Elemente de teoria informaţiei[19]

29. **(Exerciţiu cu caracter teoretic: proprietăţi dezirabile ale entropiei)**

■□ • ***CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2.2***

Prin definiţie, entropia (în sens Shannon) a unei variabile aleatoare discrete X ale cărei valori sunt luate cu probabilităţile $p_1, p_2, \ldots, p_n$ este $H(X) = -\sum_i p_i \log p_i$. Însă legătura dintre această definiţie formală şi obiectivul avut în vedere — şi anume, acela de a exprima gradul de *incertitudine* cu care se produc valorile unei astfel de variabile aleatoare — nu este foarte intuitivă.

Scopul acestui exerciţiu este de a arăta că orice funcţie $\psi_n(p_1, \ldots, p_n)$ care satisface trei proprietăţi dezirabile pentru entropie este în mod necesar de forma $-K \sum_i p_i \log p_i$ unde K este o constantă reală pozitivă. Iată care sunt aceste *proprietăţi*:

A1. **Funcţia $\psi_n(p_1, \ldots, p_n)$ este continuă în fiecare din argumentele ei şi simetrică.**
Din punct de vedere formal, în acest caz simetria se traduce prin egalitatea $\psi_n(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_n) = \psi_n(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_n)$ pentru orice $i \neq j$. Informal spus, dacă două dintre valorile care sunt luate de variabila aleatoare X (şi anume x_i şi x_j) îşi schimbă între ele probabilităţile (p_i şi respectiv p_j), valoarea entropiei lui X nu se schimbă.

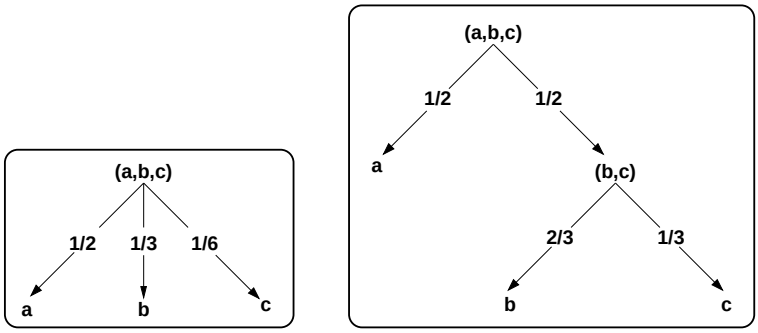
A2. **Funcţia $\psi_n(1/n, \ldots, 1/n)$ este monoton crescătoare în raport cu n.**
Altfel spus, dacă toate evenimentele sunt echiprobabile, atunci entropia creşte odată cu numărul de evenimente posibile.

A3. **Dacă faptul de a alege între mai multe evenimente posibile poate fi realizat prin mai multe alegeri succesive, atunci $\psi_n(p_1, \ldots, p_n)$ trebuie să se poată scrie/calcula ca o sumă ponderată.**
De *exemplu*, dacă evenimentele (a, b, c) se produc respectiv cu probabilităţile $(1/2, 1/3, 1/6)$, atunci acest fapt poate fi echivalat cu

- **a alege mai întâi cu probabilitate de $1/2$ între a şi (b, c),**
- **urmat de a alege între b şi c cu probabilităţile $2/3$ şi $1/3$ respectiv.**
  (A se vedea imaginile de mai jos, Encoding 1 şi Encoding 2.)

Din punct de vedere formal, proprietatea A3 impune ca, pe acest exemplu, $\psi_3(1/2, 1/3, 1/6)$ să fie egal cu $\psi_2(1/2, 1/2) + 1/2 \cdot \psi_2(2/3, 1/3)$.

[19] Observaţie importantă: În toate problemele care urmează, referitor la entropie / teoria informaţiei se va considera în mod implicit că notaţia 'log' desemnează logaritmul în baza 2. De asemenea, prin convenţie, se va considera $p \cdot \log p = 0$ pentru $p = 0$.

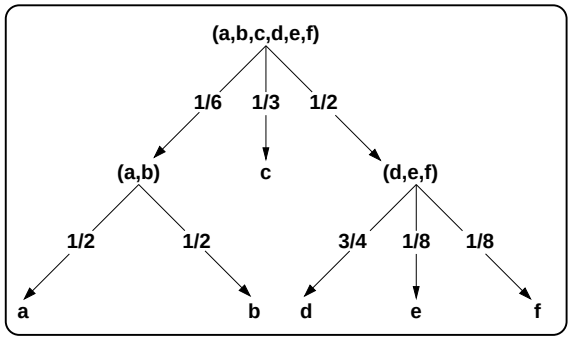


Aşadar, în acest exerciţiu vi se cere să arătaţi că dacă o funcţie de n variabile $\psi_n(p_1, \ldots, p_n)$ satisface proprietăţile A1, A2 şi A3 de mai sus, atunci există $K \in \mathbb{R}^+$ astfel încât $\psi_n(p_1, \ldots, p_n) = -K \sum_i p_i \log p_i$ unde K depinde de ψ.

***Indicaţie*:**
Veţi face rezolvarea acestei probleme în mod gradual, parcurgând următoarele puncte (dintre care primele două puncte au rolul de a vă acomoda cu noţiunile din enunţ):

a. Arătaţi că $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + \frac{1}{2} H(2/3, 1/3)$. Altfel spus, verificaţi faptul că definiţia clasică a entropiei, $H(X) = \sum_i p_i \log 1/p_i$, satisface proprietatea A3 pe *exemplul* care a fost dat mai sus.

b. Calculaţi entropia în cazul distribuţiei/„codificării" din figura de mai jos, folosind din nou proprietatea A3.



Următoarele întrebări tratează cazul particular $A(n) \overset{not.}{=} \psi(1/n, 1/n, \ldots, 1/n)$.

c. Arătaţi că

$$A(s^m) = m\, A(s) \text{ pentru orice } s, m \in \mathbb{N}^*. \tag{1}$$

Mai departe, pentru $s, m \in \mathbb{N}^*$ fixaţi, vom considera $t, n \in \mathbb{N}^*$ astfel încât $s^m \leq t^n \leq s^{m+1}$.

d. Verificaţi că, prin logaritmarea acestei duble inegalităţi şi apoi prin rearanjare, obţinem $\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n}$ pentru $s \neq 1$, şi deci

$$\mid \frac{m}{n} - \frac{\log t}{\log s} \mid \leq \frac{1}{n}. \qquad (2)$$

e. Explicaţi de ce $A(s^m) \leq A(t^n) \leq A(s^{m+1})$.

f. Combinând ultima inegalitate de mai sus cu inegalitatea (1), avem $A(s^m) \leq A(t^n) \leq A(s^{m+1}) \Rightarrow m\,A(s) \leq n\,A(t) \leq (m+1)\,A(s)$. **Verificaţi că**

$$\mid \frac{m}{n} - \frac{A(t)}{A(s)} \mid \leq \frac{1}{n} \quad \textbf{pentru} \quad s \neq 1. \qquad (3)$$

g. Combinând inegalităţile (2) şi (3), arătaţi că

$$\mid \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \mid \leq \frac{2}{n} \quad \textbf{pentru} \quad s \neq 1 \qquad (4)$$

şi, în consecinţă

$$A(t) = K \log t \quad \textbf{cu } K > 0 \textbf{ (din cauza proprietăţii A2).} \qquad (5)$$

***Observaţie*:**

Rezultatul de mai sus ($A(t) = K \log t \Leftrightarrow \psi_t(1/t, \ldots, 1/t) = Kt\frac{1}{t}\log t$) **se generalizează uşor la cazul** $\psi(p_1, \ldots, p_n)$ **cu** $p_i \in \mathbb{Q}$ **pentru** $i = 1, \ldots, n$ **(cazul** $p_i \notin \mathbb{Q}$ **nu este tratat aici):**

Considerăm o mulţime de N **evenimente echiprobabile. Fie** $\mathcal{P} = (S_1, S_2, \ldots, S_k)$ **o partiţionare a acestei mulţimi de evenimente. Notăm** $p_i = \mid S_i \mid /N$.

Propunem următoarea *codificare*: Vom alege mai întâi S_i, **una din submulţimile din partiţia** $\mathcal{P}$, **în funcţie de probabilităţile** $p_1, \ldots, p_k$. **Extragem apoi unul din elementele mulţimii** S_i, **cu probabilitate uniformă.**

Conform egalităţii (5), avem $A(N) = K\,logN$. **Folosind proprietatea A3 şi *codificarea* în doi paşi propusă mai sus, rezultă că**

$A(N) = \psi_k(p_1, \ldots, p_k) + \sum_i p_i A(\mid S_i \mid)$.

Aşadar,

$K \log N = \psi_k(p_1, \ldots, p_k) + K \sum_i p_i \log \mid S_i \mid$.

