تئوری اطلاعات، جلسه 10 : دوشنبه 5 آبان 99

* مرور جلسه قبل

* کران بالا برای طول متوسط بهترین کد preefix-free ($\bar{L}_{opt}$)

* چگونه به فشرده‌سازی به اندازه آنتروپی منبع برسیم

* نامساوی kraft برای uniquely decodable Codes

* معرفی کدینگ Huffman

مرور جلسه قبل :

$$\bar{L}(C) \triangleq \sum_{x \in \mathcal{X}} p(x)\, l(x)$$

یک کد منبع لحظه‌ای

$$\bar{L}_{opt} = \min_{C \in \text{preefix-free}} \bar{L}(C)$$

توسط نامساوی kraft توصیف می‌شود.

$$\Longrightarrow \quad H_D(X) \leqslant \bar{L}_{opt} \leqslant \bar{L}(C)$$

---

* کران بالا برای $\bar{L}_{opt}$:

یک کد C را در نظر بگیریم $\longrightarrow \bar{L}(C)$

$\hookrightarrow \bar{L}_{opt} \leqslant \bar{L}(C)$

به همین منظور کد Shannon-Fano را در نظر می‌گیریم :

اگر یک متغیر تصادفی X با توزیع $p(x)$ داریم. طول کدواژه‌ها را به صورت زیر انتخاب می‌کنیم :

$$l(x) = \left\lceil \log_D \frac{1}{p(x)} \right\rceil \quad : \quad \forall x \in \mathcal{X}$$

این مجموعه طول $l(x)$ ها در نامساوی kraft صدق می‌کنند :

$$\sum_{x \in \mathcal{X}} D^{-l(x)} = \sum_{x \in \mathcal{X}} D^{-\left\lceil \log_D \frac{1}{p(x)} \right\rceil} \leqslant \sum_{x \in \mathcal{X}} D^{-\log_D \frac{1}{p(x)}} = \sum p(x) = 1$$

تحلیل متوسط طول کد Shannon-Fano

$$\log_D \frac{1}{p(x)} \leq \left\lceil \log_D \frac{1}{p(x)} \right\rceil < \log_D \frac{1}{p(x)} + 1$$

$$\xrightarrow{\times p(x)} \xrightarrow{\sum} \quad H_D(X) \leq \underline{\bar{L}_{\text{Shannon-Fano}} < H_D(X)+1}$$

قبلا هم نشان داده بودیم.

$$\Rightarrow \quad H_D(X) \leq \bar{L}_{opt} \leq \bar{L}_{\text{Shannon-Fano}} < H_D(X)+1$$

$$\Rightarrow \quad \boxed{H_D(X) \leq \bar{L}_{opt} < H_D(X)+1}$$

---

* دو نگاه متفاوت به کدکردن یک منبع داده (iid):

نگاه ۱.

one-shot coding

$$x \in \mathcal{X}: \quad \boxed{P(x)} \longrightarrow x_1, x_2, x_3, \ldots \qquad x_i \sim P(x)$$

$$C(x_1)\; C(x_2)\; C(x_3) \ldots$$

نگاه ۲.

multi-shot coding

$$x \in \mathcal{X}: \quad \boxed{P(x)} \longrightarrow \underbrace{x_1, \ldots, x_n}_{\underline{x}_1 \in \mathcal{X}^n}, \underbrace{x_{n+1}, \ldots, x_{2n}}_{\underline{x}_2 \in \mathcal{X}^n}, \underbrace{x_{2n+1} \ldots x_{3n}}_{\underline{x}_3 \in \mathcal{X}^n}, \ldots$$

block length ← n

$$C^{(n)}(\underline{x}_1)\; C^{(n)}(\underline{x}_2)\; C^{(n)}(\underline{x}_3) \ldots$$

با نگاه ۲:

$$H_D(\overbrace{X_1, \ldots, X_n}^{\underline{X}}) \leq \bar{L}^{(n)}_{opt} < H_D(X_1, \ldots, X_n) + 1$$

$$\Rightarrow \quad nH_D(X) \leq \bar{L}^{(n)}_{opt} < nH_D(X) + 1$$

$$\Rightarrow \quad H_D(X) \leq \underline{\frac{1}{n}\bar{L}^{(n)}_{opt}} < H_D(X) + \frac{1}{n} \xrightarrow{n \to \infty} H_D(X)$$

طول متوسط مصرفی به ازای هر سمبل ورودی

⇐ اثبات قضیه فشرده‌سازی شنون.

