Teorija Informacij in sistemov, predavanje 2

Uroš Lotrič

Univerza v Ljubljani,
Fakulteta za računalništvo in informatiko

Kodi

- Vzemimo, da vsako sporočilo ustreza enemu dogodku.
- Sporočila so lahko: znaki abecede, števke 0-9, 0 in 1, ...
- **Kod** sestavljajo **kodne zamenjave**, ki so sestavljene iz znakov **kodne abecede**. Število znakov v kodni abecedi označujemo z r.
- Kodno abecedo največkrat predstavljata znaka 0 in 1, lahko je tudi večja.
- **Kod** je dodelitev kodnih zamenjav znakom osnovne abecede
- Primer: A - 0, B - 01, C - 010
- Mnogi kodi so zelo splošno uporabljani:
  - Morsejeva koda {., -, presledek}, SOS = ...- - -...
  - koda ASCII: A - 1000001, B - 1000010

2.1 Povprečna dolžina kodnih zamenjav

- Pomembna lastnost koda je dolžina kodnih zamenjav
- Kratke kodne zamenjave so bolj zaželene
  - Morse: E = ., A = .-, Y = - . - -
- Če so $\{p_1, \dots, p_n\}$ verjetnosti znakov $\{s_1, \dots, s_n\}$ osnovnega sporočila in $\{l_1, \dots, l_n\}$ dolžine prirejenih kodnih zamenjav je povprečna dolžina kodne zamenjave $= L = \sum_{i=1}^{n} p_i l_i$
- Kod je **optimalen**, če ima najmanjšo možno povprečno dolžino kodnih zamenjav - za kodiranje smo uporabili najmanjše možno število znakov
- Kod je **idealen**, če je povprečna dolžina kodnih zamenjav enaka entropiji

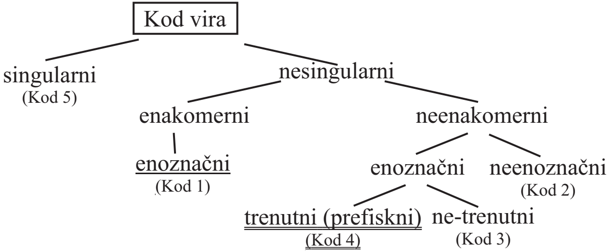
2.1 Tipi kodov 1

- osnovnim simbolom lahko priredimo poljubne kodne zamenjave
- vse kodne zamenjave niso uporabne
- singularni kodi: različnim znakom je prirejena ista kodna zamenjava
- nesingularnost ni dovolj: ne vemo kje se znak začne in kje konča
- Primeri kodov ($A = \{s_1, s_2, s_3\}$, $B = \{0, 1\}$):

A	$p(s_i)$	Kod 1	Kod 2	Kod 3	Kod 4	Kod 5
s_1	0,5	00	0	1	0	0
s_2	0,3	01	1	10	10	0
s_3	0,2	10	01	100	11	1
L		2	1.2	1.7	1.5	1

2.1 Tipi kodov 2

Razvrstitev kodov:



- Kod je **enakomeren**, če je dolžina vseh kodnih zamenjav enaka.
- Kod je **enoznačen**, če lahko poljuben niz znakov dekodiramo na en sam način.
- Kod je **trenuten**, če lahko osnovni znak dekodiramo takoj, ko sprejmemo celotno kodno zamenjavo. To velja le, če kod ne vsebuje kodne zamenjave, ki bi bila predpona kakšni drugi kodni zamenjavi.

2.1 Kodna drevesa

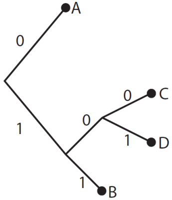
- Kodno drevo: vsako vozlišče ima največ toliko vej kot je različnih znakov v abecedi B. Veje vodijo do naslednjih vozlišč
- Vozlišča, ki predstavljajo kodne zamenjave označimo z debelo piko
- Primeri kodnih dreves za prejšnjo tabelo

2.2 Kraftova neenakost 1

- Za dolžine kodnih zamenjav $\{l_1, \dots, l_n\}$ in r znaki kodirne abecede obstaja trenutni kod, če in samo če velja

$$\sum_{i=1}^{n} r^{-l_i} \leq 1$$

- Primer: enakomerni kod za n znakov, $B = \{0, 1\}$: $n2^{-l} \leq 1 \rightarrow n \leq 2^l$, $l = 5 \rightarrow n = 32$ - število različnih kodnih zamenjav, ki jih lahko tvorimo iz 5 znakov
- Skica dokaza s kodnim drevesom:



