Parcial 1 (23 Marzo 2017), preguntas y respuestas

Informació i Seguretat (Universitat Autònoma de Barcelona)

INFORMACIÓ I SEGURETAT
23 de març de 2017

Nom i cognoms: _____ NIU: _____ Grup: _____

- Cal que justifiqueu convenientment totes les respostes
- $\log 3 = 1.58$, $\log 5 = 2.32$, $\log 7 = 2.8$

1. (1 punt, 0.5+0.5) Tenim el joc *Qui és qui?* que consta de 4 files de personatges amb 8 personatges a cada fila. Un dels jugadors escull un personatge a l'atzar i l'altre, fent preguntes amb resposta SI/NO ha d'encertar quin personatge ha escollit. Cada cop que fem una pregunta, eliminem del nostre tauler tots els personatges que no la compleixen i, amb els que queden, tornem a fer una altra pregunta.

   (a) Sabem que la meitat dels personatges són homes i la meitat dones i que cap dona té barba i només la meitat dels homes en tenen. Justifiqueu, sense calcular el valor exacte de la informació, amb quina de les següents preguntes obtenim més informació:
   - És dona?
   - Té barba?

   (b) Suposant que amb els personatges que tenim sempre podem trobar una pregunta que exactament la meitat la compleixin, justifiqueu, fent servir els coneixements de teoria de la informació, quantes preguntes hem de fer per saber quin és el personatge escollit.

Solució:

(a) Quan fem una pregunta amb resposta SI/NO, obtenim la informació màxima si les dues possibles respostes són equiprobables. En aquest cas, la resposta a la pregunta "És dona?" és SI amb probabilitat 0.5 i NO amb la mateixa probabilitat. Per tant, la informació que obtenim de la resposta és màxima. En canvi, a la pregunta "Té barba?" la probabilitat que sigui SI és 0.25 i la probabilitat que sigui NO és 0.75. Així, obtenim més informació amb la resposta de la primera pregunta.

(b) Al joc hi ha $4 \cdot 8 = 32$ personatges i s'escull un a l'atzar. L'experiència, per tant, consta de 32 successos equiprobables i té una entropia de $\log 32 = \log 2^5 = 5$ bits. Si sempre podem trobar una pregunta que exactament la meitat dels personatges la compleixin cada resposta ens aporta la màxima informació que és $H(\frac{1}{2}, \frac{1}{2}) = 1$ bit. Per tant, haurem de fer 5 preguntes per saber quin és el personatge escollit.

2. (2.5 punts, 1+0.5+0.5+0.5) Considereu una font S amb sis símbols i la següent distribució de probabilitats: $\{\frac{1}{12}, \frac{1}{12}, \frac{2}{12}, \frac{2}{12}, \frac{2}{12}, \frac{4}{12}\}$. Amb aquesta distribució de probabilitats, $H(S) = 2.41$.

   (a) Si volem construir un codi **ternari** C_1 de longitud fixa per codificar els símbols de la font, quina longitud ha de tenir aquest codi? Digueu si amb tuples o triples obtenim una longitud per símbol inferior a l'obtinguda amb el codi C_1.

   (b) Justifiqueu si pot existir o no un codi **ternari** instantani de longitud mitjana 2.3 que codifiqui els símbols de S.

   (c) Doneu un codi **binari** òptim C_2 per als missatges de S i digueu quina és la seva longitud mitjana.

   (d) Justifiqueu quin dels dos codis anteriors C_1 i C_2 és millor fent servir el concepte d'eficiència. Noteu que el primer és ternari i el segon és binari.

