Primera Tasca d'Avaluació Continuada

Javier Rozalén Sarmiento
Teoria de la Informació Clàssica i Quàntica
Universitat de Barcelona

05/11/2020

A Codificació

1. Quantes lletres té de mitjana una paraula en català?

S'ha considerat que una bona estimació de la longitud mitjana d'una paraula l'obtenim per inferència estadística, aproximant la probabilitat per la freqüència relativa. Així, s'ha escollit un llibre en català, Contes d'Andersen, s'ha comptabilitzat les paraules i les lletres, i s'ha obtingut:

nombre de paraules: 43186

nombre de lletres: 196502

Que resulta en una longitud mitjana de: $\bar{L} \approx 4,55$ lletres/paraula.

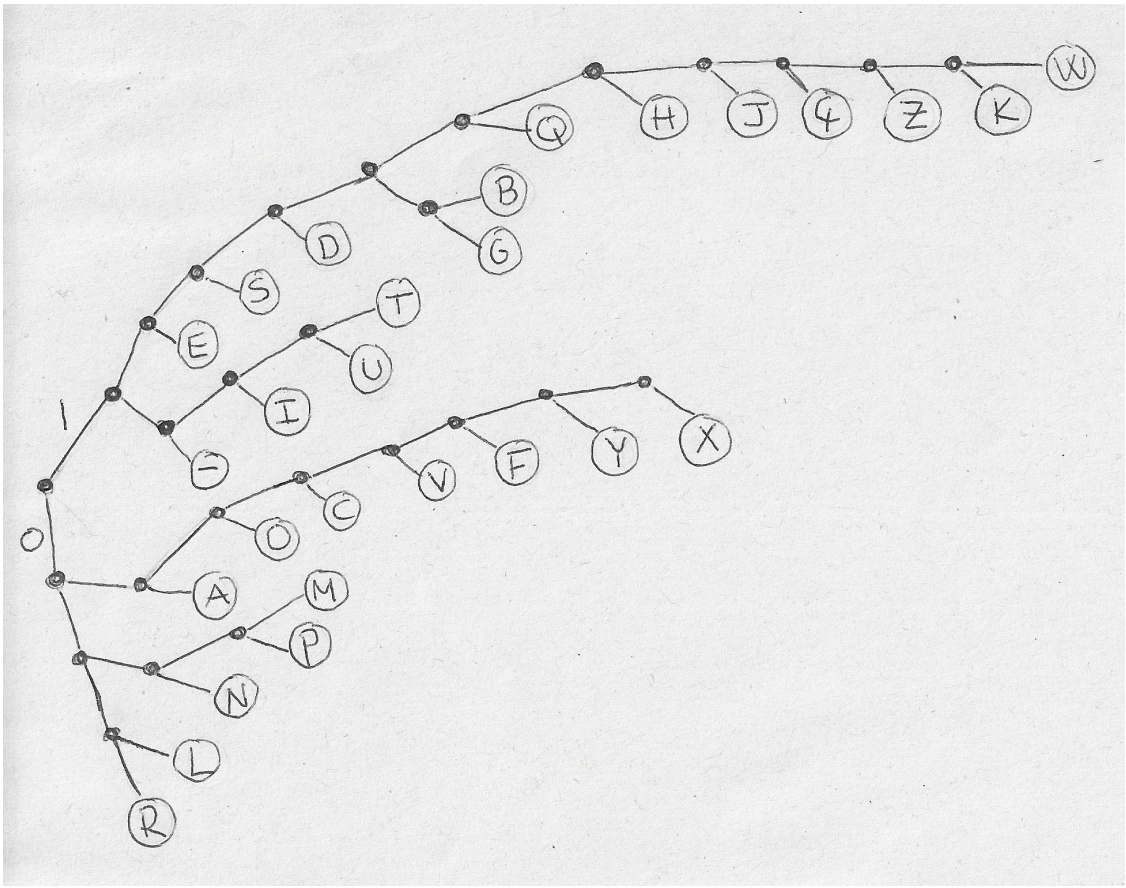
2. Doneu l'entropia de l'alfabet de la taula.

$$H(p) = -\sum_{\text{lletra}\in\text{alfabet}} p(\text{lletra}) \log_2(p(\text{lletra})) = 3,9222 \tag{1}$$

S'ha aproximat el resultat a 4 posicions decimals. El càlcul s'ha fet amb el programa Excel.

