Cvičení - Komprese dat

Michal Kupsa, Štěpán Holub

Zadání

1. Mějme náhodnou veličinu X s hodnotami v abecedě $A = \{1, 2, 3, 4\}$ a rozdělením $P_X(i) = \frac{i}{10}$, $i = 1, \dots, 4$.
   (a) Spočtěte její entropii.
   (b) Najděte pro tuto veličinu binární Shannonův kód, spočtěte jeho střední délku. Je možné některé kódové slovo zkrátit, abychom dostali prefixový kód? Pokud ano, o kolik klesne průměrná délka?
   (c) Najděte Huffmanův binární kód pro předchozí zadání a spočtěte jeho střední délku. Můžeme najít lepší prefixový kód (menší střední délku)?

2. Mějme náhodnou veličinu Y s rozložením $P_Y(i) = 0,5 - \frac{i}{10}$, $i = 1..4$.
   (a) Spočtěte její entropii.
   (b) Použijte Shannonův kód pro veličinu X z předchozího bodu a spočtěte jeho střední délku $\mathbb{E}(|f(Y)|)$. Z výsledku vysvětlete, proč lze najít lepší prefixový kód z hlediska střední délky kódu.
   (c) Najděte Huffmanův binární kód pro Y a spočtěte jeho střední délku.

3. Uvažujme i.i.d proces $X_1, X_2, \dots$, kde jednotlivé náhodné veličiny mají rozdělení stejné jako veličina X definovaná výše.
   (a) Spočtěte entropii $\mathcal{H}(X_1, X_2)$.
   (b) Uvažujme kódování po písmenkách. Definujme kód $g : A^2 \to \{0, 1\}^+$, předpisem $g(a_1a_2) = f(a_1)f(a_2)$, kde f je Huffmanův kód pro veličinu X. Jaká bude střední délka kódu $\mathbb{E}|g(X_1, X_2)|$?
   (c) Jak dlouhé zprávy musíme kódovat blokovým rozšířením popsaným výše, abychom si byli jisti, že náš výsledný blokový kód není optimální?
   (d) Najděte předpis pro délky Shannonova kódu $f : A^2 \to \{0, 1\}^+$ a spočítejte jeho střední délku. pro náhodnou veličinu (X_1, X_2).
   (e) Pokud zakódujeme všechna písmena pomocí Shannonova kódu pro X_1 a všechny dvojice písmen pomocí Shannonova kódu pro (X_1, X_2), bude takto vzniklý kód na $A \cup A^2$ prostý? Je možné to poznat z délek?

(f) Pokud zakódujeme všechna písmena pomocí Shannonova kódu pro X_1 a všechny dvojice písmen pomocí blokové konkatenace tohoto kódu, bude takto vzniklý kód na $A \cup A^2$ prostý?

(g) U obou předchozích kódů se zamyslete, zda je možné aby nastala situace $|u| < |v|$ a $|f(u)| > |f(v)|$.

(h) U obou předchozích kódů se zamyslete, zda je možné aby nastala situace $u \leq v$ a $|f(u)| > |f(v)|$.

4. Pro výše uvedený i.i.d proces $X_1, X_2, \ldots$, uvažujme prefixový kód $f : A^+ \to B^+$, který je pro slova délky n definovaný předpisem $\gamma_1(n) f_n(u)$, kde f_n je Shanonnův kód pro $(X_1, .., X_n)$.

(a) Najděte kódy pro $a \in A$.

(b) Porovnejte střední délku kódu $E|f(X_1)|$ s entropií.

(c) Zapište délky kódových slov pro $u \in A^2$.

(d) Je možné aby nastala situace $|u| < |v|$ a $|f(u)| > |f(v)|$?

Výsledky a řešení

Vzhledem k výstupní abecedě používáme všude základ logaritmu roven 2.

1. (a) $\mathcal{H}(X) = \log 5 - \frac{3}{10}\log 3 \doteq 1.85$ (přesněji asi 1.84644)
   (b) Jeden z možných Shannonových kódů psaný postupně pro vstup 1 až 4 je 1010, 100, 01, 00, se střední délkou 2.4. Tento kód není úplný, možno zkrátit na všech vstupech následovně 101, 10, 01, 00. Nyní je střední délka 2.1.
   (c) Nejlepší vzhledem ke střední hodnotě je Huffmanův kód (zas není jednoznačný) 000, 001, 01, 1, se střední délkou 1.9.

