Examen de Comunicación de Datos

10 de octubre de 2014

Nombre: _____ **D.N.I.**: _____

1. (1 punto) Una urna contiene 12 bolas idénticas en apariencia, aunque una de ellas pesa más que las demás. Para encontrarla disponemos de una balanza con dos platillos, que deberemos usar el mínimo número de veces. Consideramos dos estrategias distintas para la primera pesada:

- Poner seis bolas en cada platillo.
- Poner cuatro bolas en cada platillo.

¿Con qué estrategia obtendremos mayor información?

Llamemos X al peso de uno de los platillos respecto del otro. Con la primera estrategia X puede tomar los valores $(+,-)$ con igual probabilidad, así que $H(X) = 1$ bit; con la segunda estrategia los valores de X son $(+,-,=)$, también equiprobables, de modo que $H(X) = \log_2 3 > 1$ bits. Es mejor la segunda.

2. (1 punto) Tiene Vd. 100 000 seguidores en Twitter. Cada vez que sube un nuevo tuit sus devotos seguidores lo retuitean o responden con probabilidad p, o bien lo ignoran con probabilidad $1-p$. Si sus seguidores son estadísticamente independientes, ¿cuántos bits necesita Twitter para registrar quiénes reaccionan activamente a sus mensajes?

Registrar el activismo de los seguidores equivale a codificar $n = 100\,000$ símbolos de una fuente discreta sin memoria con función de masas de probabilidad $X \sim \{p, 1-p\}$, para lo que se necesitan, en media, $nH_2(p) = 100\,000 H_2(p)$ bits.

3. (2 puntos) Para transmitir los mensajes de una fuente discreta X podemos utilizar uno de dos canales. El primero es un canal binario con un coste de uso de 1 euro por símbolo, y el segundo es un canal octal cuyo coste de uso es de 4 euros por símbolo.

a) ¿Con qué canal el coste mínimo de transmisión de un mensaje suficientemente largo de X será menor?

b) ¿Cuál será, en cada caso, el tiempo mínimo para transmitir $n \gg 1$ símbolos de la fuente?

a) El coste por bit (unidad de información) transmitida es 1 euro en el canal binario y 4/3 en el octal, que es peor. Visto de manera equivalente: transmitir $n \gg 1$ símbolos de la fuente por el canal binario requiere $nH_2(X)$ símbolos de canal y cuesta $nH_2(X)$ euros; transmitirlos por el segundo requiere $nH_8(X) = n/3H_2(X)$ símbolos de canal y cuesta $nH_2(X)4/3$ euros.

b) $nH_2(X)/v_c$ en el primer caso y $nH_8(X)/v_c = n/3H_2(X)/v_c$ en el segundo.

4. (2 puntos) Una enfermedad rara afecta a 1 de cada 10 000 personas. Un laboratorio biomédico ha desarrollado una prueba diagnóstica que resulta positiva en 99 de cada 100 enfermos y también en dos de cada 100 personas sanas (falsos positivos).

a) ¿Qué aporta más información sobre la afectación de esa enfermedad, que el resultado de la prueba sea positivo o negativo?

b) Aunque la prueba es cara, por razones de salud pública se ha decidido aplicarla dos veces a cada sujeto, con el siguiente protocolo:

- Si el resultado es idéntico en ambas pruebas, se declara al sujeto positivo o negativo.
- Si el resultado es distinto en las dos pruebas, se declara al sujeto positivo o negativo al azar (poco ético, pero esto es un problema académico).

¿Cuánta información aporta la prueba sobre la enfermedad?

a) Sea X el estado de salud de una persona y T el resultado del test. Como

$$H(X\,|\,T=+) = H\big(\mathbb{P}(X=\text{enfermo}\,|\,T=+)\big) = H\left(\frac{99}{99+2\times 9999}\right)$$
$$H(X\,|\,T=-) = H\big(\mathbb{P}(X=\text{sano}\,|\,T=-)\big) = H\left(\frac{98\times 9999}{1+98\times 9999}\right)$$

$H(X\,|\,T=-)$ está más próxima a cero y es mejor.

