# Řešené příklady z teorie informace pro kurs *Pravděpodobnost, statistika a teorie informace* FEL ČVUT

Tomáš Kroupa
http://staff.utia.cas.cz/kroupa/
kroupa@utia.cas.cz

## Příklady

**Příklad 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.

**Příklad 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?

**Příklad 3.** 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?

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

**Příklad 5.** 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.

**Příklad 6.** 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.

**Příklad 7.** 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.

**Příklad 8.** Pro informační zdroj nad abecedou $\mathcal{X} = \{\text{a,b,c,d,e,f}\}$ byl nalezen kód $C$:

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

Je jednoznačně dekódovatelný?

**Příklad 9.** 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

*Příklad 1.* 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):

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

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.

*Příklad 2.* 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

$$\log 6 = 1 + \log 3 \doteq 2.585.$$

*Příklad 3.* 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.

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

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

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 $i = 1, \ldots, 4$ 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 $I(X;Y) = 0$.

*Příklad 5.* 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 $\log 3 \doteq 1.585$.

*Příklad 6.* 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 $Y$.

*Příklad 7.* a) Uvedený kód je optimální, neboť je instantní a délky $\ell_i$ 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:

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

*Příklad 8.* 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, je to např. kód zadaný tabulkou 1.
Ovšem kód $C$ jednoznačně dekódovatelný není, neboť zprávy *cd* a *af* jsou zakódovány shodně.

*Příklad 9.* Určíme entropie $p$ a $q$:

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

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

Tabulka 1: Jednoznačně dekódovatelný kód

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 $D(q||p)$.

## 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.