Probabilités et applications

Pascal Koiran

Mai 2003

1 Estimation

On cherche à déterminer (approximativement) un intervalle inconnu $s = [a, b] \subseteq [0, 1]$ à partir de n "observations" $(X_1, \epsilon_1), \ldots, (X_n, \epsilon_n)$. On suppose que les X_i sont indépendants et uniformément distribués dans $[0, 1]$. L'étiquette ϵ_i vaut 1 si $X_i \in [a, b]$, et vaut 0 sinon. Sur la base de ces observations on va proposer un "intervalle hypothèse" h. On note $err(h)$ la mesure de Lebesgue de la différence symmétrique $h \Delta s$. On dit que h est consistant avec les observations si $X_i \in h \Leftrightarrow \epsilon_i = 1$ pour tous les X_i.

1. Dans cette question on suppose que s appartient à un ensemble fini S d'intervalles. On suppose S connu, mais bien sûr s reste inconnu. Supposons enfin que h est produit par l'algorithme suivant : retourner un intervalle h choisi de manière arbitraire parmi les éléments de S qui sont consistants avec les observations.

   Montrez que la probabilité de retourner un h tel que $err(h) \geq \epsilon$ est majorée par $|S|(1 - \epsilon)^n$.

2. On abandonne l'hypothèse que s appartient à un ensemble fini S d'intervalles : s peut maintenant être un intervalle fermé quelquonque de $[0, 1]$. On suppose que h est produit par l'algorithme suivant : retourner $h = [\alpha, \beta]$ avec $\alpha = \min\{X_i; \epsilon_i = 1\}$ et $\beta = \max\{X_i; \epsilon_i = 1\}$.

   Montrez que la probabilité de retourner un h tel que $err(h) \geq \epsilon$ est majorée par $2(1 - \epsilon/2)^n$.

3. Reprendre les deux questions précédentes dans le cadre suivant : on ne cherche plus à déterminer des intervalles, mais des rectangles de $[0, 1]^2$ de la forme $[a, b] \times [c, d]$. Dans chacun des deux cas (le rectangle appartient à un ensemble fini connu, ou le rectangle est arbitraire) on précisera l'algorithme qui produit l'hypothèse h, et on donnera une borne sur la probabilité que $err(h) \geq \epsilon$.

4. Reprendre la première question en abandonnant l'hypothèse que les X_i sont uniformément distribués : on suppose toujours que les X_i sont indépendants et identiquement distribués, mais cette distribution commune P n'est plus nécessairement uniforme. Comment définir $err(h)$ dans ce cadre pour retrouver les résultats de la première question ? Y a t-il encore d'autres hypothèses dont vous pouvez vous passer ?

2 Géométrie algorithmique

1. Soit X un ensemble fini et $\mathcal{R}$ un ensemble de parties de X de cardinal au moins 2. On dit qu'une partie S de X est un ϵ-réseau si $S \cap R \neq \emptyset$ pour tout $R \in \mathcal{R}$ tel que $|R| > \epsilon|X|$.

   Montrez qu'il existe un ϵ-réseau de taille au plus $\ln|\mathcal{R}|/\epsilon$.

2. D'abord quelques rappels de géométrie : on appelle triangulation d'un polygone convexe P du plan un ensemble de triangles dont l'union est égale P. Par exemple, si v_1 est un sommet de P et si $v_1v_2\cdots v_rv_1$ est la suite des sommets de P obtenue en "tournant dans le sens des aiguilles d'une montre" à partir de v_1, les triangle $v_1v_2v_3$, $v_1v_3v_4$, $v_1v_4v_5$,...,$v_1v_{r-1}v_r$ forment une triangulation de P. Dans la suite on ne considérera que des triangulations de cette forme.

   Soit H un ensemble de n droites du plan. Les composantes connexes du complémentaire de l'union des droites de H sont soit bornées (ce sont alors des polygones convexes), soit des régions du plan convexes mais non bornées. Par définition "l'arrangement triangulé ΔH" est l'ensemble des triangles obtenus en triangulant toutes les régions bornées (pour simplifier on ne s'intéresse pas aux régions non bornées). On notera qu'il existe en général plusieurs arrangements triangulés pour un même H.

