Information Theory & Coding

Sept 22nd 20

Last week:

- Source Coding, codes, - injective codes
  - uniq. decodable
  - prefix-free

$$\underset{\text{prefix-free}}{c \text{ is}} \Rightarrow \underset{\text{u.d.}}{c \text{ is}} \Rightarrow \begin{cases} \text{injective} \\ c^n \text{ is also injective } \forall n \end{cases}$$

$$\not\Leftarrow \qquad \overset{?}{\Leftarrow}$$

$$\text{KraftSum}(c) = \sum_{u \in \mathcal{U}} 2^{-\text{length}(c(u))}$$

c is injective $\Rightarrow$ $KS(c) \leq \log_2(1+|\mathcal{U}|)$

c is u.d $\Rightarrow$ $\boxed{KS(c) \leq 1}$ Kraft's Inequality

c is p-f $\Rightarrow$ $KS(c) \leq 1$

if $l: \mathcal{U} \to \{0,1,2,3,\dots\}$ & $\sum_{u \in \mathcal{U}} 2^{-l(u)} \leq 1$

$\Rightarrow$ $\exists$ p-f code $c: \mathcal{U} \to \{0,1\}^*$ s.t.

$\forall u$ length$(c(u)) = l(u)$

$$\left\{\begin{array}{l} c \text{ is p.f} \\ c \text{ is u.d} \end{array}\right. \Rightarrow KS(c) \le 1$$

$$\Rightarrow \boxed{E[\text{length}(c(U))] \ge H(U)}$$

with $H(U) = \sum_{u \in \mathcal{U}} p(u) \log_2 \frac{1}{p(u)}$

Thm: for any U, $\exists$ a p.f code c s.t

$$E[\text{length}(c(U))] \le H(U) + 1$$

Pf: Let $\ell(u) = \left\lceil \log_2 \frac{1}{p(u)} \right\rceil \ge -\log_2 p(u)$

note: $2^{-\ell(u)} \le p(u)$

so $\sum 2^{-\ell(u)} \le 1$ so ℓ satisfies Kraft's Ineq

$\Rightarrow \exists$ a p.f code c with

$$\text{length}(c(u)) = \ell(u) \le \log_2 \frac{1}{p(u)} + 1$$

$$\Rightarrow \underbrace{\sum p(u) \, \text{length}(c(u))}_{E[\text{length } c(U)]} \le \sum p(u) \log_2 \frac{1}{p(u)} + \sum p(u) \cdot 1 = H(U) + 1$$

//

Corollary: suppose we have a information source (i.e., a sequence $U_1 U_2 U_3 \ldots$ of RVs $\equiv$ a stochastic process)

We can imagine constructing a code that describes n letters at a time:

$$c_n : \mathcal{U}^n \to \{0,1\}^*$$

any u.d. c_n

$$\frac{H(U_1 \ldots U_n) + 1}{n} \geq E\left[\frac{1}{n}\,\mathrm{length}\left(c_n(U_1 \ldots U_n)\right)\right] \geq \frac{1}{n} H(U_1 \ldots U_n)$$

$\exists$ a p.f. code c_n

The "most efficient code c_n" (u.d.) has

$$\frac{1}{n} H(U_1 \ldots U_n) \leq \#\text{ of bits/letter} \leq \frac{1}{n} H(U_1 \ldots U_n) + \frac{1}{n}$$

This motivates us to give $H(\,)$ a name

Def: given a discrete random variable U, taking values in $\mathcal{U}$, we define

$$H(U) = \sum_{u \in \mathcal{U}} p(u) \log_2 \frac{1}{p(u)} \quad \text{with } p(u) = \Pr(U = u).$$

as the Entropy of U.

Note $H(U) = E\left[\log_2 \frac{1}{p(U)}\right]$

Example: $\mathcal{U} = \{a, b, c\}$ with

$\Pr(U=a) = \frac{1}{2}$ $\quad \Pr(U=b) = \frac{1}{4}$ $\quad \Pr(U=c) = \frac{1}{4}$

$$\log_2 \frac{1}{p(U)} = \begin{cases} \log_2 2 & \text{when } U=a \\ \log_2 4 & \text{when } U=b \\ \log_2 4 & \text{when } U=c \end{cases}$$

$$= \begin{cases} 1 = \log_2 2 & \text{with prob } 1/2 \\ 2 = \log_2 4 & \text{with prob } 1/2 \end{cases}$$

$H(U) = \frac{3}{2}$ //

Example: U and V are RVs.