3. Determineu el codi binari associat a cada símbol de la taula segons l'algorisme de Huffman. Doneu l'estructura d'arbre que heu obtingut. Doneu la longitud mitjana en bits d'aquesta codificació. Indiqueu si aquesta longitud està dins l'interval que cabria esperar.

L'estructura d'arbre que s'ha obtingut és la següent:



A la figura s'indica un 0 i un 1 mostrant la convenció triada, a partir de la qual es poden construir els codis de tots els símbols (i comprovar-los amb la taula de més endevant). El símbol - denota l'espai en blanc. Notem que degut a les freqüències de l'alfabet hi ha certs nusos que donen lloc a diferents maneres de construir les successives ramificacions, de manera que l'estructura d'aquest arbre no és única, i aquest és una de les possibilitats.

Mostrem una taula amb els símbols de l'alfabet, la freqüència d'aparició, la codificació de Huffman i la longitud ponderada de cada símbol. També mostrem la codificació de Shannon i la longitud ponderada respectiva per cada símbol, però això s'utilitzarà al següent apartat.

Símbol	Prob.	Huffman	L ponderada	Shannon	L ponderada
(-)	0,17	100	0,51	000	0,51
E	0,13	110	0,39	001	0,39
A	0,12	010	0,36	0100	0,48
I	0,07	1010	0,28	0110	0,28
S	0,06	1110	0,24	01111	0,3
O	0,06	0110	0,24	10001	0,3
R	0,05	0000	0,2	10100	0,25
L	0,05	0001	0,2	10101	0,25
N	0,05	0010	0,2	10111	0,25
T	0,04	10111	0,2	11000	0,2
U	0,04	10110	0,2	11010	0,2
D	0,03	11110	0,15	110110	0,18
C	0,03	01110	0,15	111000	0,18
M	0,02	00111	0,1	111010	0,12
P	0,02	00110	0,1	111011	0,12
V	0,01	011110	0,06	1111000	0,07
Q	0,01	111110	0,06	1111010	0,07
B	0,009	111111	0,054	1111011	0,063
G	0,009	111110	0,054	1111100	0,063
F	0,007	0111110	0,049	11111010	0,056
H	0,005	11111110	0,04	11111100	0,04
X	0,003	01111111	0,024	111111011	0,027
Y	0,003	11111110	0,024	111111100	0,027
J	0,002	111111110	0,018	111111110	0,018
Ç	0,001	1111111110	0,01	1111111110	0,01
Z	0,0005	11111111110	0,0055	11111111110	0,0055
K	0,0003	111111111110	0,0036	111111111110	0,0036
W	0,0002	111111111111	0,0024	1111111111110	0,0026

Entenem, per 'L ponderada', $p(\text{símbol})L(\text{símbol})$. Sumant les longituds ponderades de cada símbol s'obté una longitud mitjana de: $\bar{L} \approx 3,9245$ dígits/símbol. Com cabria esperar, $H(p) < \bar{L}$. A més, donat que l'algorisme de Huffman genera codi eficient, veiem que també es compleix: $H(p) \leq \bar{L} \leq H(p) + 1$.

4. Determineu el codi binari associat a cada símbol de la taula segons l'algorisme de Shannon. Doneu la longitud mitjana en bits d'aquesta codificació. Indiqueu si aquesta longitud està dins l'interval que cabria esperar.