Prin urmare,

$$\begin{aligned} \psi_k(p_1, \ldots, p_k) &= K[\, \log N - \sum_i p_i \log \mid S_i \mid] \\ &= K[\, \log N \sum_i p_i - \sum_i p_i \log \mid S_i \mid] = -K \sum_i p_i \log \frac{\mid S_i \mid}{N} \\ &= -K \sum_i p_i \log p_i \end{aligned}$$

Răspuns:

a. Facem calculele:

$$\begin{aligned} H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) &= \frac{1}{2}\log 2 + \frac{1}{3}\log 3 + \frac{1}{6}\log 6 = \left(\frac{1}{2} + \frac{1}{6}\right)\log 2 + \left(\frac{1}{3} + \frac{1}{6}\right)\log 3 \\ &= \frac{2}{3} + \frac{1}{2}\log 3 \end{aligned}$$

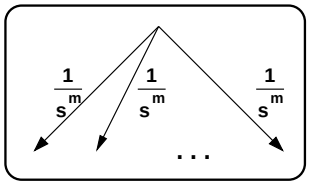
$$H\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{2}H\left(\frac{2}{3},\frac{1}{3}\right) = 1+\frac{1}{2}\left(\frac{2}{3}\log\frac{3}{2}+\frac{1}{3}\log 3\right)=1+\frac{1}{2}\left(\log 3-\frac{2}{3}\right)$$
$$= \frac{2}{3}+\frac{1}{2}\log 3$$

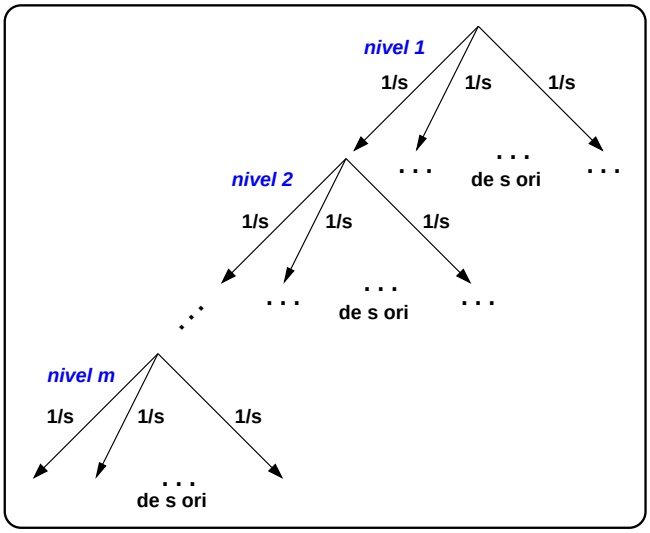
şi rezultă că $H\left(\frac{1}{2},\frac{1}{3},\frac{1}{6}\right)=H\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{2}H\left(\frac{2}{3},\frac{1}{3}\right)$.

b. Folosind proprietatea A3, entropia „codificării“ din figura dată în enunţ este:

$$\begin{aligned} H\left(\frac{1}{6},\frac{1}{3},\frac{1}{2}\right) &+ \frac{1}{6}H\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{2}H\left(\frac{3}{4},\frac{1}{8},\frac{1}{8}\right)\\ &= \frac{1}{6}\log 6+\frac{1}{3}\log 3+\frac{1}{2}\log 2+\frac{1}{6}+\frac{1}{2}\left(\frac{3}{4}\log\frac{4}{3}+\frac{2}{8}\log 8\right)\\ &= \frac{1}{6}\log 2+\frac{1}{6}\log 3+\frac{1}{3}\log 3+\frac{1}{2}+\frac{1}{6}+\frac{3}{8}(2-\log 3)+\frac{3}{8}\\ &= \frac{1}{6}+\frac{1}{2}+\frac{1}{6}+\frac{3}{4}+\frac{3}{8}+\left(\frac{1}{6}+\frac{1}{3}-\frac{3}{8}\right)\log 3\\ &= \frac{47}{24}+\frac{1}{8}\log 3 = 1.958+0.125\log 3 = 2.156 \end{aligned}$$

c. Pentru calculul lui $A(s^m)$ **se poate folosi atât o „codificare“ imediată cât şi una (des)compusă, ca în figura următoare:**





Aplicând proprietatea A3 pe „codificarea“ din figura de mai sus, partea dreaptă, avem:

$$\begin{aligned} A(s^m) &= A(s)+s\cdot\frac{1}{s}A(s)+s^2\cdot\frac{1}{s^2}A(s)+\ldots+s^{m-1}\cdot\frac{1}{s^{m-1}}A(s)\\ &= \underbrace{A(s)+A(s)+A(s)+\ldots+A(s)}_{\text{de } m \text{ ori}} = mA(s) \end{aligned}$$

d. Aplicând funcţia $\log$ **fiecărui termen al inegalităţii** $s^m\le t^n\le s^{m+1}$ **obţinem** $m\log s\le n\log t\le (m+1)\log s$. **Apoi, pentru** $s\neq 1$, **împărţind prin** $n\log s$, **rezultă:**

$$\frac{m}{n} \le \frac{\log t}{\log s} \le \frac{m}{n} + \frac{1}{n} \Rightarrow 0 \le \frac{\log t}{\log s} - \frac{m}{n} \le \frac{1}{n} \Rightarrow \left| \frac{\log t}{\log s} - \frac{m}{n} \right| \le \frac{1}{n}$$

e. Datorită proprietăţii A2 din enunţ, inegalitatea $s^m \le t^n \le s^{m+1}$ implică

$$\psi_{s^m}\left(\frac{1}{s^m}, \ldots, \frac{1}{s^m}\right) \le \psi_{t^n}\left(\frac{1}{t^n}, \ldots, \frac{1}{t^n}\right) \le \psi_{s^{m+1}}\left(\frac{1}{s^{m+1}}, \ldots, \frac{1}{s^{m+1}}\right)$$

ceea ce reprezintă exact $A(s^m) \le A(t^n) \le A(s^{m+1})$.

f. Datorită proprietăţii (1), dubla inegalitate $A(s^m) \le A(t^n) \le A(s^{m+1})$ devine $m\,A(s) \le n\,A(t) \le (m+1)\,A(s)$. Împărţind această inegalitate prin $nA(s)$, despre care se poate spune că este nenul pentru orice $s \neq 1$, rezultă:

$$\frac{m}{n} \le \frac{A(t)}{A(s)} \le \frac{m}{n} + \frac{1}{n} \Rightarrow 0 \le \frac{A(t)}{A(s)} - \frac{m}{n} \le \frac{1}{n} \Rightarrow \left| \frac{A(t)}{A(s)} - \frac{m}{n} \right| \le \frac{1}{n}$$

g. Inegalităţile duble de mai jos rescriu convenabil proprietăţile (2) şi (3):

$$-\frac{1}{n} \le \frac{m}{n} - \frac{\log t}{\log s} \le \frac{1}{n} \quad \text{şi} \quad -\frac{1}{n} \le \frac{A(t)}{A(s)} - \frac{m}{n} \le \frac{1}{n}$$

Însumându-le membru cu membru, rezultă

$$-\frac{2}{n} \le \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \le \frac{2}{n} \Rightarrow \left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \le \frac{2}{n}$$

Din inegalitatea dublă $s^m \le t^n \le s^{m+1}$ rezultă că odată cu m creşte şi n (acesta din urmă depinzând de valorile lui s, t şi m).[20] Aşadar, atunci când m tinde la infinit vom avea şi $n \to \infty$. Dacă trecem la limită inegalitatea $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \le \frac{2}{n}$ pentru $n \to \infty$, rezultă $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \to 0$, de unde avem $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| = 0$ şi deci $\frac{A(t)}{A(s)} = \frac{\log t}{\log s}$. Rezultă că $A(t) = \frac{A(s)}{\log s} \log t = K \log t$. Evident, constanta $K = \frac{A(s)}{\log s}$ nu depinde de t. Variind valorile lui t, rezultă că $A(t) = K \log t$ pentru orice $t \in \mathbb{N}^*$.

O *observaţie* finală: din inegalitatea $s^m \le t^n \le s^{m+1}$, dacă $s \neq 1$ va rezulta că de fapt şi $t \neq 1$. Atunci când pentru A se foloseşte formula clasică a entropiei, egalitatea $A(t) = K \log t$ se verifică şi în cazul $t = 1$, fiindcă $A(1) = \sum_p p \log p = 1 \log 1 = 0$, iar $K \log 1 = 0$.

30. **(Entropie, entropie corelată, entropie condiţională, câştig de informaţie: definiţii şi proprietăţi imediate)**

□ ***L. Ciortuz, pornind de la CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2***

Fie X şi Y variabile aleatoare discrete. Dăm pe scurt următoarele *definiţii*:

[20] Mai precis, este util să observaţi o succesiune de forma $s^m \le t^{n_1} \le s^{m+1} \le t^{n_2} \le s^{m+2} \le \ldots$.