U	V	prob
a	α	1/2
a	β	1/4
b	α	1/4
b	β	0

$H(UV)$

Remarks $E(f(U)) \triangleq \sum_u p(u) f(u) \triangleq \sum_{u:\, p(u) \neq 0} p(u) f(u)$

[also $\lim_{p\to 0^+} p\log p = 0$.] . Consequently

$H(UV) = 3/2$

$H(U) = \frac{3}{4}\log_2\frac{4}{3} + \frac{1}{4}\log_2 4$

$H(V) = 1/$

$H(UV)$ vs. $H(U) + H(V)$

$3/2$ $\qquad$ $\frac{3}{2}\log\frac{4}{3} + \underbrace{\frac{1}{2}\log 4}_{1}$

$\cancel{1/2}$ vs $\frac{3}{\cancel{2}}\log\frac{4}{3}$

$1 + 3\log 3$ vs $3\log 4$

$1 + \log 27$ vs $\log 64$

$\log 54$ vs $\log 64$

$\leq$

//.

Conditional entropy, Joint entropy

$$H(UV) = \sum_{u,v} p(uv) \log_2 \frac{1}{p(uv)}$$ (Joint Entropy)

$$H(U_1 \ldots U_n) = \sum_{u_1 \ldots u_n} p(u_1 \ldots u_n) \log_2 \frac{1}{p(u_1 \ldots u_n)}$$

$$= E\left[\log_2 \frac{1}{p(U_1 \ldots U_n)}\right]$$

$$\underline{p(u_1 \ldots u_n)} = \Pr(U_1 \ldots U_n = u_1 \ldots u_n)$$

$$= \Pr(U_1 = u_1, U_2 = u_2, \ldots U_n = u_n)$$

$$= \Pr(U_1 = u_1) \Pr(U_2 = u_2 | U_1 = u_1) \times$$

$$\Pr(U_3 = u_3 | U_1 = u_1, U_2 = u_2) \ldots \times$$

$$\Pr(U_n = u_n | U_1 = u_1, U_2 = u_2 \ldots U_{n-1} = u_{n-1})$$

(chain rule ~~for~~ probabilities)

$$\Rightarrow \log_2 \frac{1}{p(U_1 \ldots U_n)} = \log \frac{1}{p(U_1)} + \log \frac{1}{p(U_2 | U_1)} + \ldots$$

$$\ldots + \log_2 \frac{1}{p(U_n / U_1 \ldots U_{n-1})}$$

$$\Rightarrow H(U_1 \dots U_n) = H(U_1) + E\left[\log \frac{1}{p(U_2|U_1)}\right] + \dots + E\left[\log \frac{1}{p(U_n|U_1 \dots U_{n-1})}\right]$$

$$\stackrel{\Delta}{=} H(U_1) + H(U_2|U_1) + \dots + H(U_n|U_1 \dots U_{n-1})$$

Conditional entropies.

$$\boxed{H(U|V) = \sum_{u,v} p(uv) \log \frac{1}{p(u|v)}}$$

$$\neq \sum_{u,v} p(u|v) \log \frac{1}{p(u|v)}$$

$$= \sum_{u,v} p(v) p(u|v) \log \frac{1}{p(u|v)}$$

$$= \sum_{v} p(v) \boxed{\sum_{u} p(u|v) \log \frac{1}{p(u|v)}}$$

$$= \sum_v p(v)\, \underbrace{H(U|V=v)}$$

$$\sum_u p(u|v) \log \frac{1}{p(u|v)}.$$

$$H(U|V) = \sum_{u,v} p(uv) \log \frac{p(v)}{p(uv)}$$

$$H(U) = \sum_{u,v} p(uv) \log \frac{1}{p(u)} = \left(\sum_u p(u) \log \frac{1}{p(u)} \right)$$

$$H(U|V) - H(U) = \sum_{u,v} p(uv) \log \frac{p(u)p(v)}{p(uv)}$$

$$\leq \log \left(\sum_{u,v} \cancel{p(uv)} \frac{p(u)p(v)}{\cancel{p(uv)}} \right)$$

$$= \log 1$$

$$= 0.$$

So: $\underline{\text{Thm}}$: $H(U|V) \leq H(U)$

$=$ iff $p(uv) = p(u)p(v)$

$\equiv$ U & V indep.

Corollary : $H(UV) \leq H(U) + H(V)$

$=$ iff U & V indep.

