INFORMACIÓ I SEGURETAT
8 d'abril de 2021

Nom i cognoms (en MAJÚSCULES): _____ NIU: _____ Grup: ___

- Cal que **justifiqueu convenientment** totes les respostes.
- $\log 3 = 1.58$, $\log 5 = 2.32$, $\log 7 = 2.80$, $\log 23 = 4.52$.

1. (2.5 punts, 0.75+0.5+0.5+0.75) Disposem de dos sacs, S_1 i S_2. El primer sac conté boles taronges (t) i grogues (g) (la mateixa quantitat de cada tipus). El segon, boles grogues (g), vermelles (v), negres (n) i blaves (b) (la mateixa quantitat de cada tipus). Escollim un dels dos sacs, amb una probabilitat p que sigui el primer, i en traiem una bola. Considerem el canal amb entrada el conjunt $S = \{S_1, S_2\}$ i sortida el conjunt R dels possibles colors de les boles.

   (a) Doneu, **en funció de** p, la taula de probabilitat conjunta, les taules de probabilitats condicionades (respecte l'entrada i respecte la sortida) i la distribució del conjunt R.

   (b) Si sabem que hem obtingut una bola blava, quina és la quantitat d'incertesa (mesurada en bits) que tenim respecte el sac escollit? I si sabem que hem escollit el sac S_2, quina és la quantitat d'incertesa (mesurada en bits) que tenim respecte el color de la bola que traiem?

   (c) En el cas d'obtenir una bola groga, per quin valor de p obtenim el valor màxim d'incertesa respecte el sac escollit?

   (d) Doneu el valor de $I(R, S)$, prenent $p = \frac{1}{3}$.

   Solució:

   (a) Considerem $R = \{R_1, R_2, R_3, R_4, R_5\} = \{t, g, v, n, b\}$. La taula de probabilitats condicionades respecte l'entrada és la següent:

$p(R_j\|S_i)$	t	g	v	n	b
S_1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
S_2	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

   La taula de probabilitats conjuntes, les probabilitats de la sortida i la taula de probabilitats condicionades respecte la sortida són:

$p(R_j, S_i)$	t	g	v	n	b
S_1	$\frac{p}{2}$	$\frac{p}{2}$	0	0	0
S_2	0	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
$p(R_j)$	$\frac{p}{2}$	$\frac{1+p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$

$p(S_i\|R_j)$	t	g	v	n	b
S_1	1	$\frac{2p}{1+p}$	0	0	0
S_2	0	$\frac{1-p}{1+p}$	1	1	1

   (b) Si hem obtingut la bola blava, aleshores $H(S|R = b) = H(1, 0) = 0$. Si sabem que hem escollit el sac S_2, aleshores la quantitat d'incertesa respecte el color de la bola és $H(R|S = S_2) = H(0, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) = \log 4 = 2$ bits.

   (c) La incertesa respecte del sac escollit si sabem que hem obtingut una bola groga és $H(\frac{2p}{1+p}, \frac{1-p}{1+p})$. Si $\frac{2p}{1+p} = \frac{1}{2}$ i $\frac{1-p}{1+p} = \frac{1}{2}$, aleshores podem aconseguir el màxim d'incertesa possible. Les dues equacions donen la mateixa solució $p = \frac{1}{3}$.

   (d) Tenim que $I(S, R) = H(S) - H(S|R) = H(R) - H(R|S)$.

i. Per una banda, $H(S) = H(\frac{1}{3}, \frac{2}{3}) = \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} = \log 3 - \frac{2}{3} = 1.58 - 0.66 = 0.92$ bits. Tenim que $H(S|R=o) = H(S|R=r) = H(S|R=g) = H(S|R=b) = H(1,0) = 0$ i $H(S|R=y) = H(\frac{1}{2}, \frac{1}{2}) = 1$. A més $p(R=y) = \frac{1+\frac{1}{3}}{4} = \frac{1}{3}$. Per tant,

$$H(S|R) = \sum_{j=0}^{5} p(R_j)H(S|R_j) = \frac{1}{3}\cdot 1 = \frac{1}{3}$$

Finalment, $I(S,R) = 0.92 - 0.33 = 0.59$ bits.

ii. També es pot calcular com $H(R) - H(R|S)$, on $H(R) = H(\frac{p}{2}, \frac{1+p}{4}, \frac{1-p}{4}, \frac{1-p}{4}, \frac{1-p}{4}) = H(\frac{1}{6}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}) = \log 6 - \frac{1}{3} = 2.58 - 0.33 = 2.25$ i $H(R|S) = \frac{1}{3}\cdot H(\frac{1}{2}, \frac{1}{2}) + \frac{2}{3}H(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) = 0.33 + +0.66 \cdot 2 = 1.65$. Per tant, $I(S,R) = 2.25 - 1.65 = 0.6$

2. (2.5 punts, 0.75+1.75) Considereu la font $S_1 = \{a_1, a_2, a_3, a_4, a_5\}$, amb probabilitats $(\frac{1}{4}, \frac{1}{4}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$. Volem codificar la font per enviar les dades a través d'un canal **binari**.

