Tomáš Kroupa

1. Kolik otázek je třeba v průměru položit, abychom se dozvěděli datum narození člověka (den v roce), pokud odpovědi jsou pouze ano/ne a tázaný odpovídá pravdivě? Uvažujte, že každý 4. rok je přestupný.

Řešení:

Průměrný počet otázek je při každé strategii jejich kladení zdola omezen entropií. V tomto případě hledáme entropii náhodné veličiny X, která nabývá 366 hodnot, z toho 365 hodnot je stejně pravděpodobných a zbývající hodnota je 4×-méně pravděpodobná (přestupný den jednou za čtyři roky):

$$p_X(x) = \begin{cases} \frac{4}{4\cdot 365+1}, & x \in \{1,\dots,365\}, \\ \frac{1}{4\cdot 365+1}, & x = 366. \end{cases}$$

Z toho dostaneme entropii $H(X) \doteq 8.51$ bitů. Horší odhad nutného počtu otázek je $9 = \lceil \log 366 \rceil$, neboť stačí určit prvek množiny o 366 možných prvcích.

2. Je Vám nabídnuta účast ve hře v kostky. Máte na výběr mezi dvěma kostkami, které nejsou symetrické a mají následující pravděpodobnosti padnutí jednotlivých stran:

$$p = \left(2^{-3}, 2^{-3}, 2^{-3}, 2^{-3}, 2^{-3}, 3\cdot 2^{-3}\right),$$
$$q = \left(2^{-2}, 2^{-2}, 2^{-3}, 2^{-3}, 2^{-3}, 2^{-3}\right).$$

Preferujete-li symetrii kostky, kterou z nich si vyberete? Bude pro Vás výhodnější, když se před začátkem hry vylosuje kostka, se kterou se bude hrát?

Řešení:

Symetrii kostek budeme přirozeně hodnotit pomocí entropie. Tak dostaneme pro jednotlivé kostky

$$H(p) = 3 - \frac{3}{8}\log 3 \doteq 2.406,$$
$$H(q) = 2.5.$$

Volíme tedy druhou kostku.

Losujeme-li kostku před začátkem hry, dostaneme pravděpodobnosti popsané vektorem r, který odpovídá směsi (s koeficientem $\frac{1}{2}$) dvou původních diskrétních rozdělení popsaných pomocí p a q:

$$r = \frac{1}{2}p + \frac{1}{2}q = \left(\frac{3}{16}, \frac{3}{16}, \frac{2}{16}, \frac{2}{16}, \frac{2}{16}, \frac{4}{16}\right).$$

Potom platí
$$H(r) = \frac{25 - 3\log 3}{8} \doteq 2.531.$$
Losování kostky na začátku tedy vede k nejvyšší entropii a tudíž k nejvyšší symetrii výsledků hry. Tento způsob již zaručuje entropii blízkou symetrické kostce, jejíž entropie je maximální a rovna
$$\log 6 = 1 + \log 3 \doteq 2.585.$$

3. Odpověď X na otázku je náhodná veličina nad množinou $\Lambda = \{1, 2, 3\}$, jejíž rozdělení je popsáno vektorem pravděpodobností $({}^1\!/\!_4, {}^1\!/\!_2, {}^1\!/\!_4)$. Odpověď může být náhodně změněna, výsledek je popsán náhodnou veličinou Y. Podmíněné pravděpodobnosti $p_{Y|X}(y|x)$ jsou popsány touto tabulkou:

$p_{Y\|X}(y\|x)$	$y = 1$	$y = 2$	$y = 3$
$x = 1$	1	0	0
$x = 2$	$1/2$	0	$1/2$
$x = 3$	0	$1/2$	$1/2$

Určete:

(a) entropii veličin X a entropii veličiny Y;

(b) podmíněnou entropii veličiny X pozorované prostřednictvím veličiny Y a obráceně.

