Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

Međuispit iz predmeta **TEORIJA INFORMACIJE**, 23. studenog 2016.

Pravilo bodovanja zadataka

Netočno odgovoreni zadaci od 5 bodova donose 2 negativna boda, a netočno odgovoreni zadaci od 10 bodova donose 4 negativna boda. Neodgovoreni zadaci donose 0 bodova.

1. zadatak (5 bodova): Komunikacijskim kanalom prenosi se jako dugačak slijed poruka, pri čemu su poruke generirane iz skupa X koji sadrži četiri simbola, $X = \{x_1,...,x_4\}$. Omjer vjerojatnosti pojavljivanja poruka je $p(x_1)$: $p(x_2)$: $p(x_3)$: $p(x_4)$ = 1 : 2 : 2 : 5. Matrica uvjetnih vjerojatnosti prijelaza u kanalu zadana je kao:

$$\left[p\left(y_j \mid x_i\right)\right]=\begin{bmatrix} 0,1 & 0,2 & 0,2 & 0,5 \\ 0,1 & 0,2 & 0,5 & 0,2 \\ 0,1 & 0,5 & 0,2 & 0,2 \\ 0,5 & 0,1 & 0,2 & 0,2 \end{bmatrix}$$

Koliko iznosi entropija šuma u kanalu?

a) 1,761 bit/simbol
b) 1,671 bit/simbol
c) 0,239 bit/simbol
d) 0,329 bit/simbol
e) ništa od navedenog

Rješenje:

i) Sukladno uvjetima navedenim u zadatku vrijedi:

$$p\left(x_1\right):p\left(x_2\right):p\left(x_3\right):p\left(x_4\right)=1:2:2:5$$

Simboli koji se pojavljuju na izlazu iz izvorišta moraju sačinjavati potpuni vjerojatnosni skup.

Iz toga slijedi: $p(x_1) = 0,1$, $p(x_2) = p(x_3) = 0,2$ i $p(x_4) = 0,5$ tj.

$$\left[p\left(x_i\right)\right]=\begin{bmatrix} 0,1 & 0,2 & 0,2 & 0,5 \end{bmatrix}$$

Matrica združenih vjerojatnosti računa se prema poznatom izrazu:

$$\left[p(x_i,y_j)\right]=\left[p(x_i)p(y_j \mid x_i)\right]=\begin{bmatrix} 0,01 & 0,02 & 0,02 & 0,05 \\ 0,02 & 0,04 & 0,1 & 0,04 \\ 0,02 & 0,1 & 0,04 & 0,04 \\ 0,25 & 0,05 & 0,1 & 0,1 \end{bmatrix}$$

Entropiju šuma izračunamo prema izrazu:

$$H(Y|X) = -\sum_{i=1}^{4}\sum_{j=1}^{4} p\left(x_i, y_j\right) \log_2 p\left(y_j \mid x_i\right) \left[\frac{\text{bit}}{\text{simbol}}\right]$$

$$H(Y|X) = 1{,}761 \frac{\text{bit}}{\text{simbol}}$$

2. zadatak (5 bodova): Diskretno bezmemorijsko izvorište generira simbole iz skupa $X = \{x_1, x_2, x_3, x_4\}$. Vjerojatnosti pojavljivanja simbola su sljedeće: $p(x_1) = 0{,}4$, $p(x_2) = 0{,}3$, $p(x_3) = 0{,}2$ i $p(x_4) = 0{,}1$. Izračunajte količinu informacije koja se prenosi u poruci $x_1x_2x_1x_2$.

a) 6,70 bit/poruka
b) 7,386 bit/poruka
c) 7,703 bit/poruka
d) 6,118 bit/poruka
e) ništa od navedenog

