Classifier avec des arbres

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

Durée de l'épreuve: 3 heures 30 minutes

Session 2006

ATTENTION !

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

Important.
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***.

Introduction

Ce sujet explore quelques algorithmes pour classifier des données à l'aide d'arbres, en cherchant à optimiser divers critères (éloignement, poids, feuilles, etc).

1 Préambule

Considérons la suite d'entiers (u_n) définie pour $0 \leqslant n \leqslant 100000$ par :

$$u_n = \begin{cases} \textit{votre } \mathtt{u_0} \textit{ (à reporter sur votre fiche)} & \text{si } n = 0 \\ (15\,991 \times u_{n-1}) \mod 65\,539 & \text{si } n \geqslant 1 \end{cases}$$

On définit la matrice M de taille $m \times m$ suivante : $M_{i,j} = M_{j,i} = u_{(i-1)m+j}$ pour $1 \leqslant i < j \leqslant m$, $M_{i,i} = 0$ pour $1 \leqslant i \leqslant m$. Dans toute la suite du problème, il vous sera demandé de répondre aux questions pour $m \in \{5, 50, 150\}$.

Question 1 *Les coefficients de la partie triangulaire supérieure de M sont-ils tous distincts pour* ***a)*** *$m = 5$,* ***b)*** *$m = 50$, et* ***c)*** *$m = 150$? Quel est la somme maximale d'une ligne de M pour* ***d)*** *$m = 5$,* ***e)*** *$m = 50$, et* ***f)*** *$m = 150$?*

On note Δ_k la différence maximale, en valeur absolue, de deux éléments d'un même sous-carré de taille k. Formellement, $\Delta_k = \max\{|M_{i',j'} - M_{i,j}|, 1 \leqslant i < i' \leqslant i + k - 1 \leqslant m, 1 \leqslant j < j' \leqslant j + k - 1 \leqslant m\}$

Question 2 *Que vaut Δ_k pour* ***a)*** *$k = 2$ et $m = 5$,* ***b)*** *$k = 5$ et $m = 50$, et* ***c)*** *$k = 10$ et $m = 150$?*

On aura besoin plus loin (partie 3) de la matrice M' symétrique, de diagonale nulle, et dont la partie triangulaire supérieure est définie par :
- $M'_{i,i+1} = 32000$ pour $1 \leqslant i < m$
- $M'_{1,m} = 32000$
- $M'_{i,j} = M_{i,j}$ si $2 \leqslant j - i < m - 1$ et $M_{i,j} = 0 \mod 3$
- $M'_{i,j} = 0$ sinon.

On définit Δ'_k pour M' comme on a défini Δ_k pour M.

Question 3 *Combien y a t-il de coefficients non nuls dans M' pour* ***a)*** *$m = 5$,* ***b)*** *$m = 50$, et* ***c)*** *$m = 150$? Quelle est la plus petite valeur de k telle que $\Delta'_k = \Delta'_{k+1}$? pour* ***d)*** *$m = 5$,* ***e)*** *$m = 50$, et* ***f)*** *$m = 150$?*

2 Arbre de proximité

On veut classifier m éléments, chacun correspondant à une colonne de la matrice M. On génère l'arbre binaire $\mathcal{A}_m$ des feuilles à la racine, comme suit (voir Figure 1) :
- initialisation : m sommets feuilles $C_i = \{i\}$, $1 \leqslant i \leqslant m$, chacun correspondant à un indice de colonne de M

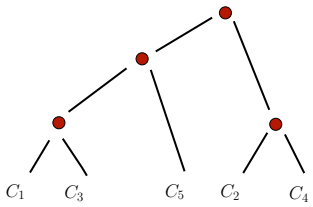

FIG. 1 – Un exemple d'arbre de proximité.