اشاره داخل پرانتز : $X \sim \text{Bernoulli}(0.9)$

$$P(X=0) = 0.1$$
$$P(X=1) = 0.9$$

$$(X_1, \ldots, X_n) \in \mathcal{X}^n$$

تعداد همه رشته های ممکن: 2^n

ولی در واقع اکثر رشته هایی که از این منبع تولید می شوند آنهایی هستند که تقریبا $\frac{n}{10}$ صفر دارند و $\frac{9n}{10}$ یک دارند. ← رشته های نوعی (typical seq.)

چند مثال : فرض کنید $\mathcal{X}=\{A,B,C\}$, $P(A)=P(B)=P(C)=\frac{1}{3}$

و $D=\{0,1\}$

$\Leftarrow H(X) \simeq 1.58$

- کد شنون-فانو : $l(x)=\lceil \log_2 \frac{1}{P(x)} \rceil \leftarrow C_{SF}=\{00,01,10\}$ → $\bar{L}_{SF}=2$

- کد Huffman : (بهینه) $\leftarrow C_{Huff}=\{0,10,11\} \leftarrow \bar{L}_{Huff}=1.67$

- کد Huffman حالت‌ها با $n=2$: $\mathcal{X}^2=\{AA, AB, AC, BA, BB, BC, CA, CB, CC\} \rightarrow C^{(2)}_{Huff}=\{000, 001, 010, 011, 100, 101, 110, 1100, 1101\}$

$\Rightarrow \bar{L}^{(2)}_{Huff}=1.61$

- multi-shot Huffman:
  - $n=3 \rightarrow 1.60$
  - $n=4 \rightarrow 1.60$
  - $n=5 \rightarrow 1.59$

- multi-shot Shannon-Fano:
  - $n=2 \rightarrow 2$
  - $n=3 \rightarrow 1.67$
  - $n=4 \rightarrow 1.75$
  - $n=5 \rightarrow 1.60$

* تا الان تمرکز فقط بر روی کدهای لحظه‌ای بود.

* اثبات نامساوی kraft برای کدهای یکتا دیکود شونده.

قضیه: مجموعه طول‌های یک کد یکتا دیکود شونده D-ary در نامساوی زیر صدق می‌کند

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1$$

و به طور عکس هر مجموعه طولی که در این نامساوی صدق کند متناظر با یک کد یکتا دیکود شونده است.

اثبات: $\text{Input } x_1 \cdots x_n \longrightarrow C(x_1) \cdots C(x_n)$

$$x_1 \cdots x_n \neq y_1 \cdots y_m \longrightarrow C(x_1) \cdots C(x_n) \neq C(y_1) \cdots C(y_m)$$

$$A \triangleq \sum_{x \in \mathcal{X}} D^{-l(x)}$$

$$A^r = \left(\sum_{x_1 \in \mathcal{X}} D^{-l(x_1)}\right)\left(\sum_{x_2 \in \mathcal{X}} D^{-l(x_2)}\right) \cdots \left(\sum_{x_r \in \mathcal{X}} D^{-l(x_r)}\right)$$

$$= \sum_{x_1, \ldots, x_r \in \mathcal{X}^r} D^{-l(x_1)} \cdots D^{-l(x_r)}$$

$$= \sum_{\underline{x} \in \mathcal{X}^r} D^{-\boxed{l(\underline{x})}} \longrightarrow = l(x_1) + \cdots + l(x_r)$$

$$= \sum_{i=1}^{r l_{max}} a(i) \, D^{-i}$$

$$\boxed{l_{max} = \max_{x \in \mathcal{X}} l(x)}$$

$a(i)$ ← تعداد رشته‌های منبع $\underline{x} \in \mathcal{X}^r$ که به یک رشته با طول i در خروجی نگاشته شده است.

می دانیم رشته های ورودی به طول i حداکثر D^i تا هستند. $\leftarrow$ $a(i) \leqslant D^i$

$$\Rightarrow \quad A^r = \sum_{i=1}^{rl_{max}} a(i) D^{-i} \leqslant \sum_{i=1}^{rl_{max}} D^i \cdot D^{-i} = rl_{max}$$

$$\Rightarrow \quad A \leqslant \underbrace{(rl_{max})^{1/r}}_{\xrightarrow{r \to \infty} 1} \quad : \forall r \geqslant 1$$

$$\Rightarrow \quad A \leqslant 1$$

نامساوی کرافت برای کدهای یکتا دیکدشدنی ✓

داخل پرانتز:

$$\alpha = (rl_{max})^{1/r}$$

$$\rightarrow \ln\alpha = \frac{1}{r}\left[\ln r + \ln l_{max}\right] \xrightarrow{r \to \infty} 0$$

$$\Rightarrow \alpha \xrightarrow{r \to \infty} 1$$