Postupak rješavanja:

$$I\left(x_i\right) = -\log_2 p\left(x_i\right) \ \text{bit/simbol}$$

$$I\left(x_1x_2x_1x_2\right) = -\sum_{i=1}^{4} \log p\left(x_i\right) = -\log_2\left(p\left(x_1\right)\cdot p\left(x_2\right)\cdot p\left(x_1\right)\cdot p\left(x_2\right)\right)$$

$$I\left(x_1x_2x_1x_2\right) = -\log_2\left(0{,}4\cdot 0{,}3\cdot 0{,}4\cdot 0{,}3\right)$$

$$I\left(x_1x_2x_1x_2\right) = 6{,}118 \ \text{bit/poruka}$$

3. zadatak (5 bodova): Poruka „aaaaaaaaaa*“ (navodnici nisu dio poruke) kodira se algoritmom LZ77 tako da je maksimalna duljina posmičnog prozora 1 simbol, a prozora za kodiranje 10 simbola. Koliko uređenih trojki (*pomak, duljina, sljedeći simbol*) generira navedeni algoritam kako bi kodirao poruku?

a) jednu
b) dvije
c) tri
d) četiri
e) ništa od navedenog

Rješenje:

U postupku kodiranja dobiju se dvije trojke: (0,0,a) i (1,9,*).

4. zadatak (5 bodova): Dan je skup simbola $\{s_1, s_2, \ldots, s_m\}$ s pripadajućim vjerojatnostima pojavljivanja: $p(s_i) = p_i$, $i = 1, \ldots, m$. Simboli su jednoznačno kodirani prefiksnim kodom. Ako je $m = 6$ i ako su duljine kodnih riječi zadane kao $l_1 = l_2 = 1$, $l_3 = l_5 = 2$, $l_4 = l_6 = 3$, odredite najmanji broj simbola abecede prefiksnog koda.

a) 3

b) 2

c) 4

d) 1

Postupak rješavanja:

$$\sum_{i=1}^{m} d^{-l_i} \leq 1 \rightarrow \frac{1}{d}+\frac{1}{d}+\frac{1}{d^2}+\frac{1}{d^3}+\frac{1}{d^2}+\frac{1}{d^3} \leq 1$$

$$2d^2+2d+2-d^3 \leq 0$$

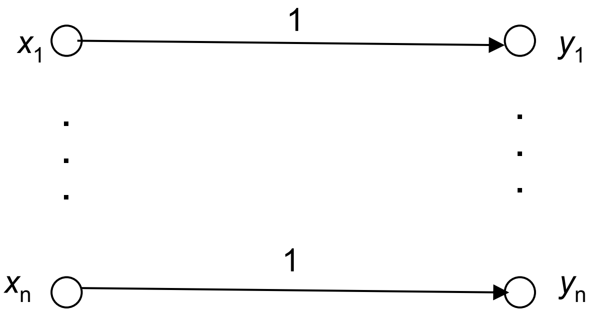
Očito za d = 1 kod nema smisla. Za d = 2 (binarni kod) nejednakost nije zadovoljena, a za d = 3 je. Dakle, najmanji broj simbola abecede prefiksnog koda je 3.

6. zadatak (5 bodova): Informacijski izvor generira simbole iz abecede koju čini osam različitih simbola. Izvor svake milisekunde generira jedan simbol. Simboli se prenose diskretnim informacijskim kanalom. Od svih mogućih kanala odaberite onaj koji ima maksimalan kapacitet i odredite koliko je maksimalno bita informacije moguće prenijeti takvim kanalom u jedinici vremena.

a) 2000 bit/s
b) 1,98 bit/s
c) 3000 bit/s
d) 3 bit/s
e) ništa od navedenog

Postupak rješavanja:

Maksimalna količina informacije prenosi se kanalom bez šuma.



$$\left[P(Y|X)\right]=\begin{bmatrix}1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & 1\end{bmatrix}$$

U takvom kanalu maksimalna transinformacija, tj. kapacitet kanala se postiže kada su svi simboli na ulazu kanala jednako vjerojatni. U tom slučaju $H(X)$ iznosi 3 bit/simbol. S obzirom na brzinu slanja simbola, kanalom će biti moguće prenijeti $H(X)$ bit/simbol svake milisekunde što ukupno daje 3000 bit informacije po sekundi.