Solució:

   (a) Per construir un codi de longitud fixa, necessitem $6 \leq 3^L$, és a dir, $L \geq \log_3(6) = 1 + \log_3(2) = 1.63$, o bé $L \geq \frac{\log_2(6)}{\log_2(3)} = \frac{1+1.58}{1.58} = 1.63$. Per tant $L = 2$. Si agrupem els símbols de dos en dos, necessitem $6^2 \leq 3^L$, és a dir, $L \geq 2\log_3(6) = 3.26$ o $L \geq 2\frac{\log_2(6)}{\log_2(3)} = 3.26$; per tant $L = 4$ i no millorem la longitud per símbol del codi C_1, ja que és la mateixa $L/2 = 2$. Si agrupem de tres en tres, necessitem $6^3 \leq 3^L$, és a dir, $L \geq 3\log_3(6) = 4.9$ o $L \geq 3\frac{\log_2(6)}{\log_2(3)} = 4.9$; per tant $L = 5$ i sí millorem la longitud per símbol que és $L/3 = 5/3 = 1.67 < 2$. Per tant, a partir d'agrupar els símbols de tres en tres o més, obtenim una longitud per símbol inferior a 2.

   (b) Pel primer teorema de Shannon, tenim que $\bar{L} \geq \frac{H(S)}{\log 3}$. Calculem primer $H(S)$:

$$H(S) = \frac{2}{12}\log 12 + \frac{3}{6}\log 6 + \frac{1}{3}\log 3 = \frac{2}{12}(2 + \log 3) + \frac{3}{6}(1 + \log 3) + \frac{1}{3}\log 3 =$$

$$= \frac{4+6}{12} + (\frac{2+6+4}{12})\log 3 = \frac{10}{12} + \log 3 = 0.83 + 1.58 = 2.41.$$

   Per tant, com que $2.3 > H(S)/\log 3 = 1.5$, sí que pot existir un codi ternari instantani (òptim o no) amb longitud mitjana 2.3, però no ho podem assegurar.

   (c) Aplicant l'algorisme de Huffman, obtenim el següent codi binari òptim

$$C = \{110, 010, 111, 011, 00, 01\}.$$

   La longitud mitjana és $\bar{L} = 2 \cdot \frac{4}{12} + 2 \cdot \frac{2}{12} + 3(\frac{2+2+1+1}{12}) = \frac{30}{12} = 2.5$.

   (d) El codi ternari C_1 té longitud mitjana $\bar{L}_1 = 2$, per tant l'eficiència és $\eta_1 = \frac{H(S)}{\bar{L}_1 \cdot \log 3} = \frac{1.5}{2} = 0.75$. En canvi, el codi binari C_2 té longitud mitjana superior $\bar{L}_2 = 2.5$, però l'eficiència és $\eta_2 = \frac{H(S)}{\bar{L}_2 \cdot \log 2} = \frac{2.4}{2.5} = 0.9$. Per tant, el codi C_2 és millor que el C_1.

3. (2 punts, 0.5+0.5+0.5+0.5) Sigui $\{A_1, A_2\}$ el conjunt d'entrades i $\{B_1, B_2, B_3, B_4\}$ el de sortides d'un canal amb matriu de probabilitats condicionades

$$\begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} \end{pmatrix}.$$

Suposem que les probabilitats dels símbols d'entrada són $\{\frac{1}{4}, \frac{3}{4}\}$.

(a) Quina és la probabilitat de que la sortida sigui B_1?

(b) Quina és la informació mitjana de la entrada si sabem que la sortida és B_1?

(c) Quina és la informació mitjana de la entrada si sabem la sortida?

(d) Quina és la informació mútua de l'entrada i la sortida?

Solució: Calculem les probabilitats condicionades $p(B_j|A_i)$ (que venen donades per la matriu del canal), les conjuntes $p(A_i, B_j)$ i les probabilitats condicionades $p(A_i|B_j)$.

$p(B_j\|A_i)$	B_1	B_2	B_3	B_4
A_1	$\frac{1}{2}$	0	$\frac{1}{2}$	0
A_2	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$

$p(A_i, B_j)$	B_1	B_2	B_3	B_4
A_1	$\frac{1}{8}$	0	$\frac{1}{8}$	0
A_2	$\frac{3}{16}$	$\frac{3}{16}$	0	$\frac{3}{8}$
$p(B_j)$	$\frac{5}{16}$	$\frac{3}{16}$	$\frac{1}{8}$	$\frac{3}{8}$

$p(A_i\|B_j)$	B_1	B_2	B_3	B_4
A_1	$\frac{2}{5}$	0	1	0
A_2	$\frac{3}{5}$	1	0	1

(a) La probabilitat de que la sortida sigui B_1 és $p(B_1) = \frac{5}{16}$.