Pf : $\underbrace{H(UV)} = H(V) + \underbrace{H(U|V)}$ //

Note the following

$$H(U) - H(U|V) = \underline{H(U) + H(V) - H(UV)}$$

$$= H(V) - H(V|U)$$

$$\triangleq I(U;V)$$

Def : given two Random Variables U & V we define the mutual Information $I(U;V)$ as

$$\begin{aligned} I(U;V) &= H(U) - H(U|V) \\ &= H(V) - H(V|U) \\ &= H(U) + H(V) - H(UV). \\ &= I(V;U) \end{aligned}$$

Observations:

① if $U_1 U_2 U_3 \dots$ are iid

$$H(U_1 \dots U_n) = H(U_1) + H(U_2 | U_1) + \dots$$
$$+ \dots + H(U_n | U_1 \dots U_{n-1})$$
$$= H(U_1) + H(U_2) + \dots + H(U_n)$$

independence

$$= n H(U_1) \quad \text{identical distrib.}$$

$$\Rightarrow \frac{1}{n} H(U_1 \dots U_n) = H(U_1)$$

Then the most efficient (u.d) code c_n^* for $U_1 \dots U_n$

has $0 \le \frac{1}{n} E[\text{length } c_n^*(U_1 \dots U_n)] - H(U_1) \le \frac{1}{n}$

$$\Rightarrow \lim_{n \to \infty} \frac{1}{n} E[\text{length } c_n^*(U_1 \dots U_n)] = H(U_1)$$

if $(U_1 V_1)$, $(U_2 V_2)$, ... $(U_n V_n)$ are iid

$$\Rightarrow \lim_{n\to\infty} E\left[\frac{1}{n}\,\text{length}\ C^*_{n,UV}(U_1V_1U_2V_2\ldots U_nV_n)\right] = H(UV)$$

$$H(UV)$$

$$\frac{1}{n}E\,\text{len}\left(C^*_{n,UV}(U_1V_1\ldots U_nV_n)\right)$$

$$\le \frac{1}{n}E\,\text{len}\left(C^*_{n,U}(U_1\ldots U_n)\right) \to H(U)$$

$$+ \frac{1}{n}E\,\text{len}\left(C^*_{n,V}(V_1\ldots V_n)\right) \to H(V)$$

$$\Rightarrow \quad H(UV) \le H(U) + H(V).$$

Also: to encode $U_1V_1U_2V_2\ldots U_nV_n$, we could have first described $V_1V_2\ldots V_n$ using a good code for V, will take $H(V)$ bits/letter, then describe $U_1,U_2\ldots U_n$ conditional on $V_1\ldots V_n$, this will take $\approx H(U|V)$ bits/letter, total length

$H(V) + H(U|V)$ bits/letter $= H(UV)$.

In words $\underset{u}{\overset{I(U;V)}{H(U) - H(U|V)}}$ is measuring the saving in bits/letter to describe U when V is known.

So far:

– Entropy $H(U) = \sum_u p(u) \log \frac{1}{p(u)}$

$$H(UV) = \sum_{uv} p(uv) \log \frac{1}{p(uv)}$$

$$H(U_1 \ldots U_n) = \sum \text{ arbitrary}$$

– Chain Rule: $H(U_1 \ldots U_n) = H(U_1) + H(U_2|U_1) + \ldots + H(U_n|U_1 \ldots U_{n-1})$

– Cond. Entropy: $H(U|V) = \sum_{u,v} p(uv) \log \frac{1}{p(u|v)}$.

$$\left.\begin{array}{l} - \; H(UV) \leq H(U) + H(V) \\ \equiv \; H(U|V) \leq H(U) \\ \equiv \; I(U;V) \geq 0. \end{array}\right\} \text{equality iff } U \text{ \& } V \text{ indep.}$$

Conditional Mutual Information:

$$I(U;V) = H(U) + H(V) - H(UV) = H(U) - H(U|V)$$
$$= H(V) - H(V|U)$$

the natural generalization is

$$I(U;V|W) = H(U|W) + H(V|W) - H(UV|W)$$
$$= H(UW) - H(W) + H(VW) - H(W) - H(UVW) + H(W)$$
$$= H(UW) + H(VW) - H(UVW) - H(W)$$
$$= H(U|W) - H(U|WV)$$
$$= H(V|W) - H(V|UW)$$

Thm: $I(U;V|W) \geq 0$.

equality if U, V are indep conditional on W.

$\equiv$ [$U - W - V$ form a Markov chain.]