Curs 5-1: Entropia Shannon

Conf.dr. Maria Jivulescu

Departamentul de Matematică
UPT

Recapitulare: Variabile aleatoare discrete

O varb. aleatoare este descrisă prin tabelul de distribuție

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}, \quad \sum_{i=1}^{n} p_i = 1$$

unde: x_i-valorile lui X, iar , $p_i = P(X = x_i), 0 \leq p_i \leq 1, \sum_{i=1}^{n} p_i = 1$

sau p.m.f.(probability mass function)

$$p(x) = \begin{cases} p_1, & \text{daca } x = x_1 \\ p_2, & \text{daca } x = x_2 \\ \ldots & \\ p_n, & \text{daca } x = x_n \\ 0, & \text{in rest} \end{cases}$$

Informatia Shannon a unui eveniment

Ne intereseaza să cuantificam câta informație se găsește într-un mesaj.
Putem calcula cantitatea de informatie care se afla intr-un eveniment folosind probabilitatea acestui eveniment.
Unitatea de masurare a informatiei este **bit**=binary information unit.
Se defineste **informatia Shannon** (information content) a evenim. A:

$$I(A) = \log_2 \frac{1}{p(A)} = -\log_2 p(A)$$

- Daca $p(A) = 1$, atunci informatia $I(A) = 0$, pentru ca nu exista nicio surpriza=incertitudine (evenimentul va avea loc cu certitudine).
- Fie cazul aruncarii unei monede corecte (p(Cap)=p(Pajura)=1/2). $I(Cap) = -\log_2(1/2) = 1$ bit.
- Fie cazul aruncarii unei monede false, unde $p(Cap) = 0.1$. Atunci $I(Cap) \approx 3.322$, deci este nevoie de 3.32 biti pentru a transmite informatia legata de evenimentul Cap(aceasta are loc mai rar, in comparatie cu cazul monedei corecte).

Informatia Shannon a unui eveniment

- Aruncarea unui zar cu 6 fete.
  Probabilitatea de a obtine una din fețele zarului este $1/6 < 1/2$ (in comparatie cu aruncarea monedei si obtinerea fetei cu Cap). Deci ne asteptam sa avem nevoie de mai multi biti pentru a putea trimite acest eveniment, avand loc mai multa incertitudine(surpriza).
  Intr-adevar, $I(A) = 2.5$ biti $> I(Cap) = 1$ bit.
- Extragerea unei carti dintr-un teanc de 52 de cari de joc.
  - Extragerea oricarei carti are prob $1/52$, deci $I(A) = \log_2 52 \approx 5.7$ biti
  - Daca ai informatia ca s-a extras o carte cu forma "inima", de cati biti ai nevoie acum?
    $I(A) = \log_2(52/13) = 2$biti.
  - Daca ai informatia ca s-a extras o carte cu numarul 7, de cati biti ai nevoie acum?
    $I(A) = \log_2(52/4) \approx 3.7$ biti

Information content

Continutul informational =masura a incertitudinii (gradul de supriza) unui eveniment Observatii generale:

- evenimente cu probabilitate mica au mai mult continut informational;
- evenimentele cu probabilitate mare, aduc informatie mai puțina.

In general, observam ca relatia dintre probabilitatea unui eveniment si informatia continuta de eveniment nu este liniara si, de fapt, evenimentele cu probabilitate mica, aduc mai multa surprize si "cară" mai multa informatie!

Informatia Shannon a unei variabile aleatoare

Vom calcula informatia continuta intr-o v.a. X, $p = (p_i)_i$. Aceasta informatie se numeste *entropia informatiei (entropia Shannon)*.
Entropia Shannon cuantifica informatia medie continuta in mesaj.
A fost introdusa de catre Claude E. Shannon, in 1948, in lucrarea "A mathematical Theory of Communication" prin formula

$$H_2(X) \equiv H(p_1, \ldots, p_n) := \sum_i p_i \log_2 \frac{1}{p_i} = -\sum_i p_i \log_2 p_i \qquad (1)$$

unde $p_i = P(X = x_i)$, iar $\frac{1}{p_i}$–"surpriza=incertitudinea" ca variabila X ia valoarea x_i (Prin conventie, $0 \log 0 \equiv 0$).
Obs: Entropia este o masura a incertitudinii legate de rezultatele unui experiment aleator, descris printr-o variabila aleatoare. Cu cat o variabila aleatoare este mai apropiata de distributia uniforma , cu atat incertitudinea legata de valorile variabilei este mai mare.
Pentru cazul $p_i = 1/n$, avem maxim: $H(X) = \log_2 n$.