- construction : $m-1$ sommets internes, chacun correspondant à l'union ensembliste de deux sommets déjà construits. On note $|C|$ le cardinal d'un sommet C. L'éloignement de deux sommets C et C' est $d(C,C') = \frac{1}{|C|.|C'|}\sum_{i\in C}\sum_{j\in C'} M_{i,j}$. A chaque étape, on choisit de fusionner deux sommets d'éloignement minimal.
Pour caractériser complètement l'algorithme, il faut spécifier quelle paire de sommets est choisie en cas d'égalité des éloignements : parmi les paires (C,C') telle que $d(C,C')$ est minimal, on fusionnera le sommet C qui contient le plus petit indice et le sommet C' qui contient le plus petit indice n'appartenant pas à C.
La profondeur dans l'arbre $\mathcal{A}_m$ du sommet racine est 0, et la profondeur d'un autre sommet est celle de son père plus 1.

Question 4 *Que vaut $d(C,C')$ à la dernière étape de l'algorithme (création du sommet racine) pour pour* ***a)*** $m=5$, ***b)*** $m=50$, *et* ***c)*** $m=150$ *? Quelle est la profondeur du sommet C_1 pour* ***d)*** $m=5$, ***e)*** $m=50$, *et* ***f)*** $m=150$ *? Quelle est la somme des profondeurs des feuilles de l'arbre $\mathcal{A}_m$ pour* ***g)*** $m=5$, ***h)*** $m=50$, *et* ***i)*** $m=150$ *?*
★ *Vous présenterez à l'oral l'algorithme que vous avez utilisé, ainsi que sa complexité.*

3 Arbres couvrants

On considère la matrice M définie en partie 1. M est symétrique et de diagonale nulle, de taille $m\times m$. On associe à M le graphe non-orienté $G(M)=(V,E)$ défini par la donnée :
- d'un ensemble $V=\{v_1,v_2,\ldots,v_m\}$ de m sommets
- d'un ensemble E de paires de sommets, appelées arêtes, défini comme suit :
$(v_i,v_j)\in E \Leftrightarrow M_{i,j}\neq 0$. Le poids d'une arête $(v_i,v_j)\in E$ est $M_{i,j}$. Une arête $(v_i,v_j)\in E$ est caractérisée par un unique triplet $(M_{i,j},i,j)$ avec $i<j$; on dit alors que v_i est le premier sommet de E, et v_j le deuxième. L'ordre canonique sur les arêtes est l'ordre lexicographique sur les triplets $(M_{i,j},i,j))$: pour cet ordre une arête en précède une autre si elle de poids inférieur, ou de même poids et de premier sommet d'indice plus petit, ou encore de mêmes poids et premier sommet mais de deuxième sommet d'indice plus petit. Cet ordre jouera un rôle important dans la suite.

Question 5 *On trie les $|E|$ arêtes de $G(M)=(V,E)$ selon l'ordre canonique (croissant). Quel est le triplet correspondant à l'arête numéro $(|E| \text{ div } 2)$ pour* ***a)*** $m=5$, ***b)*** $m=50$, *et* ***c)*** $m=150$ *? Ici $(|E| \text{ div } 2)$ est le quotient de la division euclidienne de $|E|$ par 2.*

Deux sommets v_i et v_j sont *voisins* s'ils sont reliés par une arête, i.e. si $M_{i,j}\neq 0$. Il existe un chemin d'extrémités v_i et v_j s'il existe une suite de $p+1$ sommets distincts

$$w_{k_0}=v_i, v_{k_1}, v_{k_2}, \ldots, v_{k_{p-1}}, v_{k_p}=v_j$$

tels que $v_{k_{q-1}}$ et v_{k_q} sont voisins pour $1 \leqslant q \leqslant p$. Un cycle est un chemin dont les deux extrémités son confondues. Une composante connexe de $G(M)$ est une classe d'équivalence de la relation "chemin" : deux sommets sont dans la même composante connexe si et seulement s'il existe un chemin qui les relie.
Un arbre couvrant de $G(M)$ est un graphe $(V, \mathcal{A})$ où $\mathcal{A} \subset E$, qui ne comprend qu'une seule composante connexe et qui est sans cycle. Ceci équivaut à la donnée d'un ensemble $\mathcal{A}$ de $m-1$ arêtes de E tel qu'il existe un et un seul chemin dans le graphe $(V, \mathcal{A})$ entre toute paire de sommets de V : en effet s'il n'y a pas de chemin il y a plusieurs composantes connexes, et s'il y en a deux, il y a un cycle.
On définit de même le graphe $M'(G) = (V, E')$ associé à la matrice M', et on note $(V, \mathcal{A}')$ un arbre couvrant de ce graphe. On étudie dans la suite plusieurs arbres couvrants de $G(M)$ et de $G(M')$.