Řešení:

Zřejmě lze přímo určit $H(X)$ a $H(Y|X)$. Platí
$$H(X) = \tfrac{1}{4} \cdot 2 + \tfrac{1}{2} \cdot 1 + \tfrac{1}{4} \cdot 2 = \tfrac{3}{2} \quad \text{a} \quad H(Y|X) = \tfrac{1}{4} \cdot 0 + \tfrac{1}{2} \cdot 1 + \tfrac{1}{4} \cdot 1 = \tfrac{3}{4}.$$
Ke stanovení $H(Y)$ a $H(X|Y)$ je třeba nejprve určit sdružené rozdělení náhodného vektoru (X, Y). To získáme jako $p_{XY} = p_{Y|X} p_X$:

$p_{XY}(x, y)$	$y = 1$	$y = 2$	$y = 3$
$x = 1$	$\frac{1}{4}$	0	0
$x = 2$	$\frac{1}{4}$	0	$\frac{1}{4}$
$x = 3$	0	$\frac{1}{8}$	$\frac{1}{8}$

Z tabulky rovnou vidíme, že $H(Y) = \frac{1}{2} \cdot 1 + \frac{1}{8} \cdot 3 + \frac{3}{8} \log \frac{8}{3} \doteq 1.41$. Z řetězcového pravidla snadno určíme $H(X|Y)$ jako
$$H(X|Y) = H(X, Y) - H(Y) = H(X) + H(Y|X) - H(Y) \doteq 0.84.$$

4. Uvažujme abecedu $\Lambda = \Omega = \{0, 1\}$. Náhodná veličina X má binomické rozdělení $\text{Bi}(1, {}^3\!/\!_4)$, kde $P[X = 1] = {}^3\!/\!_4$. Šum je vyjádřen veličinou Z mající rozdělení $\text{Bi}(1, {}^1\!/\!_{10})$, kde $P[Z = 1] = {}^1\!/\!_{10}$. Položme $Y = X \oplus Z$, přičemž $\oplus$ je součet modulo 2. Určete vzájemnou informaci veličin X a Y.

Řešení:

Platí $I(X; Y) = H(Y) - H(Y|X)$. Stanovme nejprve $H(Y|X)$. Z definice Y dostaneme následující tabulku podmíněných pravděpodobností:

$p_{Y\|X}(y\|x)$	$y = 0$	$y = 1$
$x = 0$	$9/10$	$1/10$
$x = 1$	$1/10$	$9/10$

Z té plyne $H(Y|X) = \frac{1}{4}H(Y|X = 0) + \frac{3}{4}H(Y|X = 1) = H(Y|X = 0) \doteq 0.47$. Dále spočítáme p_Y jako marginální rozdělení pomocí p_{XY}:

$p_{XY}(x, y)$	$y = 0$	$y = 1$
$x = 0$	$9/40$	$1/40$
$x = 1$	$3/40$	$27/40$

Dostaneme $H(Y) \doteq 0.88$ a tudíž $I(X; Y) \doteq 0.41$. Jelikož $H(X) = 0.81$, lze usoudit, že při přenosu se ztratí část informace o X díky šumu Z.

5. Rozhodněte, zda kód je jednoznačně dekódovatelný (JD) nebo instantní.

   (a) 1,01,001,000

   (b) 1,101,10,01

   (c) 01,11,00,001

   Pokud je kód JD a nikoli instantní, lze nalézt instatní kód stejné střední délky?

Řešení:

(a) je instantní. (b) není JD: řetězec 101 lze parsovat jako (101) nebo jako (10)(1). (c) není instantní, ale je JD. Při dekódování je totiž nutné pouze rozlišit 3. a 4. kódové slovo: pokud je počet jednotkových bitů mezi řetězcem 001 . . . a nejbližším bitem 0 sudý, potom 001 parsujeme jako (00)(1 . . . , jinak jako (001) . . .

Ke kódu (c) snadno nalezneme instantní kód, jehož slova mají stejnou délku. Stačí uvážit, že délky kódových slov 2,2,2,3 splňují Kraftovu nerovnost,

$$2^{-2} + 2^{-2} + 2^{-2} + 2^{-3} \leq 1,$$

a proto lze nalézt odpovídající instantní kód. Je to např. kód 01, 11, 00, 101.

6. Pro zadaný informační zdroj X nalezněte Shannonův kód a Huffmanův kód:

$$p_X(a) = 0.1,\ p_X(b) = 0.15,\ p_X(c) = 0.05,\ p_X(d) = 0.4,\ p_X(e) = 0.3.$$

Řešení:

Například:

	a	b	c	d	e
C_S	1010	100	10111	00	01
C_H	0110	010	0111	1	00

Srovnáním délek kódových slov ihned plyne, že Huffmanův kód má menší střední kódovou délku. Přesvědčte se o tom výpočtem!