- **Entropia variabilei X:**
  $H(X) \stackrel{def.}{=} -\sum_i P(X = x_i) \log P(X = x_i) \stackrel{not.}{=} E_X[-\log P(X)]$.
  Prin *convenţie*, dacă $p(x) = 0$ atunci vom considera $p(x) \log p(x) = 0$.
- **Entropia condiţională specifică a variabilei Y în raport cu valoarea x_k a variabilei X:**
  $H(Y \mid X = x_k) \stackrel{def.}{=} -\sum_j P(Y = y_j \mid X = x_k) \log P(Y = y_j \mid X = x_k)$
  $\stackrel{not.}{=} E_{Y|X=x_k}[-\log P(Y \mid X = x_k)]$.
- **Entropia condiţională medie a variabilei Y în raport cu variabila X:**
  $H(Y \mid X) \stackrel{def.}{=} \sum_k P(X = x_k) H(Y \mid X = x_k) \stackrel{not.}{=} E_X[H(Y \mid X)]$.
- **Entropia corelată a variabilelor X şi Y:**
  $H(X,Y) \stackrel{def.}{=} -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(X = x_i, Y = y_j)$
  $\stackrel{not.}{=} E_{X,Y}[-\log P(X,Y)]$.
- **Informaţia mutuală a variabilelor X şi Y, numită de asemenea *câştigul de informaţie* al variabilei X în raport cu variabila Y (sau invers):**
  $MI(X,Y) \stackrel{not.}{=} IG(X,Y) \stackrel{def.}{=} H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
  (Observaţie: ultima egalitate de mai sus are loc datorită rezultatului de la punctul c de mai jos.)

Arătaţi că:

a. $H(X) \geq 0$. **În particular,** $H(X) = 0$ **dacă şi numai dacă variabila** X **este constantă.**

b. $H(Y \mid X) = -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(Y = y_j \mid X = x_i)$.

c. $H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$.
Mai general: $H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1})$ **(regula de înlănţuire).**

Răspuns:

a. Este uşor să arătăm că $H(X) = -\sum_i P(X = x_i) \log P(X = x_i) \geq 0$.

Ştim că $\log x \leq 0$ **pentru** $\forall x \leq 1$ **şi** $\log x \geq 0$ **pentru** $\forall x \geq 1$. **De asemenea ştim că** $P(X = x_i) \in [0,1]$ **(fiind o probabilitate). Aşadar,**

$$H(X) = -\sum_i P(X = x_i) \log P(X = x_i) = \sum_i \underbrace{P(X = x_i)}_{\geq 0} \underbrace{\log \frac{1}{P(X = x_i)}}_{\geq 0} \geq 0$$

Pentru a arăta că $H(X) = 0$ **dacă şi numai dacă** X **este constantă vom demonstra că ambele implicaţii au loc:**

„⇒“ Presupunem că $H(X) = 0$, **adică** $\sum_i P(X = x_i) \log \frac{1}{P(X = x_i)} = 0$. **Datorită faptului că fiecare termen din această sumă este mai mare sau egal cu** 0, **rezultă că** $H(X) = 0$ **doar dacă pentru** $\forall i$, $P(X = x_i) = 0$ **sau** $\log \frac{1}{P(X = x_i)} = 0$, **adică dacă pentru** $\forall i$, $P(X = x_i) = 0$ **sau** $P(X = x_i) = 1$. **Cum însă** $\sum_i P(X = x_i) = 1$ **rezultă că există o singură valoare** x_1 **pentru** X **astfel**

încât $P(X = x_1) = 1$, iar $P(X = x) = 0$ pentru orice $x \neq x_1$. Altfel spus, variabila aleatoare discretă X este constantă.[21]

„⇐“ Presupunem că variabila X este constantă, ceea ce înseamnă că X ia o singură valoare x_1, cu probabilitatea $P(X = x_1) = 1$. Prin urmare, $H(X) = -1 \cdot \log 1 = 0$.

b. Pentru a demonstra egalitatea cerută vom porni de la definiţia lui $H(Y \mid X)$ şi apoi vom efectua câteva transformări elementare:

$$\begin{aligned}
H(Y \mid X) &= \sum_i P(X = x_i) H(Y \mid X = x_i) \\
&= \sum_i P(X = x_i) \left[-\sum_j P(Y = y_j \mid X = x_i) \log P(Y = y_j \mid X = x_i) \right] \\
&= -\sum_i \sum_j \underbrace{P(X = x_i) P(Y = y_j \mid X = x_i)}_{= P(X = x_i, Y = y_j)} \log P(Y = y_j \mid X = x_i) \\
&= -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(Y = y_j \mid X = x_i)
\end{aligned}$$

c. În primul rând, trebuie să demonstrăm că

$$H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

Din definiţia entropiei corelate ştim că: $H(X, Y) = -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(X = x_i, Y = y_j)$. **Vom aplica mai întâi regula de multiplicare,** $P(X, Y) = P(X) \cdot P(Y \mid X)$**, după care vom transforma logaritmul produsului în sumă de logaritmi. Pentru claritatea demonstraţiei vom nota prescurtat** $p(x_i) = P(X = x_i)$, $p(x_i, y_j) = P(X = x_i, Y = y_j)$, **etc.**

$$\begin{aligned}
H(X, Y) &= -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j) \\
&= -\sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) \log[p(x_i) \cdot p(y_j \mid x_i)] \\
&= -\sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) [\log p(x_i) + \log p(y_j \mid x_i)] \\
&= -\sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) \log p(x_i) - \sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) \log p(y_j \mid x_i) \\
&= -\sum_i p(x_i) \log p(x_i) \cdot \underbrace{\sum_j p(y_j \mid x_i)}_{=1} - \sum_i p(x_i) \sum_j p(y_j \mid x_i) \log p(y_j \mid x_i) \\
&= H(X) + \sum_i p(x_i) H(Y \mid X = x_i) = H(X) + H(Y \mid X)
\end{aligned}$$

Pentru a demonstra egalitatea $H(X, Y) = H(Y) + H(X \mid Y)$, se procedează analog, înlocuind $p(x_i, y_j)$ nu cu $p(x_i) \cdot p(y_j \mid x_i)$, ci cu $p(y_i) \cdot p(x_j \mid y_i)$.

[21] Mai corect spus, X este constantă pe tot domeniul de definiţie, evenual cu excepţia unei mulţimi de probabilitate 0.

Pentru cazul general

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1})$$

vom folosi regula de înlănţuire de la variabile aleatoare

$$P(X_1, \ldots, X_n) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid X_1, X_2) \cdot \ldots \cdot P(X_n \mid X_1, \ldots, X_{n-1}),$$

precum şi scrierea entropiei sub formă de medie, $H(X) = E\left[\log \frac{1}{P(X)}\right]$**:**

$$\begin{aligned} H(X_1, \ldots, X_n) &= E\left[\log \frac{1}{p(x_1, \ldots, x_n)}\right] \\ &= -E_{p(x_1,\ldots,x_n)}[\log p(x_1, \ldots, x_n)] \\ &= -E_{p(x_1,\ldots,x_n)}[\log p(x_1) + \log p(x_2 \mid x_1) + \ldots + \log p(x_n \mid x_1, \ldots, x_{n-1})] \\ &= -E_{p(x_1)}[\log p(x_1)] - E_{p(x_1,x_2)}[\log p(x_2 \mid x_1)] - \ldots \\ &\quad - E_{p(x_1,\ldots,x_n)}[\log p(x_n \mid x_1, \ldots, x_{n-1})] \\ &= H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1}) \end{aligned}$$

31. **(Entropie, entropie condiţională, câştig de informaţie: exemplificare [clasică])**

■□ • ○ ***CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 2***

Problema aceasta se referă la aruncarea a două zaruri perfecte, cu 6 feţe.

a. Calculează distribuţia probabilistă a sumei numerelor de pe cele două feţe care au fost obţinute/„observate“ în urma aruncării zarurilor.

În continuare, suma aceasta va fi asimilată cu o variabilă aleatoare, notată cu S**.**

b. Cantitatea de *informaţie* obţinută (sau: *surpriza* pe care o resimţim) la „observarea“ producerii valorii x **a unei variabile aleatoare** X **oarecare este prin *definiţie***

$$\mathit{Information}(P(X = x)) = \mathit{Surprise}(P(X = x)) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x).$$

Această cantitate este exprimată (numeric) în *biţi de informaţie*.

Cât de surprins vei fi atunci când vei „observa“ $S = 2$**, respectiv** $S = 11$**,** $S = 5$ **şi** $S = 7$**? (Vei exprima de fiecare dată rezultatul în biţi. Puteţi folosi** $\log_2 3 = 1.584962501$**.)**

c. Calculează entropia variabilei S**.**

d. Să presupunem acum că vei arunca aceste două zaruri pe rând, iar la aruncarea primului zar se obţine numărul 4. Cât este entropia lui S **în urma acestei „observaţii“? S-a pierdut, ori s-a câştigat informaţie în acest proces? Calculează cât de multă informaţie (exprimată în biţi) s-a pierdut ori s-a câştigat.**

Răspuns:

a. Redăm distribuţia lui S (uşor de calculat) în următorul tabel:

