Plus Proche Voisin

Épreuve pratique d'algorithmique et de programmation
Concours commun des Écoles normales supérieures

Durée de l'épreuve: 3 heures 30 minutes

Juin/Juillet 2018

ATTENTION !

N'oubliez en aucun cas de recopier votre u_0
à l'emplacement prévu sur votre fiche réponse

Important.

Il vous a été donné un numéro u_0 qui servira d'entrée à vos programmes. Les réponses attendues sont généralement courtes et doivent être données sur la fiche réponse fournie à la fin du sujet. À la fin du sujet, vous trouverez en fait deux fiches réponses. La première est un exemple des réponses attendues pour un $\widetilde{u_0}$ particulier (précisé sur cette même fiche et que nous notons avec un tilde pour éviter toute confusion !). Cette fiche est destinée à vous aider à vérifier le résultat de vos programmes en les testant avec $\widetilde{u_0}$ au lieu de u_0. Vous indiquerez vos réponses (correspondant à votre u_0) sur la seconde et vous la remettrez à l'examinateur à la fin de l'épreuve.

En ce qui concerne la partie orale de l'examen, lorsque la description d'un algorithme est demandée, vous devez présenter son fonctionnement de façon schématique, courte et précise. Vous ne devez en aucun cas recopier le code de vos procédures !

Quand on demande la complexité en temps ou en mémoire d'un algorithme en fonction d'un paramètre n, on demande l'ordre de grandeur en fonction du paramètre, par exemple: $O(n^2)$, $O(n \log n)$,...

Il est recommandé de commencer par lancer vos programmes sur de petites valeurs des paramètres et de ***tester vos programmes sur des petits exemples que vous aurez résolus préalablement à la main ou bien à l'aide de la fiche réponse type fournie en annexe***. Enfin, il est recommandé de lire l'intégralité du sujet avant de commencer afin d'effectuer les bons choix de structures de données dès le début.

1 Préliminaires

Dans cet exercice on s'intéresse à rechercher dans un ensemble de vecteurs réels $\mathcal{X}$ de dimension d l'élément le plus proche, au sens du carré de la distance euclidienne, d'un vecteur requête $\mathbf{y} \in \mathcal{Y}$. Ce problème, dit de la "recherche du plus proche voisin", est un classique de l'informatique qui apparaît dans un grand nombre de domaines, notamment en apprentissage machine.

2 Définitions

On rappelle que pour deux entiers naturels a et b, $a \bmod b$ désigne le reste de la division entière de a par b.
Étant donné u_0, on définit par récurrence

$$\forall t \in \mathbb{N}, u_{t+1} = 19\ 999\ 999 u_t \bmod 19\ 999\ 981\ .$$

À partir de trois entiers positifs d, n, et m, on en déduit n vecteurs $(\mathbf{x}_i)_{0 \leq i < n}$ de dimension d, et m vecteurs $(\mathbf{y}_j)_{0 \leq j < m}$ de dimension d :

$$\forall 0 \leq k < d, \begin{cases} \mathbf{x}_i[k] = u_{id+k} \bmod 1\ 000, \\ \mathbf{y}_j[k] = u_{(n+j)d+k} \bmod 1\ 000, \end{cases}$$

où $\mathbf{x}_i[k]$ (respectivement $\mathbf{y}_j[k]$) désigne la k-ième coordonnée de $\mathbf{x}_i$ (respectivement $\mathbf{y}_j$). On notera dorénavant $\mathcal{X} = \{\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}\}$ et $\mathcal{Y} = \{\mathbf{y}_0, \ldots, \mathbf{y}_{m-1}\}$. Par ailleurs, on notera $\mathbf{z}$ un vecteur quelconque dans la suite de l'énoncé.

L'entier u_0 vous est donné, et doit être recopié sur votre fiche réponse avec vos résultats. Une fiche réponse type vous est donnée en exemple, et contient tous les résultats attendus pour une valeur de u_0 différente de la votre (notée $\tilde{u}_0$). Il vous est conseillé de tester vos algorithmes avec $\tilde{u}_0$.

Question 1 *Indiquer les valeurs de* $\mathbf{x}_{33}[13]$ *et de* $\mathbf{y}_{123}[43]$ *dans les cas suivants :* ***a)*** *Pour* $n = 1\ 000$, $m = 10\ 000$ *et* $d = 128$, ***b)*** *Pour* $n = 1\ 000$, $m = 10\ 000$ *et* $d = 200$, ***c)*** *Pour* $n = 100$, $m = 10\ 000$ *et* $d = 128$.