Entropia Shannon.

$H(X)$ reprezinta numarul mediu de biti necesari pentru a reprezenta un eveniment asociat unei distributii de probabilitate.
Observatii:

- cea mai mica entropie este asociata cu o v.a. ce are o singura valoare, pe care o ia cu probabilitate 1 (X este determinista). Avem $H(X) = 0$.
- cea mai mare entropie este asociata unei v.a. unde valorile variabilei au aceeasi probabilitate de a fi luate (distributia uniforma).
- cazul aruncarii unui zar; atunci, $H(X) = \frac{1}{6} \sum \log_2(6) \approx 2.58$.
- in general, daca $X \sim Unif(m)$, atunci $H(X) = \log_2 m$.

Proprietati matematice ale entropiei Shanon

- $H(X) \geq 0$;
- $H(X)$-functie concava: $H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)$;
- $H(X)$ este invarianta la permutari ale realizarilor unei variabile aleatoare (depinde de probabilitati, nu de valoarea realizarilor variabilei X);
- $H(X) = 0 \Leftrightarrow X$ este o variabila determinista: $D_X = \{a\}, p(X = a) = 1$;
- $H(X) \leq \log |D_X|$ (cu egalitate in cazul distrb. uniforme).

Entropia binara

Fie o variabila aleatoare $X \sim Bernoulli(p)$, i.e. $X = \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix}$

Entropia Shannon

$$H(X) = -p \log_2(p) - (1-p) \log_2(1-p)$$

Semnificatia entropiei

Entropia (in biti) ne spune cantitatea medie de informatie (in biti) ce trebuie furnizata pentru a rezolva incertitudinea legata de rezultatul unei experiente.
$H(X)$ este marginea inferioara a numarului de biti ce trebuie, in medie, folositi pentru a codifica un mesaj.

- daca trimitem mai putini biti in medie, atunci receptorul va avea invertitudine legata de mesaj;
- daca trimitem mai multi biti in medie, se va pierde din capacitatea canalului de comunicatie, prin trimiterea de biti de care nu avem nevoie;
- atingerea limitei inferioare (a entropiei) este un tel pentru o codare eficienta (cel putin din perspectiva compresiei informatiei).

Teorema Shannon

Teorema Shannon: exista o schema de compresare a unui mesaj astfel incat informatia produsa de sursa poate fi stocata folosind $H(X)$ biti pe fiecare simbol de sursa.
Deci, daca informatia este comprimata mai mult de valoarea data de $H(X)$, exista, cu probabilitate mare, riscul de eroare la momentul recompunerii mesajului original.

Aplicatii in Data compression

Exemplu: se considera o v.a. X ce ia valorile din alfabetul cu 4 litere $\{A, B, C, D\}$ cu probabilitatile $\{1/2, 1/4, 1/8, 1/8\}$.

- Fara data compression, vom coda fiecare din cele 4 simboluri cu 2 biti; de exemplu A=00,B=01, C=10, D=11. Presupunem ca fiecare simbol are aceeasi sansa de aparitie $(1/4)$. Atunci, $H(X) = \sum \frac{1}{4} \log 4 = 2$, deci vom avea nevoie de 2 biti pentru un simbol.
- Daca analizam densitatea de probabilitate, atunci o strategie de data compression ar fi sa folosim coduri scurte pentru simboluri ce apar cu probabilitate mai mare si coduri mai lungi pentru simboluri ce apar cu probabilitate mai mica; de exemplu, A=0,B=10, C=110, D=111. Lungimea medie a unui cod este: $\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4}$
- Calculand $H(X)$ avem ca $H(X) = \frac{7}{4}$.
  Deci, sunt necesar $7/4 \approx 1.75$ biti pentru a transmite un simbol.

Daca (X, Y) vector aleator, atunci $H(X, Y)$ este entropia Shannon asociata vectorului aleator discret (X, Y) si se defineste:

$$H(X, Y) := -\sum_{x,y} p(x, y) \log_2 p(x, y)$$

Au loc inegalitatile Shannon:

$$H(x), H(y) \leq H(X, Y) \leq H(X) + H(Y)$$

Informatia mutuala

$$I(X : Y) := H(X) + H(Y) - H(X, Y)$$

$I(X : Y)$ masoara corelația (legatura) dintre X și Y;
$I(X : Y) = 0 \Leftrightarrow X, Y$ independente.

