\cdot 6$$

$$= \frac{1}{6}(1+2+3+4+5+6) = \frac{1}{6}\cdot 21 = \frac{7}{2} = 3.5$$

$$\Omega_2 = \Omega_1 \times \Omega_1 = \left\{ \begin{matrix} (1,1), (1,2), \dots, (1,6), \\ (2,1), (2,2), \dots, (2,6), \\ \vdots \qquad\qquad \vdots \\ (6,1), (6,2), \dots, (6,6) \end{matrix} \right\} \quad P((i,j)) = \frac{1}{36} = \frac{1}{|\Omega_2|}$$

the uniform probability space again

$$\Omega_2 \xrightarrow{X} \mathbb{R} \qquad X(i,j) = i+j = \text{total of the dice}$$

$$\mathbb{E}X = \frac{1}{6}(1+1) + \frac{1}{6}(1+2) + \dots + \frac{1}{6}(6+6)$$

$$= \frac{1}{36}\cdot 2 + \frac{2}{36}\cdot 3 + \frac{3}{36}\cdot 4 + \frac{4}{36}\cdot 5 + \frac{5}{36}\cdot 6 + \frac{6}{36}\cdot 7 = \frac{252}{36} = 7$$

$$\frac{1}{36}\cdot 12 + \frac{2}{36}\cdot 11 + \frac{3}{36}\cdot 10 + \frac{4}{36}\cdot 9 + \frac{5}{36}\cdot 8$$

Entropy of a sample space (§2.2)

In 1948, Claude Shannon tried to quantify how much information we acquire when we are told the outcome ω_i of a sampling from a probability space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$ having probabilities $p_1, \ldots, p_m$ so $P(\omega_i) = p_i$.

We will eventually call this the Shannon entropy

$$H(\Omega) = H(p_1, p_2, \ldots, p_m) \text{ of } \Omega$$

The idea is to first define/normalize by saying the self-information $I(\omega_i)$ for an outcome heads/tails of a fair coinflip

$$\Omega = \{\underset{\omega_1}{\text{heads}}, \underset{\omega_2}{\text{tails}}\}$$

$$P(\text{heads}) = \tfrac{1}{2} = p_1 \qquad P(\text{tails}) = \tfrac{1}{2} = p_2$$

is $I(\text{heads}) := I(\text{tails}) := 1 \text{ bit}$

Then if one did 2 coin flips, each outcome would have $P(\omega_i) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = P(\text{heads, heads})$
$= P(\text{heads, tails})$
and should have twice the self-information, that is, $I(\omega_i) = 2$ bits.

Similarly k coin flips have outcomes ω_i with all $P(\omega_i) = \underbrace{\frac{1}{2} \cdot \frac{1}{2} \cdots \frac{1}{2}}_{k \text{ times}} = \frac{1}{2^k}$

and should have $I(\omega_i) = k$ bits

$$= -\log_2\left(\frac{1}{2^k}\right) = -\log_2(p_i)$$

This motivates the choice that...

DEF'N: An outcome ω_i in Ω having $Pr(\omega_i) = p_i$ has self-information $I(\omega_i) := -\log_2(p_i)$ and the (Shannon) entropy/information for Ω is the expected value EI of the self-information:

$$H(\Omega) := H(p_1, p_2, \dots, p_m)$$
$$= -p_1 \log_2(p_1) - \dots - p_m \log_2(p_m)$$
$$= -\sum_{i=1}^{m} p_i \log_2(p_i) \quad \text{in bits.}$$

EXAMPLES

(1) $\Omega = \{A, B, C, D, E\}$ with

$$(p_1, p_2, p_3, p_4, p_5) = \left(\tfrac{1}{2}, \tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{8}\right)$$

has $H(\Omega) = \frac{1}{2}\cdot 1 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3$

$$= \frac{1}{2} + \frac{12}{8} = \frac{1}{2} + \frac{3}{2} = 2$$

(2) $\Omega = \{\omega_1, \dots, \omega_m\}$ with uniform distribution

$$(p_1, \dots, p_m) = \left(\tfrac{1}{m}, \dots, \tfrac{1}{m}\right)$$

has $I(\omega_i) = -\log_2\left(\frac{1}{m}\right) = \log_2(m) \quad \forall i$

and $H(\Omega) = H\left(\frac{1}{m}, \dots, \frac{1}{m}\right) = \frac{1}{m}\log_2(m) + \dots + \frac{1}{m}\log_2(m)$

$$= \log_2(m)$$

e.g.

$$H\left(\tfrac{1}{2}, \tfrac{1}{2}\right) < H\left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right) < H\left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}\right) < \dots$$

$= \log_2(2)$ $= 1$ bit; $= \log_2(3)$ bits; $= \log_2(4)$ $= 2$ bits

---

Q: Why is $H(\Omega)$ always nonnegative?

How well does $H(\Omega)$ capture the notion of the information conveyed by knowing the outcome ω_i from a sampling of Ω? A supporting result...

THEOREM (Roman THM 1.1.1) Any function $H(p_1, \dots, p_m)$ defined for all sequences $(p_1, \dots, p_m)$ with $p_i \in [0,1]$, $\sum_{i=1}^{m} p_i = 1$

having these properties

(i) H is continuous as a function of the p_i

(ii) $H\left(\frac{1}{n}, \dots, \frac{1}{n}\right) < H\left(\frac{1}{n+1}, \dots, \frac{1}{n+1}\right)$ for all $n = 1, 2, \dots$

(iii) $H(\underbrace{p_1, \dots, p_r}_{\text{let } p = p_1 + \dots + p_r}, \underbrace{q_1, \dots, q_s}_{\text{let } q = q_1 + \dots + q_s}) =$

$$H(p,q) + p\,H\left(\frac{p_1}{p}, \dots, \frac{p_r}{p}\right) + q\,H\left(\frac{q_1}{q}, \dots, \frac{q_s}{q}\right)$$

must be of the form

$$H(p_1, \dots, p_m) = -\sum_{i=1}^{m} p_i \log_b(p_i)$$

CONVENTION: $p \log(p) = 0$ if $p = 0$

for some choice of base $b > 1$

NOTE: This pins $H(\Omega)$ down up to a multiple: $\log_b(p) = \frac{\log_2(p)}{\log_2(b)}$

(Shannon picked $b = 2$)