On appelle plus proche voisin de $\mathbf{y}_j$ dans $\mathcal{X}$ le vecteur $\mathbf{x}$ défini par :

$$\mathbf{x} = \arg\min_{\mathbf{x}_i \in \mathcal{X}} \left(\|\mathbf{x}_i - \mathbf{y}_j\|_2\right)^2 ,$$

c'est-à-dire le vecteur de $\mathcal{X}$ le plus proche au sens du carré de la distance Euclidienne du vecteur $\mathbf{y}_j$. Notons qu'il est possible que plusieurs $\mathbf{x}_i$ soient simultanément les plus proches voisins du vecteur $\mathbf{y}_j$. On choisira alors systématiquement celui d'indice le plus faible dans le reste de cet énoncé.
Cette notion est étendue à celle de k-ième plus proche voisin, en retirant de $\mathcal{X}$ un à un $k-1$ fois le plus proche voisin, puis en calculant le plus proche voisin sur l'ensemble restant.

Enfin, on ne cherchera pas à vérifier si tous les vecteurs $\mathbf{x}_i$ sont distincts deux à deux. Dans un tel cas, seul celui d'indice le plus faible pourra être considéré un plus proche voisin d'un vecteur donné.

3 Recherche naïve

Pour trouver les plus proches voisins d'un vecteur $\mathbf{y}_j$, on propose dans un premier temps de réaliser une recherche exhaustive. Elle consiste à calculer pour tout $0 \leq i < n$ les distances $(\|\mathbf{x}_i - \mathbf{y}_j\|_2)^2$ puis à identifier ensuite les plus proches voisins. On qualifie cet algorithme de naïf. On répète ce procédé indépendamment pour tous les vecteurs de $\mathcal{Y}$.

Question à développer pendant l'oral 1 *Estimer (en utilisant la notation $\mathcal{O}$) la complexité de cet algorithme en fonction des paramètres du problème (n, m et d).*

Question 2 *Pour $n = 100\ 000$, $m = 1$ et les valeurs de d suivantes, déterminer l'indice du plus proche voisin dans $\mathcal{X}$ de y_0 :* ***a)*** $d = 1$, ***b)*** $d = 32$, ***c)*** $d = 128$.

Question à développer pendant l'oral 2 *Justifier du fait que tout algorithme permettant de trouver les plus proches voisins de m vecteurs requêtes a une complexité au moins de l'ordre de $\Omega(md)$ (on rappelle ici que la notation Ω est la notation duale de $\mathcal{O}$: $u(n) = \Omega(v(n)) \Leftrightarrow \exists w \in \mathbb{R}^{\mathbb{N}}, u = vw \wedge \exists m \in \mathbb{R}^{*+}, \exists n_0, \forall n \geq n_0, w(n) \geq m$).*

On appelle représentativité de $\mathbf{x}_i$ dans $\mathcal{X}$ et pour $\mathcal{Y}$ le nombre de fois que $\mathbf{x}_i$ est le plus proche voisin d'un vecteur de $\mathcal{Y}$ (on rappelle pour la dernière fois que si un vecteur $\mathbf{y}_j$ est à distance minimale de plusieurs vecteurs dans $\mathcal{X}$, seul celui d'indice le plus faible est considéré).

Question 3 *Pour $n = 100$, $m = 10\ 000$ et $d = 32$, calculer les représentativités de :* ***a)*** $\mathbf{x}_{12}$, ***b)*** $\mathbf{x}_{34}$, ***c)*** $\mathbf{x}_{56}$.

Soit $\mathbf{z}$ un vecteur de $\{0, \ldots, 999\}^d$. On note $r(\mathbf{z})$ la représentativité de $\mathbf{z}$ dans $\mathcal{X}_\mathbf{z} = \mathbf{x}_0, \ldots, \mathbf{x}_{n-1}, \mathbf{z}$ et pour $\mathcal{Y}$. En d'autres termes, on s'intéresse à la représentativité de $\mathbf{z}$ lorsqu'il est ajouté en tant que dernier élément (d'indice n) dans $\mathcal{X}$.

Question 4 *Pour $d = 1$, déterminer $\max_{\mathbf{z} \in \{0,\ldots,999\}^d} r(\mathbf{z})$ pour les paramètres suivants :* ***a)*** $n = 3$, $m = 1000$, ***b)*** $n = 3$, $m = 10\ 000$, ***c)*** $n = 3$, $m = 100\ 000$.

Les parties 4, 5 et 6 peuvent être résolues de façon indépendante.

4 Graphe des k plus proches voisins

On appelle graphe des k plus proches voisins associé à $\mathcal{X} = \{\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}\}$ le graphe $G = \langle \{0, \ldots, n-1\}, E \rangle$ où E est défini comme suit : $(u, v) \in E$ si et seulement si $\mathbf{x}_u$ est l'un des k plus proches voisins de $\mathbf{x}_v$ dans $\mathcal{X} - \{\mathbf{x}_v\}$ ou $\mathbf{x}_v$ est l'un des k plus proches voisins de $\mathbf{x}_u$ dans $\mathcal{X} - \{\mathbf{x}_u\}$.