El codi binari de cada símbol segons l'algorisme de Shannon el trobem a la taula de la pregunta anterior. Pel que fa a la longitud mitjana s'ha obtingut: $\bar{L} \approx 4,4657$. Fent una comprovació anàloga a l'anterior veiem: $H(p) \leq \bar{L} \leq H(p) + 1$, de manera que sí és com cabria esperar, donat que l'algorisme de Shannon també proporciona codi eficient.

B Detecció/Correcció d'errors

Suposeu que utilitzem com a codi binari de l'alfabet de la taula, la posició del símbol en la taula expressada en 5 bits, es a dir (espai blanc) = 00000, E = 00001, A = 00010, ... Suposeu també que per enviar la informació utilitzeu un canal sorollós on el 0 (1) enviat es pot rebre com a 1 (0) amb probabilitat p.

1. Quina és la capacitat d'informació del canal en funció de p?

Tal i com hem vist a teoria, quan el canal és binari és útil fer la següent manipulació:

$$I(X,Y) = H(Y) - H(Y|X) = H(Y) - \sum_{i=0}^{1} P(X=i)H(Y|X=i) = H(Y) - \sum_{i=0}^{1} P(X=i)H(p)$$

Ara, donat que Y és una variable aleatòria binària tenim la restricció $H(Y) \leq 1$, i com que la capacitat d'informació maximitza la informació mútua $I(X,Y)$, tindrem:

$$C_I(p) = 1 - H(p) = 1 + (1-p)\log_2(1-p) + p\log_2(p)$$

2. Utilitzant 5 bits, quin és el nombre màxim de símbols distingibles que podeu enviar en funció de p? Quin rang de p ens permetria distingir els 28 símbols del nostre alfabet al ser enviats pel canal en qüestió?

El nombre màxim de missatges distingibles ve donat per la següent expressió:

$$M_{\text{màx}} = 2^{nC_I}$$

on n és el nombre de bits que té cada missatge/símbol. En el nostre cas, $n = 5$ i C_I és la capacitat d'informació del canal calculada a l'apartat anterior. Substituïnt:

$$M_{\text{màx}} = 2^{5[1+(1-p)\log_2(1-p)+p\log_2(p)]}$$

Si volem trobar el rang de p que permet distingir 28 símbols (és el cas del nostre alfabet) n'hi ha prou amb imposar $M_{\text{màx}} = 28$:

$$28 = 2^{5[1+(1-p)\log_2(1-p)+p\log_2(p)]} \implies \frac{\log_2(28)}{5} - 1 = (1-p)\log_2(1-p) + p\log_2(p)$$

Si es resol aquesta equació numèricament s'obtenen dues solucions degut a la simetria de la funció entropia. Naturalment, ens quedem amb la més propera a 0, que en aquest cas és: $p \approx 0,0041142$. Si el soroll és major ja no podrem distingir de manera segura 28 símbols[1], però si és menor, sí, de manera que el resultat que hem donat és la fita superior de p, i.e., $p \leq 0,0041142$. El resultat és raonable si considerem que aquest canal sense soroll accepta

[1]De fet, sí que podríem, per $p \geq 1 - 0,0041142$ (simetria), tot i que a tals nivells de soroll i sense saber realment quins bits han canviat, és impràctic.

32 missatges diferents, i estem exigint-ne 28, que és una fracció prou alta, de manera que n'hi ha prou amb poc soroll per evitar un grau tan alt de distingibilitat.

3. En funció de p, quina és la probabilitat de què a l'enviar un símbol de l'alfabet (5 bits) a través d'aquest canal, no el rebem correctament?