S	2	3	4	5	6	7	8	9	10	11	12
$P(S)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

b. Conform definiţiei date, vom avea:

$$\begin{aligned}
\mathit{Information}(S=2) &= -\log_2(1/36) = \log_2 36 = 2\log_2 6 = 2(1+\log_2 3)\\
&= 5.169925001 \text{ biţi}\\
\mathit{Information}(S=11) &= -\log_2 2/36 = \log_2 18 = 1+2\log_2 3 = 4.169925001 \text{ biţi}\\
\mathit{Information}(S=5) &= -\log_2 4/36 = \log_2 9 = 2\log_2 3 = 3.169925001 \text{ biţi}\\
\mathit{Information}(S=7) &= -\log_2 6/36 = \log_2 6 = 1+\log_2 3 = 2.584962501 \text{ biţi}
\end{aligned}$$

c. Conform definiţiei pentru entropie (vezi problema 30), $H(S)$ este media ponderată (cu ajutorul probabilităţilor) a „surprizelor“ / cantităţilor de informaţie produse la „observarea“ tuturor valorilor variabilei S. Făcând calculele, vom obţine:

$$\begin{aligned}
H(S) &= -\sum_{i=1}^{n} p_i \log p_i\\
&= -\left(2\cdot\frac{1}{36}\log\frac{1}{36}+2\cdot\frac{2}{36}\log\frac{2}{36}+2\cdot\frac{3}{36}\log\frac{3}{36}+2\cdot\frac{4}{36}\log\frac{4}{36}+\right.\\
&\quad \left.2\cdot\frac{5}{36}\log\frac{5}{36}+\frac{6}{36}\log\frac{6}{36}\right)\\
&= \frac{1}{36}(2\log_2 36+4\log_2 18+6\log_2 12+8\log_2 9+10\log_2\frac{36}{5}+6\log_2 6)\\
&= \frac{1}{36}(2\log_2 6^2+4\log_2 6\cdot 3+6\log_2 6\cdot 2+8\log_2 3^2+10\log_2\frac{6^2}{5}+6\log_2 6)\\
&= \frac{1}{36}(40\log_2 6+20\log_2 3+6-10\log_2 5)\\
&= \frac{1}{36}(60\log_2 3+46-10\log_2 5) = 3.274401919 \text{ biţi.}
\end{aligned}$$

d. Distribuţia variabilei S condiţionată de observarea feţei 4 la prima aruncare este:

S	2	3	4	5	6	7	8	9	10	11	12
$P(S\|...)$	0	0	0	1/6	1/6	1/6	1/6	1/6	1/6	0	0

În consecinţă, folosind definiţia entropiei condiţionale specifice (vezi de asemenea problema 30), vom avea:

$$H(S|\mathit{First\text{-}die\text{-}shows}\text{-}4) = -6\cdot\frac{1}{6}\log_2\frac{1}{6} = \log_2 6 = 2.58 \text{ biţi,}$$

ceea ce înseamnă că se obţine următorul *câştig* de informaţie:

$$IG(S;\mathit{First\text{-}die\text{-}shows}\text{-}4) = H(S) - H(S|\mathit{First\text{-}die\text{-}shows}\text{-}4) = 3.27 - 2.58 = 0.69 \text{ biţi.}$$

Altfel spus, atunci când ni se comunică faptul că la aruncarea celor două zaruri primul dintre ele produce faţa 4, această informaţie va reduce ulterior entropia variabilei S (sau, am putea spune, „surpriza“ medie provocată de valorile ei) cu 0.69 biţi.

32. **(Probabilităţi marginale, entropii, entropii condiţionale medii)**

■□ • ***CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 3***

Un doctor trebuie să pună un diagnostic unui pacient care are simptome de răceală (C, de la engl. cold). Factorul principal pe care doctorul îl ia în considerare pentru a elabora diagnosticul este timpul, adică starea vremii de afară (T). Variabila aleatoare C ia două valori, *yes* şi *no*, iar variabila aleatoare T ia 3 valori: *sunny* (însorit), *rainy* (ploios) şi *snowy* (foarte rece, să zicem). Distribuţia corelată a celor două variabile este dată în tabelul următor:

	$T = sunny$	$T = rainy$	$T = snowy$
$C = no$	0.30	0.20	0.10
$C = yes$	0.05	0.15	0.20

a. Calculaţi probabilităţile marginale $P(C)$ şi $P(T)$.

***Sugestie*: Folosiţi formula $P(X = x) = \sum_y P(X = x; Y = y)$. De exemplu,**

$$P(C = no) = P(C = no, T = sunny) + P(C = no, T = rainy) + P(C = no, T = snowy).$$

b. Calculaţi entropiile $H(C)$ şi $H(T)$.

c. Calculaţi entropiile condiţionale medii $H(C|T)$ şi $H(T|C)$.

Răspuns:

a. Folosind formula dată, vom obţine: $P_C = (0.6, 0.4)$ şi $P_T = (0.35, 0.35, 0.30)$.

b. Aplicând definiţia pentru entropie (vezi problema 30), rezultă:

$$\begin{aligned} H(C) &= 0.6 \log \frac{5}{3} + 0.4 \log \frac{5}{2} = \log 5 - 0.6 \log 3 - 0.4 = 0.971 \text{ biţi} \\ H(T) &= 2 \cdot 0.35 \log \frac{20}{7} + 0.3 \log \frac{10}{3} \\ &= 0.7(2 + \log 5 - \log 7) + 0.3(1 + \log 5 - \log 3) \\ &= 1.7 + \log 5 - 0.7 \log 7 - 0.3 \log 3 = 1.581 \text{ biţi.} \end{aligned}$$

c. Aplicând definiţia pentru entropie condiţională medie (vezi de asemenea

problema 30), vom avea:

$$\begin{aligned}
&H(C|T)\\
&= 0.35 \cdot H\left(\frac{0.30}{0.30+0.05}, \frac{0.05}{0.30+0.05}\right) + 0.35 \cdot H\left(\frac{0.20}{0.20+0.15}, \frac{0.15}{0.20+0.15}\right) +\\
&\quad 0.30 \cdot H\left(\frac{0.10}{0.10+0.20}, \frac{0.20}{0.20+0.10}\right)\\
&= \frac{7}{20} \cdot H\left(\frac{6}{7}, \frac{1}{7}\right) + \frac{7}{20} \cdot H\left(\frac{4}{7}, \frac{3}{7}\right) + \frac{3}{10} \cdot H\left(\frac{1}{3}, \frac{2}{3}\right)\\
&= \frac{7}{20} \cdot \left(\frac{6}{7}\log\frac{7}{6} + \frac{1}{7}\log 7\right) + \frac{7}{20} \cdot \left(\frac{4}{7}\log\frac{7}{4} + \frac{3}{7}\log\frac{7}{3}\right) + \frac{3}{10} \cdot \left(\frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2}\right)\\
&= \frac{7}{20} \cdot \left(\log 7 - \frac{6}{7} - \frac{6}{7}\log 3\right) + \frac{7}{20} \cdot \left(\log 7 - \frac{8}{7} - \frac{3}{7}\log 3\right) + \frac{3}{10} \cdot \left(\log 3 - \frac{2}{3}\right)\\
&= \frac{7}{10}\log 7 - \left(\frac{3}{10} + \frac{4}{10} + \frac{2}{10}\right) - \log 3\left(\frac{6}{20} + \frac{3}{20} - \frac{3}{10}\right)\\
&= \frac{7}{10}\log 7 - \frac{3}{20}\log 3 - \frac{9}{10} = 0.82715 \text{ biţi.}
\end{aligned}$$

Similar,

$$\begin{aligned}
&H(T|C)\\
&= 0.60 \cdot H\left(\frac{0.30}{0.30+0.20+0.10}, \frac{0.20}{0.30+0.20+0.10}, \frac{0.10}{0.30+0.20+0.10}\right) +\\
&\quad 0.40 \cdot H\left(\frac{0.05}{0.05+0.15+0.20}, \frac{0.15}{0.05+0.15+0.20}, \frac{0.20}{0.05+0.15+0.20}\right)\\
&= \frac{3}{5} \cdot H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) + \frac{2}{5} \cdot H\left(\frac{1}{8}, \frac{3}{8}, \frac{1}{2}\right)\\
&= \frac{3}{5}\left(\frac{1}{2} + \frac{1}{3}\log 3 + \frac{1}{6}(1+\log 3)\right) + \frac{2}{5}\left(\frac{1}{8}\cdot 3 + \frac{3}{8}(3-\log 3) + \frac{1}{2}\right)\\
&= \frac{3}{5}\left(\frac{2}{3} + \frac{1}{2}\log 3\right) + \frac{2}{5}\left(2 - \frac{3}{8}\log 3\right)\\
&= \frac{6}{5} + \frac{3}{20}\log 3 = 1.43774 \text{ biţi.}
\end{aligned}$$

33. (Calcularea entropiei unei variabile aleatoare continue: cazul distribuţiei exponenţiale)

■□ ***CMU, 2011 spring, R. Rosenfeld, HW2, pr. 2.c***

Pentru o distribuţie de probabilitate continuă P, entropia se defineşte astfel:

$$H(P) = \int_{-\infty}^{+\infty} P(x)\log_2\frac{1}{P(x)}dx$$

Calculaţi entropia *distribuţiei* continue *exponenţiale* de parametru $\lambda > 0$. Vă reamintim că definiţia acestei distribuţii este următoarea:

$$P(x) = \begin{cases} \lambda e^{-\lambda x}, & \textbf{dacă } x \geq 0;\\ 0, & \textbf{dacă } x < 0. \end{cases}$$

***Indicaţie*: Dacă $P(x) = 0$, veţi presupune că $-P(x)\log P(x) = 0$.**

Răspuns:

Dat fiind faptul că funcţia P se anulează pe intervalul $(-\infty, 0)$, este natural ca mai întâi să „rupem" intervalul de integrare pentru $\int_{-\infty}^{\infty} P(x)\log_2 \frac{1}{P(x)}\,dx$ în două: $(-\infty, 0]$ şi $[0, \infty)$. Aşadar,

$$\begin{aligned} H(P) &= \int_{-\infty}^{0} P(x)\log_2 \frac{1}{P(x)}\,dx + \int_{0}^{\infty} P(x)\log_2 \frac{1}{P(x)}\,dx \\ &\overset{\text{def. } P}{=} \int_{-\infty}^{0} 0\log_2 0\,dx + \int_{0}^{\infty} \lambda e^{-\lambda x}\log_2 \frac{1}{\lambda e^{-\lambda x}}\,dx \end{aligned}$$

Prima dintre aceste două ultime integrale este 0, conform *indicaţiei* din enunţ. Pentru a putea calcula cea de-a doua integrală (în expresia căreia apare numărul e), vom schimba baza logaritmului, şi anume vom trece din baza 2 în baza e (baza logaritmului natural, $\ln$).[22]

Prin urmare,

$$\begin{aligned} H(P) &= \frac{1}{\ln 2}\int_0^{\infty} \lambda e^{-\lambda x}\ln\frac{1}{\lambda e^{-\lambda x}}\,dx = \frac{1}{\ln 2}\int_0^{\infty} \lambda e^{-\lambda x}\left(\ln\frac{1}{\lambda} + \ln\frac{1}{e^{-\lambda x}}\right)\,dx \\ &= \frac{1}{\ln 2}\int_0^{\infty} \lambda e^{-\lambda x}\left(-\ln\lambda + \ln e^{\lambda x}\right)\,dx \\ &= \frac{1}{\ln 2}\int_0^{\infty} \lambda e^{-\lambda x}\left(-\ln\lambda + \lambda x\right)\,dx \\ &= \frac{1}{\ln 2}\int_0^{\infty} \lambda e^{-\lambda x}(-\ln\lambda)\,dx + \frac{1}{\ln 2}\int_0^{\infty} \lambda e^{-\lambda x}\lambda x\,dx \\ &= \frac{-\ln\lambda}{\ln 2}\int_0^{\infty} \lambda e^{-\lambda x}\,dx + \frac{\lambda}{\ln 2}\int_0^{\infty} \lambda e^{-\lambda x} x\,dx \\ &= \frac{\ln\lambda}{\ln 2}\int_0^{\infty} \left(e^{-\lambda x}\right)'\,dx - \frac{\lambda}{\ln 2}\int_0^{\infty} \left(e^{-\lambda x}\right)' x\,dx \end{aligned}$$

Prima integrală se rezolvă foarte uşor:

$$\int_0^{\infty} \left(e^{-\lambda x}\right)'\,dx = e^{-\lambda x}\Big|_0^{\infty} = e^{-\infty} - e^0 = 0 - 1 = -1$$

Pentru a rezolva cea de-a doua integrală se poate folosi *formula de integrare prin părţi*:

$$\int_0^{\infty} \left(e^{-\lambda x}\right)' x\,dx = e^{-\lambda x}x\Big|_0^{\infty} - \int_0^{\infty} e^{-\lambda x}x'\,dx = e^{-\lambda x}x\Big|_0^{\infty} - \int_0^{\infty} e^{-\lambda x}\,dx$$

Integrala definită $e^{-\lambda x}x\Big|_0^{\infty}$ nu se poate calcula direct (din cauza conflictului $0\cdot\infty$ care se produce atunci când lui x i se atribuie valoarea-limită ∞), ci se calculează folosind *regula lui l'Hôpital*:

$$\lim_{x\to\infty} xe^{-\lambda x} = \lim_{x\to\infty}\frac{x}{e^{\lambda x}} = \lim_{x\to\infty}\frac{x'}{\left(e^{\lambda x}\right)'} = \lim_{x\to\infty}\frac{1}{\lambda e^{\lambda x}} = \frac{1}{\lambda}\lim_{x\to\infty} e^{-\lambda x} = e^{-\infty} = 0,$$

[22] Pentru aceasta, vom folosi formula $\log_a b = \frac{\log_c b}{\log_c a}$, valabilă pentru orice $a > 0$, $b > 0$ şi $c > 0$, cu $a \neq 1$ şi $c \neq 1$. În calculele care urmează vom folosi şi alte formule de la logaritmi.

deci

$$e^{-\lambda x}x\big|_0^\infty = 0 - 0 = 0.$$

Integrala $\int_0^\infty e^{-\lambda x}\,dx$ se calculează uşor:

$$\int_0^\infty e^{-\lambda x}\,dx = -\frac{1}{\lambda}\int_0^\infty \left(e^{-\lambda x}\right)'\,dx = -\frac{1}{\lambda}\,e^{-\lambda x}\big|_0^\infty = -\frac{1}{\lambda}(0-1) = \frac{1}{\lambda}$$

Prin urmare,

$$\int_0^\infty \left(e^{-\lambda x}\right)' x\,dx = 0 - \frac{1}{\lambda} = -\frac{1}{\lambda},$$

ceea ce conduce la rezultatul final:

$$H(P) = \frac{\ln\lambda}{\ln 2}(-1) - \frac{\lambda}{\ln 2}\left(-\frac{1}{\lambda}\right) = -\frac{\ln\lambda}{\ln 2} + \frac{1}{\ln 2} = \frac{1-\ln\lambda}{\ln 2}.$$

34. **(Entropia relativă: definiţie şi proprietăţi elementare; exprimarea câştigului de informaţie cu ajutorul entropiei relative)**

■□ ○ ***prelucrare de Liviu Ciortuz după CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.2***

Entropia relativă **sau** ***divergenţa Kullback-Leibler*** **(KL) a unei distribuţii p în raport cu o altă distribuţie q — ambele distribuţii fiind discrete — se defineşte astfel:**

$$KL(p||q) \stackrel{def.}{=} -\sum_{x\in X} p(x)\log\frac{q(x)}{p(x)}$$

Din perspectiva teoriei informaţiei, divergenţa KL specifică numărul de *biţi adiţionali* care sunt necesari în medie pentru a transmite valorile variabilei X atunci când presupunem că aceste valori sunt distribuite conform distribuţiei („model“) q, dar în realitate ele urmează o altă distribuţie, p.[23]

a. Demonstraţi că $KL(p||q) \geq 0$ şi apoi arătaţi că egalitatea are loc dacă şi numai dacă $p = q$.[24]

***Indicaţie*:**
Pentru a demonstra punctul acesta puteţi folosi *inegalitatea lui Jensen*:
Dacă $\varphi : \mathbb{R} \to \mathbb{R}$ este o funcţie convexă, atunci pentru orice $t \in [0,1]$ şi orice $x_1, x_2 \in \mathbb{R}$ urmează $\varphi(tx_1 + (1-t)x_2) \leq t\varphi(x_1) + (1-t)\varphi(x_2)$. Dacă φ este funcţie strict convexă, atunci egalitatea are loc doar dacă $x_1 = x_2$.
Mai general, pentru orice $a_i \geq 0,\ i = 1,\ldots,n$ cu $\sum_i a_i \neq 0$ şi orice $x_i \in \mathbb{R},\ i =$

[23] *Atenţie*: Divergenţa KL nu este o măsură de *distanţă* între două distribuţii probabiliste, fiindcă în general ea nu este simetrică ($KL(p||q) \neq KL(q||p)$). Pentru „simetrizare“, se consideră $JSD(p||q) = \frac{1}{2}(KL(p||q) + KL(q||p))$, care se numeşte *divergenţa Jensen-Shannon*.
Măsura

$$\begin{aligned} d(X,Y) &\stackrel{def}{=} H(X,Y) - IG(X;Y) = H(X) + H(Y) - 2IG(X;Y) \\ &= H(X \mid Y) + H(Y \mid X), \end{aligned}$$

cunoscută ca *variaţia informaţiei*, este o măsură de distanţă (metrică), adică este nenegativă, simetrică, implică egalitatea indiscernabililor şi satisface inegalitatea triunghiului.

[24] Mai general, $KL(p||q)$ este cu atât mai mică cu cât „asemănarea“ dintre distribuţiile p şi q este mai mare.