(a) Quina longitud té un codi de longitud fixa per codificar la font? Quina és la longitud d'un codi de longitud fixa si volem codificar els elements de S en 2-tuples? En quin dels dos casos la longitud per símbol codificat és menor?

(b) Doneu un codi binari òptim i la seva longitud mitjana. Aquest codi té eficiència 1? En cas afirmatiu, doneu l'entropia de la font. En cas negatiu, digueu quina hauria de ser la distribució de probabilitats per tal que el codi òptim obtingut tingui eficiència 1.

Solució:

(a)
- Per tal de codificar 5 elements, el codi de longitud fixa ha de complir que $2^L \leq 5$ i, per tant, $L = 3$.
- En el cas de les 2-tuples, tenim que $2^L \leq 5^2 = 25$. Per tant, $L = 5$.
- En el cas de les 2-tuples tenim que la longitud per símbol és $\frac{L}{2} = \frac{5}{2} = 2,5$. Així, la longitud per símbol és menor si codifiquem 2-tuples.

(b)
- Apliquem l'algorisme de Huffamn per obtenir un codi òptim. Un dels codis que es poden obtenir amb aquest algorisme és, per exemple, $\{10, 01, 00, 111, 110\}$.
- La longitud mitjana és $\bar{L} = 2\cdot 2\cdot\frac{1}{4} + 2\cdot\frac{1}{6} + 2\cdot 3\cdot\frac{1}{6} = 1 + \frac{1}{3} + 1 = 2.33$.
- No pot ser tenir eficiència 1 ja que per a cada símbol i la probabilitats no és 2^{-L_i}, on L_i és la longitud de la paraula codi. De fet, les probabilitats no són potències de 2. Per tal que el codi anterior tingui eficiència 1, la distribució de probabilitats ha de ser $(2^{-2}, 2^{-2}, 2^{-2}, 2^{-3}, 2^{-3}) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$.

3. (2.5 punts, 0.75+0.5+1.25) Hem sofert una emboscada en un dels laboratoris secrets de *Umbrella Corporations* situats en Raccoon city[1]. Hem descobert la seqüència genètica que és capaç de destruir el virus que fa que aquestes criatures, ara descontrolades, continuïn vives. La cadena d'ADN està en una memòria USB comprimida:

`(0,0,A),(0,0,C),(2,5,A),(0,0,G),(2,1,T),(3,1,C),(8,2,T),(6,4,C).`

(a) Descomprimiu el missatge i recupereu la cadena d'ADN

(b) Calculeu el percentatge de compressió de la codificació anterior suposant que les posicions es codifiquen amb 4 bits i les bases nitrogenades de l'ADN $\{A, C, G, T\}$ amb 2 bits. (**Nota:** Podeu deixar el resultat en format de fracció.)

(c) Abans que els *Tyrant* ens agafin, hem de complir amb la missió i enviar la cadena a la base. Comprimiu la cadena amb el mètode LZ78 i decidiu quina de les dues compressions és la més eficient per a tal d'enviar-la a la base (considereu el nombre de bits necessaris per codificar l'índex segons el valor més alt obtingut en la codificació i considereu 2 bits per codificar la base nitrogenada).

Solució:

(a) La cadena original serà:

Input	*Seqüència*
`(0,0,A)`	`A`
`(0,0,C)`	`AC`
`(2,5,A)`	`ACACACAA`
`(0,0,G)`	`ACACACAAG`
`(2,1,T)`	`ACACACAAGAT`
`(3,1,C)`	`ACACACAAGATGC`
`(8,2,T)`	`ACACACAAGATGCCAT`
`(6,4,C)`	`ACACACAAGATGCCATTGCCC`

(b) La cadena original ocupa 21 caràcters, per tant fan falta $21 \times 2 = 42$ bits. La cadena comprimida ocupa $8 \cdot (4+4+2) = 80$. Per tant, $R = \frac{80}{42} = \frac{40}{21}$ i el percentatge de compressió és $(1-R) \cdot 100 = (1-\frac{40}{21}) \cdot 100 = \frac{-1900}{21} = -90.47\%$.