b) Sea T' el resultado de duplicar el test con el protocolo dado. La información que aporta la prueba sobre la enfermedad es ahora $I(X;T') = H(T') - H(T'\,|\,X)$ donde

$$H(T'\,|\,X) = \frac{1}{10\,000}H\Big(\underbrace{(99/100)^2 + \frac{1}{2}\times 2\times\frac{99}{100}\times\frac{1}{100}}_{p_1}\Big) + \frac{9999}{10\,000}H\Big(\underbrace{(2/100)^2 + \frac{1}{2}\times 2\times\frac{98}{100}\times\frac{2}{100}}_{p_2}\Big)$$

y

$$H(T') = H\Big(\underbrace{\frac{1}{10\,000}p_1 + \frac{9999}{10\,000}p_2}_{p_3}\Big).$$

Observe que

$$p_1 = \mathbb{P}(T'=+\,|\,X=\text{enfermo})$$
$$p_2 = \mathbb{P}(T'=+\,|\,X=\text{sano})$$
$$p_3 = \mathbb{P}(T'=+).$$

5. (2 puntos) La capacidad del canal con matriz de probabilidades de transición

$$\begin{bmatrix} 1-p & p & 0 \\ (1-p)/2 & p & (1-p)/2 \\ 0 & p & 1-p \end{bmatrix}$$

se obtiene cuando los símbolos 1 y 3 son equiprobables.

a) Calcule la capacidad.

b) Calcule $H(X\,|\,Y)$ si las entradas son equiprobables. Simplifique todo lo que pueda.

a) $H(Y\,|\,X) = H(p) + \overline{\alpha}q\log 2$, con $\overline{\alpha} = 1-\alpha$, $\alpha/2 = \mathbb{P}(X=1)$ y $q = 1-p$. Además

$$H(Y) = H(q/2, p, q/2) = H(p) + q\log 2$$

luego $I(X;Y) = H(Y) - H(Y\,|\,X) = \alpha q\log 2$, que es máxima si $\alpha = 1$. La capacidad es $C = q = 1-p$ bits, igual que la de un canal binario con borrado.

b) Para entradas equiprobables

$$H(X\,|\,Y) = qH(1/3) + p\log 3.$$

6. (1 punto) Se tiene una fuente discreta que emite símbolos con la siguiente distribución de probabilidades: $\mathbf{p} = (\frac{1}{3}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{27}, \frac{1}{27}, \frac{1}{27})$. Se desea utilizar un alfabeto de codificación ternario de símbolos $\{0,1,2\}$.

a) ¿Es factible obtener un código que tenga eficiencia unidad codificando los símbolos de uno en uno?

b) ¿Existe un código instantáneo con 1 palabras de longitud 1, 4 palabras de código de longitud 2 y 4 palabras de código con longitud 3? Demuestre su respuesta matemáticamente.

a) Sí: $p_i = 3^{-\ell_i}$ para $\ell_i = 1$ o $\ell_i = 2$ o $\ell_i = 3$. Esta fuente admite un código ternario óptimo.

b) Sí: como $3^{-1} + 4 \times 3^{-2} + 4 \times 3^{-3} = 25/27 < 1$, en virtud del teorema de Kraft es posible construir un código instantáneo con palabras de esas longitudes.

7. (1 punto) Se tiene una fuente discreta que emite símbolos del siguiente alfabeto $F = \{a, b, c, d, e, f, g, h, i\}$ y cuyo vector de probabilidades es el siguiente: $\mathbf{p} = \{0{,}3, 0{,}25, 0{,}1, 0{,}1, 0{,}08, 0{,}07, 0{,}06, 0{,}03, 0{,}01\}$. Se desea utilizar un alfabeto de codificación cuaternario de símbolos $\{0, 1, 2, 3\}$.

a) Aplique el algoritmo de Huffman para obtener las palabras de código para cada uno de los símbolos de fuentes.

b) ¿Cuál es la longitud media de las palabras de código obtenidas? ¿Cuál es la eficiencia del código?

Un código compacto es, por ejemplo, éste (asigne las etiquetas como quiera):

0,3	0,25	0,1	0,1	0,08	0,07	0,06	0,03	0,01

que tiene 3 palabras de longitud 1, 3 de longitud 2 y 3 de longitud 3, y cuya longitud es $L = 1{,}45$. La eficiencia es $\approx 93{,}3\,\%$.