ORIE 4742 - Info Theory and Bayesian ML

Lecture 4: Source Coding

February 4, 2020

Sid Banerjee, ORIE, Cornell

entropy and information

rv X taking values $\mathcal{X} = \{a_1, a_2, \ldots, a_k\}$, with pmf $\mathbb{P}[X = a_i] = p_i$

Shannon's entropy function

- outcome $X = a_i$ has *information content*: $h(a_i) = \log_2\left(\frac{1}{p_i}\right)$
- random variable X has *entropy*: $H(X) = \mathbb{E}[h(X)] = \sum_{i=1}^{k} p_i \log_2\left(\frac{1}{p_i}\right)$

- only depends on distribution of X (i.e., $H(X) = H(p_1, p_2, \ldots, p_k)$)
- $H(X) \geq 0$ for all X
- if $X \perp\!\!\!\perp Y$, then $H(X, Y) = H(X) + H(Y)$
  where joint entropy $H(X, Y) \triangleq \sum_{(x,y)} p(x, y) \log_2 1/p(x, y)$
- if $X \sim$ uniform on $\mathcal{X}$, then $H(X) = \log_2 |\mathcal{X}|$; else, $H(X) \leq \log_2 |\mathcal{X}|$

source coding

the source coding problem

suppose we are given a database $D = (X_1 X_2 \ldots X_n)$, where each X_i is a letter in an alphabet $\mathcal{X}$, generated iid according to $X_i \sim \{p_1, p_2, \ldots, p_k\}$

$H(X) = \frac{3}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + \frac{1}{32}\cdot 5 + \frac{1}{32}\cdot 6$

Eg $\mathcal{X} = \{a, b, c, d, e, f, g, h\}$ $= \frac{79}{32} \approx 2.5$

$P = \{\underbrace{1/4, 1/4, 1/4}_{2\text{bits}}, \underbrace{1/8}_{3}, \underbrace{1/16}_{4}, \underbrace{1/32}_{5}, \underbrace{1/64, 1/64}_{6}\}$

D = (a a c d a b g f a b b c a b .)

$n = 1$ (ie, $D = X_1$)

- 'Naive' encoding = Use 3 bits ($a = 000$, $b = 001$, ..., $h = 111$)
- Want to decode with $\geq 75\%$ prob. use 2 bits

a - 00

b - 01

c - 10

defgh - 11

the source coding problem

suppose we are given a database $D = (X_1 X_2 \ldots X_N)$, where each X_i is a letter in an alphabet $\mathcal{X}$, generated iid according to $X_i \sim \{p_1, p_2, \ldots, p_k\}$

lossless compression

compress every database D into a *codeword* $L = \phi(D)$ such that we can exactly recover $\hat{D} = \phi^{-1}(L)$ $= D$

δ-lossy compression $L = \phi(D)$ defined only for $D \in \mathcal{S}_\delta$ s.t. $\mathbb{P}[\mathcal{S}_\delta] \geq 1 - \delta$

$\mathbb{P}[\phi^{-1}(L) = D] \geq 1-\delta$

the source coding problem

suppose we are given a database $D = (X_1 X_2 \ldots X_N)$, where each X_i is a letter in an alphabet $\mathcal{X}$, generated iid according to $X_i \sim \{p_1, p_2, \ldots, p_k\}$

lossless compression

compress every database D into a *codeword* $L = \phi(D)$ such that we can exactly recover $D = \phi^{-1}(L)$

Shannon's source coding theorem

if X has entropy $H(X)$, then can compress $D = (X_1 X_2 \ldots X_n)$ into a codeword $L = \phi(D)$ of expected size $|L| = n\ell$ bits, such that

$$H(X) \leq \ell < H(X) + \frac{1}{n}$$

$\left(\Rightarrow nH(X) \leq L \leq nH(X)+1 \right)$

moreover, no lossless encoder ϕ has expected codeword size $< nH(X)$

Mackay's bent coin lottery

A coin with $p_1 = 0.1$ will be tossed $N = 1000$ times.

The outcome is $\mathbf{x} = x_1 x_2 \ldots x_N$.

e.g., $\mathbf{x}$ = 000001001000100...00010

You can buy any of the 2^N possible tickets for £1 each, before the coin-tossing.

If you own ticket $\mathbf{x}$, you win £1,000,000,000.

Q To have a 99% chance of winning, at lowest possible cost, which tickets would you buy?

- And how many tickets is that?

Express your answer in the form $2^{(\cdots)}$.

Lottery tickets available

2^N {
0000000000.....00000
0000000000.....00001
0000000000.....00010
0000000000.....00011
0000000000.....00100
0000000000.....00101
0000000000.....00110
0000000000.....00111
⋮
0010000001.....01000
⋮
1111111111.....11101
1111111111.....11110
1111111111.....11111

Mackay's bent coin lottery: warmup

what if you could buy only one ticket?