7. zadatak (5 bodova): Huffmanovim binarnim kodom potrebno je kodirati n simbola ($n \in \mathbf{N}$) s vjerojatnostima pojavljivanja $p_1, p_2, \dots, p_n$ ($p_i \neq 0$, $i = 1, \dots, n$). Koliko iznosi najveća duljina kodne riječi pridružena nekom simbolu uzimajući u razmatranje sve moguće razdiobe vjerojatnosti pojavljivanja simbola?

a) n bit
b) $n-2$ bita
c) $n-1$ bita

d) $n/2 + 1$ bita

e) ništa od navedenog

Rješenje:

Najveća duljina kodne riječi pridružena je simbolu s najmanjom vjerojatnosti pojavljivanja i koji je u svakom koraku kodiranja obuhvaćen Huffmanovim binarnim stablom.

Dakle, pretpostavimo skup od n simbola x_1, x_2, x_3,..., x_n s vjerojatnostima pojavljivanja $p_1 \geq p_2 \geq p_3 \geq ... \geq p_{n-1} \geq p_n$.

Odredimo duljinu kodne riječi pridruženu simbolu x_n.

Korak	Nadsimbol i uvjet
1.	$p_n + p_{n-1} < p_{n-3}$
2.	$(p_n + p_{n-1}) + p_{n-2} < p_{n-4}$
3.	$[(p_n + p_{n-1}) + p_{n-2}] + p_{n-3} < p_{n-5}$
...	...
n-2.	$[(p_n + p_{n-1}) + p_{n-2}] + p_{n-3} + ... + p_2 < 1$
n-1.	$[(p_n + p_{n-1}) + p_{n-2}] + p_{n-3} + ... + p_2 + p_1 = 1$

Dakle: $l(x_n)$ = n - 1 bit

8. zadatak (5 bodova): Informacijski izvor generira simbole hrvatske abecede. Slijed se kodira kodom LZW. Prije početka kodiranja u memoriju kodera pohranjeni su zapisi oblika (indeks, znak) za sva slova abecede i za znakove koji se u tekstu mogu pojaviti. Razmotrimo poseban slučaj (npr. izvor je u kvaru) kad izvor generira isključivo slovo **a**. Ako izvor izgenerira točno 4951 slovo **a**, koliko će pri tome biti novih zapisa oblika (indeks, riječ) u rječniku? **Pojašnjenje:** svaki slijed sastavljen od uzastopnih slova **a**, npr. **aaa**, a koji nije od ranije sadržan u rječniku, smatra se jednim novim zapisom u rječnik, oblika (indeks, riječ).

a) 100

b) 99

c) 4950

d) 4951

e) ništa od navedenog

Postupak rješavanja:

Ako izvor kodira slijed uzastopnih slova **a** moguće je zamijetiti sljedeću pravilnost:

1) kad izvor kodira prvo slovo **a** (na samom početku dugačkog slijeda), ne zapisuje ništa u rječnik jer za slovo **a** već postoji zapis u rječniku;

2) zatim uzima sljedeće slovo **a** i u rječnik zapisuje simbol **aa** te njemu pripadajući indeks (to je jedan novi zapis);

3) zatim uzme to drugo slovo **a** i spaja ga s naredna dva slova **a** te u rječnik zapisuje **aaa**

4) itd.

Ako to prikažemo tablicom, vidljiva je sljedeća pravilnost:

iteracija i	broj uzastopnih slova **a** N_i(**a**)	ukupan broj novih zapisa u rječniku x_i
1	1	0
2	2	1
3	4	2
4	7	3
5	11	4
6	16	5