Question à développer pendant l'oral 3 *Montrer qu'un graphe des k plus proches voisins est un graphe non orienté.*

Question 5 *Pour chaque jeu de paramètres donné, trouver la plus petite valeur de k pour laquelle le graphe des k plus proches voisins associé à $\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}$ est connexe*[1] *:* ***a)*** $n = 100$, $d = 128$, ***b)*** $n = 200$, $d = 30$, ***c)*** $n = 300$, $d = 26$, ***d)*** $n = 1000$, $d = 1$.

Question à développer pendant l'oral 4 *Présenter votre algorithme et donner sa complexité.*

On appelle attracteur un sommet du graphe des k plus proches voisins étant voisin de tous les autres sommets.

Question 6 *Pour les jeux de paramètres suivants, déterminer le plus petit entier k pour lequel le graphe des k plus proches voisins associé à $\mathcal{X}$ admet au moins un attracteur :* ***a)*** $n = 100$, $d = 128$, ***b)*** $n = 300$, $d = 300$, ***c)*** $n = 500$, $d = 1000$.

Question à développer pendant l'oral 5 *Pour $d = n$, donner un exemple de $\mathcal{X}$ pour lequel le plus petit entier k demandé à la question 6 est 1. Même question avec le plus petit entier k qui vaut $n - 1$.*

On appelle clique dans un graphe $G = \langle V, E \rangle$ un sous-ensemble S de sommets tels que $\forall u, v \in S, u \neq v \Rightarrow (u, v) \in E$.

Question 7 *Pour chacun des graphes des k plus proches voisins obtenus à partir de $\mathcal{X}$ avec les paramètres suivants, indiquer la cardinalité de sa plus grande clique :* ***a)*** $n = 100$, $d = 128$, $k = 15$, ***b)*** $n = 100$, $d = 200$, $k = 20$, ***c)*** $n = 100$, $d = 222$, $k = 30$.

Question à développer pendant l'oral 6 *Présenter votre algorithme et démontrer qu'il rend un résultat correct.*

5 Faible dimensionalité

Question 8 *Dans le cas où $d = 2$, déterminer l'indice de l'élément de $\mathcal{X}$ de représentativité maximale (en cas d'égalité, on donnera uniquement le plus petit indice) dans les cas suivants :* ***a)*** $n = 100\,000$, $m = 10\,000$, ***b)*** $n = 1\,000\,000$, $m = 100\,000$, ***c)*** $n = 1\,000\,000$, $m = 1\,000\,000$.

Question à développer pendant l'oral 7 *Présenter votre algorithme et donner sa complexité.*

1. On rappelle qu'un graphe connexe est un graphe pour lequel toute paire de sommets peut être reliée par un chemin.

6 Vecteurs binaires

À partir de maintenant et pour le reste de l'énoncé, on remplacera les vecteurs générés habituellement en appliquant $x \mapsto x \pmod 2$ à toutes les coordonnées. Ainsi :

$$\forall 0 \leq k < d, \begin{cases} \mathbf{x}_i[k] = u_{id+k} \bmod 2, \\ \mathbf{y}_j[k] = u_{(n+j)d+k} \bmod 2, \end{cases}$$

Question 9 *Déterminer l'indice de l'élément de $\mathcal{X}$ de représentativité maximale (en cas d'égalité, on donnera uniquement le plus petit indice) dans les cas suivants :* ***a)*** $n = 1\,000$, $d = 10$, $m = 1\,000\,000$, ***b)*** $n = 10\,000$, $d = 10$, $m = 1\,000\,000$, ***c)*** $n = 1\,000\,000$, $d = 20$, $m = 1\,000\,000$.

Question à développer pendant l'oral 8 *Présenter votre algorithme.*

Fiche réponse type: Plus Proche Voisin

$\widetilde{u_0}$: 42

Question 1

a) (861,137)

b) (119,636)

c) (861,26)

Question 2

a) 3539

b) 69487

c) 15952

Question 3

a) 12

b) 35

c) 33

Question 4

a) 305

b) 2867

c) 28407

Question 5

a) 2

b) 2

c) 2

d) 9

Question 6

a) 89

b) 238

c) 427

Question 7

a) 5

b) 7

c) 9

Question 8

a) 3096

b) 25908

c) 287

Question 9

a) 8

b) 174

c) 1290

Fiche réponse: Plus Proche Voisin

Nom, prénom, u_0: .....

Question 1

a)

b)

c)

Question 2

a)

b)

c)

Question 3

a)

b)

c)

Question 4

a)

b)

c)

Question 5

a)

b)

c)

d)

Question 6

a)

b)

c)

Question 7

a)

b)

c)

Question 8

a)

b)

c)

Question 9

a)

b)

c)