

Ülesanded tõenäosusest (täielik salastus, entroopia mõiste)

9.märts, 2006

Ülesanded

Ülesanne 1. Kujutlege järgmist telemängu, milles on võimalik võita auhind, juhul kui osaleja õigesti ära arvab, millises kolmest (esialgu suletud) kastist auhind asub. Auhind on igas kastis tõenäosusega $\frac{1}{3}$. Mängu käik on järgmine:

- Mängija valib ühe kolmest suletud kastist, mis jääb suletuks.
- Mängujuht avab ühe mittevalitud kastidest, kus auhinda ei ole (vähemalt üks mittevalitud kastidest peab olema tühi, sest auhind on vaid ühes kolmest kastist).
- Mängujuht annab mängijale võimaluse oma valikut muuta, st otsustada teise suletud kasti kasuks.

Koduseid ettevalmistusi tehes vaatlete kolme strateegiat antud mängu mängimiseks: a) alati jääda esialgse valiku juurde, b) alati muuta oma valikut, c) visata münti. Arvutage auhinna saamise tõenäosus kõigi kolme strateegia (a,b,c) korral.

Ülesanne 2. Olgu meil järgmine krüptosüsteem, milles avatekst X omandab väärtusi hulgast $\{x_1, x_2, x_3\}$, võti K väärtusi hulgast $\{k_1, k_2, k_3\}$ ja krüptogramm Y väärtusi hulgast $\{y_1, y_2, y_3, y_4\}$. Krüpteerimisreeglid on esitatud järgmise tabelina:

E	k_1	k_2	k_3
x_1	y_1	y_2	y_3
x_2	y_2	y_4	y_1
x_3	y_3	y_1	y_4

Eeldame, et võti K ja avatekst X on sõltumatud, kusjuures tõenäosused on järgmised: $p(x_1) = 0.4$, $p(x_2) = p(x_3) = 0.3$, $p(k_1) = p(k_2) = 0.3$ ja $p(k_3) = 0.4$. Leia $p(x | y_4)$ iga $x \in \{x_1, x_2, x_3\}$ korral.

Ülesanne 2A. Tõesta, et kui võti k on ühtlase jaotusega juhuslik suurus, siis nihkesiffer valemiga $y = E_k(x) = x + k \pmod{26}$ on täielikult salastav, st iga $x, y \in \{0, \dots, 25\}$ korral $p(x | y) = p(x)$.

Ülesanne 3. Leia eelmises ülesandes toodud suuruste kontekstis entroopia $H[X]$ ja tingimuslik entroopia $H[X | Y]$.

Ülesanne 4. Leia Huffmani puu ja vastavad koodid järgmisele juhuslikule suurusele X väärtuste hulgaga $\{x_1, \dots, x_7\}$ ja tõenäosustega

$$p_1 = 0.12, p_2 = 0.08, p_3 = 0.11, p_4 = 0.17, p_5 = 0.22, p_6 = 0.1, p_7 = 0.2.$$

Leia keskmine koodi pikkus ℓ ja Shannoni entroopia $H[X]$.

Ülesanne 5. Juhuslik suurus X valitakse ühtlase jaotusega hulgast $\{1, 2, \dots, 16\}$. Juhuslik suurus Y arvutatakse juhuslikust suurusest X valemi $Y = X^2 \pmod{17}$ järgi. Leia juhusliku suuruse Y Shannoni entroopia $H[Y]$.

Ülesanne 6* (iseseisvaks uurimiseks). Millistel tingimustel langeb Huffmani koodi keskmine pikkus täpselt kokku Shannoni entroopiaga?

Lahendused

Ülesanne 1. Olgu E sündmus, et esimesena valitud kastis on auhind ja W olgu sündmus, et mängija võidab auhinna. Selge, et sündmuse E tõenäosus on $P[E] = \frac{1}{3}$.

Strateegia a) korral valikut ei muudeta, mistõttu $P[W] = P[E] = \frac{1}{3}$.

Strateegia b) korral muudetakse valikut alati. Seega, kui esimesena valitud kastis oli auhind, otsustatakse lõpuks tühja kasti kasuks. Kui aga esimesena valitud kastis auhinda ei olnud, on lõpuks valitud kast alati auhinnaga. Seega, $P[W] = 1 - P[E] = \frac{2}{3}$.

