Master CSI 1 2016-2017

Théorie de l'information : DS du 18 octobre 2016

Master Sciences et Technologies, mention Mathématiques ou Informatique, parcours Cryptologie et Sécurité informatique

Responsable : Gilles Zémor

Durée : 1h30. Sans document. Les exercices sont indépendants.

– Exercice 1. Soit Σ l'ensemble des transpositions sur l'ensemble à quatre éléments $\{1, 2, 3, 4\}$. On rappelle qu'une transposition permute deux éléments de l'ensemble. On considère le quadruplet $X = [X_1, X_2, X_3, X_4]$ obtenu à partir de $[1, 2, 3, 4]$ en lui appliquant une transposition choisie uniformément dans Σ. Par exemple la transposition $(1, 2)$ produit $X = [2, 1, 3, 4]$ et la transposition $(1, 3)$ produit $X = [3, 2, 1, 4]$.

a) Calculer $H(X_i)$, $i = 1, 2, 3, 4$.

b) Les variables X_1, X_2 sont-elles indépendantes ? Calculer $I(X_1, X_2)$.

c) Que vaut $H(X_3|X_1, X_2)$?

– **Solution.**

a) Écrire les six valeurs de (X_1, X_2, X_3, X_4) fait apparaître que $P(X_1 = 1) = 1/2$ et $P(X_1 = 2) = P(X_1 = 3) = P(X_1 = 4) = 1/6$. On en déduit que

$$H(X_1) = \frac{1}{2}\log_2 2 + \frac{3}{6}\log_2 6 = \frac{1}{2} + \frac{1}{2}(\log_2 2 + \log_2 3) = 1 + \frac{1}{2}\log_2 3 \approx 1.79.$$

On a $H(X_i) = H(X_1)$ pour $i = 2, 3, 4$.

b) Il est clair que $P(X_1 = 1, X_2 = 1) = 0$ alors que $P(X_1) \neq 0$ et $P(X_2) \neq 0$, les variables X_1 et X_2 ne peuvent donc pas être indépendantes.

La variable (X_1, X_2) prend six valeurs distinctes, chacune avec probabilité $1/6$, en d'autres termes (X_1, X_2) suit une loi uniforme et $H(X_1, X_2) = \log_2 6 = \log_2 2 + \log_2 3$. Donc

$$I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2) = 2 + \log_2 3 - 1 - \log_2 3 = 1.$$

c) La variable X_3 est entièrement déterminée par (X_1, X_2), donc

$$H(X_3|X_1, X_2) = 0.$$

– Exercice 2. Donner un exemple de variables aléatoires X et Y, Y prenant ses valeurs dans $\mathcal{Y}$, telles que $H(X|Y=y) > H(X)$ pour un certain $y \in \mathcal{Y}$.

– **Solution.** Considérons le couple (X,Y) qui prend les valeurs $(0,0),(1,0),(1,1)$ avec comme probabilités respectives $1/4, 1/4, 1/2$. On voit que $P(X=0|Y=0) = P(X=1|Y=0) = 1/2$, donc $H(X|Y=0) = 1$. Mais la loi de X n'est pas uniforme, donc $H(X) < H(X|Y=0)$.

– Exercice 3. Montrer que pour toute variable aléatoire X et toute fonction f définie sur l'ensemble des valeurs prises par X, on a $H(f(X)) \leqslant H(X)$.

– **Solution.** On a $H(f(X)) \leqslant H(f(X)) + H(X|f(X))$ car une entropie conditionnelle est toujours positive, donc $H(f(X)) \leqslant H(f(X),X)$. Mais $H(f(X),X) = H(X) + H(f(X)|X) = H(X)$ car $f(X)$ étant entièrement déterminée par X on a $H(f(X)|X) = 0$.

– Exercice 4. Soit C un code préfixe pour lequel l'inégalité de Kraft est une égalité. Montrer que tout mot de $\{0,1\}^*$ soit est le préfixe d'un mot de C, soit a pour préfixe un mot de C.

– **Solution.** Supposons qu'il existe un mot $z \in \{0,1\}$ qui ne soit ni préfixe ni suffixe d'un mot de C. Alors en ajoutant z à C on obtient un nouveau code préfixe. Mais alors la quantité

$$\sum_{x \in C} 2^{-\ell(x)}$$

augmente et l'inégalité de Kraft n'est plus vérifiée. Un tel z n'existe donc pas.

– Exercice 5. Donner un exemple de loi $p = (p_1,p_2,p_3,p_4,p_5)$ d'une variable prenant ses valeurs dans un ensemble à cinq éléments, pour laquelle l'algorithme de Huffman peut donner trois codes différents de distributions des longueurs

$$(1,2,3,4,4),\ (1,3,3,3,3) \text{ et } (2,2,2,3,3).$$

Pouvez-vous caractériser l'ensemble de ces lois p ?

