Cryptographie et Sécurité
Série 3 : Entropie

02 Octobre 2019

A rendre exclusivement sur papier, au plus tard au début de la séance du **09 Octobre 2019 et avant 17h15**.

Toutes vos réponses doivent être justifiées.

Définitions de l'Entropie

Soit une source d'information qui est réprésentée par une variable aléatoire discrète X comportant n symboles, chaque symbole x_i ayant une probabilité p_i d'apparaître. Alors l'entropie de la source, notée $H(X)$, est définie comme :

$$H(X) = -\sum_{i=1}^{n} p_i log_2(p_i) = \sum_{i=1}^{n} p_i log_2(\frac{1}{p_i})$$

L'entropie est maximale lorsque toutes les probabilités p_i sont égales. Cela correspond à l'intuition d'une incertitude maximale.

L'entropie jointe pour deux variables aléatoires X et Y , notée $H(X, Y)$ est simplement l'entropie pour toutes les paires x,y possibles :

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

Enfin, l'entropie conditionnelle de X étant donné Y, est définie comme suit :

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

L'entropie conditionnelle est très importante en Cryptographie. Par exemple, la valeur $H(Plaintext|Ciphertext)$ nous donne une idée de l'incertitude qu'on a sur le texte clair lorsqu'on observe le texte chiffré.

Définition du Codage de Huffman

Le codage de Huffman à pour but d'attribuer les bits de façon efficace aux variables que l'on veut coder. Le but est de construire un arbre binaire de la façon suivante : trier les variables par fréquence, puis fusionner les deux moins fréquentes en une nouvelle variables, et réitérer jusqu'à n'avoir plus que deux variables, que l'on trie.

Exemple : Soit les variables A,B,C,D avec probabilités d'apparition respectives 0.2, 0.5, 0.2, 0.1 :

- On ordonne les variables de la plus probable à la moins probable : B, A, C, D.
- On met les deux dernières ensembles : $E = C, D$; $P(E) = 0.3$.
- On trie à nouveau : B, E, A.
- On fusionne les deux dernières : $F = E, A$;$P(F) = 0.5$.
- Dernier tri : B, F.

On obtient donc le codage suivant :

- $B = 0$, $F = 1$.
- Puis on décompose $F = E, A$: $E = 10$, $A = 11$.
- Puis on décompose $E = C, D$: $C = 100$, $D = 101$.

On obtient donc $B = 0$, $A = 11$, $C = 100$, $D = 101$. L'idée est de réduire au maximum l'expression des éléments les plus fréquents.

On peut calculer l'efficacité d'un codage de la manière suivante :

Si X est une variable aléatoire qui représente les A,B,C,D définis précédemment avec leurs probabilités, l'efficacité E est définie comme le rapport de l'entropie et de la taille moyenne pour exprimer X (nombre moyen de bits):

$$E = \frac{H(X)}{\sum_{x \in X} P(x) \cdot nombredebitspourencoderx}$$

Exercice 1 : Entropie

- Soit X une variable aléatoire prenant les valeurs 0 et 1, avec probabilité p_i de donner 0 (et $1 - p_i$ de donner 1, évidemment). Dessinez le graphique de l'entropie selon p_i, pour $p_i = 0$, $p_i = 0.1$, $p_i = 0.2$, ..., $p_i = 0.9$, $p_i = 1$.
- Que signifierait une entropie nulle ? Et une entropie infinie ? D'après les définitions, ces deux cas sont-ils possibles ?

Exercice 2 : Calculs d'Entropie

Soit le système de cryptage symétrique, assez simple, suivant : l'ensemble des plaintexts $P = \{m_1, m_2, m_3\}$, l'ensemble des ciphertexts $C = \{1, 2, 3, 4, 5\}$, et l'ensemble des clés $K = \{k_1, k_2, k_3\}$, avec les règles de chiffrement suivantes :

	m_1	m_2	m_3
k_1	3	2	1
k_2	4	5	2
k_3	1	4	3

On suppose les clés équiprobables, et les plaintexts qui suivent ces probabilités d'apparitions :

$$p(m_1) = \frac{1}{4} \quad p(m_2) = \frac{3}{20} \quad p(m_3) = \frac{6}{10}$$

Calculez les entropies suivantes :

- $H(P)$.
- $H(K)$.
- $H(C)$.
- $H(P|C)$.

Exercice 3 : Codage de Huffman

Soit un générateur aléatoire qui génère les caractères A et B (caractères ASCII 8-bit) avec probabilité 0.3 et 0.7 respectivement. Soit $S^2 = \{AA, AB, BA, BB\}$ la source qui émet les caractères deux à deux (chaque caractère émis par S avec les probabilités définies ci-dessus).

Calculez :

- $H(S)$.
- $H(S^2)$.
- Calculez le codage de Huffman pour S^2.
- Comparez l'efficacité du codage de Huffman avec l'efficacité du codage initial de S^2.
- Quel est l'intêret principal du codage de Huffman ?

Exercice 4 : Entropie d'un Flux Biaisé

Soit un générateur "aléatoire" G pas très performant, qui génére des 0 et des 1, avec les probabilités $P(0) = 0.5 + \delta$ et $P(1) = 0.5 - \delta$.

On propose de créer un autre générateur aléatoire A comme suit : on génère deux bits avec G. Si la séquence est 00 ou 11, on l'ignore et on recommence. Si la séquence est 01, on génère un 0. Si la séquence est 10, on génère un 1.

- Calculez la probabilité de chaque paire (00, 11, 01, 10) d'être générée par G.
- Calculez la probabilité de 0 et de 1 avec le nouveau générateur A.
- Calculez combien de bits G doit générer pour que A génère x bits.
- Quel est l'avantage de cette méthode ? Quel est l'inconvénient ?
- Que se passe t-il si l'algorithme utilise les paires superposées (première paire avec bits 1 et 2, puis deuxième paire avec bits 2 et 3, puis avec bits 3 et 4, etc) ?