Strateegia c) korral lisandub arutlustesse teine (sõltumatu) juhuslik sündmus – mündivise. Tähistame M sündmust, et mündivise soovitas valikut muuta. Eeldatavasti $P[M] = \frac{1}{2}$. Võita saab kahel (teineteist välistaval) juhul:

- Esimesel korral valiti auhinnaga kast ja mündivise soovitas valikut mitte muuta.
- Esimesel korral valiti tühi kast ja mündivise soovitas valikut muuta.

Seega, $P[W] = P[E] \cdot (1 - P[M]) + (1 - P[E]) \cdot P[M] = \frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{2}$.

Ülesanne 2. Kasutame valemit

$$\begin{aligned} p(x | y_4) &= \frac{p(x)}{p(y_4)} p(y_4 | x) \\ &= \frac{p(x) \cdot \sum_{k \in \{k_1, k_2, k_3\}} p(k) \cdot P[E_k(x) = y_4]}{\sum_{x \in \{x_1, x_2, x_3\}} \sum_{k \in \{k_1, k_2, k_3\}} p(x) \cdot p(k) \cdot P[E_k(x) = y_4]}. \end{aligned}$$

Arvutades saame, et $p(y_4) = p(x_2) \cdot p(k_2) + p(x_3) \cdot p(k_3) = 0.3^2 + 0.3 \cdot 0.4 = 0.21$ ja seega

$$p(x_1 | y_4) = 0, \quad p(x_2 | y_4) = \frac{0.3^2}{0.21} \approx 0.4286, \quad p(x_3 | y_4) = \frac{0.3 \cdot 0.4}{0.21} \approx 0.5714.$$

Ülesanne 2A. Kasutades seost $p(y) \cdot p(x | y) = p(y | x) \cdot p(x)$ ja valemeid $p(y) = \sum_x p(x) \cdot p(y | x)$ ja $p(y | x) = \sum_k p(k) \cdot [E_k(x) = y]$ saame ¹ esmalt, et $p(y | x) = \frac{1}{26}$, sest iga paari (x, y) korral leidub parajasti üks võti k nii

¹Siin tähistab $[\]$ nm. *Iversoni sümbolit*, st $[E_k(x) = y] = \begin{cases} 1, & \text{kui } E_k(x) = y; \\ 0, & \text{kui } E_k(x) \neq y. \end{cases}$

et $E_k(x) = y$, sest sobivat (ja unikaalset) võtit saab alati arvutada valemi $k = y - x \pmod{26}$ põhjal. Seega ka $p(y) = \sum_x p(x) \cdot p(y | x) = \frac{1}{26} \cdot \sum_x p(x) = \frac{1}{26}$, sest $\sum_x p(x) = 1$. Nüüd saamegi esimesena toodud võrrandist seose $p(x | y) = p(x)$.

Ülesanne 3. Kõigepealt arvutame entroopia

$$H[X] = - \sum_x p(x) \cdot \log_2 p(x) = -0.4 \cdot \log_2 0.4 - 0.3 \cdot \log_2 0.3 - 0.3 \cdot \log_2 0.3 \approx 1.57$$

Tingimusliku entroopia arvutamiseks kasutame definitsiooni:

$$H[X | Y] = \sum_y p(y) \cdot H[X | y] = - \sum_y p(y) \cdot \sum_x p(x | y) \cdot \log_2 p(x | y).$$

Selleks tuleb arvutada tõenäosused $p(y)$ ja $p(x | y)$. Saame,

$$\begin{aligned} p(y_1) &= \sum_{x,k} p(x, k) \cdot [E_k(x) = y_1] = \sum_{x,k} p(x) \cdot p(k) \cdot [E_k(x) = y_1] \\ &= p(x_1) \cdot p(k_1) + p(x_2) \cdot p(k_3) + p(x_3) \cdot p(k_2) \\ &= 0.3 \cdot 0.4 + 0.3 \cdot 0.4 + 0.3 \cdot 0.3 = 0.33 \\ p(y_2) &= 0.4 \cdot 0.3 + 0.3 \cdot 0.3 = 0.21 \\ p(y_3) &= 0.4 \cdot 0.4 + 0.3 \cdot 0.3 = 0.25 \\ p(y_4) &= 0.21. \end{aligned}$$

Tingimuslike tõenäosuste arvutamisel saame:

$p(x y)$	y_1	y_2	y_3	y_4
x_1	$\frac{0.3 \cdot 0.4}{0.33} \approx 0.3636$	$\frac{0.3 \cdot 0.4}{0.21} \approx 0.5714$	$\frac{0.4 \cdot 0.4}{0.25} = 0.64$	0
x_2	$\frac{0.3 \cdot 0.4}{0.33} \approx 0.3636$	$\frac{0.3 \cdot 0.3}{0.21} \approx 0.4286$	0	$\frac{0.3 \cdot 0.3}{0.21} \approx 0.4286$
x_3	$\frac{0.3 \cdot 0.3}{0.33} \approx 0.2727$	0	$\frac{0.3 \cdot 0.3}{0.25} \approx 0.36$	$\frac{0.3 \cdot 0.4}{0.21} \approx 0.5714$

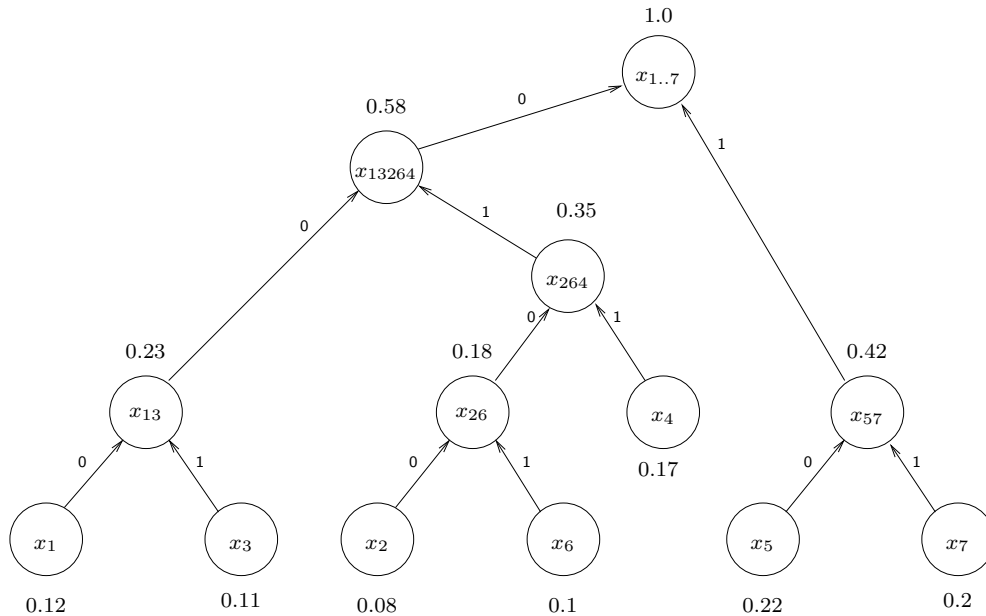
Arvutame tingimuslikud entroopiad:

$$\begin{aligned} H[X | y_1] &\approx -2 \cdot 0.3636 \cdot \log_2 0.3636 - 0.2727 \cdot \log_2 0.2727 \\ &\approx 1.0614 + 0.5112 \approx 1.573 \\ H[X | y_2] &\approx 0.5137 \\ H[X | y_3] &\approx 0.9427 \\ H[X | y_4] &\approx 0.5137. \end{aligned}$$

Lõpuks arvutame tingimusliku entroopia kui keskväärtuse:

$$H[X | Y] \approx 0.33 \cdot 1.573 + 0.21 \cdot 0.5137 + 0.25 \cdot 0.9427 + 0.21 \cdot 0.5137 \approx 0.97.$$

Ülesanne 4. Huffmani puu tuleb järgmine:



Elementide x_1, \dots, x_7 koodid on seega (vastavalt) 000, 0100, 001, 011,10, 0101, ja 11. Koodi keskmine pikkus tuleb

$$\ell = 0.12 \cdot 3 + 0.08 \cdot 4 + 0.11 \cdot 3 + 0.17 \cdot 3 + 0.22 \cdot 2 + 0.1 \cdot 4 + 0.2 \cdot 2 = 2.76$$

ja Shannoni entroopia:

$$\begin{aligned} H[X] &= -0.12 \cdot \log_2 0.12 - 0.08 \cdot \log_2 0.08 - 0.11 \cdot \log_2 0.11 - \\ &\quad - 0.17 \cdot \log_2 0.17 - 0.22 \cdot \log_2 0.22 - 0.1 \cdot \log_2 0.1 - 0.2 \cdot \log_2 0.2 \\ &\approx 0.3671 + 0.2915 + 0.3503 + 0.4346 + 0.4806 + 0.3322 + 0.4644 \\ &\approx 2.72. \end{aligned}$$