3.1 Arbre couvrant de poids minimal

Le poids d'un arbre couvrant $(V, \mathcal{A})$ est défini comme la somme des poids des arêtes de $\mathcal{A}$. On admettra que l'algorithme suivant renvoie un arbre couvrant de poids minimal :
- trier les arêtes de E selon l'ordre canonique (croissant) et les numéroter, dans cet ordre, $\{e_1, e_2, \ldots, ..., e_{|E|}\}$
- initialisation : $\mathcal{A} \leftarrow \{e_1\}$
- pour $i = 2$ à $|E|$ faire
  si $(V, \mathcal{A} \cup \{e_i\})$ est sans cycle alors $\mathcal{A} \leftarrow \mathcal{A} \cup \{e_i\}$ (ajouter l'arête e_i)
- renvoyer $(V, \mathcal{A})$

Question 6 *Quel est le poids de la dernière arête ajoutée à l'arbre $\mathcal{A}$ dans l'algorithme avec la matrice M, pour* ***a)*** *$m = 5$,* ***b)*** *$m = 50$, et* ***c)*** *$m = 150$? Quel est le poids total de l'arbre $\mathcal{A}'$ obtenu avec la matrice M' pour* ***d)*** *$m = 5$,* ***e)*** *$m = 50$, et* ***f)*** *$m = 150$?*

★ *Vous présenterez à l'oral l'algorithme que vous avez utilisé, ainsi que sa complexité.*

3.2 Arbre couvrant de degré minimal

Le degré d'un sommet est le nombre de ses voisins. On note $\mathsf{deg}_F(v)$ le degré d'un sommet v dans le graphe (V, F) où $F \subset E$. Le degré $\mathsf{deg}_{\mathcal{A}}$ d'un arbre couvrant $(V, \mathcal{A})$ est le maximum des degrés $\mathsf{deg}_{\mathcal{A}}(v_i)$ des m sommets dans l'arbre. Trouver un arbre couvrant de degré minimal est un problème difficile. On utilise des algorithmes gloutons ou procédant à des échanges locaux, en espérant qu'ils donneront de bons résultats.

Le premier algorithme glouton cherche à diminuer le degré maximal d'un sommet dans l'arbre courant : - partir d'un arbre couvrant initial $(V, \mathcal{A}_0)$: $\mathcal{A} \leftarrow \mathcal{A}_0$. On prendra $\mathcal{A}_0$ comme l'arbre de poids minimal obtenu à la partie 3.1
- tant qu'il existe trois sommets (u, v, w) tels que $(u, v) \in E \setminus \mathcal{A}$, $(u, w) \in \mathcal{A}$, $(v, w) \in \mathcal{A}$, $\mathsf{deg}_{\mathcal{A}} = \mathsf{deg}_{\mathcal{A}}(w) \geqslant \max(\mathsf{deg}_{\mathcal{A}}(u), \mathsf{deg}_{\mathcal{A}}(v)) + 2$, faire $\mathcal{A} \leftarrow \mathcal{A} \cup \{(u, v)\} \setminus \{(u, w)\}$
Pour caractériser complètement l'algorithme, il faut spécifier quel triplet de sommets est choisi à chaque étape, s'il y a plusieurs candidats possibles : parmi ceux-ci, on prendra le triplet (u, v, w) tel que l'arête (u, v) est la plus petite pour l'ordre canonique, et on supprimera la plus petite arête (u, w) ou (v, w) possible.