Sigui I l'esdeveniment *el missatge és incorrecte*, i sigui C l'esdeveniment *el missatge és correcte*. Si p és la probabilitat que es flipi un bit, $1-p$ és la probabilitat que es mantingui, de manera que $P(C) = (1-p)^5$. D'altra banda, donat que I∪C=Ω, on Ω denota l'espai mostral, podem escriure:

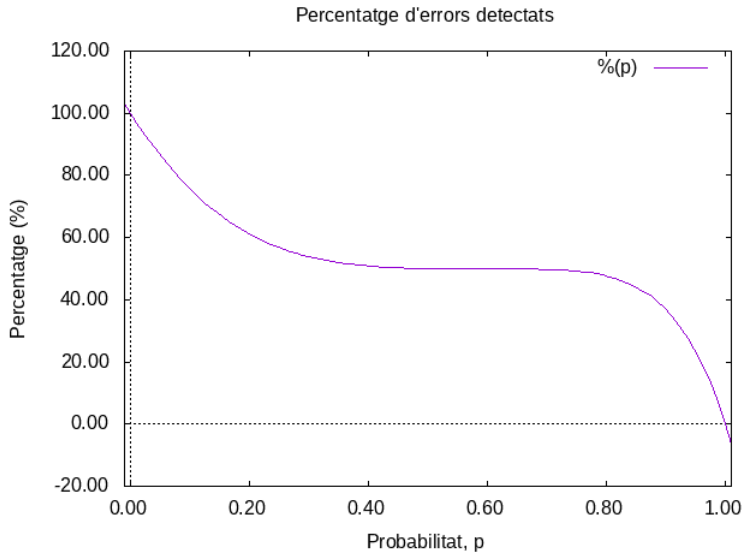
$$P(I) = 1 - P(C) = 1 - (1-p)^5$$

4. Si afegim un bit de paritat a l'enviar cada símbol (5 bits pel símbol + 1 bit de paritat), quin percentatge dels símbols que no es reben correctament poden ser detectats com erronis? Feu la gràfica d'aquest percentatge en funció de p.

Per fer aquest càlcul dividirem la probabilitat de detectar un missatge erroni entre la probabilitat de que el missatge sigui, en efecte, erroni. La segona probabilitat l'hem calculat a l'apartat 3, i és: $P(I) = 1 - (1-p)^5$. Pel càlcul del numerador considerem: si canvia un nombre parell de bits, independentment de quins, no detectarem l'error per paritat, de manera que excloem aquests casos del comptatge. D'altra banda, si canvia un nombre senar de bits sí ho detectarem, i això es pot donar de $\binom{6}{k}$ maneres, on k és el nombre de bits flipats. Hi ha una excepció a aquest raonament, i és que quan canvia només un bit hem d'excloure la possibilitat que es tracti del bit de paritat. Amb tot això s'obté:

$$\%(p) = 100 \cdot \frac{5p(1-p)^5 + 20p^3(1-p)^3 + 6p^5(1-p)}{1-(1-p)^5}$$

La gràfica de $\%(p)$:



5. Doneu els resultats numèrics dels dos apartats anteriors (B.3 i B.4) tant per a $p = 0,01$ com per a $p = 0,1$.

Considerant $p = 0,1$:

$$P(I) \approx 0,40985; \qquad \%(p) \approx 75,6707$$

Considerant $p = 0,01$:

$$P(I) \approx 0,0490; \qquad \%(p) \approx 97,0597$$

6. Per aquest canal sorollós que estem considerant, doneu una alternativa de codificació per l'alfabet de la taula i/o mètode de detecció/correcció d'errors, que millori la detecciò dels errors de transmissió. Per aquesta alternativa proposada, determineu quin percentatge dels símbols que no es reben correctament poden ser detectats com erronis. Feu la gràfica d'aquest percentatge en funció de p i compareu-lo amb els resultats de l'apartat de B.4.