$1, \ldots, n$, **avem** $\varphi\left(\frac{\sum_i a_i x_i}{\sum_j a_j}\right) \leq \frac{\sum_i a_i \varphi(x_i)}{\sum_j a_j}$. **Dacă** φ **este strict convexă, atunci egalitatea are loc doar dacă** $x_1 = \ldots = x_n$. **Evident, rezultate similare pot fi formulate şi pentru funcţii concave.**

b. Câştigul de informaţie poate fi definit ca fiind entropia relativă dintre distribuţia corelată observată a lui X **şi** Y **pe de o parte, şi produsul distribuţiilor marginale** p_X **şi** p_Y **pe de altă parte:**

$$IG(X,Y) \stackrel{def.}{=} KL(p_{X,Y} \,||\, (p_X\, p_Y)) = -\sum_x \sum_y p(x,y) \log\left(\frac{p(x)p(y)}{p(x,y)}\right)$$

Arătaţi că această nouă definiţie a câştigului de informaţie este echivalentă cu definiţia dată anterior (vezi problema 30). Cu alte cuvinte, arătaţi că

$$KL(p_{X,Y} \,||\, (p_X\, p_Y)) = H[X] - H[X \mid Y]$$

Observaţie: **Din noua definiţie introdusă mai sus pentru câştigul de informaţie, rezultă imediat că**

$$\begin{aligned} IG(X,Y) &= \sum_y p(y) \sum_x p(x \mid y) \log \frac{p(x \mid y)}{p(x)} = \sum_y p(y) KL(p_{X|Y} \,||\, p_X) \\ &= E_Y[KL(p_{X|Y} \,||\, p_X)] \end{aligned}$$

ceea ce înseamnă că $IG(X,Y)$ **poate fi văzută ca o medie (în raport cu distribuţia lui** Y**) a divergenţei KL dintre distribuţia condiţională a lui** X **în raport cu** Y **pe de o parte, şi distribuţia lui** X **pe de altă parte.**

c. Arătaţi că $IG(X,Y) \geq 0$ **pentru orice variabile aleatoare discrete** X **şi** Y**. În particular,** $IG(X,Y) = 0$ **dacă şi numai dacă** X **şi** Y **sunt independente.**

Răspuns:

a. Vom dovedi inegalitatea $KL(p||q) \geq 0$ **folosind inegalitatea lui Jensen, în expresia căreia vom înlocui** φ **cu funcţia convexă** $-\log_2$**, pe** a_i **cu** $p(x_i)$ **şi pe** x_i **cu** $\frac{q(x_i)}{p(x_i)}$**. (Pentru convenienţă, în cele ce urmează vor renunţa la indicele variabilei** x**.) Vom avea:**

$$\begin{aligned} KL(p \,||\, q) &\stackrel{def.}{=} -\sum_x p(x) \log \frac{p(x)}{q(x)} \\ &\stackrel{Jensen}{\geq} -\log(\sum_x p(x)\frac{q(x)}{p(x)}) = -\log(\underbrace{\sum_x q(x)}_{1}) = -\log 1 = 0 \end{aligned}$$

Vom demonstra acum că $KL(p||q) \geq 0 \Leftrightarrow p = q$**.**
Egalitatea $p(x) = q(x)$ **implică** $\frac{q(x)}{p(x)} = 1$**, deci** $\log \frac{q(x)}{p(x)} = 0$ **pentru orice** x**, de unde rezultă imediat** $KL(p||q) = 0$**.**
Pentru a demonstra implicaţia inversă, se ţine cont că în inegalitatea lui Jensen are loc egalitatea doar în cazul în care $x_i = x_j$ **pentru orice** i **şi** j**. În cazul de**

faţă, această condiţie se traduce prin faptul că raportul $\frac{q(x)}{p(x)}$ este acelaşi pentru orice valoare a lui x. Ţinând cont că $\sum_x p(x) = 1$ şi $\sum_x p(x)\frac{q(x)}{p(x)} = \sum_x q(x) = 1$, rezultă că $\frac{q(x)}{p(x)} = 1$ sau, altfel spus, $p(x) = q(x)$ pentru orice x, ceea ce înseamnă că distribuţiile p şi q sunt identice.

b. Vom folosi regula de multiplicare, şi anume $p(x, y) = p(x \mid y)p(y)$:

$$\begin{aligned}
KL(p_{XY} \,||\, (p_X\, p_Y)) &\overset{def.\ KL}{=} -\sum_x \sum_y p(x,y) \log \left(\frac{p(x)p(y)}{p(x,y)} \right) \\
&= -\sum_x \sum_y p(x,y) \log \left(\frac{p(x)p(y)}{p(x \mid y)p(y)} \right) = -\sum_x \sum_y p(x,y)[\log p(x) - \log p(x \mid y)] \\
&= -\sum_x \sum_y p(x,y) \log p(x) - \left(-\sum_x \sum_y p(x,y) \log p(x \mid y) \right) \\
&= -\sum_x \log p(x) \underbrace{\sum_y p(x,y)}_{=p(x)} - H[X \mid Y] \\
&= H[X] - H[X \mid Y] = IG(X,Y)
\end{aligned}$$

c. Atât inegalitatea că $IG(X, Y) \geq 0$ cât şi echivalenţa $IG(X, Y) = 0 \Leftrightarrow X$ şi Y sunt independente constituie consecinţe imediate ale punctelor b şi a discutate mai sus.

35. **(Informaţia mutuală, aplicaţie: selecţia de trăsături)**

□ ***CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 6***

În tabelul alăturat se dă un set de opt observaţii/instanţe, reprezentate ca tupluri de valori ale variabilelor aleatoare binare de „intrare“ X_1, X_2, X_3, X_4, X_5 şi ale variabilei aleatoare binare de „ieşire“ Y.

Am dori să reducem spaţiul de trăsături $\{X_1, X_2, X_3, X_4, X_5\}$ folosind o metodă de selecţie de tip *filtru*.

X_1	X_2	X_3	X_4	X_5	Y
0	1	1	0	1	0
1	0	0	0	1	0
0	1	0	1	0	1
1	1	1	1	0	1
0	1	1	0	0	1
0	0	0	1	1	1
1	0	0	1	0	1
1	1	1	0	1	1

a. Calculaţi informaţia mutuală $MI(X_i, Y)$ pentru fiecare i.

b. Ţinând cont de rezultatul de la punctul precedent, alegeţi cel mai mic subset de trăsături în aşa fel încât cel mai bun clasificator antrenat pe acest spaţiu (redus) de trăsături să fie cel puţin la fel de bun ca şi cel mai bun clasificator antrenat pe întreg spaţiul de trăsături. Justificaţi alegerea pe care aţi făcut-o.

Răspuns:

a. Pentru calculul informaţiei mutuale putem folosi formula din problema 34:

$$MI(X,Y) = \sum_x \sum_y p_{X,Y}(x,y) \log \left(\frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} \right)$$

Probabilităţile marginale, estimate în sensul verosimilităţii maxime (MLE), sunt:

	P_{X_1}	P_{X_2}	P_{X_3}	P_{X_4}	P_{X_5}	P_Y
0	0.5	3/8	0.5	0.5	0.5	0.25
1	0.5	5/8	0.5	0.5	0.5	0.75

iar probabilităţile corelate sunt:

X_i	Y	$P_{X_1,Y}$	$P_{X_2,Y}$	$P_{X_3,Y}$	$P_{X_4,Y}$	$P_{X_5,Y}$
0	0	1/8	1/8	1/8	1/4	0
0	1	3/8	1/4	3/8	1/4	1/2
1	0	1/8	1/8	1/8	0	1/4
1	1	3/8	1/2	3/8	1/2	1/4

Se poate observa că X_1 şi Y sunt independente, deci $MI(X_1,Y) = 0$, conform problemei 34.c. Similar, $MI(X_3,Y) = 0$. În rest, efectuând calculele obţinem $MI(X_2,Y) = 0.01571$, $MI(X_4,Y) = 0.3113$ şi $MI(X_5,Y) = 0.3113$.

b. La selecţia de trăsături vom alege acele trăsături X_i care au informaţie mutuală nenulă în raport cu Y. Acestea sunt X_2, X_4 şi X_5. Celelalte două trăsături, X_1 şi X_3 sunt independente în raport cu Y.

Totuşi, inspectând datele, observăm că dacă vom selecta doar trăsăturile X_2, X_4 şi X_5 vom avea două instanţe (vezi prima şi ultima linie din tabel) care au aceleaşi trăsături ($X_2 = 1, X_4 = 0, X_5 = 1$) dar au etichete/ieşiri diferite: $Y = 0$, respectiv $Y = 1$. Aşadar, vom adăuga la setul de trăsături selectate anterior şi variabila X_1, care va permite dezambiguizarea în cazul acestor două instanţe.

***Observaţie*: Deşi $MI(X_1,Y) = 0$, nu rezultă că $MI(X_1,X_j) = 0$ pentru $j \in \{2,4,5\}$. Aceasta explică de ce adăugarea variabilei X_1 la setul $\{X_2, X_4, X_5\}$ poate aduce un câştig de informaţie.**

36. (Entropia corelată: forma particulară a relaţiei de „înlănţuire“ în cazul variabilelor aleatoare independente)

□ ***prelucrare de Liviu Ciortuz după CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 7.b***

Demonstraţi că dacă X şi Y sunt variabile aleatoare independente discrete, atunci $H(X,Y) = H(X) + H(Y)$.

Este adevărată şi reciproca acestei afirmaţii? Adică, atunci când are loc egalitatea $H(X,Y) = H(X) + H(Y)$ rezultă că variabilele X şi Y sunt independente?

Răspuns:

Conform problemei 30.c, *formula de înlănţuire* a entropiilor pentru cazul general (adică, indiferent dacă X şi Y sunt sau nu independente) este:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{8}$$

Conform definiţiei câştigului de informaţie (vezi problema 30),

$$IG(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{9}$$

De asemenea, conform problemei 34.c,

$$IG(X,Y) = 0 \Leftrightarrow X \text{ şi } Y \text{ sunt independente.} \tag{10}$$

Din relaţiile (9) şi (10) rezultă că

$$H(Y) = H(Y|X) \Leftrightarrow X \text{ şi } Y \text{ sunt independente.} \tag{11}$$

Aşadar, dacă X şi Y sunt independente, coroborând relaţiile (11) şi (8) vom avea $H(X,Y) = H(Y) + H(X)$.

Invers, dacă $H(X,Y) = H(X) + H(Y)$, din relaţia (8) rezultă că $H(Y) = H(Y|X)$, ceea ce implică faptul că X şi Y sunt independente, conform relaţiei (11).

