

Seminar 0x01

Cristian Rusu

1 Scopul seminarului

În acest seminar vom rezolva niște probleme care implică:

- conceptul de informație și entropia lui Shannon;
- codarea cu dimensiune variabilă a datelor;
- detectarea și corectarea erorilor.

2 Exerciții

1. Avem un pachet de cărți de joc (52 de cărți). Luăm cărțile pentru prima dată afară din pachet. Câtă informație avem în acest moment despre cărți? Amestecăm cărțile aleator (aproso, cum facem asta, algoritmic, eficient?). Câtă informație avem acum în pachetul de cărți? (folosiți și aproximarea lui Stirling pentru calcularea rezultatului)
2. Se dă o urnă în care avem 5 bile roșii și 3 bile albastre. Ni se spune că cineva extrage o bilă din urnă și aceasta este albastră. Cerințe:
 - (a) câtă informație primim în urma acestei observații?
 - (b) care a fost entropia urnei înainte de extragere? care este entropia urnei după extragere?
 - (c) continuați să calculați entropia presupunând că extragem pas cu pas toate bilele albastre;
 - (d) similar cu cerința anterioară pentru bilele roșii (începând cu urna inițială presupunem că extrageți rând pe rând fiecare bilă roșie și calculați entropia la fiecare pas);
3. Se dau 12 bile. 11 dintre ele au aceeași greutate, iar una este mai ușoară sau mai grea decât toate celelalte. Aveți la dispoziție o balanță. Folosind un număr minim de cântăriri, găsiți bila diferită.
4. Aveți în Curs 0x01 un slide despre A. Turing. Calculați entropia textului (nivel de caracter/simbol, fără grupări) din slide: prima dată cu și apoi fără diacritice. Comparați cele două valori calculate.
5. Se dă un număr natural x pe N biți. Câtă informație am câștigat dacă:
 - (a) ni se spune despre x că are exact două valori "1" în reprezentarea sa binară;
 - (b) ni se spune despre x că are exact $N/2$ valori "1" în reprezentarea sa binară;
 - (c) ni se spune despre x că are o secvență continuă de $N/4$ biți de "1" în reprezentarea sa binară (restul biților sunt "0");
 - (d) ni se spune despre x că are MSB setat la "1";
 - (e) ni se spune despre x că este impar;
 - (f) ni se spune despre x că este o putere a lui 2;
 - (g) ni se spune despre x că care primii $N/2$ valori din reprezentarea sa binară sunt setați la "0";
 - (h) ni se spune despre x că este un număr prim (aici doar o estimare aproximativă este posibilă);
 - (i) ni se spune despre x că are în reprezentarea sa binară un număr par de biți setați la "1";
 - (j) ni se spune că $x = 42$.

- (a) calculați cifra de verificare p pentru codul ISBN-10: 0-306-40615 p ;
- (b) arătați că orice cod ISBN poate corecta orice eroare de o cifră;
- (c) arătați că orice cod ISBN poate detecta orice interschimbare a două cifre consecutive din cod;
- (d) care este distanța minimă pentru acest cod?
- (e) ce se întâmplă dacă schimbăm regula pentru cifra de verificare astfel încât $\sum_{i=1}^9 x_i + p = 0 \bmod 11$? Putem detecta/corecta erori la fel de bine?
- (f) verificați voi codul ISBN-13¹.
9. Considerăm utilizarea literelor în limba engleză. Conform Wikipedia² cele 26 de litere (“a” la “z”) apar cu frecvențele: 8.20%, 1.50%, 2.80%, 4.30%, 13%, 2.20%, 2%, 6.10%, 7%, 0.15%, 0.77%, 4%, 2.40%, 6.70%, 7.50%, 1.90%, 0.10%, 6%, 6.30%, 9.10%, 2.80%, 0.98%, 2.40%, 0.15%, 2% și 0.07%.
- Cerințe:
- (a) câți biți am folosi pentru a coda alfabetul dat folosind șiruri binare cu dimensiune fixă?
- (b) calculați entropia acestui alfabet;
- (c) care ar fi entropia dacă toate literele ar avea aceeași frecvență? Care este relația cu entropia calculată pe alfabetul dat?
- (d) aveți posibilitatea de a înlătura din alfabetul de mai sus o literă. Eliminați litera astfel încât să minimizați/maximizați entropia noului alfabet (puteți să scrieți un program care să verifice, dar trebuie să înțelegeți cum actualizați toate probabilitățile atunci când eliminați o literă);
- (e) în Anexa 1 puteți verifica codul Morse. Considerăm că un punct ocupă un bit iar o linie ocupă doi biți de spațiu în memorie. Cerințe:
- ce legătură observați între probabilitățile literelor și lungimea lor în cod Morse?
 - decodați mesajul: .- ... -.-. . - - —
 - mesajele codate în codul Morse se pot decoda unic? Puteți să decodați mesajul de mai sus dacă nu aveți spațiile după fiecare literă?
 - calculați lungimea medie pentru codarea alfabetului de mai sus în codul Morse, comparați cu entropia calculată anterior.
10. Avem un alfabet cu două simboluri: simbolul “A” este codat pe un singur bit și are probabilitatea de apariție p și simbolul “B” este codat pe doi biți și are probabilitatea de apariție $1-p$ (bineînțeles).
- Cerințe:
- (a) calculați lungimea medie a mesajelor cu alfabetul de mai sus;
- (b) definim rata entropiei ca fiind entropia supra lungimea medie a mesajelor. Găsiți p pentru care rata entropiei este minimizată/maximizată.
11. Considerăm că avem un alfabet cu N simboluri. Am discutat la curs despre faptul că entropia este maximizată dacă avem “incertitudine maximă” deci avem un alfabet în care simbolurile au aceeași probabilitate de apariție (adică $p_i = 1/N$ pentru $i = 1, \dots, N$ – aceasta se numește distribuția uniformă). Demonstrați matematic această afirmație.
12. Avem un alfabet cu N simboluri x_i , fiecare cu probabilitatea de apariție p_i . Considerăm că simbolurile sunt sortate în ordinea descrescătoare a probabilităților de apariție: adică $p_1 \geq p_2 \geq \dots \geq p_N$. Codarea acestor simboluri se realizează în felul următor: simbolul x_i este codat cu șirul binar $00 \dots 001$ (de $i-1$ ori bitul 0, apoi bitul 1). Sunt relevante valorile p_i în acest caz? Care este lungimea medie a unui mesaj codat în acest fel? Comparați această valoare cu entropia. Care este avantajul principal al acestei codări?

¹https://en.wikipedia.org/wiki/International_Standard_Book_Number

²https://en.wikipedia.org/wiki/Letter_frequency

Anexa 1

A	• —	U	• • —
B	— • • •	V	• • • —
C	— • — •	W	• — —
D	— • •	X	— • • —
E	•	Y	— • — —
F	• • — •	Z	— — • •
G	— — •		
H	• • • •		
I	• •		
J	• — — —		
K	— • —	1	• — — — —
L	• — • •	2	• • — — —
M	— —	3	• • • — —
N	— •	4	• • • • —
O	— — —	5	• • • • •
P	• — — •	6	— • • • •
Q	— — • —	7	— — • • •
R	• — •	8	— — — • •
S	• • •	9	— — — — •
T	—	0	— — — — —

Codul Morse (sursa: wikipedia).