EISTI. **Année scolaire 2008-2009**

DÉPARTEMENT " INFORMATIQUE "

THÉORIE DE L'INFORMATION

Série d'exercices N^{o}2

PARTIE I. ENTROPIE D'UNE SOURCE. DÉFINITIONS ET PROPRIÉTÉS.

Exercice 1 (Calcul d'entropie. Exemple.)**.** Soit une source d'alphabet $\Omega = \{1, 2, 3, 5, 4\}$. Calculer son entropie pour les distributions de probabilités suivantes.

1. $P_1 = \{0.2, 0.2, 0.2, 0.2, 0.2\}$
2. $P_2 = \{0.05, 0.05, 0.05, 0.05, 0.8\}$
3. $P_3 = \{0.1, 0.2, 0.3, 0.15, 0.25\}$

Solution de l'exercice 1

Dans tous les cas on applique la définition de l'entropie :

$$H(X) = -\sum_{i=1}^{5} p_i \log(p_i)$$

1. $H(X_1) = -5 * 0.2 * \log(0.2) = -\log\left(\frac{1}{5}\right) = \log(5) \simeq 2.32.$
2. $H(X_2) = -4 * 0.05 * \log(0.05) - 0.8\log(0.8) \simeq 1.12.$
3. $H(X_2) = -0.1\log 0.1 - 0.2\log 0.2 - 0.3\log 0.3 - 0.15\log 0.15 - 0.25\log 0.25 \simeq 2.23.$

On remarquera que dans le premier cas l'entropie est la plus grande. *C'est une propriété générale de l'entropie : elle atteint son maximum lorsque tous les symboles d'un alphabet donné de taille n sont équiprobables. Elle est alors égale à* $\log(n)$.

Exercice 2.

1. On lance une pièce dont les deux cotés sont identiques : pile. Quelle est l'entropie associée à cette expérience ?
2. On lance un dé équilibré à 6 faces. Quelle est l'information moyenne apportée par l'observation de la parité du résultat ?
3. Un jeu de cartes contient 3 piques, 4 trèfles, 2 cœurs et 1 carreau. On tire une carte au hazard. Quelle est l'entropie de l'observation de la couleur de la carte ?

PARTIE II. ENTROPIE D'UN COUPLE "ÉMETTEUR-RÉCEPTEUR".

Exercice 3. Nous reprenons ici l'exemple du TD1 (voir exercice 1). Soient X et Y deux variables aléatoires prenant leurs valeurs respectivement dans $\Omega_X = \{x_1, x_2, x_3\}$ et $\Omega_Y = \{y_1, y_2\}$ et ayant la matrice de probabilités conjointes suivante

$$P(X,Y) = \begin{array}{c|cc} & y_1 & y_2 \\ \hline x_1 & 0.25 & 0 \\ x_2 & 0.1 & 0.3 \\ x_3 & 0.1 & 0.25 \end{array}$$

Nous avons établi pour ce couple de variables aléatoires les distributions de probabilité suivantes :

Distribution marginale de X.

$$p(x_1) = 0.25 \quad p(x_2) = 0.1 + 0.3 = 0.4 \quad p(x_3) = 0.1 + 0.25 = 0.35$$

Distribution marginale de Y.

$$p(y_1) = 0.25 + 0.10 + 0.1 = 0.45 \quad p(y_2) = 0.3 + 0.25 = 0.55$$

Les distributions conditionnelles

$$P(X|Y) = \begin{pmatrix} \frac{5}{9} & 0 \\ \frac{2}{9} & \frac{6}{11} \\ \frac{2}{9} & \frac{5}{11} \end{pmatrix} \text{ et } P(Y|X) = \begin{pmatrix} 1 & 0 \\ \frac{1}{4} & \frac{3}{4} \\ \frac{2}{7} & \frac{5}{7} \end{pmatrix}$$

1. Calculer $H(X)$,$H(Y)$,$H(X,Y)$,$H(X|Y)$, $H(Y|X)$.
2. On définit l'information mutuelle de X et Y ou encore le gain d'information

$$I(X|Y) = H(X) - H(X|Y)$$

cette quantité représente la diminution de l'incertitude sur X lorsqu'on a observé Y. Calculer $I(X|Y)$ et $I(Y|X)$. Verifier la propriété de symétrie de l'information mutuelle :

$$I(X|Y) = I(Y|X)$$

PARTIE III. ENTROPIE : UN JEU D'ESPION !

Le chiffrement de César consiste à décaler l'alphabet de k postions de façon cyclique et de remplacer chaque lettre d'un massage clair par une lettre correspondante de l'alphabet décalé. La clé secrète de ce chiffre est un entier $1 \leq k \leq 25$ qui représente le décalage de l'alphabet. Nous allons dans un premier temps étudier la sécurité de ce chiffre et ensuite apprendre la méthode d'analyse des fréquences qui a permis de la casser.