Idea 1 – any ticket with 10% '1's and 90% '0s'

Idea 2 – $\underbrace{00\ldots0}_{9}\,1\,\underbrace{00\ldots0}_{9}\,1\,\ldots$ $\qquad \mathbb{P}[\text{win}] = (0.1)^{100}(0.9)^{900}$

$\Uparrow$

Idea 3 – $00\ldots0$ $\qquad \mathbb{P}[\text{win}] = (0.9)^{1000}$

$$\mathbb{P}[\text{ticket } i \text{ wins}] = (0.1)^{\#\text{ of ones}}(0.9)^{\#\text{ of zeros}}$$

Mackay's bent coin lottery: warmup

what if you could buy k tickets?

buy tix with $\leq r$ ones such that $\boxed{\sum_{i=0}^{r} \binom{1000}{i} \leq k}$

# of tix with i ones

$$\underbrace{\mathbb{P}[\text{win with } k \text{ tix}]}_{\text{want this } \geq 0.99} = \sum_{i=0}^{r} (0.1)^i (0.9)^{1000-i} \binom{1000}{i}$$

$\Rightarrow$ Want min r s.t. $\sum_{i=0}^{r} (0.1)^i (0.9)^{1000-i} \binom{1000}{i} \geq 0.99$

recall: two useful facts

- counting via binary entropy for $N \in \mathbb{N}$, $k \leq N$: $\binom{N}{k} \approx 2^{NH_2(k/N)}$
- Chebyshev's inequality for any rv. X with mean $\mathbb{E}[X]$, finite variance $\sigma^2 > 0$, and any $k > 0$: $\mathbb{P}[|X - \mathbb{E}[X]| \geq k\sigma] \leq \frac{1}{k^2}$

• Choose $\underline{r = N\left(p + k \frac{\sqrt{p(1-p)}}{\sqrt{N}}\right)}$ buy all tix with $\leq r$ '1's

$\Rightarrow \mathbb{P}[\text{winning}] \geq \underbrace{\mathbb{P}[\text{true \# of '1's} \leq r]}$ $\quad \mathbb{P}[\text{not winning}]$

$\geq 1 - \frac{1}{k^2} = 1 - \delta$

$\frac{\sum X_i}{N} \approx N\left(p, \frac{\sigma}{\sqrt{N}}\right)$

$\Rightarrow$ want $k \approx \frac{1}{\sqrt{\delta}}$

Eg. for 75% prob, need $k = 2$
99% prob, need $k = 10$

Mackay's bent coin lottery: solution

If $p = 0.1$, $N = 1000$ then this is ≈ 0.1

Suggested soln — buy all tix with $\leq N\left(p + 10\sqrt{\frac{p(1-p)}{N}}\right)$ ones

$\Rightarrow$ Guaranteed we win with prob ≥ 0.99

How many tix did we buy? Let $r = N\left(p + 10\sqrt{\frac{p(1-p)}{N}}\right)$

$$= \sum_{i=0}^{r} 2^{N H_2\left(\frac{i}{N}\right)} \approx 2^{N H_2\left(\frac{r}{N}\right)}$$

$$= 2^{N H_2\left(p + \frac{10\sqrt{p(1-p)}}{\sqrt{N}}\right)}$$

$$\approx 2^{N H_2(p)}$$

(last term in summation $\gg$ sum of all other terms)

(lossy) source coding theorem for binary sources

given $X^N = (X_1 X_2 \dots X_N)$, where each $X_i \sim$Bernoulli(p)

δ-lossy compression

$L = \phi(X^N)$ defined only for $X^N \in \mathcal{S}_\delta$ s.t. $\mathbb{P}[\mathcal{S}_\delta] \geq 1 - \delta$

Good tix (typical seq)

00 . . . 00

00 . . . 01 '1'

$(\leq \eta \text{ '1's})$

01 . . . 0 η '1s'

Bad tix

$(> \eta \text{ 1s})$ $\leftarrow 2^{N(1 - H_2(p))}$

but $\mathbb{P}[\text{bad tix}] \leq \delta$

$\leftarrow \approx 2^{N H_2\left(p + \frac{\sigma}{\sqrt{\delta N}}\right)} \Rightarrow N H_2\left(p + \frac{\sigma}{\sqrt{\delta N}}\right)$ bits

$\mathbb{P}[\text{good tix}] \geq 1 - \delta$

(lossy) source coding theorem for binary sources

given $X^N = (X_1 X_2 \dots X_N)$, where each $X_i \sim$Bernoulli(p)

δ-lossy compression

$L = \phi(X^N)$ defined only for $X^N \in \mathcal{S}_\delta$ s.t. $\mathbb{P}[\mathcal{S}_\delta] \geq 1 - \delta$

- δ-sufficient subset S_δ: smallest subset of $\{0, 1\}^N$ s.t. $\mathbb{P}[\mathcal{S}_\delta] \geq 1 - \delta$
- essential information content in X^N: $H_\delta(X^N) \triangleq \log_2 |S_\delta|$

(lossy) source coding theorem for binary sources

given $X^N = (X_1 X_2 \dots X_N)$, where each $X_i \sim$Bernoulli(p)

δ-lossy compression

$L = \phi(X^N)$ defined only for $X^N \in \mathcal{S}_\delta$ s.t. $\mathbb{P}[\mathcal{S}_\delta] \geq 1 - \delta$

- δ-sufficient subset S_δ: smallest subset of $\{0,1\}^N$ s.t. $\mathbb{P}[\mathcal{S}_\delta] \geq 1 - \delta$
- essential information content in X^N: $H_\delta(X^N) \triangleq \log_2 |S_\delta|$

Shannon's source coding theorem (lossy version)

if X has entropy $H(X)$, then for any $\epsilon > 0$ and $0 < \delta < 1$, there exists N_0 s.t. for all $N > N_0$, we have

$$\left| \frac{H_\delta(X^N)}{N} - H(X) \right| \leq \epsilon$$

for lossless - $L(D) = \begin{cases} \overset{\text{good}}{0} + NH_2(p) \text{ bits} \xleftarrow{1-\delta} 1 + NH_2(p) \text{ bits} \\ \overset{\text{bad}}{1} + D \xleftarrow{\delta} 1 + N \text{ bits} \end{cases}$