Broj uzastopnih slova **a** (lijevi stupac tablice), N_i(**a**), se povećava svaki puta za ukupan broj novih zapisa, x_i, tj. vrijedi: $N_i(\mathbf{a}) - N_{i-1}(\mathbf{a}) = x_i$, za $i \geq 2$. Iz toga možemo zaključiti da za ukupan broj uzastopno generiranih slova **a**, N_M, vrijedi:

$$N_M = \sum_{i=2}^{M} x_i + N_1 = \sum_{i=1}^{M-1} i + 1 = 1 + \frac{M \cdot (M-1)}{2}$$

$$M^2 - M - 2(N_M - 1) = 0$$

$$M_{1,2} = \frac{1 \pm \sqrt{1 + 8(N_M - 1)}}{2}$$

Očito je da samo jedno rješenje ima smisla, pa konačno vrijedi:

$$M = \frac{1 + \sqrt{1 + 8(N_M - 1)}}{2}$$

Dakle, ako je $N_M = 4951$, tada je $M = 100$. Konačno, na kraju slijeda od 4951 slova **a** broj novih zapisa, x_M, je za jedan manji od M te iznosi 99, što je ujedno točan odgovor na ovaj zadatak.

9. zadatak (10 bodova): Neka su X_1 i X_2 diskretne slučajne varijable koje poprimaju vrijednosti iz skupova $\{1, 2, ..., m\}$, odnosno $\{m + 1, m + 2, ..., m + n\}$, $m, n \in \mathbb{N}$, te neka su njihove pripadajuće razdiobe vjerojatnosti p_{X_1}, odnosno p_{X_2}. Neka slučajna varijabla X definirana je kao

$$\mathrm{X} = \begin{cases} X_1 \text{ s vjerojatnošću } \alpha \\ X_2 \text{ s vjerojatnošću } 1 - \alpha \end{cases}$$

Odredite maksimalnu vrijednost entropije $H(X)$ u ovisnosti o parametru α.

a) $log_2 \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}}$

b) $\frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}}$

c) $log_2 \frac{1-\alpha}{\alpha} + H(X_1) - H(X_2)$

d) $log_2\left(2^{H(X_1)} + 2^{H(X_2)}\right)$

e) ništa od navedenog

Rješenje:

i)

Radi lakoće zapisa stavimo: $p_{X_1} = p_1$ i $p_{X_2} = p_2$. Koristeći da je $p_X(x) = \alpha p_1(x)$ za $x \in X_1$, odnosno $p_X(x) = (1-\alpha)p_2(x)$ za $x \in X_2$ dobivamo

$$H(X) = -\sum_{i=1}^{m} \alpha p_1(x=i) log_2\, \alpha p_1(x=i) - \sum_{i=m+1}^{m+n} (1-\alpha)p_2(x=i) log_2\,(1-\alpha)p_2(x=i)$$

$$= -\alpha \sum_{i=1}^{m} p_1(x=i)[log_2\alpha + log_2\, p_1(x=i)] - (1-\alpha) \sum_{i=m+1}^{m+n} p_2(x=i)[log_2(1-\alpha) + log_2\, p_2(x=i)]$$

$$= -\alpha log_2\alpha - (1-\alpha)log_2(1-\alpha) + \alpha H(X_1) + (1-\alpha)H(X_2)\ \frac{\text{bit}}{\text{simbol}}$$

ii)

Maksimum $H(X)$ u ovisnosti o α dobivamo iz

$$\frac{dH(X)}{d\alpha} = log_2 \frac{1-\alpha}{\alpha} + H(X_1) - H(X_2) = 0$$

što daje

$$\frac{1-\alpha}{\alpha} = 2^{H(X_2)-H(X_1)}$$

odnosno

$$\alpha = \frac{1}{1 + 2^{H(X_2)-H(X_1)}} = \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}}.$$

Zamjenom prethodnog izraza u $H(X)$ dobivamo

$$H(X) \le \frac{2^{H(X_1)}}{2^{H(X_1)}+2^{H(X_2)}}\left(H(X_1) - log_2 \frac{2^{H(X_1)}}{2^{H(X_1)}+2^{H(X_2)}}\right) + \frac{2^{H(X_2)}}{2^{H(X_1)}+2^{H(X_2)}}\left(H(X_2) - log_2 \frac{2^{H(X_2)}}{2^{H(X_1)}+2^{H(X_2)}}\right)$$

što nakon sređivanja daje

$$H(X) \leq log_2\left(2^{H(X_1)} + 2^{H(X_2)}\right)$$