Shannon Entropy

Def (Support) X: random variable (discrete)

$$\mathrm{supp}(X) = \{x \mid P_X(x) > 0\} \quad \text{where } P_X(x) = P(X=x)$$

Def (Entropy) X: random var. w/ finite supp

$$H(X) = \sum_{x \in \mathrm{supp}(X)} -P_X(x) \log_2 P_X(x)$$

Thm (Shannon)

φ is a coding function of Y if $\varphi : \mathrm{supp}(Y) \xrightarrow[\text{inj}]{} \{0,1\}^*$,

L = length of $\varphi(Y)$.

If $X_1, \cdots, X_n$ iid, φ is a coding fcn of $X_1, \cdots X_n$,

Then (1) $\forall \varphi$, $\mathbb{E} L \geq nH(X) + o(n)$

(2) $\exists \varphi$ s.t. $\mathbb{E} L \leq nH(X) - o(n)$

pf sketch

Ex
$$\begin{cases} \text{If } Y \sim \mathrm{unif}(1, \cdots, r), \text{ then} \\ (1)\ \mathbb{E} L \geq \log_2 r - 2 \\ (2)\ \exists \varphi \text{ s.t. } \mathbb{E} L \leq \log_2 r \end{cases}$$

For general X, let $\mathrm{supp}(X) = \{x_1, \cdots, x_k\}$ and $p_i = P_X(x_i)$

Then whp $\#\{j \mid X_j = x_i\} = (1 \pm o(1)) np_i$

Thus, $X_1, \cdots, X_j$ is "almost" uniform on $\binom{n}{np_1 \cdots np_k}$ elements.

$$\log_2 \binom{n}{np_1, \cdots np_k} = \log_2(n!) - \sum_i \log_2((np_i)!)$$

$$= n\log_2 n - n\log_2 e + o(n) - \sum_i \left(np_i \log_2(np_i) - np_i \log_2 e + o(n)\right)$$

$$= n\left(\Sigma - p_i \log_2 p_i\right) + o(n)$$

$$= nH(X) + o(n)$$

Remark $H(X)$ "measures" the expected # bits need to encode X

Def (Conditional Entropy)

X, Y: random var w/ finite supp.

$$H(X|Y) = H(X,Y) - H(Y) \quad (1)$$

$$= \mathbb{E}_y H(X|Y=y) \quad (2)$$

$$= \sum_{x,y} -P_{X,Y}(x,y)\log_2 \frac{P_{X,Y}(x,y)}{P_Y(y)} \quad (3)$$

pf $(1)=(3)$: $H(Y) = \sum_y -P_Y(y)\log_2 P_Y(y)$

$$= \sum_{xy} -P_{XY}(x,y)\log_2 P_Y(y)$$

$$H(X,Y) = \sum_{xy} -P_{XY}(x,y)\log_2 P_{XY}(x,y)$$

$(2)=(3)$: $(3) = \sum_{xy} -P_Y(y)\left(\frac{P_{XY}(xy)}{P_Y(y)}\log_2 \frac{P_{XY}(xy)}{P_Y(y)}\right)$

Prop (Chain rule)

$$H(X_1, \cdots, X_n) = H(X_1|X_2 \cdots X_n) + H(X_2|X_3 \cdots X_n) + \cdots + H(X_n)$$

Prop (non-negativity)

$$H(X) \ge 0 \quad , \quad H(X|Y) \ge 0$$

Prop (uniform bound)

$H(X) \le \log_2 |supp(X)|$, "=" iff X uniform

pf $-x\log_2 x$ is concave, so by Jensen

$$\frac{\sum - P_X(x)\log P_X(x)}{K} \le -\frac{1}{K}\log_2\frac{1}{K} \quad \text{where } K = |supp(X)|$$

Prop (Drop conditioning)

- $H(X|Y) \le H(X)$, "=" iff X, Y independent
- $H(X|YZ) \le H(X|Z)$ "=" iff X, Y conditionally independent given Z
  (i.e. $(X|Z=z), (Y|Z=z)$ indep $\forall z$)

pf $$H(X|Y) = \sum_x \underbrace{\sum_y P_Y(y)}_{\mathbb{E}_y} \left(-\frac{P_{XY}(xy)}{P_Y(y)} \log_2 \frac{P_{XY}(xy)}{P_Y(y)} \right)$$

$$\overset{\text{Jensen}}{\le} \sum_x \left(-\left(\sum_y P_Y(y) \cdot \frac{P_{XY}(xy)}{P_Y(y)} \right) \log_2 \left(\sum_y P_Y(y) \frac{P_{XY}(xy)}{P_Y(y)} \right) \right)$$

$$= \sum_x - P_X(x)\log_2 P_X(x) = H(X)$$

$$H(X|YZ) = \mathbb{E}_z H(X|Y, Z=z) \le \mathbb{E}_z H(X|Z=z) = H(X|Z)$$

Cor (subaddivity)

- $H(X,Y) \le H(X) + H(Y)$ "=" iff X, Y indep
- $H(X,Y|Z) \le H(X|Z) + H(Y|Z)$ "=" iff X, Y C.I. given Z

