Probabilités et applications à l'algorithmique, TD numéro 7
Entropie de Shannon et codage, théorie des questionnaires
Natacha Portier
lundi 18 avril 2002

Question 1 : l'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.

Question 1,2

Montrer que ce code est préfixe.

Question 1,3

Montrer que la longueur moyenne L vérifie $H_2(p) \leq L \leq H_2(p) + 1$.

Question 1,4

Ce codage est-il optimal ?

Question 2 : théorie des questionnaires

Imaginons que nous avons un objet pris parmi les x_i et que nous cherchons à l'identifier. Pour cela nous posons des questions auquelles les réponses possibles sont des lettres de l'alphabet $A = \{\alpha_1, ..., \alpha_D\}$. Les questions sont posées les unes après les autres et dépendent des réponses déjà obtenues. Nous avons donc un questionnaire $Q = \{Q_m / m \in A^*\}$. La question Q_m appliquée à l'objet x_j donne la réponse $\alpha_{m,j}$, ce qu'on note : $Q_m(x_j) = \alpha_{m,j}$. Le questionnaire Q appliqué à x_i fournit donc la suite de réponses $Q(x_i) = (a_{1,i}, a_{2,i}, a_{3,i}, ...)$ avec $a_{k,i} = \alpha_{a_{1,i} ... a_{k-1,i}, i}$. Une règle d'arrêt est un sous-ensemble T de A^*. Le couple (Q, T) où Q est un questionnaire et T une règle d'arrêt est un identificateur. Il lui correspond un codage h définit par : si $Q(x_i) = (a_{1,i}, a_{2,i}, a_{3,i}, ...)$ et $l_i = \inf\{n | a_{1,i}...a_{n,i} \in T\}$ alors $h(x_i) = a_{1,i}...a_{l_i,i}$. Un identificateur est admissible si h est bien défini et injectif.

Question 2,1

Soit (Q,T) un identificateur admissible. Montrer que h est préfixe.

Question 2,2

Soit h un codage préfixe. Donner un identificateur admissible (Q,T) correspondant au codage h.

La longueur moyenne du code h est le nombre moyen de questions à poser pour identifier un objet choisi au hasard selon la distribution p parmi les x_i. L_{inf} est donc le nombre moyen minimum de questions à D réponses que l'on doit poser pour identifier un objet tiré au sort selon la distribution p parmi k objets distinguables entre eux.

Question 2,3 : un problème de billes

On nous présente 15 billes rigoureusement identiques d'aspect et identiques en poids sauf pour l'une d'elles qui est légèrement plus lourde (un centième de poids en plus). Comment faire pour retrouver le plus rapidement possible en moyenne la bille de poids différent ? Et dans le pire des cas ? Et si on a douze billes et qu'on ne sait pas si la fausse est plus lourde ou plus légère ? Et si on le sait ? Et avec une balance qui peut juste renseigner sur l'éventuelle égalité de deux poids ?

Question 3 : H est bien une quantité d'information

Qu'est-ce qui justifie que H soit appelé quantité d'information plutôt que L_{inf} ?

Soit $p = (p_1, ..., p_k)$ la distribution de probabilité d'entropie H des lettres $x_1, ..., x_k$. Soit n un entier. Calculer la distribution de probabilité $p^{(n)}$ des mots de longueur n et son entropie $H(p^{(n)})$. Soit $L_{inf}^{(n)}$ la longueur moyenne du codage optimal pour $p^{(n)}$. Donner un encadrement de la "quantité d'information par symbole" $L_{inf}^{(n)}/n$ et conclure.

Question 4 : entropie

Les deux questions suivantes sont indépendantes. La base logarithmique utilisée pour le calcul de l'entropie est quelconque.

Question 4,1

Soit X une variable aléatoire qui prend les valeurs $x_1, ..., x_n$ avec la distribution finie de probabilités $p = (p_1, ..., p_n)$ et Y une variable aléatoire indépendante de X qui prend les valeurs $y_1, ..., y_l$ avec la distribution finie de probabilités $q = (q_1, ..., q_l)$. On appelle

$p \otimes q$ la distribution de probabilité du couple (X, Y). Calculer l'entropie de $p \otimes q$ en fonction de celles de p et q.

Question 4,2

Soit $p = (p_1, p_2)$ une distribution finie de probabilités. Donner les valeurs de p qui minimisent et maximisent son entropie H. Généraliser à une distribution $p = (p_1, ..., p_n)$. (par convention , $H(p_1, ..., p_{n-1}, 0) = H(p_1, ..., p_{n-1})$).