Una opció un tant pedestre però que ja millora significativament l'eficiència és afegir 3 bits de paritat que analitzin exclusivament els 4 primers bits del missatge (Hamming(7,4)) i descuidar el 5è bit. Tenim:

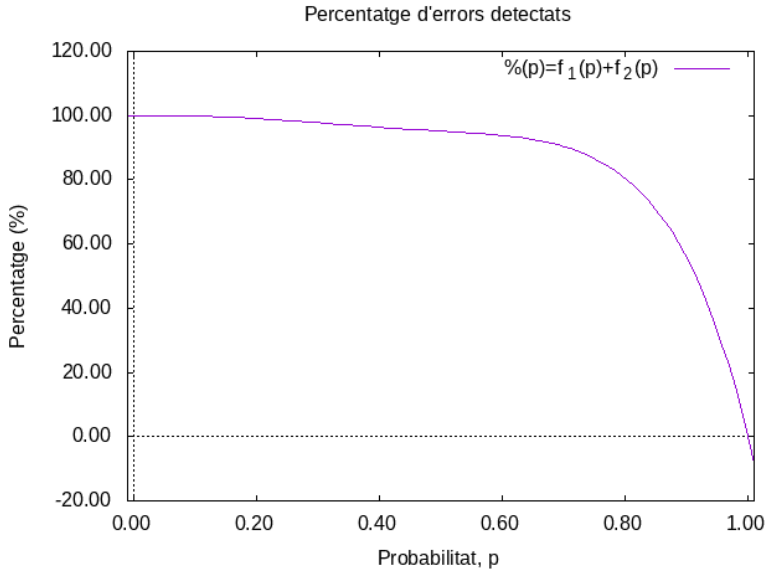
$$f_1(p) = 100 \cdot \frac{5p(1-p)^7 + 25p^2(1-p)^6 + 55p^3(1-p)^5 + 70p^4(1-p)^4}{1-(1-p)^5}$$

$$f_2(p) = 100 \cdot \frac{56p^5(1-p)^3 + 28p^6(1-p)^2 + 8p^7(1-p)}{1-(1-p)^5}$$

El percentatge és la suma d'aquestes dues funcions (ho escrivim així per qüestions de longitud):

$$\%(p) = f_1(p) + f_2(p)$$

Cal parar atenció a certs casos especials en el càlcul de la funció $\%(p)$ en què no es detectaria cap error o bé només canviarien bits de paritat, per això el numerador no és una binomial. La gràfica:



Només mirant la gràfica ja es veu que el percentatge es manté gairebé al seu valor màxim en tot el rang de p. Això ja és el que un s'espera mentre calcula el numerador de $\%(p)$, doncs realment és possible detectar quasi bé tots els errors (corregir-los ja no és el mateix). També observem que, contràriament a la funció de l'apartat B.4, la primera derivada és aproximadament nul·la a p petites, i això el fa bastant més interessant, doncs els valors típics de p no acostumen a allunyar-se del 0, de manera que no només tenim un millor percentatge de detecció en un rang més gran de p, sinó que el tenim, també, en el rang més típic de p. Numèricament:

$$\%(0,01) \approx 99,99982; \qquad \%(0,1) \approx 99,85541$$

C Simulacions

Les simulacions s'han fet amb el llenguatge de programació Python (versió 3.X).

1. Feu un programa gue generi símbols amb la freqüència donada en la taula i que els codifiqui en paraules-codi de longitud donada per l'algorisme de Huffman (A.3). Comproveu que la longitud mitjana és la predita en A.3

Per calcular la longitud mitjana s'ha generat 200000 símbols amb la freqüència corresponent a cada un i s'ha aproximat, com és costum, la probabilitat per la freqüència d'aparició. S'ha obtingut, en una determinada execució del codi, $\bar{L} = 3,922885$.

2. Feu un programa que generi símbols codificats en 5 bits + 1 bit de paritat i que simuli el canal sorollós de l'apartat B amb $p = 0,1$. Compareu els resultats de les simulacions amb els resultats corresponents obtinguts en B.5

Per una determinada execució del codi s'obté:

Per $p = 0,1$, $k = 500000$:

$$P(I) = 0,409186; \qquad \%(p) = 75,68$$

Per $p = 0,01$, $k = 500000$:

$$P(I) = 0,049118; \qquad \%(p) \approx 97,04$$

on k denota el nombre de símbols generats en l'execució. Són resultats molt propers als calculats teòricament a B.5.