ENS Cachan - UPMC

Introduction à la Physique Statistique LP357
2014-2015

TD 4 : Entropie et compression de données

En 1948, tandis qu'il travaillait aux Laboratoires Bell, l'ingénieur en génie électrique Claude Shannon formalisa mathématiquement la nature statistique de "l'information manquante" dans les signaux des lignes téléphoniques. Pour ce faire, il développa le concept général d'entropie et d'information. Dans ce td, nous allons étudier cette entropie (qui se trouve être la même, à une constante multiplicative près, que celle de Gibbs) et ses liens avec la compression de données.

1 Surprise, incertitude et information

Considérez une variable aléatoire X qui peut prendre une valeur parmi N possibles dans un alphabet χ qui contient L lettres. La probabilité de prendre une valeur $X = x_i \in \chi$ est p_i. Pour cet exercice, considérons par exemple un alphabet de quatre lettres $\chi = \{A, B, C, D\}$ avec les probabilités $p_A = 1/2$, $p_B = 1/4$, $p_C = 1/8$, $p_D = 1/8$.

▷ **1-1** Nous tirons une valeur $X = x_i$. Plus cette valeur est improbable, plus nous sommes surpris de la voir apparaître. Définissons donc la "surprise" comme $\log_2 1/p_i$. Appelons, avec Shannon, la surprise moyenne "l'incertitude" $H[X]$. Comment s'écrit l'incertitude en général ? Que vaut-elle dans notre exemple ?

▷ **1-2** Montrez que l'incertitude $H[X]$ satisfait $0 \leq H[X] \leq \log_2 L$; que la valeur 0 est atteinte si et seulement si une seule lettre a une probabilité non nulle ; et que la valeur $\log_2 L$ est atteinte si et seulement si toutes les lettres sont équiprobables.

2 Entropie de Shannon et questions OUI/NON

Nous voulons maintenant répondre à la question : à quel point est-il difficile de deviner la valeur d'une variable aléatoire ?

▷ **2-1** J'ai tiré au hasard l'une des lettres A,B,C,D avec les probabilités précédentes. Vous avez le droit de poser des questions, du type "Cette lettre est-elle dans l'ensemble $\{x_1, x_2, \ldots\}$?", auquelles je répondrai par oui ou par non. Votre but est de trouver le plus vite possible la lettre tirée. Combien de questions devez vous poser en moyenne avant de découvrir la lettre que j'ai tirée ? Comparez ce nombre avec l'entropie de l'alphabet.

▷ **2-2** Considérons un nouvel alphabet de 2^b lettres équiprobables (avec b entier). Combien des questions est-il nécessaire de poser en moyenne ? Comparez avec l'entropie du système.

▷ **2-3** (Théorème de Shannon (1948)) Considérons enfin un alphabet général avec A lettres x_i, $i = 1 \ldots A$ avec probabilités p_i. Comment généraliser la stratégie précédente ? Combien des questions devez vous poser en moyenne pour découvrir la lettre tirée ? Pour simplifier, on fera l'hypothèse qu'il soit toujours possible de partager les lettres en deux groupes tels que la probabilité de chacun des deux groupes est $1/2$.

3 Compression de données

Vous avez un fichier avec N valeurs aléatoires $x_i \in \chi$, l'alphabet χ contenant L lettres. Ce fichier prend une place de $N \log_2 L$ bits dans votre disque dur.

▷ **3-1** Adoptons la stratégie suivante, en utilisant les questions OUI/NON. On écrit un fichier où, au lieu de garder chaque lettre, on garde, pour chaque lettre, les réponses aux question de l'exemple précédent avec OUI= 1 et NON=0. Quelle est la taille de ce nouveau fichier ? Quel est le facteur de compression (rapport entre la taille du fichier avant et après compression) ?

▷ **3-2** Comparez ce résultat avec les performances de votre programme de compression favori (zip, rar ...). Pour cela, utiliser les fichiers ci-joints, qui sont des suites de 10^4 caractères Ascii [1] tirés au hasard avec les alphabet suivant :

(i) *uniform.txt* : caractères tirés au hasard parmi L'ENSEMBLE des 256 caractères possibles.
(ii) *half.txt* : caractères tirés au hasard parmi LA MOITIE (soit 128) des caractères possibles.
(iii) *abcd.txt* : caractères tirés au hasard parmi ABCD.
(iv) *abcd2.txt* : caractères tirés au hasard parmi ABCD avec les probabilités de l'exercice 2.

4 L'entropie de l'anglais

▷ **4-1** Les fréquences des lettres dans l'anglais sont disponibles sur :
http ://en.wikipedia.org/wiki/Letter_frequencies. Calculez le rapport de compression de l'anglais (négligez les espaces, majuscules et autres caractères spéciaux.)

▷ **4-2** Comparez vos résultats aux meilleurs compresseurs disponibles :
http ://en.wikipedia.org/wiki/Hutter_Prize. D'où vient la différence observée ?

▷ **4-3** Estimez (et commentez vos résultats) l'entropie de l'anglais grace à l'experience numérique :
http ://www.math.psu.edu/dlittle/java/informationtheory/entropy/index.html

1. On rappelle que les caractères Ascii sont codés sur 8 bits et prennent 256 valeurs possibles, voir *http ://fr.wikipedia.org/wiki/ASCII*