(b) La informació mitjana de la entrada si sabem que la sortida és B_1 és $H(A|B_1)$. Considerant els valors de les taules de l'apartat anterior, tenim que $H(A|B_1) = H(\frac{2}{5}, \frac{3}{5}, 0, 0) = \frac{2}{5}\log(5/2) + \frac{3}{5}\log(5/3) = \frac{2}{5}(\log 5 - 1) + \frac{3}{5}(\log 5 - \log 3) = \log 5 - \frac{2}{5} - \frac{3}{5}\log 3 \approx 2.32 - 0.4 - 0.6 \cdot 1.58 \approx 0.97$.

(c) La informació mitjana de la entrada si sabem quin és el símbol de sortida és $H(A|B)$. Tenim que $H(A|B) = \sum_{j=1}^{4} p(B_j)H(A|B_j)$. Per l'apartat anterior $H(A|B_2) \approx 0.97$. A més $H(A|B_2) = H(A|B_3) = H(A|B_4) = 0$. Per tant, $H(A|B) = \frac{5}{16} \cdot 0.97 \approx 0.3$.

També es pot resoldre sabent que $H(A|B) = \sum_{i=1}^{2}\sum_{j=1}^{4} p(A_i, B_j)\log p(A_i|B_j)$. En aquest cas, $H(A|B) = \frac{1}{8}\log(5/2) + \frac{3}{16}\log(5/3) = \frac{1}{8}(2.32 - 1) + \frac{3}{16}(2.32 - 1.58) = 0.165 + 0.138 \approx 0.3$.

(d) La informació mútua és $I(A, B) = H(A) - H(A|B)$. Tenim que $H(A) = H(\frac{1}{4}, \frac{3}{4}) = \log 4 - \frac{3}{4}(\log 3) \approx 2 - 1.185 \approx 0.815$. Per tant, $I(A, B) \approx 0.81 - 0.30 \approx 0,51$.

La informació mútua també és pot calcular com $I(A, B) = H(B) - H(B|A)$. Tenim que $H(B) = H(\frac{5}{16}, \frac{3}{16}, \frac{1}{8}, \frac{3}{8}) = \frac{5}{16}(4 - \log 5) + \frac{3}{16}(4 - \log 3) + \frac{3}{8} + \frac{3}{8}(3 - \log 3) \approx 1.88$. Tenim que $H(B|A) = \sum_{i=1}^{2} p(A_i)H(B|A_i) = \frac{1}{4}H(\frac{1}{2}, \frac{1}{2}) + \frac{3}{4}H(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}) = \frac{1}{4} + \frac{3}{4} \cdot 1.5 = 1.375$. Per tant, $I(A, B) \approx 1.88 - 1.385 \approx 0.5$.

4. (2 punts, 1+1)

   (a) Considereu la cadena: "ABBCBBCBBCA".

      i. Comprimiu la cadena fent servir LZ77 amb les següents mides del diccionari i del buffer:
         - $D = 3$, $B = 3$;
         - $D = 15$, $B = 15$.

      ii. Considereu que cada caràcter ocupa 8 bits i la mida del diccionari i del buffer determinen el nombre de bits de cada índex. Doneu (en forma de fracció) les taxes de compressió obtingudes en els dos casos anteriors.

   (b) Sigui $S = \{A, B, C\}$ amb distribució de probabilitats $p_A = 0.5$, $p_B = 0.25$ i $p_C = 0.25$.

      i. Descodifiqueu la seqüència rebuda 3872, que ha estat codificada fent servir codificació aritmètica, sabent que la cadena codificada té longitud 2.

      ii. Doneu una altra seqüència que codifiqui la mateixa cadena i que sigui més eficient.

Solució:

(a) La codificació de la cadena amb les diferents mides del diccionari i el buffer és

- $(0, 0, A), (0, 0, B), (1, 1, C), (3, 3, B), (3, 2, A)$;
- $(0, 0, A), (0, 0, B), (1, 1, C), (3, 6, A)$.