Exercice 4 (Rendons à César ce qui est à César : son chiffre !)**.**

1. Montrer que pour les messages de longueur $l = 1$ le chiffre de César est parfaitement sûr au sens de Shannon.

2. Montrer que pour les messages de longueur $l \geq 2$ le chiffre n'est plus parfaitement sûr. Pour cela analysez l'exemple suivant. Soient le message $m = AB$ et le chiffré $c = DM$. Montrer que $P[M = m|C = c] = 0$ tandis que $P[M = m] \neq 0$.

Exercice 5 (Cryptanalyse du chiffre de César)**.** La méthode d'analyse des fréquences a été inventé par le savant arabe AL-Kindi au IX-ème siècle. On suppose que l'on connaît la langue du texte clair et que l'on dispose du message chiffré. Dans le ca de chiffre de César on cherche à déterminer le paramètre k, clé du chiffre. La méthode consiste à comparer l'histogramme d'occurrences des caractères du chiffré avec la table des fréquences d'occurrence des caractères de la langue du texte clair. Voici la table des fréquences de la langue française.

Lettre	Fréquence %	Lettre	Fréquence %
A	8.4	N	7.13
B	1.06	O	5.26
C	3.03	P	3.01
D	4.18	Q	0.99
E	17.26	R	6.55
F	1.12	S	8.08
G	1.27	T	7.07
H	0.92	U	5.74
I	7.34	V	1.32
J	0.31	W	0.04
K	0.05	X	0.45
L	6.01	Y	0.3
M	2.96	Z	0.12

1. Votre mission, si vous l'acceptez, consiste à déchiffrer le message secret de votre binôme. Chacun de vous va composer un message de son choix. Pour le chiffrer, utiliser `http://www.bibmath.net/crypto/substi/cryptcesar.php3`l'applet java sur ce site. Envoyez le chiffré obtenu à votre voisin (par e-mail ou chat). Ensuite chacun cherchera à trouver la clé et le message clair par analyse fréquentielle.
2. Et maintenant, déchiffrez ceci :

UHGCHNK. GHNL OHNL IKHIHLHGL MKHBL PTZHGL IHNK ITKMBK TN STGBSBUTK. OHMKX TOBHG OT ITKMBK ET HN OHNL OHNEXS. NG OKTB OTNMHNK OXNM MHNCHNKL OHEXK ATNM IHNK OHBK LT IKHBX. NG SHFUB FTKVATBM XG SBZSTZTGM LNK ET KHNMX.

PARTIE IV. EXERCICES SUPPLÉMENTAIRES

Exercice 6 (Retour au problème de la fausse pièce)**.** Supposons que nous avons 12 pièces toutes identiques sauf une de poids différent. Nous ne savons pas si la fausse pièce est plus lourde ou plus légère que les autres.

On dispose d'une balance à deux plateaux. **L'objectif** est de déterminer avec le moins de pesées possible la fausse pièce **et** si elle plus légère ou plus lourde.

Nous avons établi la dernière fois qu'il y avait $24 = 2 \times 12$ réponses possibles. Elles sont toutes équiprobables. Soit X la variable aléatoire associée.

Dans cette exercice nous allons voir comment la notion d'entropie permet de choisir la stratégie optimale de pesées. Nous allons nous intéresser au choix de la première pesée. Le principe pourra ensuite être appliqué récursivement.

Nous avons déjà vu qu'il était inutile de peser 6 pièces contre 6. On a alors à choisir parmi les possibilités suivantes : 5 contre 5, 4 contre 4, 3 contre 3, 2 contre 2 et 1 contre 1.

1. Quelle est l'entropie $H(X)$? (utiliser l'équiprobabilité pour donner la réponse).
2. On associe à la première pesée une variable aléatoire Y qui prend trois valeurs : $\{G, D, E\}$. Pour chaque choix de nombre de pièces à peser calculer $H(X|Y)$ et ensuite le gain d'information apporté en observant Y : $I(X|Y) = H(X) - H(X|Y)$. En déduire le meilleur choix : celui qui maximise le gain d'information.

   Indication. On peut utiliser l'équiprobabilité des réponses pour calculer l'entropie conditionnelle moyenne $H(X|Y)$. En effet, l'observation de Y partage l'ensemble de valeurs possibles de X en trois sous-ensembles : A, B et C. Etant donné que les distributions conditionnelles de X sachant A (ou sachant B ou C) sont toujours uniformes, les entropies correspondantes seront $H(X|Y = G) = \log(Card(A))$, $H(X|Y = D) = \log(Card(A))$, $H(X|Y = E) = \log(Card(A))$. Alors on a

$$H(X|Y) = P(Y = G)H(X|Y = G) + P(Y = D)H(X|Y = D) + P(Y = E)H(X|Y = E)$$

   Prenons par exemple le cas d'une pesée de 2 pièces contre 2. La balance penchera vers la gauche ($Y = G$) dans 4 cas sur 24 (soit l'une des pièces à gauche est plus lourde, soit l'un des pièces à droite est plus légère). Elle penchera à droite dans 4 autres cas. Elle restera donc en équilibre dans 16 cas restants. L'entropie conditionnelle moyenne est

$$H(X|Y) = \frac{4}{24}\log 4 + \frac{16}{24}\log 16 + \frac{4}{24}\log 4 = \frac{90}{24} \simeq 3.33$$