(4)

pf $H(X,Y|Z) = H(X|Y,Z) + H(Y|Z)$

$\leq H(X|Z) + H(Y|Z)$

Cor (submodularity)

$$H(X,Y,Z) + H(Z) \leq H(X,Z) + H(Y,Z)$$

pf $\underset{\|}{H(X,Y|Z) + 2H(Z)} \leq \underset{\|}{H(X|Z) + H(Y|Z) + 2H(Z)}$

Prop Suppose f is a deterministic function

- $H(f(X)|X) = 0$
- $H(X, f(X,Y) | Y) = H(X|Y)$
- $H(X|Y, f(Y)) = H(X|Y)$
- $H(f(X)|Y) \leq H(X|Y)$
- $H(X|f(Y)) \geq H(X|Y)$

pf (1) $= \mathbb{E}_x H(f(X)|X=x) = 0$

(2) $H(X, f(XY)|Y) = \underbrace{H(f(XY)|XY)}_{= 0} + H(X|Y)$

(3) $H(X|Y f(Y)) = H(X f(Y)|Y) - H(f(Y)|Y)$

$= H(X|Y)$

(4) $H(f(X)|Y) = H(X f(X)|Y) - H(f(X)|Y)$

$\leq H(X|Y)$

(5) $H(X|Y) = H(X|Y f(Y)) \leq H(X|f(Y))$

Bit-wise info

Q What's the max # of sets $A_1, \cdots, A_m \subseteq 2^{[n]}$

s.t. all bits are unbalanced ?

$$\#\{j \mid i \in A_j\} = \frac{m}{3} \quad \forall i \in [n]$$

Easy for balanced: if $\#\{j \mid i \in A_j\} = \frac{m}{2}$ $\forall i$

then max $m = 2^n$ by taking all sets.

A Construction: $\binom{[n]}{\frac{n}{3}} \geq \frac{1}{n+1} 2^{h(\frac{1}{3})n}$, $h(p) = H(\mathrm{Bern}(p))$

$$= -p\log_2 p - (1-p)\log_2(1-p)$$

Upper bound: Let $\mathbb{1}_{A_j}$ be the indicator vector of A_j

$$\underset{\{0,1\}^n}{\underset{\cap}{(X_1, \cdots, X_n)}} = \mathrm{unif}(\mathbb{1}_{A_j}, j \in [m])$$

$$H(X_1, \cdots, X_n) = \log m$$

$$\wedge|$$

$$H(X_1) + \cdots + H(X_n) = n\,h(\tfrac{1}{3}) \Rightarrow m \leq 2^{h(\frac{1}{3})n}$$

Union-closed Conjecture

Conjecture If $\mathcal{F} = \{A_1, \dots, A_m\} \subseteq 2^{[n]}$ is union-closed, i.e.

$$A_i \cup A_j \in \mathcal{F} \quad \forall i, j$$

then $\exists i$ s.t. $\#\{j \mid i \in A_j\} \geq \frac{m}{2}$

Construction $\mathcal{F} = 2^{[n]}$, each i is in $\frac{m}{2}$ sets from $\mathcal{F}$

Thm (Gilmer, many ppl)

$\exists i$ s.t. $\#\{j \mid i \in A_j\} \geq cm$

Gilmer : $c = 0.01$

many ppl : $c = \frac{3-\sqrt{5}}{2} \approx 0.381966$

Current best : $c \approx 0.38271$

Pf Idea Unif X, Y from $\mathcal{F}$

Indicator vector $(X_1, \dots, X_n), (Y_1, \dots, Y_n)$

Note that $X \cup Y$ is supported on $\mathcal{F}$

$\Rightarrow H(X \cup Y) \leq H(X)$

Goal If $\forall i$, $\#\{j \mid i \in A_j\} < cm$, then $H(X \cup Y) > H(X)$

pf $\cdot$ $X_{<i} = (X_1, \dots, X_{i-1})$

$$H(X \cup Y) = \sum_i H((X \cup Y)_i \mid (X \cup Y)_{<i}) \quad \text{chain rule}$$

$$\geq \sum_i H((X \cup Y)_i \mid X_{<i}, Y_{<i})$$

⑦

$$H(X) = \sum_i H(X_i \mid X_{<i}) = \sum_i H(X_i \mid X_{<i}, Y_{<i})$$

<u>WTS</u> $H((X \cup Y)_i \mid X_{<i}, Y_{<i}) \geq H(X_i \mid X_{<i}, Y_{<i})$ --- ①

Let $p = \mathbb{P}(X_i = 1 \mid X_{<i}, Y_{<i})$

$q = \mathbb{P}(Y_i = 1 \mid X_{<i}, Y_{<i})$

- ① $\Leftrightarrow$ $\mathbb{E}(h(p+q-pq)) \geq \mathbb{E}\, h(p)$

  where $\mathbb{E}$ over $p, q \sim \mu \times \mu$, μ given by $X_{<i}$

- $\#\{j \mid i \in A_j\} < cm \Leftrightarrow \mathbb{E} p < c$

For $i=1$, $p=q$ are deterministic, and

$h(2p-p^2) > h(p)$ is strict

(Need $p < c = \frac{3-\sqrt{5}}{2}$ here)

<u>WTS</u> $p, q \sim \mu$ indep, $\mathbb{E} p \leq c$. Then $\mathbb{E}(h(p+q-pq)) \geq \mathbb{E}\, h(p)$

μ on $[0,1]$

Olympiad problem !!