7. Nalezněte (binární) Huffmanův kód pro informační zdroj X popsaný pravděpodobnostmi

$$\left(\frac{1}{3}, \frac{1}{5}, \frac{1}{5}, \frac{2}{15}, \frac{2}{15}\right).$$

Určete, zda je nalezený Huffmanův kód optimální i pro zdroj Y s pravděpodobnostmi

$$\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right),$$

a zdůvodněte proč.

Řešení:

Huffmanův kód pro zdroj X:

$$\{00, 10, 11, 010, 011\}.$$

Ovšem jiný Huffmanův kód pro zdroj X dostaneme bitovou inverzí ve všech kódových slovech:

$$\{11, 01, 00, 101, 100\}.$$

Chceme ukázat, že Huffmanův kód s délkami kódových slov

$$2, 2, 2, 3, 3$$

je optimální i pro zdroj Y s rovnoměrnou pravděpodobnostní funkcí. Ten má entropii

$$H(Y) = \log 5 \doteq 2.32.$$

Střední délka uvažovaného Huffmanova kódu je pro tuto rovnoměrnou pravděpodobnostní funkci

$$3 \cdot 2 \cdot \frac{1}{5} + 2 \cdot 3 \cdot \frac{1}{5} = \frac{12}{5}.$$

Střední délka libovolného instantního kódu s délkami kódových slov $\ell_1, \ldots, \ell_5$ pro Y však musí splňovat nerovnost

$$\frac{1}{5} \sum_{i=1}^{5} \ell_i \geq H(Y),$$

čemuž vyhovuje každý instantní kód, jehož délky kódových slov splňují

$$\sum_{i=1}^{5} \ell_i \geq 12.$$

Jelikož je součet délek slov Huffmanova kódu pro X právě 12, je tento kód optimální i pro zdroj Y.

8. Mějme kód s těmito čtyřmi kódovými slovy:

$$0, 10, 110, 111.$$

a) Rozhodněte, zda je tento kód optimální pro nějaký informační zdroj, případně popište alespoň dva informační zdroje, pro něž je optimální. b) Odstraníme-li z posledního kódového slova první bit, bude takto upravený kód jednoznačně dekódovatelný? Odpověď zdůvodněte.

Řešení:

a) Uvedený kód je optimální, neboť je instantní a délky ℓ_i jeho kódových slov splňují

$$\ell_i = -\log 2^{-n_i}, \quad n_i \in \mathbb{N}$$

přičemž

$$\sum_{i=1}^{4} 2^{-n_i} = 1.$$

Volbou $n_1 = 1$, $n_2 = 2$, $n_3 = n_4 = 3$ tak dostaneme optimání kód pro informační zdroj popsaný pravděpodobnostmi

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right).$$

Uvedený kód je ovšem Huffmanův (a tudíž optimální) i např. pro zdroj popsaný pravděpodobnostmi

$$\left(\frac{1}{2}, \frac{3}{8}, \frac{1}{16}, \frac{1}{16}\right).$$

b) Výsledný kód nemůže být jednoznačně dekódovatelný, neboť délky jeho kódových slov nesplňují Kraftovu nerovnost:

$$2^{-1} + 2 \cdot 2^{-2} + 2^{-3} \not\leq 1.$$

9. Pro informační zdroj nad abecedou $\mathcal{X} = \{$a,b,c,d,e,f$\}$ byl nalezen kód C:

a	001
b	1001
c	0010
d	1110
e	1010
f	01110

Je jednoznačně dekódovatelný?

Řešení:

Pro délky kódových slov

$$3, 4, 4, 4, 4, 5$$

ověříme Kraftovu nerovnost:

$$2^{-3} + 4 \cdot 2^{-4} + 2^{-5} = \frac{13}{32} < 1.$$

Tedy existuje jednoznačně dekódovatelný binární kód s takto zadaným délkami kódových slov:

a	000
b	0010
c	0011
d	0100
e	0101
f	01100

Ovšem původní kód C jednoznačně dekódovatelný není, neboť např. zprávy cd a af jsou zakódovány shodně.