2. (a),(c) Stejně jako pro X z bodu (1), jen kódová slova budou v opačném pořadí.
   (b) Shannonův kód f pro X z předchozího bodu má střední hodnotu $\mathbb{E}(|f(Y)|) = 3.1$. že se dá kód zlepšit, lze vysvětlit více způsoby:
   - jako výše lze kódová slova zkrátit,
   - střední délka kódu přesahuje entropii o více než o jedna,
   - kód nectí zásadu, že jevy s větší pravděpodobností mají kratší kód; lze ho tedy zlepšit prohozením kódových slov,
   - nejpracnější postup (a zde zbytečný) pak je najít Shannonův kód (nebo dokonce Huffmanův) a ukázat, že jsou lepší.

3. (a) Z i.i.d. plyne $\mathcal{H}(X_1, X_2) = 2\mathcal{H}(X) \doteq 3.69$.
   (b) Z linearity střední hodnoty plyne $\mathbb{E}|g(X_1, X_2)| = 2\mathbb{E}|f(X)| = 3.8$. (Mimochodem, zde není důležitá nezávislost).
   (c) Rozdíl mezi entropií X a zakódováním pomocí Huffmanova kódu je více než 0.05. n-tá konkatenace má střední délku rovnou $n\mathbb{E}|f(u)|$, přičemž kóduje veličinu $(X_1, X_2, \ldots, X_n)$ o entropii $n \cdot \mathcal{H}(X)$. Je to $n = 20$, kdy bude rozdíl entropie a střední délky kódu větší než 1. V tu chvíli je jasné, že Shannonův a Huffmanův kód přímo pro n-tice bude lepší.
   (d) Uspořádejme vstupy lexikograficky: 11, 12, 13, 14, 21, 22,...,44. Délky kódových slov jsou pak postupně 7,6,6,5, 6,5,5,4, 6,5,4,4, 5,4,4,3. Střední délka je 4.31.
   (e) Pokud zkombinujeme Shannonův kód pro písmenka a Shannonův kód pro digramy, může se stát, že písmenko bude mít přiřazené stejné kódové slovo, jako digram, protože existují digramy, které jsou zhruba stejně pravděpodobné jako nějaké písmeno. Pokud bychom se tomu chtěli vyhnout, mohli bychom zkoušet různé páry Shannonových kódů a kombinovat. Většinou to nepůjde.

   Lze nicméně poměrně snadno rozhodnout, zda lze pro písmena a digramy najít Shannonovy kódy, jejichž kombinace bude dokonce prefixový kód. Stačí vzít všechny délky kódových slov pro písmenka a digramy, tedy $4+16 = 20$ délek a spočítat pro ně „Kraftův součet“. Pokud

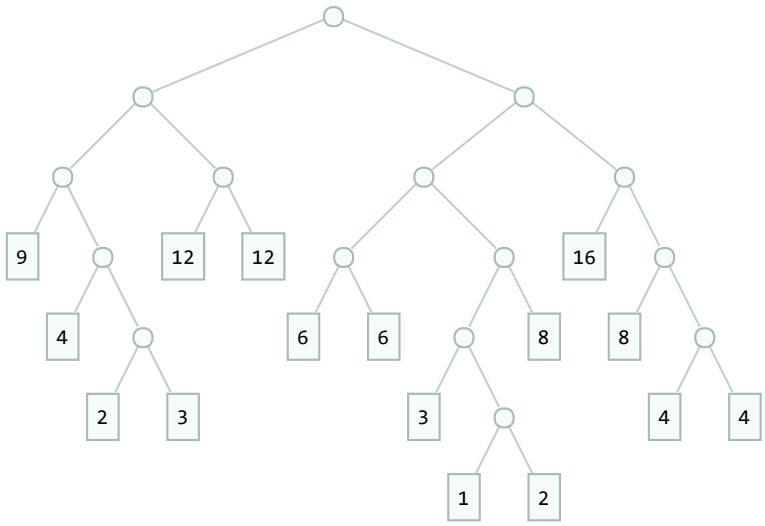
je menší roven jedné, pak to lze a víme i jak, pokud je součet větší než jeden, pak to nelze.

(f) Sjednocením konkatenací dostáváme prostý kód na $A \cup A^2 \cup \ldots \cup A^n$, pro libovolné n, protože blokové rozšíření prefixového kódu je prefixové (viz skripta).