   On s'intéresse au problème de localisation de point : étant donné un point x qui n'appartient à aucune des n droites, on souhaite déterminer à quelle région il appartient. Cela revient à déterminer sa position par rapport à chacune des droites ("au dessus" ou "en dessous" ; pour une droite verticale, "à gauche" ou "à droite"). Pour cela un algorithme naif effectuerait les n tests l'un après l'autre. Le but de cette question est de construire à partir de H une structure de données qui permettra de résoudre ce problème plus rapidement. Pour simplifier on ne considérera que les entrées x situées à l'intérieur d'un triangle abc fixé, et on supposera que les droites ab, bc et ca appartiennent à H. Ceci pour éviter d'avoir à traiter le cas des régions non bornées.

   Le principe de fonctionnement de la structure de données est le suivant. Au premier niveau on choisit un sous-ensemble T de H qui contient les trois droites ab, bc et ca, puis on construit l'arrangement triangulé ΔT. Etant donné un point x, on détermine à quel triangle de ΔT il appartient.

   Expliquez comment on peut trouver ce triangle en temps $O(|T|)$.

3. Montrez qu'il existe un T de taille $O(\log n)$ tel que l'intérieur d'aucun triangle de ΔT n'est intersecté par plus de $n/2$ droites de H.

   Indication : Soit Θ l'ensemble des triangles dont chaque sommet est l'intersection de deux droites de H. On associe à chaque élément θ de Θ l'ensemble R_θ des droites de H qui intersectent l'intérieur de θ. On pourra s'intéresser à la famille $(R_\theta)_{\theta\in\Theta}$ de parties de H.

4. Si on choisit T comme ci-dessous, en quoi cela nous aide t-il à localiser x ? Expliquez comment fonctionnent les niveaux suivants de la structure de données. On ne demande pas de décrire la structure dans les moindres

détails, mais d'expliquer son principe de fonctionnement et l'avantage par rapport à l'algorithme naif de localisation.

5. Soit $S = (s_i)_{1 \leq i \leq s}$ une suite d'éléments de X. Pour $R \in \mathcal{R}$, on note R_S le nombre d'indices i tels que $s_i \in R$.
   On dit que S est une ϵ-approximation si
   $$\left| \frac{R_S}{s} - \frac{|R|}{|X|} \right| \leq \epsilon$$
   pour tout $R \in \mathcal{R}$.
   Quel lien y a t-il entre les notions d'ϵ-réseau et d'ϵ-approximation ?
6. Montrez qu'il existe une suite de longueur s qui est une ϵ-approximation avec $s = O(\ln |\mathcal{R}|/\epsilon^2)$.
   On pourra utiliser sans démonstration l'inégalité suivante : Soit $p \in [0, 1]$ et $Y_1, \ldots, Y_n$ une suite de variables aléatoires indépendantes telles que $Y_i = 1 - p$ avec probabilité p et $Y_i = -p$ avec probabilité $1 - p$.
   Alors $\Pr[|Y| > a] < 2e^{-2a^2/n}$ pour tout $a > 0$, où $Y = Y_1 + \cdots + Y_n$.

3 Marche au hasard

Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires indépendantes telles que

$$P[X_i = -1] = P[X_i = 1] = 1/2.$$

On note $S_n = \sum_{i=1}^{n} X_i$ pour $n \geq 1$, et $S_0 = 0$.

1. Soit N le premier instant tel que $S_N > 0$. Montrez que pour $|s| < 1$, $2E(s^N) = s(1 + E(s^N)^2)$.
   Indication : Dans le cas où $X_1 = -1$, on pourra s'intéresser au premier instant N_1 tel que $S_{N_1} = 0$.
2. Soit ϕ_n la probabilité que $N = n$. Donnez une formule close pour la fonction génératrice $\Phi(s)$ des ϕ_n.
3. Soit R l'instant du premier retour en 0 : c'est le premier instant $n > 0$ tel que $S_n = 0$. Soit f_n la probabilité que $R = n$. Quelle est la série génératrice $F(s)$ des f_n ? On pourra chercher un lien entre $F(s)$ et $\Phi(s)$.
4. Calculer la probabilité u_n que $S_n = 0$. En déduire que la série génératrice $U(s)$ des u_n est égale à $1/\sqrt{1 - s^2}$.
5. En s'appuyant sur une relation entre $U(s)$ et $F(s)$, en déduire que u_{2n} est égal à la probabilité que $R > 2n$.