10. Informační zdroje nad abecedou $\mathcal{X} = \{1, \ldots, 5\}$ jsou popsány dvěma pravděpodobnostními funkcemi p a q. Uvažujme binární kódy C_1 a C_2 pro tyto zdroje:

x	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
1	1/2	1/2	0	0
2	1/4	1/8	10	100
3	1/8	1/8	110	101
4	1/16	1/8	1110	110
5	1/16	1/8	1111	111

Ověřte, že kód C_1 je optimální pro zdroj s pravděpodobnostní funkcí p a kód C_2 je optimální pro zdroj s pravděpodobnostní funkcí q. Pokud užijeme kód C_1 pro zdroj popsaný pomocí q, jaké chyby se dopustíme?

Řešení:

Určíme entropie p a q:

$$H(p) = 1.875,$$
$$H(q) = 2.$$

Avšak střední délka kódu C_1 je 1.875, střední délka kódu C_2 je 2: oba kódy jsou optimální. Střední délka $L_q(C_1)$ kódu C_1 použitého na zdroj s pravděpodobnostní funkcí q je

$$\frac{1}{2} + (2+3) \cdot \frac{1}{8} + 2 \cdot 4 \cdot \frac{1}{8} = 2.125.$$

Kód C_1 tedy není pro q optimální: za nevhodné kódování zaplatíme chybou o velikosti rozdílu

$$L_q(C_1) - H(q) = 0.125,$$

což je právě hodnota informační divergence $D(q||p)$.

11. Markovský řetězec nad abecedou $\Lambda = \{a, b, c\}$ je popsán těmito pravděpodobnostmi přechodu: pravděpodobnosti přechodů ze stavu a i ze stavu b jsou rovnoměrné, přechod ze stavu c do c nikdy nemůže nastat, ostatní podmíněné pravděpodobnosti jsou shodné. Určete rychlost entropie odpovídajícího markovského zdroje informace.

Řešení:

Ze zadání dostaneme matici přechodu:

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Rychlost entropie markovského zdroje splňuje $H((X_n)_{n\in\mathbb{N}}) = H(X_2|X_1)$. Stačí tedy určit

$$H(X_2|X_1) = \sum_{x_1\in\{a,b,c\}} p(x_1)\cdot H(X_2|x_1),$$

kde $p = (p(a), p(b), p(c))$ je stacionární rozdělení řetězce a $H(X_2|x_1)$ jsou podmíněné entropie jednotlivých řádku matice $\mathbf{P}$. Platí

$$H(X_2|a) = H(X_2|b) = \log 3, \quad H(X_2|c) = \log 2 = 1.$$

Stacionární rozdělení nalezneme řešením soustavy

$$p\cdot\mathbf{P} = p, \quad p(a)+p(b)+p(c) = 1.$$

Dostaneme $p = (0.375, 0.375, 0.25)$. Proto

$$H(X_2|X_1) = 2\cdot 0.375\log 3 + 0.25\cdot 1 = 0.75\log 3 + 0.25 \doteq 1.435.$$

12. Určete rychlost entropie markovského zdroje informace s abecedou $\mathcal{X} = \{a, b, c\}$, jehož matice přechodu je

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Jaká je maximální rychlost entropie libovolného markovského zdroje s abecedou $\mathcal{X}$? Odpověď zdůvodněte.

Řešení:

Nejprve nalezneme stacionární rozdělení $\mathbf{p} = (p_1, p_2, p_3)$ Markovova řetězce zadaného maticí $\mathbf{P}$. Řešíme soustavu lineárních rovnic

$$\mathbf{p} = \mathbf{pP},$$
$$p_1 + p_2 + p_3 = 1.$$

Jejím řešením je vektor $\mathbf{p} = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)$. Informační zdroj $X_1, X_2, \ldots$ je markovský, pokud pro jeho počáteční rozdělení $\mathbf{p}(0)$ platí $\mathbf{p}(0) = \mathbf{p}$. Potom je rychlost entropie tohoto zdroje rovna střední podmíněné entropii $H(X_2|X_1)$. Tedy

$$H(X_2|X_1) = \sum_{i=0}^{2} p_i(0)\cdot H(X_2|X_1 = i) = \frac{2}{5}\cdot H\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{2}{5}.$$

Maximální rychlost entropie markovského zdroje pro tento zdroj je maximum entropie na tříprvkové množině, neboť $H(X_2|X_1)$ je konvexní kombinací podmíněných entropií na tříprvkových množinách. V našem případě tedy $\log 3 \doteq 1.585$.

13. Nalezněte střední délku Huffmanova kódu pro informační zdroj generující řetězec

$$aaababbcabc$$

Určete, jaké úspory dosáhneme použitím blokového Huffmanova kódování s délkou bloku 2.