Entropia conditionata.

Dată variabila aleatoare conditionata $(X|Y = y)$, entropia acestei variabile este

$$H(X|Y = y) := -\sum_x p_{X|Y}(x|y) \log_2 p_{X|Y}(x|y)$$

Definim entropia conditionata

$$H(X|Y) = \sum_y p_Y(y) H(X|Y = y) = -\sum_{x,y} p_{(X,Y)} \log(p_{X|Y}(x|y))$$

Avem:

$$H(X|Y) = H(X, Y) - H(Y),\ H(Y|X) = H(X, Y) - H(X)$$

Deci, $H(X) \geq H(X|Y)$, cu egalitate dacă X și Y sunt variabile aleatoare independente.

Idee: Conditionarea nu creste entropia unei variabile aleatoare.

Exemplu

Se considera o sursă care transmite un string periodic de biti:

$$\ldots 01001001001 \ldots$$

Sa se calculeze entropia sursei, tinand cont de bitii care au fost receptati anterior.
Vom analiza urmatoarele cazuri:

- nu se cunoste niciun simbol al alfabetului; avem: $P_X(x = 0) = \frac{2}{3}$, $P_X(x = 1) = \frac{1}{3}$, deci $H(X) \approx 0.918$ biti
- un simbol este cunoscut; in acest caz, fie Y informatia data de cunosterea simbolului; vom calcula $H(X|Y) = \sum_y p_Y(y) H(X|Y = y)$, unde $p_Y(y)$ este probabilitatea ca simbolul cunoscut sa se afle in sir.

Exemplu

Vom descrie mai jos $p_{X|Y}(x|y)$, folosind observația că, dacă simbolul anterior este 1, atunci urmatorul este 0, de vreme ce nu exista 11 in sir.

$p_{X\|Y}(x\|y)$	$x=0$	$x=1$
$y=0$	$1/2$	$1/2$
$y=1$	1	0

Entropia condiționata are valoarea:
$H(X|Y) = -\frac{2}{3}H(X|Y) - \frac{1}{3}H(X|Y=1) = \frac{2}{3}$ biti. Observam ca daca se primeste informatie de la sursa, entropia descreste. Altfel spus, primirea unui bit de informatie a scazut "incertitudinea" sursei X.

Exemplu

Cazul in care doi biti de informatie sunt cunoscuti.
In acest caz, ar trebui ce simbol urmeaza dupa cei doi. Se va observa acest lucru in entropia conditionata. Distributia probabilitatii conditionate este:

$p_{X\|Y}(x\|y)$	$x=0$	$x=1$
$y=00$	0	1
$y=01$	1	0
$y=10$	1	0
$y=11$	-	-

Avem $H(X|Y=y)=0, \forall y$, deci nu a mai ramas nicio incertitudine.

Informatia mutuala conditionata.

Pentru trei variabile aleatoare X, Y, Z informatia mutuala conditionata se defineste ca fiind:

$$I(X : Y|Z) := H(X|Z) + H(Y|Z) - H(X, Y|Z)$$

Avem: $I(X : Y|Z) = 0 \Leftrightarrow X$ și Y sunt conditional independente relativ la Z.

Entropia relativa

Entropia relativa (divergenta Kullback-Leibler) masoara cat de diferita este distributia $p(x)$ de distributia de probabilitate $q(x)$.
Definitie: Fie $\mathcal{X}$ o multime finita si $p(x)$ și $q(x)$ două distributii de probabilitate pe $\mathcal{X}$. Entropia relativa este:

$$D(p(x)||q(x)) := \sum_x p(x) \log(\frac{p(x)}{q(x)}) = -H(X) - \sum_x p(x) \log q(x)$$

Convenim $-p(x) \log 0 = +\infty$, dacă $p(x) > 0$.
Proprietati:

- $D(p(x)||q(x)) \geq 0$ (egalitate iff $p(x) = q(x), \forall x \in \mathcal{X}$).
- In consecinta, $H(X) \leq \log |\mathcal{X}|$ (egalitate daca X este distribuita uniform.)
- $I(X : Y) = D(p_{(X,Y)}(x, y)||p_X(x)p_Y(y))$