(c) La codificació de la cadena d'ADN amb el mètode LZ78 és:

	Dicc	*Codi*		*Dicc*	*Codi*
0	`null`				
1	`A`	`(0,A)`	6	`AT`	`(1,T)`
2	`C`	`(0,C)`	7	`G`	`(0,G)`
3	`AC`	`(1,C)`	8	`CC`	`(2,C)`
4	`ACA`	`(3,A)`	9	`ATT`	`(6,T)`
5	`AG`	`(1,G)`	10	`GC`	`(7,C)`
			11	`CC`	`(2,C)`

A l'últim pas també és admissible el valor Dicc `CCC` i Codi `(8,Null)`. Prenent l'ultima tuple $(2, C)$, tenim que el valor més alt dels índexs és 7 i per tant necessitem 3 bits per representar els índexs. El missatge ocupa $11 \cdot (3+2) = 55$ bits. Aleshores, $R = \frac{55}{42}$ i és més eficient considerar la compresió LZ78 (encara que és millor no considerar cap).

[1] Exercici ambientat en l'univers de Resident Evil.

4. (2.5 punts, 0.5+0.5+0.5+1)

(a) Quantes matrius de probabilitats condicionades diferents es poden donar en el cas d'un canal sense soroll, on el conjunt d'entrades té cardinal 7?

(b) Si A, B representen el conjunt d'entrades i sortides d'un canal del que coneixem la matriu de probabilitats condicionades. Podem assegurar en un canal determinista que el valor de $I(A, B)$ només depèn de la distribució inicial de probabilitats de la variable aleatòria A? I si el canal no és determinista? **Justifiqueu la resposta.**

(c) Sigui $\{A_1, A_2, A_3\}$ el conjunt d'entrades i $\{B_1, B_2, B_3\}$ el conjunt de sortides d'un canal, amb matriu de probabilitats condicionades:

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{2}{3} \end{pmatrix},$$

i. Doneu una regla de descodificació per màxima versemblança (MV).

ii. Considereu la distribució inicial $\{\frac{1}{2}, \frac{1}{2}, 0\}$, doneu una regla de descodificació per mínima probabilitat d'error (MPE) i calculeu la probabilitat mitjana d'error fent servir aquesta regla de descodificació

Solució:

(a) Volem comptar quantes matrius podem escriure que tinguin exactament un "1" en cada columna. Una matriu d'aquestes vindrà determinada si coneixem una seqüència amb els valors $\{1, 2 \ldots, 7\}$ ordenada de totes les maneres possibles. Cada seqüència ens indicarà en quina fila hem d'escriure un "1", respectivament a la primera columna, segona columna, etc. Per exemple $\{1, 3, 5, 6, 4, 7, 2\}$ ens donaria la matriu

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

La quantitat de seqüències que podem escriure és el nombre de permutacions que podem fer amb 7 elements, o sigui $7! = 5040$.

(b) La matriu de probabilitats de la sortida condicionades respecte l'entrada sempre està fixada pel canal. En tots els cassos, si coneixem la distribució de probabilitats de A, a partir de la matriu de probabilitats condicionades del canal podem calcular les probabilitats de B i llavors podem calcular $I(A, B) = H(B) - H(B|A)$.

(c) i. Una regla per descodificar per descodificació per màxima versemblança és aquella que a cada B_j li assigna un A_i tal que $P(B_j|A_i)$ sigui màxima. En el nostre cas podem fer:

$$\begin{aligned} B_1 &\longrightarrow A_1(\text{ o } A_3) \\ B_2 &\longrightarrow A_2 \\ B_3 &\longrightarrow A_3 \end{aligned}$$

ii. Si la distribució de probabilitats inicial és $\{\frac{1}{2}, \frac{1}{2}, 0\}$, aleshores la matriu de probabilitats conjuntes és la següent:

$p(A_i \cap B_j)$	B_1	B_2	B_3
A_1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
A_2	0	$\frac{1}{3}$	$\frac{1}{6}$
A_3	0	0	0

Una regla per descodificar per la mínima probabilitat d'error és aquella que a cada B_j li assigna un A_i tal que $P(A_i|B_j)$ sigui màxima (o, també, que sigui màxima la probabilitat conjunta $P(A_i \cap B_j)$). En el nostre cas podem fer:

$$\begin{aligned} B_1 &\longrightarrow A_1 \\ B_2 &\longrightarrow A_2 \\ B_3 &\longrightarrow A_1(\text{ o } A_2) \end{aligned}$$

La probabilitat mitjana d'error en la descodificació és $1 - \sum_j P(f(B_j), B_j) = 1 - \frac{1}{6} - \frac{1}{3} - \frac{1}{6} = \frac{1}{3}$.