Řešení:

Při použití Huffmanova kódu o délce bloku 1 máme tuto tabulku:

	a	b	c
$p(x)$	$\frac{5}{11}$	$\frac{4}{11}$	$\frac{2}{11}$
C_H	0	11	10

Při použití bloku délky 2 dostaneme tento Huffmanův kód:

	aa	ab	ba	ac	ca	bb	bc	cb	cc
$p(x,y)$	$\frac{25}{121}$	$\frac{20}{121}$	$\frac{20}{121}$	$\frac{10}{121}$	$\frac{10}{121}$	$\frac{16}{121}$	$\frac{8}{121}$	$\frac{8}{121}$	$\frac{4}{121}$
C_H^2	01	110	111	000	1011	100	0011	1010	0010

Střední délky kódů vztažené na 1 znak zdrojové abecedy:

$$L(C_H) = \frac{17}{11} \doteq 1.545, \qquad \frac{L(C_H^2)}{2} \doteq \frac{3.041}{2} = 1.521.$$

Tedy dosáhneme komprese lepší o

$$\left(1 - \frac{1.521}{1.545}\right) \times 100\,\% = 1.55\,\%.$$

14. Při vysílání dvouprvkové abecedy $\{\cdot, -\}$ se zkreslí 8% teček a 2% čárek. Zpráva, ve které se tečky a čárky vyskytují rovnočetně, obsahuje 154 bitů informace. Kolik z nich uvedený informační kanál v průměru přenese správně?

Řešení:

Rozdělení informačního zdroje X s abecedou $\{\cdot, -\}$ je na vstupu popsáno rovnoměrnou pravděpodobnostní funkcí p_X. Podmíněné pravděpodobnosti $p_{Y|X}(y|x)$, $x, y \in \{\cdot, -\}$, vyjadřující možnou záměnu symbolu jsou dány maticí

$$\begin{pmatrix} 0.92 & 0.08 \\ 0.02 & 0.98 \end{pmatrix}.$$

Nejprve určíme zpětné podmíněné pravděpodobnosti $p_{X|Y}(x|y)$ pomocí Bayesova vzorce:

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{\sum\limits_{x' \in \{\cdot,-\}} p_{Y|X}(y|x')p_X(x')}.$$

Potom vzájemnou informaci:

$$I(X;Y) = H(X) - H(X|Y) = 0.725.$$

Ve zprávě o 154 bitech se tak za předpokladu nezávislosti vysílaných znaků zachová přibližně $154 \cdot 0.725 = 112$ bitů.

15. Informační zdroj X vysílá znaky z abecedy $\mathcal{X} = \{a_1, a_2, a_3, a_4\}$ s pravděpodobnostmi $p_X(a_i)$, $i = 1, \dots, 4$. Na výstupu kanálu je pozorován informační zdroj Y nad stejnou abecedou $\mathcal{X}$. Chybovost přenosu je popsána podmíněnými pravděpodobnostmi $p_{Y|X}(a_j|a_i)$, $i, j = 1, \dots, 4$. Pokud jsou podmíněné entropie $H(Y|X = a_i)$, $i = 1, \dots, 4$ maximální, jaká je vzájemná informace X a Y?

Řešení:

Maximum podmíněné entropie $H(Y|X = a_i)$ se pro každé $i = 1, \dots, 4$ nabývá pro rovnoměrnou podmíněnou pravděpodobnostní funkci:

$$p_{Y|X}(a_j|a_i) = \frac{1}{4}, \quad j = 1, \dots, 4.$$

Proto platí

$$H(Y|X) = \sum_{i=1}^{4} p_X(a_i) \cdot H(Y|X = a_i) = \log 4 \cdot \sum_{i=1}^{4} p_X(a_i) = 2.$$

Rozdělení zdroje Y je též rovnoměrné, neboť pro každé $i = 1, \dots, 4$ platí

$$p_Y(a_i) = \sum_{j=1}^{4} p_{Y|X}(a_i|a_j)p_X(a_j) = \frac{1}{4}\sum_{j=1}^{4} p_X(a_j) = \frac{1}{4}.$$

Tudíž $H(Y) = \log 4 = 2$ a proto $I(X;Y) = 0$.

Reference

[1] J. Adámek. *Stochastické procesy a teorie informace - úlohy.* Vydavatelství ČVUT, 1989.

[2] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.