Question 7 *Quel est le poids de la dernière arête ajoutée à l'arbre $\mathcal{A}$ dans l'algorithme avec la matrice M, pour* **a)** $m = 5$, **b)** $m = 50$, *et* **c)** $m = 150$ *? Quel est le degré de l'arbre $\mathcal{A}'$ obtenu avec la matrice M' pour* **d)** $m = 5$, **e)** $m = 50$, *et* **f)** $m = 150$ *?*

★ *Vous présenterez à l'oral l'algorithme que vous avez utilisé, ainsi que sa complexité.*

Le deuxième algorithme procède à des échanges d'arêtes plus élaborés :
- partir d'un arbre couvrant initial $(V, \mathcal{A}_0)$: $\mathcal{A} \leftarrow \mathcal{A}_0$. On prendra $\mathcal{A}_0$ comme l'arbre de poids minimal obtenu à la partie 3.1
- tant qu'on peut trouver trois sommets (u, v, w) tels que $(u, v) \in E \setminus \mathcal{A}$, et que w est un sommet sur l'unique chemin de u à v dans $\mathcal{A}$ vérifiant $\mathsf{deg}_{\mathcal{A}} = deg_{\mathcal{A}}(w) \geqslant \max(\mathsf{deg}_{\mathcal{A}}(v_i), \mathsf{deg}_{\mathcal{A}}(v_j)) + 2$, alors faire $\mathcal{A} \leftarrow \mathcal{A} \cup \{(u, v)\} \setminus \{(w, w')\}$, où w' est le sommet précédent ou suivant w sur l'unique chemin de u à v dans $\mathcal{A}$

Comme auparavant, il faut spécifier quel triplet de sommets est choisi à chaque étape, s'il y a plusieurs candidats possibles : parmi ceux-ci, on prendre le triplet (u, v, w) tel que l'arête (u, v) est la plus petite pour l'ordre canonique, et on supprimera la plus petite arête (w, w') possible.

Question 8 *Quel est le poids de la dernière arête ajoutée à l'arbre $\mathcal{A}'$ dans l'algorithme avec la matrice M', pour* **a)** $m = 5$, **b)** $m = 50$, *et* **c)** $m = 150$ *? Quel est le degré de l'arbre $\mathcal{A}$ obtenu avec la matrice M pour* **d)** $m = 5$, **e)** $m = 50$, *et* **f)** $m = 150$ *?*

★ *Vous présenterez à l'oral l'algorithme que vous avez utilisé, ainsi que sa complexité.*

3.3 Arbre couvrant feuillu

Une feuille est un sommet n'ayant qu'un voisin. On note $\mathsf{feuille}(\mathcal{A})$ le nombre de feuilles d'un arbre couvrant $(V, \mathcal{A})$. Trouver un arbre couvrant ayant un nombre maximal de feuilles est un problème difficile. On utilise l'algorithme suivant, en espérant qu'il donnera un bon résultat :
- partir d'un arbre couvrant initial $(V, \mathcal{A}_0)$: $\mathcal{A} \leftarrow \mathcal{A}_0$. On prendra $\mathcal{A}_0$ comme l'arbre de poids minimal obtenu à la partie 3.1
- tant qu'il existe deux arêtes e et e' avec $e \in E \setminus \mathcal{A}$, $e' \in \mathcal{A}$, e' appartenant à l'unique cycle de $(V, \mathcal{A} \cup \{e\})$ et $\mathsf{feuille}(\mathcal{A} \cup \{e\} \setminus \{e'\}) > \mathsf{feuille}(\mathcal{A})$, faire $\mathcal{A} \leftarrow \mathcal{A} \cup \{e\} \setminus \{e'\}$. Pour caractériser complètement l'algorithme, il faut spécifier quelle paire d'arêtes est choisie à chaque étape, s'il y a plusieurs candidats possibles : parmi celles-ci, on prendre l'arête e la plus petite pour l'ordre canonique, puis l'arête e' la plus petite possible.

Question 9 *Quel est le poids de la dernière arête ajoutée à l'arbre $\mathcal{A}$ dans l'algorithme avec la matrice M, pour* **a)** $m = 5$, **b)** $m = 50$, *et* **c)** $m = 150$ *? Quel est le le nombre de feuilles de l'arbre $\mathcal{A}'$ obtenu avec la matrice M' pour* **d)** $m = 5$, **e)** $m = 50$, *et* **f)** $m = 150$ *?*

★ *Vous présenterez à l'oral l'algorithme que vous avez utilisé, ainsi que sa complexité.*