Ülesanne 5. Et funktsiooni $f(x) = x^2 \pmod{17}$ määramispiirkond $\{1, \dots, 16\}$ on suhteliselt väike, siis arvutame funktsiooni $f(x)$ tabeli:

x	$f(x)$
1	1
2	4
3	9
4	16
5	8
6	2
7	15
8	13
9	13
10	15
11	2
12	8
13	16
14	9
15	4
16	1

Tabelist on näha, et suurusel Y on kaheksa erinevat võimalikku väärtust: $Y = \{1, 2, 4, 8, 9, 13, 15, 16\}$, kusjuures igal väärtusel on täpselt kaks originaali hulgas $X = \{1, \dots, 16\}$. Et X on ühtlase jaotusega (s.t. kõik tõenäosused võrdsed $\frac{1}{16}$), siis iga $y \in Y$ tõenäosus on $\frac{2}{16} = \frac{1}{8}$. Seega, entroopia tuleb:

$$H[Y] = 8 \cdot \frac{1}{8} \log_2 \frac{1}{\frac{1}{8}} = \log_2 8 = \mathbf{3} .$$

Ülesanne 6.* Olgu X juhuslik suurus tõenäosusjaotusega (p_1, \dots, p_n) . Olgu (w_1, \dots, w_n) mingi prefiksivaba kood, kusjuures

$$-\sum_i p_i \cdot \log_2 p_i = H[X] = \ell = \sum_i p_i \cdot \|w_i\|,$$

kus $\|w_i\|$ tähistab koodsõna $w_i \in \{0, 1\}^*$ pikkust (bittide arvu). Teisendades keskmise koodi pikkuse avaldist järgmisel viisil

$$\ell = \sum_i p_i \cdot \log_2 2^{\|w_i\|} = \sum_i p_i \cdot \log_2 \frac{1}{2^{-\|w_i\|}},$$

saame järgmise võrduse (mis on ekvivalentne väitega, et koodi keskmine pikkus langeb kokku Shannoni entroopiaga):

$$0 = \ell - H[X] = \sum_i p_i \cdot [\log_2 p_i + \log_2 \frac{1}{2^{-\|w_i\|}}] = \sum_i p_i \cdot \log_2 \frac{p_i}{2^{-\|w_i\|}}.$$

Et Krafti võrratuse tõttu $\sum_i 2^{-\|w_i\|} \leq 1$, siis Kullback-Liebleri võrratusest tulenevalt kehtib viimane võrdus parajasti siis, kui

$$p_i = 2^{-\|w_i\|} \tag{1}$$

iga $i \in \{1, \dots, n\}$ korral. Siit tuleneb, et tõenäosusjaotuse (p_1, \dots, p_n) mingi prefiksivaba koodi (w_1, \dots, w_n) keskmine pikkus saavutab teoreetilise miinimumi (Shannoni entroopia) parajasti siis, kui $p_i = 2^{-\|w_i\|}$.

Lõpuks jääb üle näidata, et kui tõenäosusjaotuse (p_1, \dots, p_n) kõik tõenäosused p_i on kahe negatiivsed astmed (st $p_i = 2^{-k_i}$), siis Huffmani algoritm konstrueerib tingimusi (1) rahuldava prefiksivaba koodi. Olgu $p_i = 2^{-k_i}$, kus $k_1 \leq \dots \leq k_n$. Väite tõestuseks kasutame induktsiooni n järgi. Väide ilmselt kehtib $n = 1$ korral, sest siis on tõenäosusjaotus triviaalne ($p_1 = 1$). Oletame, et väide kehtib $n-1$ korral ja vaatleme tõenäosusruumi (p_1, \dots, p_n) , kus $p_i = 2^{-k_i}$. On selge, et $k_{n-1} = k_n = \max\{k_1, \dots, k_n\}$, sest vastasel korral ei saa olla $\sum_i p_i = 1$. Seega on ka $p_{n-1} + p_n = 2^{-k_n+1}$ kahe negatiivne aste, mistõttu rakendub induktsiooni eeldus.