(g) V případě kódů z bodu (e) i (f) může obecně nastat situace, kdy digram ab má kratší kód než písmenko c. V takovém případě musí být součin pravděpodobností písmen v digramu větší než pravděpodobnost písmena c a to tak, aby se to projevilo při zaokrouhlování logaritmu při určování délek pro Shannonův kód. V našem konkrétním případě, nejpravděpodobnější digram 44 má pravděpodobnost $16/100$ a délku kódu 3, nejméně pravděpodobné písmenko 1 má pravděpodobnost $1/10$ a délku kódu 4. V případě kódování z bodu e) bude mít tedy písmeno delší kód než digram. V případě bodu f) je to už jinak. Délka blokového kódu nejpravděpodobnějšího digramu 44 je $2+2$. To už není méně než délka kódu nejméně pravděpodobného písmene 1. (Pokud by ovšem byly pravděpodobnosti více vychýleny, pak by popisovaná situace mohla nastat i pro blokový kód.)

(h) Tato situace nastat nemůže ani v obecném případě jiného rozdělení. Pokud je u prefixem v pak je pravděpodobnost u větší nebo rovna pravděpodobnosti v, tedy délka Shannonova kódu pro u je menší nebo rovna délce Shannonova kódu pro v. Tím jsme ošetřili kód z bodu e). Pokud bychom pro v uvažovali blokové rozšíření Shannonova kódu, pak je důsledek očividný, neboť kód $f^*(u) = f(u)$ pak bude přímo prefixem kódu $f^*(v)$.

Pro zajímavost, graf Huffmanova kódu pro (X_1, X_2) vypadá (např.) takto (pravděpodobnosti v procentech):



Tabulka (zaokrouhlených) normalizovaných středních délek

$$\frac{\mathbb{E}\left(\left|f\left(X_{[0..n)}\right)\right|\right)}{n}$$

Shannonova a Huffmanova kódu pro n nezávislých opakování X vypadá takto:

n	Shannon	Huffman
1	2.4	1.9
2	2.155	1.865
3	1.96133	1.859
4	1.98708	1.85323
5	1.97254	1.85347
6	1.92483	1.85143
7	1.92851	1.85111
8	1.89376	1.85065
9	1.9025	1.84955

Zajímavé je, že Huffmanův kód pro pět iterací je v přepočtu na jedno použití horší než pro čtyři iterace: součiny čtveřic pravděpodobností se shodou okolností aproximují mocninami dvojky relativně lépe než součiny pětic. Podobně pro Shannonův kód mezi šesti a sedmi iteracemi a osmi a devíti iteracemi. Tento jev je u Shannonova kódu vzhledem k „bezmyšlenkovitému“ zaokrouhlování nahoru méně překvapivý.

4. (a) Pro vstupy 1,2,3,4 jsou kódová slova po řadě 1011010, 101100, 10101, 10100. Je to původní Shannonův kód s prefixem $\gamma_1(1) = 101$.

   (b) Střední délka kódu stoupla o 3 (délka každého kódového slova takto stoupla) na hodnotu 5.4. Porovnáním s entropií 1.85 je vidět, že je kód značně neoptimální.

   (c) Použijeme dřívější výsledek a přičteme všude délku prefixu $\gamma_1(2) = 100$, t.j. 3, k délkám 7,6,6,5, 6,5,5,4, 6,5,4,4, 5,4,4,3, původního Shannonova kódu pro digramy.

   (d) Může to nastat, pokud má nějaké kratší slovo výrazně menší pravděpodobnost než slovo delší. Taková situace nastane kdykoliv existují dvě písmenka o různé pravděpodobnosti (pro jednoduchost nenulové pravděpodobnosti). Řekněme, že $P(a) > P(b) > 0$. Uvažujme $u = a^{n+1}$ a $v = b^n$. Informační obsahy budou s n růst lineárně, t.j.

   $$\mathcal{I}_{X_1,X_2,\ldots,X_{n+1}}(a^n) = (n+1)\mathcal{I}_X(a), \qquad \mathcal{I}_{X_1,X_2,\ldots,X_n}(a^n) = n\mathcal{I}_X(b)$$

   Rozdíl délek kódových slov $|f_n(v)| - |f_{n+1}(u)|$ půjde tedy do nekonečna lineárním tempem. Naopak $\gamma_1(n)$ má logaritmický růst. Tedy i po zaokrouhlení bude pro dostatečně velké n,

   $$|\gamma_1(n) f_n(b^n)| > |\gamma_1(n+1) f_{n+1}(a^{n+1})|.$$