Extension de la segmentation sur un tableau de données à entrées aléatoires

Jean-Pascal ABOA YAPO

Laboratoire Ceremade, Universite Paris IX Dauphine (e-mail : aboa@pi.ceremade.dauphine.fr)

Nous présentons dans ce travail deux méthodes de segmentation sur un tableau de données à entrées aléatoires. Dans ce type de tableau, l'élément réel habituel du modèle tabulaire classique est remplacé par une variable aléatoire de carré intégrable. La nature du tableau induit une extension de la notion de question permettant le partitionnement. Nous proposons deux méthodes basées sur deux types de question, l'une aléatoire, l'autre "symbolique". A partir d'une fonction d'impureté définie sur les nœuds de l'arbre binaire, on généralise ainsi le critère de Gini [Breiman, 84] et la notion d'entropie [Quinlan, 86] (pour l'extension de l'algorithme CART [Breiman, 84]).

Approche gibbsienne des mesures gaussiennes sur R^{Z^d} : de la statique à la dynamique.

Olivier GARET

Université de Lille I, F-59655 Villeneuve d'Ascq (e-mail : garet@jacta.univ-lille1.fr)

En théorie des probabilités, l'approche gibbsienne consiste à décrire la structure de systèmes infinis de variables aléatoires par les interactions qui les lient. Nous nous intéressons ici à l'étude d'interactions quadratiques pour des variables à valeurs réelles indexées par Z^d.

L'approche traditionnelle (D.L.R.) consiste en la détermination des mesures dont les lois conditionnelles sur des volumes finis de Z^d coïncident avec une spécification donnée au préalable, ici de type gaussien. Prolongeant les travaux de Dobrushin, nous donnons des conditions nécessaires et suffisantes simples et explicites pour l'existence et l'unicité d'une solution – appelée mesure de Gibbs – au sein d'une classe de mesures à support "tempéré".

Une autre approche des mesures de Gibbs est de les considérer comme mesures d'équilibres, *i.e.* limites en temps d'un système stochastique associé. Nous montrons que la non-unicité des mesures de Gibbs gaussiennes se traduit par une influence de la condition initiale sur le comportement asymptotique du système gradient associé, tandis qu'il y a ergodicité du système en cas d'unicité de la mesure de Gibbs.

Plus précisément, nous exhibons un algorithme stochastique à pas décroissants permettant de construire chaque mesure de Gibbs extrémale comme limite asymptotique pour une certaine condition initiale déterministe. De surcroît, nous déterminons une partie du bassin d'attraction de chaque mesure de Gibbs extrémale.

Martingales et grandes déviations pour les arbres binaires de recherche.

Jean JABBOUR

Laboratoire LAMA, UVSQ (Email: jabbour@math.uvsq.fr)

The study of the binary search trees (b.s.t.) is an important problem related to the analysis of algorithms. These b.s.t's are the trees generated by the Quicksort algorithm in order to sort n numbers x_1, x_2, ..., x_n. The algorithm proceeds by inserting successively x_1, x_2, ..., x_n as labels of the nodes of an infinite binary tree in the following way : you compare the number x_j that you wish to insert

with the root label, if there is no label (which happens for $j = 1$) you put x_j as the root label; if x_j is less (resp. greater) than the root label then you restart the procedure with the lefthand (resp. righthand) son as the root; you go on in this way until reaching an unlabelled node, which will receive x_j as a label. Thus, the x_j are sorted from left to right in an increasing order. Furthermore, if you want to know whether a number x is a label of the tree, you just have to go through the tree as if to insert the label x until either finding an unlabelled node (an unsuccessful search) or finding a label equal to x (a successful search).

The precise definition of the b.s.t. generated by this algorithm is the binary tree formed by the labelled nodes (called the "nodes" of the b.s.t.) and their unlabelled sons (called "leaves"). By convention, the b.s.t. generated in the case $n = 0$ is the "empty tree" in the sense that it has no node and just one leaf at the root. In the study, the x_j are treated as random variables. The generated b.s.t. T_n depends only on the order of the x_j, that is the permutation $\sigma_n \in \mathcal{S}_n$ satisfying $x_{\sigma_n(1)} < x_{\sigma_n(2)} < \dots < x_{\sigma_n(n)}$. Thus, T_n is also a random variable and its law depends only on the law of σ_n; as usual we suppose it to be an uniform law over $\mathcal{S}_n$.

Concerning these variables, the main results so far established are the following :

$$\frac{h_n}{\log n} \xrightarrow{P} c \quad \text{and} \quad \frac{h'_n}{\log n} \xrightarrow{P} c' \tag{1}$$

$$\frac{k_n - 2\log n}{\sqrt{2\log n}} \xrightarrow{\mathcal{L}} \mathcal{N}([0,1]) \tag{2}$$

where $c \simeq 4.311$ and $c' \simeq 0.373$ are universal constants respectively lower and greater than 1 of the equation $x \log \frac{2e}{x} = 1$.

Our purpose is to get deeper results. Our main improvement concerns (1) : an almost sure large deviation result is established on the "empirical measure" of the leaves.

Une application des modèles d'urnes aux bases de données : la formule de Yao et ses extensions.

Laurent NEMIROVSKI

Laboratoire Prism, UVSQ (Email: Laurent.Nemirovski@prism.uvsq.fr)

Nous montrons comment la formule bien connue de Yao, qui donne le nombre moyen de blocs à lire lors d'une requête, connaissant la taille du résultat de cette requête, rentre dans le cadre des allocations aléatoires, et peut être modélisée par un modèle d'urnes adéquat. Ceci nous permet tout d'abord d'obtenir une information complète sur la loi de probabilité du nombre de blocs sélectionnés, et ensuite d'envisager différentes extensions; nous traitons ici le cas où les blocs peuvent avoir des capacités distinctes.