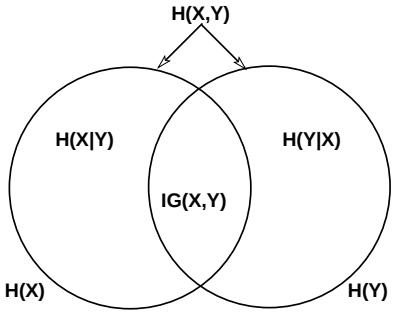
***Observaţie*: Din egalităţile (8) şi (9), rezultă**

$$H(X,Y) = H(X) + H(Y) - IG(X,Y).$$

Conform proprietăţii de ne-negativitate a câştigului de informaţie ($IG(X,Y) \geq 0$; vezi problema 34.c), rezultă

$$H(X,Y) \leq H(X) + H(Y).$$

Această ultimă relaţie, precum şi relaţiile (8) şi (9) sunt ilustrate în figura alăturată.



37. **(Cross-entropie: definiţie, o proprietate (ne-negativitatea) şi un exemplu simplu de calculare [a valorii cross-entropiei])**

□ • ○ ***CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 3.c***

***Cross-entropia* a două distribuţii p şi q, desemnată prin *CH(p,q)*, reprezintă numărul mediu de biţi necesari pentru a codifica un eveniment dintr-o mulţime oarecare de posibilităţi, atunci când schema de cod[ific]are folosită se bazează pe o distribuţie de probabilitate dată q, în loc să se bazeze pe distribuţia „adevărată“ p. În cazul în care distribuţiile p şi q sunt discrete, această noţiune se defineşte formal astfel:**

$$CH(p,q) = -\sum_x p(x) \log q(x).$$

În cazul distribuţiilor continue, definiţia se obţine/construieşte prin analogie:

$$CH(p,q) = -\int_X p(x)\,\log q(x)\,dx.$$

a. Poate oare cross-entropia să ia valori negative? Faceţi o demonstraţie sau daţi un contra-examplu.

b. În multe experimente, pentru a stabili calitatea diferitelor ipoteze/modele, se procedează la evaluarea/compararea lor pe un set de date. Să presupunem că, urmărind să faci predicţia *funcţiei de probabilitate* asociate unei anumite *variabile aleatoare* care are 7 valori posibile, ai obţinut (printr-un procedeu oarecare) două *modele* diferite, iar *distribuţiile de probabilitate* prezise de către aceste două modele sunt respectiv:

$$q_1 = \left(\frac{1}{10}, \frac{1}{10}, \frac{1}{5}, \frac{3}{10}, \frac{1}{5}, \frac{1}{20}, \frac{1}{20}\right) \text{ şi } q_2 = \left(\frac{1}{20}, \frac{1}{10}, \frac{3}{20}, \frac{7}{20}, \frac{1}{5}, \frac{1}{10}, \frac{1}{20}\right).$$

Să zicem că pentru evaluare foloseşti un set de date caracterizat de următoarea distribuţie *empirică*:

$$p_{empiric} = \left(\frac{1}{20}, \frac{1}{10}, \frac{1}{5}, \frac{3}{10}, \frac{1}{5}, \frac{1}{10}, \frac{1}{20}\right).$$

Calculează cross-entropiile $CH(p_{empiric}, q_1)$ şi $CH(p_{empiric}, q_2)$.
Care dintre aceste două modele va conduce la o cross-entropie mai mică? Putem oare garanta că acest model este într-adevăr [cel] mai bun? Explică/justifică răspunsul [pe care l-ai] dat.

Răspuns:

a. Nu, cross-entropia nu poate lua valori negative. Iată cum demonstrăm:
Ştim că pentru orice funcţii de probabilitate p şi q şi pentru orice x (care aparţine domeniului de valori al unei variabile aleatoare care are o astfel de distribuţie de probabilitate), valorile $p(x)$ şi $q(x)$ satisfac inegalităţile $0 \leq p(x) \leq 1$ and $0 \leq q(x) \leq 1$. Inegalitatea $q(x) \leq 1$ implică faptul că $\log q(x) \leq 0$. Din $0 \leq p(x)$ şi $-\log q(x) \geq 0$, rezultă că $0 \leq -p(x)\log q(x)$. În consecinţă, suma tuturor acestor termeni va fi de asemenea mai mare sau egală cu 0, deci cross-entropia nu poate fi niciodată negativă.

***Observaţie importantă*: Spre deosebire de entropie (vezi problema 68), cross-entropia nu este mărginită. Ea poate creşte la infinit; vezi cazul când pentru o anumită valoare x sunt adevărate simultan relaţiile $p(x) \neq 0$ şi $q(x) = 0$.**[25]

[25] Mai precis, $\lim_{q(x)\to+0}(-p(x)\cdot\log_2 q(x)) = -p(x)(-\infty) = +\infty$.

b. Facem calculele, folosind formula cross-entropiei:

$$CH(p_{empiric}, q_1) =$$
$$-\left(\frac{1}{20}\log_2\frac{1}{10} + \frac{1}{10}\log_2\frac{1}{10} + \frac{1}{5}\log_2\frac{1}{5} + \frac{3}{10}\log_2\frac{3}{10} + \frac{1}{5}\log_2\frac{1}{5} + \frac{1}{10}\log_2\frac{1}{20} + \frac{1}{20}\log_2\frac{1}{20}\right) = \frac{3}{20}\log_2 10 + \frac{2}{5}\log_2 5 + \frac{3}{10}\log_2\frac{10}{3} + \frac{3}{20}\log_2 20 =$$
$$\frac{3}{20}\log_2 2\cdot 5 + \frac{2}{5}\log_2 5 + \frac{3}{10}\log_2\frac{2\cdot 5}{3} + \frac{3}{20}\log_2 2^2\cdot 5 =$$
$$\left(\frac{3}{20} + \frac{3}{10} + 2\cdot\frac{3}{20}\right) + \left(\frac{3}{20} + \frac{2}{5} + \frac{3}{10} + \frac{3}{20}\right)\log_2 5 - \frac{3}{10}\log_2 3 =$$
$$\frac{3}{4} + \log_2 5 - \frac{3}{10}\log_2 3 = 2.596439345 \text{ biţi}$$

$$CH(p_{empiric}, q_2) =$$
$$-\left(\frac{1}{20}\log_2\frac{1}{20} + \frac{1}{10}\log_2\frac{1}{10} + \frac{1}{5}\log_2\frac{3}{20} + \frac{3}{10}\log_2\frac{7}{20} + \frac{1}{5}\log_2\frac{1}{5} + \frac{1}{10}\log_2\frac{1}{10} + \frac{1}{20}\log_2\frac{1}{20}\right) =$$
$$\frac{1}{10}\log_2 20 + \frac{1}{5}\log_2 10 + \frac{1}{5}\log_2\frac{20}{3} + \frac{3}{10}\log_2\frac{20}{7} + \frac{1}{5}\log_2 5 =$$
$$\frac{1}{10}\log_2 2^2\cdot 5 + \frac{1}{5}\log_2 2\cdot 5 + \frac{1}{5}\log_2\frac{2^2\cdot 5}{3} + \frac{3}{10}\log_2\frac{2^2\cdot 5}{7} + \frac{1}{5}\log_2 5 =$$
$$\left(2\cdot\frac{1}{10} + \frac{1}{5} + 2\cdot\frac{1}{5} + 2\cdot\frac{3}{10}\right) + \left(\frac{1}{10} + 3\cdot\frac{1}{5} + \frac{3}{10}\right)\log_2 5 - \frac{1}{5}\log_2 3 - \frac{3}{10}\log_2 7 =$$
$$\frac{7}{5} + \log_2 5 - \frac{1}{5}\log_2 3 - \frac{3}{10}\log_2 7 = 2.562729118 \text{ biţi.}$$

Aşadar, distribuţia $p_{empiric}$ are o cross-entropie mai mică în [raport cu] modelul q_2. Este deci rezonabil să afirmăm că alegerea modelului q_2 este mai bună.

Totuşi, nu putem garanta că acest model este întotdeauna cel mai bun, fiindcă aici lucrăm cu o distribuţie „empirică", iar distribuţia „adevărată" nu neapărat se reflectă în mod complet/perfect în această distribuţie empirică.

De obicei, *bias-ul de eşantionare* (engl., sampling bias), precum şi *insuficienţa datelor de antrenament* vor contribui la lărgirea „spaţiului" care diferenţiază distribuţia adevărată de distribuţia empirică. Prin urmare, în practică, atunci când concepem un [astfel de] experiment de evaluare a mai multor distribuţii probabiliste, trebuie să avem permanent în minte faptul acesta şi, dacă este posibil, să folosim tehnici care reduc/minimizează aceste riscuri.

38. **(Inegalitatea lui Gibbs: un caz particular; comparaţie între valorile entropiei şi ale cross-entropiei)**

□ ***Liviu Ciortuz, 2012, după www.en.wikipedia.org***

Fie $P = \{p_1, \ldots p_n\}$ o distribuţie de probabilitate discretă.

a. Arătaţi că pentru orice distribuţie de probabilitate $Q = \{q_1, \dots q_n\}$ are loc inegalitatea:

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i} p_i \log_2 q_i$$

Altfel spus, $H(P) \leq CH(P,Q)$, unde $H(P)$ este entropia distribuţiei P, iar $CH(P,Q)$ este *cross-entropia* lui P în raport cu Q.

b. Arătaţi că în formula de mai sus egalitatea are loc dacă şi numai dacă $p_i = q_i$ pentru $i = 1, \dots, n$.