– **Solution.**

La loi $(2/5, 1/5, 1/5, 1/10, 1/10)$ convient, de même que la loi $(3/8, 1/4, 1/8, 1/8, 1/8)$. Supposons $p_1 \geqslant p_2 \geqslant p_3 \geqslant p_4 \geqslant p_5$. Si l'algorithme de Huffman donne les trois distributions des longueurs ci-dessus, c'est que les codes associés sont tous optimaux, ce qui veut dire qu'ils ont des longueurs moyennes égales. On peut donc écrire

$$\begin{aligned}\bar{\ell} &= p_1 + 2p_2 + 3p_3 + 4p_4 + 4p_5 \\ &= p_1 + 3(p_2+p_3+p_4+p_5) \\ &= 2(p_1+p_2) + 3(p_3+p_4+p_5)\end{aligned}$$

Si on pose $p_2 = x$ et $p_3 = y$, on en déduit donc que $p_1 = x + y$ et $p_4 + p_5 = p_2 = x$. De $p_1 + p_2 + p_3 + p_4 + p_5 = 1$, on déduit

$$3x + 2y = 1.$$

L'inégalité $p_3 \leqslant p_2$ impose $y \leqslant x$ et les inégalités $p_5 \leqslant p_4 \leqslant p_3$ imposent $(p_4+p_5)/2 \leqslant p_3$, soit $x/2 \leqslant y$. En écrivant $y = 1/2 - 3x/2$ on déduit de ces inégalités que

$$\frac{1}{5} \leqslant x \leqslant \frac{1}{4}.$$

Finalement, l'ensemble des lois qui conviennent est l'ensemble des lois de la forme :

$$p_1 = \frac{1}{2} - \frac{x}{2}, p_2 = x, p_3 = \frac{1}{2} - \frac{3x}{2}, p_4 = \frac{x}{2} + t, p_5 = \frac{x}{2} - t$$

où

$$\begin{aligned} \frac{1}{5} \leqslant & x \leqslant \frac{1}{4} \\ 0 \leqslant & t \leqslant \frac{1}{2} - 2x \end{aligned}$$

– Exercice 6. Soit X une variable aléatoire à valeurs entières. La variable B est également une variable à valeurs entières, de plus indépendante de X. Soit $Y = X + B$. On suppose que X peut être retrouvée sans ambiguïté à partir de Y. Montrer dans ce cas que $H(Y) = H(X) + H(B)$.

– **Solution.** Comme X peut être retrouvée à partir de Y, on a $H(X|Y) = 0$, donc

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

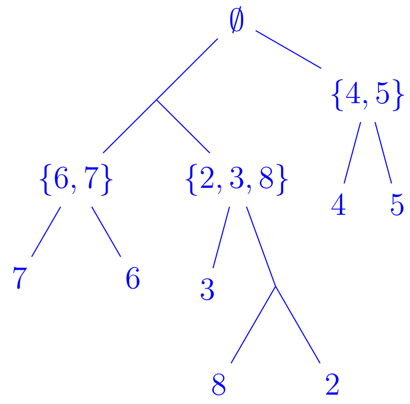
Comme l'application $(X, B) \mapsto (X, X + B)$ est une bijection, on a $H(X, Y) = H(X, B)$, donc

$$H(Y) = H(X, B).$$

Or X et B sont supposées indépendantes, donc $H(X, B) = H(X) + H(B)$.

– Exercice 7. Un joueur A réalise une variable $Z = X + Y$ où X et Y sont deux variables indépendantes de même loi uniforme dans l'ensemble $\{1, 2, 3, 4\}$. Un joueur B doit découvrir la valeur de Z en posant des questions dont la réponse est «oui» ou «non». Une procédure est dite optimale si elle permet au joueur B de poser une suite de questions successives dont les réponses déterminent entièrement Z, et telle que le nombre moyen de questions soit minimum.

Donner une procédure optimale pour déterminer Z et calculer le nombre moyen de questions associé.



– **Solution.** Une procédure n'est pas autre chose qu'un arbre binaire donc chaque sommet qui n'est pas une feuille représente une question du type «est-ce que Z appartient à tel ensemble de valeurs ?». La procédure s'identifie donc à un code préfixe, donc le nombre de questions moyen est juste la longueur moyenne du code. Pour trouver une procédure optimale il suffit donc d'appliquer l'algorithme de Huffman. Il y a plusieurs arbres possibles dans le cas présent, l'un d'entre eux est représenté ci-dessus. La première question est donc «est-ce que $Z \in \{4,5\}$?». Le nombre moyen de questions est donc :

$$\overline{q} = 2\left(\frac{4}{16}+\frac{3}{16}\right)+3\left(\frac{2}{16}+\frac{2}{16}+\frac{3}{16}\right)+4\left(\frac{1}{16}+\frac{1}{16}\right)=\frac{43}{16}$$