Master CSI 1 2014-2015

Théorie de l'information : DS du 20 octobre 2014

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

Responsable : Gilles Zémor

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

– Exercice 1. Soit p une loi $p = (p_1, p_2, \dots, p_m)$. Soit u la loi uniforme sur m objets. Montrer que $H(p) = \log_2 m - \mathrm{D}(p\,||\,u)$.

– **Solution.**

On a :

$$\begin{aligned}
\mathrm{D}(p\,||\,u) &= \sum_{i=1}^{m} p_i \log(mp_i) \\
&= \left(\sum_i p_i\right) \log m + \sum_i p_i \log_i \\
&= \log m - H(p).
\end{aligned}$$

D'où

$$H(p) = \log m - \mathrm{D}(p\,||\,u).$$

– Exercice 2. On considère une variable aléatoire X de loi uniforme dans l'ensemble $\{1, 2, 3, 4\}$. Puis, après avoir observé la valeur de X, on créé une variable Y, de loi uniforme dans l'ensemble des entiers i avec $X \leqslant i \leqslant 4$. Calculer $H(Y|X)$, $H(X|Y)$, $H(X,Y)$, $H(X)$ et $H(Y)$.

– **Solution.**

Comme X est uniforme on a $H(X) = 2$.

$X = i$ étant fixé, on a $H(Y|Y = i)$ qui est l'entropie d'une loi uniforme sur un ensemble à $5 - i$ éléments. On a donc

$$\begin{aligned}
H(Y|X) &= \sum_{i=1}^{4} P(X = i) H(Y|X = i) \\
&= \frac{1}{4}(\log 4 + \log 3 + \log 2) \\
&= \frac{3}{4} + \frac{1}{4}\log 3 \approx 1.15.
\end{aligned}$$

On a déduit $H(X,Y) = H(X) + H(Y|X) = \frac{11}{4} + \frac{1}{4}\log 3 \approx 3.15$.

Pour calculer $H(Y)$ il faut trouver la loi de Y. On obtient :

$$\begin{aligned}
P(Y=1) &= P(X=1)P(Y=1|X=1) = \frac{1}{16}\\
P(Y=2) &= P(X=1)P(Y=2|X=1) + P(X=2)P(Y=2|X=2) = \frac{7}{48}\\
P(Y=3) &= P(X=1)P(Y=3|X=1) + P(X=2)P(Y=3|X=2)\\
&\quad + P(X=3)P(Y=3|X=3) = \frac{1}{4}\frac{1}{4} + \frac{1}{4}\frac{1}{3} + \frac{1}{4}\frac{1}{2} = \frac{13}{48}\\
P(Y=4) &= P(X=1)P(Y=4|X=1) + P(X=2)P(Y=4|X=2)\\
&\quad + P(X=3)P(Y=4|X=3) + P(X=4)P(Y=4|X=4) = \frac{25}{48}.
\end{aligned}$$

D'où :

$$H(Y) = \frac{3}{48}\log\frac{48}{3} + \frac{7}{48}\log\frac{48}{7} + \frac{13}{48}\log\frac{48}{13} + \frac{25}{48}\log\frac{48}{25} \approx 1.66.$$

Enfin, on en déduit

$$H(X|Y) = H(X,Y) - H(Y) \approx 1.5.$$

– EXERCICE 3.

a) Construire un code préfixe avec comme distribution des longueurs

$$1, 3, 3, 3, 5, 5.$$

b) Quelles sont toutes les distributions des longueurs possibles d'un code préfixe associé à une loi sur 5 symboles ?

– **Solution.**

a) Par exemple : $C = \{0, 100, 101, 110, 11110, 11111\}$.

b) Ce sont tous les $\{\ell_1, \ldots, \ell_5\}$ avec

$$\sum_{i=1}^{5} \frac{1}{2^{\ell_i}} \leqslant 1$$

plus précisément ce sont toutes les distributions

$$\{1, 2, 3^+, 4^+, 4^+\}, \{1, 3^+, 3^+, 3^+, 3^+\}, \{2^+, 2^+, 2^+, 3^+, 3^+\}$$

où les longueurs sont ordonnées dans le sens croissant et où ℓ^+ signifie «ℓ ou plus élevé».

– Exercice 4. Soit $p = (p_1, \dots, p_m)$ une loi où on a supposé les p_i ordonnés $p_1 \geqslant p_2 \geqslant \dots \geqslant p_m$.

a) Montrer que si $p_1 < 1/3$, alors tous les mots d'un code de Huffman associé à p sont de longueur au moins 2.

b) Donner un exemple de loi $p_1 \geqslant \dots \geqslant p_m$ telle que $p_1 > 1/3$ et le code de Huffman associé à p a tous ses mots de longueur au moins 2.