***Observaţie*: În formula din enunţ, în locul bazei 2 pentru logaritm poate fi folosită orice bază supraunitară.**

***Indicaţie*: Dacă în inegalitatea dată se trece termenul din partea stângă în partea dreaptă, obţinem $0 \leq \sum_{i=1}^{n} p_i \log_2 p_i - \sum_i p_i \log_2 q_i \Leftrightarrow 0 \leq -\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i}$. Puteţi face legătura dintre expresia din partea dreaptă a acestei ultime inegalităţi şi definiţia *entropiei relative* (numită de asemenea *divergenţa Kullback-Leibler*, vedeţi problema 34) şi apoi să folosiţi proprietăţile entropiei relative.**

Răspuns:

Expresia $-\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i}$ la care s-a ajuns în *Indicaţie* este exact divergenţa Kullback-Leibler dintre distribuţiile P şi Q. Formal, scriem acest lucru astfel: $KL(P||Q) = -\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} = CH(P,Q) - H(P)$.

a. La problema 34.a am demonstrat inegalitatea $KL(P||Q) \geq 0$, care are loc pentru orice distribuţii probabiliste discrete P şi Q. Aceasta este exact proprietatea de care avem nevoie pentru a justifica inegalitatea dată în enunţ la acest punct ($H(P) \leq CH(P,Q)$).

b. Tot la problema 34.a s-a demonstrat că $KL(P||Q)$ are valoarea 0 dacă şi numai dacă distribuţiile P şi Q sunt identice. În contextul nostru, această proprietate se transpune imediat sub forma $H(P) = CH(P,Q) \Leftrightarrow p_i = q_i$ pentru $i = 1, \dots, n$.

Elemente de teoria informaţiei

65. **(Probabilităţi marginale, entropii, entropii condiţionale medii)**

• ◦ *CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 1.d*

Echipa de fotbal american The Steelers (din Pittsburgh) va juca în cupa Superbowl *XLV* contra echipei The Green Bay Packers. Pregătindu-şi meciul, ei (fotbaliştii echipei The Steelers) se gândesc să-şi definească strategia de joc în funcţie de doi factori majori:

- **dacă jucătorul Ben Roethlisberger va fi (sau nu) accidentat la vremea meciului (Injured = yes / no), şi**
- **cum anume va fi vremea (Weather = foggy / rainy / clear sky).**

Iată distribuţia corelată a acestor două tipuri de evenimente:

	Weather = foggy	rainy	clear sky	P(Injured)
Injured = no	0.1	0.25	0.35	
Injured = yes	0.05	0.1	0.15	
P(Weather)				

a. Pornind de la distribuţia corelată dată, completaţi ultima linie şi ultima coloană din tabelul de mai sus cu valorile corespunzătoare distribuţiilor marginale $P(Weather)$ şi $P(Injured)$.

b. Calculaţi entropiile *H(Weather)* şi *H(Injured)*.

c. Calculaţi entropiile condiţionale medii *H(Injured | Weather)* şi *H(Weather | Injured)*.

Veţi putea folosi următoarele aproximări: $\log_2 3 = 1.585$, $\log_2 5 = 2.322$, $\log_2 7 = 2.807$, $\log_2 11 = 3.459$ **şi** $\log_2 13 = 3.700$.

66. **(O margine superioară pentru valorea entropiei unei variabile aleatoare discrete)**

■ □ ∗ *CMU, 2003 fall, T. Mitchell, A. Moore, HW1, pr. 1.1*

***Comentariu*: La problema 30 s-a demonstrat că entropia oricărei variabile aleatoare discrete este ne-negativă ($H(X) \geq 0$).[27] La acest exerciţiu veţi demonstra — tot pentru cazul discret — că există şi o margine superioară pentru $H(X)$.**

Aşadar, fie X o variabilă aleatoare discretă care ia n valori şi urmează distribuţia probabilistă P. Conform definiţiei, entropia lui X este

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \log_2 P(X = x_i).$$

Arătaţi că $H(X) \leq \log_2 n$.

***Sugestie*: Puteţi folosi inegalitatea $\ln x \leq x - 1$ care are loc pentru orice $x > 0$.**

[27] Extensia acestei proprietăţi la cazul variabilelor aleatoare continue este imediată.

67. **(Entropia corelată: forma particulară a relaţiei de „înlănţuire“ în cazul variabilelor aleatoare independente)**

□ • ○ ∗ *prelucrare de Liviu Ciortuz după CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 7.b*

La problema 36 s-a demonstrat că dacă X şi Y sunt variabile aleatoare independente, atunci are loc egalitatea $H(X,Y) = H(X)+H(Y)$ şi reciproc. Demonstraţia a folosit proprietăţi/rezultate obţinute în problemele anterioare (30 şi 34).

Implicaţia directă (X şi Y independente $\Rightarrow H(X,Y) = H(X) + H(Y)$) se poate obţine însă şi în mod direct, pornind de la definiţia independenţei variabilelor aleatoare. Vă cerem să faceţi astfel demonstraţia acestei implicaţii. Veţi trata mai întâi cazul variabilelor aleatoare discrete şi apoi cazul variabilelor aleatoare continue.

68. **(Entropie corelată şi condiţionată: formula de „înlănţuire“ condiţională)**

□ • ○ ∗ *CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 4.b*

Demonstraţi că proprietatea

$$H(X,Y|A) = H(Y|A) + H(X|Y,A)$$

este adevărată pentru oricare 3 variabile aleatoare discrete X, Y şi A.

Explicaţi în mod intuitiv, într-o singură frază, care este semnificaţia proprietăţii de mai sus.

69. **(O proprietate a câştigului de informaţie: ne-negativitatea)**

■ □ ∗ *Liviu Ciortuz*

Definiţia *câştigului de informaţie* (sau: a *informaţiei mutuale*) al unei variabile aleatoare X în raport cu o altă variabilă aleatoare Y este $IG(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$.[28] La problema 34 s-a demonstrat — pentru cazul în care X şi Y sunt discrete — că $IG(X,Y) = KL(P_{X,Y}||P_X\,P_Y)$, unde KL desemnează *entropia relativă* (sau: *divergenţa Kullback-Leibler*), P_X şi P_Y sunt distribuţiile variabilelor X şi, respectiv, Y, iar $P_{X,Y}$ este distribuţia corelată a acestor variabile. Tot la problema 34 s-a arătat că divergenţa KL este întotdeauna ne-negativă. În consecinţă, $IG(X,Y) \geq 0$ pentru orice X şi Y.

La acest exerciţiu vă cerem să demonstraţi inegalitatea $IG(X,Y) \geq 0$ în manieră directă, adică fără a apela la divergenţa Kullback-Leibler.

***Sugestie*: Puteţi folosi următoarea formă a inegalităţii lui Jensen:**

$$\sum_{i=1}^{n} \alpha_i \log x_i \leq \log\left(\sum_{i=1}^{n} \alpha_i x_i\right)$$

unde baza logaritmului se consideră supraunitară, $\alpha_i \geq 0$ pentru $i = 1, \ldots, n$ şi $\sum_{i=1}^{n} \alpha_i = 1$.

[28] Vezi problema 30.

70. **(Cross-entropia — o aplicaţie: selecţia modelelor probabiliste)**

□ • ○ ***CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 10***

Avem un zar măsluit (engl., unfair die). Probabilităţile [reale] de apariţie pentru fiecare dintre feţele de la 1 la 6 sunt date de distribuţia

$$P_{true} = (0.08, 0.55, 0.15, 0.12, 0.05, 0.05).$$

Fără să cunoască acest fapt, două persoane, identificate cu A şi respectiv B, ne sugerează următoarele *modele* de probabilitate pentru zarul măsluit:

$$P_A = (0.07, 0.14, 0.24, 0.24, 0.05, 0.26)$$
$$P_B = (0.25, 0.13, 0.21, 0.03, 0.11, 0.27)$$

a. Elaboraţi câteva idei relativ la cum am putea măsura/determina care dintre aceste două modele este mai bun.

b. Calculaţi cross-entropiile $CH(P_{true}, P_A)$, $CH(P_{true}, P_A)$ şi $CH(P_{true}, P_{true})$. Presupunând că alegem ca măsură/mijloc de evaluare a modelelor cross-entropia, care dintre cele două modele (P_A şi P_B) credeţi că este mai bun?

71. **(Proprietăţi ale entropiei: Adevărat sau Fals?)**

□ ∗ ***CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 2.a.1***
CMU, 2008 fall, Eric Xing, final exam, pr. 1.4
CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 7

Stabiliţi dacă următoarele propoziţii sunt adevărate sau false.

a. Entropia nu este negativă.

b. $H(X, Y) \geq H(X) + H(Y)$ pentru orice două variabile aleatoare X şi Y.

c. Dacă X şi Y sunt variabile aleatoare independente, atunci $H(X, Y) = H(X) + H(Y)$.