MIM1 - Probabilités et applications- TD 6

Emmanuelle Lebhar
elebhar@ens-lyon.fr

Entropie de Shannon et codage
7 mars 2005

Exercice 1 (Entropie de Shannon et codage)
Soit $A = \{\alpha_1, ..., \alpha_D\}$ un alphabet $M = \{x_1, ..., x_k\}$ un ensemble d'objets appelés messages. Un codage est une application de M dans A^* qui à un message x_i associe son code c_i de longueur l_i.

1. **Inégalité de Kraft.** Montrer que si le codage h a la propriété du préfixe alors
$$\sum_{i=1}^{k} D^{-l_i} \leq 1.$$
2. Supposons l'inégalité précédente satisfaite. Montrer que pour tout alphabet A de D lettres et tout ensemble de k messages il existe un codage h ayant la propriété du préfixe et dont les codes sont de longueurs (l_i).
3. **Borne de Shannon sur la longueur moyenne d'un code préfixe.** Supposons maintenant que les messages proviennent d'une source aléatoire, le message x_i ayant la probabilité p_i. La longueur moyenne de codage est alors définie par :
$$L(h) = \sum_{i=1}^{k} p_i l_i.$$
Soit L_{inf} la plus petite de ces longueurs pour un code préfixe. Nous allons estimer L_{inf}. Shannon définit l'entropie d'un événement comme la quantité de surprise $-\log p_i$ qu'aurait un observateur lorsqu'il découvre la réalisation de cet événement. Plus cet événement est improbable plus l'observateur sera surpris. Si on fait la moyenne sur tous les événements possibles on obtiendra l'entropie du système :
$$H_D(p) = -\sum_{i=1}^{k} p_i \log_D p_i.$$
C'est la quantité d'information relative à la distribution $p = (p_1, ..., p_k)$.
Montrer que
$$H \leq L_{inf} \leq H + 1.$$
Indications : Chercher L_{inf}, c'est chercher les l_i^* entiers qui minimisent $L(h)$ en satisfaisant la contrainte de Kraft. On pourra commencer par montrer que le minimum éventuel vérifie $\sum_{i=1}^{k} D^{-l_i^*} = 1$. Chercher ensuite les l_i^* à l'aide, par exemple, d'un changement de variable.

Exercice 2 (Algorithme de Shannon)

On fixe $D = 2$ et on suppose $p_1 \geq p_2 \geq ... \geq p_k$. On pose $q_s = \sum_{i=1}^{s-1} p_i$. La méthode de Shannon consiste à écrire pour chaque i un développement dyadique de q_i :

$$q_i = \frac{a_1(i)}{2^1} + \frac{a_2(i)}{2^2} + \frac{a_3(i)}{2^3} + ...$$

et à prendre pour code de x_i : $a_1(i)...a_{l_i}(i)$ avec $l_i = \lceil -\log_2 p_i \rceil$.

1. Donner le code obtenu pour $p_1 = 0,27$, $p_2 = 0,23$, $p_3 = 0,2$, $p_4 = 0,15$, $p_5 = 0,1$ et $p_6 = 0,05$. Calculer L et la comparer à l'entropie.
2. Montrer que ce code est préfixe.
3. Montrer que la longueur moyenne L vérifie $H_2(p) \leq L \leq H_2(p) + 1$.
4. Ce codage est-il optimal ?

Exercice 3 (Algorithme de Fano)

Fixons $D = 2$ et supposons $p_1 \geq p_2 \geq ... \geq p_k$. On regroupe les u premiers objets où u est le plus petit entier tel que

$$p_1 + ... + p_u \geq \frac{1}{2}$$

Nous obtenons ainsi une partition M de l'ensemble des objets en deux sous-ensemble M_0 et M_1. Soient π_0 et π_1 les probabilités respectives de ces ensembles. Nous appliquons à M_0 l'algorithme de dichotomie pour obtenir une partition $M_0 = M_{00} + M_{01}$ où $M_{00} = \{x_1, ..., x_v\}$ et v est le plus petit entier tel que $p_1 + ... + p_v \geq \pi_0/2$. On applique à M_1 ce même algorithme et on obtient la partition $M_1 = M_{10} + M_{11}$. On continue l'algorithme ainsi de suite jusqu'à ce qu'on ne puisse plus appliquer la dichotomie, ce qui se produit quand un ensemble M_a ne comporte plus qu'un élément et a est alors le code de cet élément.

1. Refaire les questions 1,2, et 3 de l'exercice précédent.

Exercice 4 (Algorithme mystère)

Donner un algorithme pour obtenir un codage préfixe optimal. Le faire tourner sur l'exemple de la question 1 de l'exercice 2 et comparer la longueur moyenne L obtenue avec l'entropie de la distribution.