– **Solution.**

a) S'il y a un mot de longueur 1, il est forcément associé à la probabilité la plus forte, soit p_1. Ceci implique que p_1 est choisi en dernier par l'algorithme de Huffman. Mais à l'avant-dernière étape on doit donc avoir trois probabilités $p_1 \geqslant p_2' \geqslant p_3'$. Mais alors

$$\frac{1}{3} \geqslant p_1 \geqslant p_2' > p_3'$$

contredit $p_1 + p_2' + p_3' = 1$.

b) La loi $(\frac{1}{3}, \frac{2}{9}, \frac{2}{9}, \frac{2}{9})$ donne forcément la distribution des longueurs $(2, 2, 2, 2)$. De même que la loi $(\frac{1}{3} + 3\varepsilon, \frac{2}{9} - \varepsilon, \frac{2}{9} - \varepsilon, \frac{2}{9} - \varepsilon)$ pour ε suffisamment petit.

– Exercice 5. Soit X une variable de loi uniforme prenant ses valeurs dans l'ensemble $\mathcal{X} = \{000, 001, 010, 100\}$. Une variable Y est créée à partir de $X = X_1X_2X_3$ on supprimant aléatoirement un symbole de X, pour donner X_2X_3 avec probabilité 1/3, X_1X_3 avec probabilité 1/3, ou X_1X_2 avec probabilité 1/3. La variable Y prend donc ses valeurs dans l'ensemble $\mathcal{Y} = \{00, 01, 10\}$.

Calculer $H(Y)$ et l'information mutuelle $I(X, Y)$. Que vaut $H(X|Y)$?

– **Solution.**

Il s'agit de commencer par trouver la loi de Y. On a :

$$\begin{aligned}
P(Y = 00) &= P(X = 000)P(Y = 00|X = 000) + P(X = 001)P(Y = 00|X = 001) \\
&= P(X = 010)P(Y = 00|X = 010) + P(X = 100)P(Y = 00|X = 100) \\
&= \frac{1}{4} + \frac{1}{4}\left(\frac{1}{3} + \frac{1}{3} + \frac{1}{3}\right) = \frac{1}{2}.
\end{aligned}$$

Un calcul similaire donne $P(Y = 01) = P(Y = 10) = \frac{1}{4}$. On en déduit

$$H(Y) = \frac{1}{2}\log 2 + 2\frac{1}{4}\log 4 = \frac{3}{2}.$$

Calculons maintenant $H(Y|X)$. Conditionné par $X = 000$, il n'y a plus d'incertitude sur Y et $H(Y|X = 000) = 0$. Conditionné par chacune des trois autres valeurs de X, Y prend deux valeurs avec une loi $(\frac{1}{3}, \frac{2}{3})$ ou bien trois valeurs avec une loi uniforme et on trouve :

$$H(Y|X) = P(X = 001)h\left(\frac{1}{3}\right) + P(X = 010)\log 3 + P(X = 100)h\left(\frac{1}{3}\right)$$

soit $H(Y|X) = \frac{3}{4}\log 3 - \frac{1}{3} \approx 0.85$.

On en déduit $I(X,Y) = H(Y) - H(Y|X) \approx 0.65$.

Enfin on peut en déduire $H(X|Y)$ par

$$H(X|Y) = H(X) - I(X,Y) = 2 - I(X,Y) \approx 1.35.$$

– Exercice 6. On sait qu'une rare bactérie nocive se trouve dans le vin provenant d'un certain vignoble qu'on cherche à localiser. Des études préliminaires restreignent l'ensemble des vignobles possibles à un ensemble de six vignobles. Pour ces six vignobles, numérotés de 1 à 6, la probabilité d'être le porteur de la bactérie est donnée par le 6-uple $(p_1, p_2, \ldots, p_6) = (8/23, 6/23, 4/23, 2/23, 2/23, 1/23)$. On applique maintenant des tests à des échantillons consistant en un mélange de vins de différents vignobles.

a) Donner une borne inférieure sur l'espérance du nombre de tests nécessaire pour déterminer le vignoble infecté.

b) Par quel mélange de vins faut-il commencer le premier test afin de minimiser le nombre de tests ?

– **Solution.**

a) Il s'agit de déterminer X est le numéro du vignoble porteur de la bactérie. L'information donnée par X est $H(X)$. Chaque test n'admet que deux réponses possibles, il apporte donc au maximum un bit d'information. L'entropie $H(X)$ est donc une borne inférieure sur le nombre de tests nécessaire pour déterminer X. Le nombre de tests sera en moyenne supérieur à $H(X)$ si les tests rapportent moins d'un bit d'information. Le calcul donne $H(X) \approx 2.28$. C'est une borne inférieure sur le nombre moyen de tests.

b) Il faut que le premier test rapporte le plus d'information possible : pour cela il faut que son issue soit la plus incertaine possible, et donc que la probabilité d'un test positif soit la plus proche possible de 1/2. On peut mélanger 1 et 3 ; ou bien 2, 3 et 4.