La cadena inicial ocupa $8 \cdot 11 = 88$ bits. En el primer cas, necessitem 2 bits per codificar cada índex. Per tant, cada 3-tupla ocupa $2 + 2 + 8 = 12$ bits. En total, la cadena comprimida ocupa $5 \cdot 12 = 60$ bits i la taxa de compressió és $R = \frac{60}{88} = \frac{15}{22} = 0.68$ bpb. En el segon cas, necessitem 4 bits per cada índex i cada 3-tupla ocupa $4 + 4 + 8 = 16$. La taxa de compressió és $R = \frac{4 \cdot 16}{88} = \frac{8}{11} = 0.72$ bpb.

(b) La taula de probabilitats i intervals serà:

Car.	*Prob.*	*Interval*
`A`	0.5	[0.0,0.5)
`B`	0.25	[0.5,0.75)
`C`	0.25	[0.75,1)

Considerem el valor $c_0 = 0.3872$. Com $c_0 \in [0, 0.5)$, el primer caràcter codificat és A. Per obtenir el segon caràcter, calculem $c_1 = \frac{c_0 - 0}{0.5} = 2c_0 = 0.7744 \in [0.75, 1)$, que es correspon a C. Per tant, la cadena codificada és AC. Quan codifiquem la cadena AC, l'interval corresponent a A és $[0, 0.5)$ i el corresponent a AC és $[0.375, 0.5)$. Per tant, podem prendre el valor $0.4 \in [0.375, 0.5)$ que té menys xifres decimals i una codificació més eficient de la cadena seria 4.

5. (2.5 punts, 1.5+0.75+0.25) Sigui $\{A_1, A_2, A_3\}$ el conjunt d'entrades i $\{B_1, B_2, B_3\}$ el de sortides d'un canal, amb matriu de probabilitats condicionades:

$$\begin{pmatrix} 0 & \frac{2}{5} & \frac{3}{5} \\ \frac{3}{5} & 0 & \frac{2}{5} \\ \frac{2}{5} & \frac{3}{5} & 0 \end{pmatrix}.$$

(a) Quina seria la probabilitat mitjana d'error descodificant a mínima probabilitat d'error (MPE) si la distribució inicial de probabilitats és $(\frac{1}{6}, \frac{1}{6}, \frac{2}{3})$? Doneu també la funció de descodificació a MPE.

(b) Digueu de quin tipus de canal es tracta i quina és la seva capacitat.

(c) Doneu una distribució inicial que faci que s'assoleixi la capacitat.

Solució:

(a) La matriu de probabilitats conjuntes és la següent:

$$\begin{pmatrix} 0 & 2/30 & 3/30 \\ 3/30 & 0 & 2/30 \\ 8/30 & 12/30 & 0 \end{pmatrix}.$$

Fixant-nos en els valors màxims a cada columna obtenim la següent funció de descodificació a mínima probabilitat d'error:

$$\begin{aligned} B_1 &\longrightarrow A_3 \\ B_2 &\longrightarrow A_3 \\ B_3 &\longrightarrow A_1 \end{aligned}$$

Aleshores, la probabilitat mitjana d'error en la descodificació és

$$\overline{p}_e = 1 - \frac{8}{30} - \frac{12}{30} - \frac{3}{30} = 1 - \frac{23}{30} = \frac{7}{30} = 0.23.$$

(b) El canal és completament simètric, per tant podem calcular la capacitat fent servir la fórmula $C = \log 3 - H$, on $H = H(\frac{2}{5}, \frac{3}{5}) = \frac{2}{5}\log(5/2) + \frac{3}{5}\log(5/3) = \frac{2}{5}(\log 5 - 1) + \frac{3}{5}(\log 5 - \log 3) = \log 5 - \frac{2}{5} - \frac{3}{5}\log 3 = 2.32 - 0.4 - 0.6 \cdot 1.58 = 0.97$. Per tant, $C = \log 3 - 0.97 = 0.61$.

(c) En aquest tipus de canal, la capacitat s'assoleix quan els valors de la sortida són equiprobables i en un canal totalment simètric això es dona si i només si els valors de l'entrada són equiprobables; o sigui amb distribució de probabilitats $\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$.