2.3 Povprečna dolžina zamenjav in entropija 1

$$H_r(X) \leq L \rightarrow \frac{H(X)}{\log r} \leq L$$

- Dokažimo: $H(X) - L \log r \leq 0$ (uporabimo tangento na logaritem in Kraftovo neenakost)
- Najkrajše povprečne kodne zamenjave imamo, če je $H_r(X) = L \rightarrow l_i = -\log_r p_i$
- zgornje je res samo, če je za vsak i l_i celo število, drugače $l_i = \lceil -\log_r p_i \rceil$
- kljub zaokrožitvi imajo bolj verjetni znaki krajše kodne zamenjave

2.3 Povprečna dolžina zamenjav in entropija 2

$$L < H_r(X) + 1 \rightarrow L < \frac{H(X)}{\log r} + 1$$

- Dokaz za zgornjo mejo (upoštevamo zaokrožitev navzgor)
- **omejitev povprečne dolžine kodnih zamenjav**

$$H_r(X) \leq L < H_r(X) + 1$$

$$\frac{H(X)}{\log r} \leq L < \frac{H(X)}{\log r} + 1$$

Kod je **gospodaren**, če je L znotraj zgornjih mej.

- **učinkovitost koda**

$$\eta = \frac{H(x)}{L \log r} \quad , \quad 0 < \eta \leq 1$$

2.3 Shannonov prvi teorem

- Prva ideja: krajše kodne zamenjave za bolj verjetne znake (pravkar obdelali)
- Druga ideja: združevanje znakov v večje, sestavljene znake in kodiranje sestavljenih znakov
  - $(X, X, \dots, X) = X^n$, $H(X^n) = nH(X)$
  - $H(X^n) \leq L_n < H(X^n) + 1 \rightarrow nH(X) \leq L_n < nH(X) + 1$
  - $H(X) \leq L_n/n < H(X) + 1/n$
  - zgornja meja povprečne dolžine kodne zamenjave na znak L_n/n se v limiti $n \rightarrow \infty$ približuje $H(X)$
- Teorem: Za nize neodvisnih znakov dolžine n (X^n) obstajajo kodi za katere velja

$$\lim_{n \rightarrow \infty} \frac{L_n}{n} = H(X)$$

  pri čemer je $H(X)$ entropija vira X.
- Posledica: z večanjem n se zgornja meja lahko poljubno približa entropiji
- Cena: kompleksno kodiranje in zakasnitev pri dekodiranju

3.1 Shannonov kod 1

- znake razvrstimo po padajočih verjetnostih
- število znakov v vsaki kodni zamenjavi določimo iz enačbe $l_k = \lceil -\log_r p_k \rceil$
- za vse simbole izračunamo kumulativne verjetnosti $P_k = \sum_{i=1}^{k-1} p_i$
- P_k pretvorimo v bazo r. Kodno zamenjavo predstavlja prvih l_k znakov necelega dela števila

3.1 Shannonov kod 2

- Primer 1: $\{0.5, 0.25, 0.125, 0.125\}$
  $L = 7/4, H = 7/4, \eta = 1$
- Primer 2: $\{0.39, 0.19, 0.16, 0.13, 0.13\}$
  $L = 2.61, H = 2.17, \eta = 0.83$
- Kod ni najboljši: za verjetnosti $\{0.9, 0.1\}$ dobimo namesto kodnih zamenjav dolžine 1, eno dolžine 1 in drugo dolžine 4!

3.2 Fanojev kod

- znake razvrstimo po padajočih verjetnostih
- Znake razdelimo v r čim bolj enako verjetnih skupin
- Vsaki skupini priredimo enega od r znakov kodne abecede
- Deljenje ponovimo na vsaki od skupin. Ponavljamo, dokler je mogoče
- Primer 1: $\{0.5, 0.25, 0.125, 0.125\}$, $L = 7/4, H = 7/4, \eta = 1$
- Primer 2: $\{0.39, 0.19, 0.16, 0.13, 0.13\}$, $L = 2.26, H = 2.17, \eta = 0.96$