UNIVERSITÉ CAEN NORMANDIE

UFR des Sciences

Université de Caen Normandie
Département mathématiques-informatique
2018–2019, 2e session session
27 juin 2019
Valentin Lemière, Grégory Bonnet, Bruno Crémilleux, Bruno Zanuttini

L3 informatique, L3 mathématiques

Examen

Unité M.MIM5E2 : Aide à la décision et intelligence artificielle
2 h — Tous documents autorisés

Chaque candidat doit, au début de l'épreuve, porter son nom dans le coin de la copie qu'il cachera par collage après avoir été pointé. Il devra en outre porter son numéro de place sur chacune des copies, intercalaires, ou pièces annexées.

Cet énoncé comporte 3 pages.

1 Fouille de données : arbres de décision (2 points)

Pour chacun des 2 énoncés ci-dessous (notés E1 et E2), indiquez si chacune des propositions (a) à (f) est vraie ou fausse pour l'énoncé correspondant.

Important : pour chaque proposition, justifiez en 1 à 3 lignes votre réponse.

Question 1 (énoncé E1) *« Un arbre de décision construit automatiquement à partir d'un fichier de données est... » :*

- *(a) une structure sous forme arborescente qui résume le jeu de données*
- *(b) un nouveau débouché de l'exploitation forestière*
- *(c) un ensemble de règles indépendantes issues du jeu de données*
- *(d) une structure décrivant précisément tous les exemples du jeu de données*
- *(e) un ensemble de règles où chaque règle conclut sur une classe*
- *(f) une méthode d'intelligence artificielle qui date de plus de 30 ans.*

Question 2 (énoncé E2) *« Dans les arbres de décision, le gain d'information est ... » :*

- *(a) une mesure pour déterminer les données qui rapportent le plus d'argent*
- *(b) un critère de sélection d'un attribut du jeu de données pour construire l'arbre*
- *(c) un critère de sélection d'un attribut du jeu de données pour construire l'arbre le plus complet possible*
- *(d) un critère qui vise à réduire le désordre des données suivant la classe au fur et à mesure de la construction de l'arbre*
- *(e) un critère de sélection de données pour mettre en évidence les données les plus informatives*
- *(f) un critère développé par le Groupement d'Achats des Investisseurs de Normandie (GAIN) pour déterminer les attributs les plus informatifs d'un jeu de données.*

2 Fouille de données : extraction de motifs fréquents (4 points)

Eloane et Alaric disposent de données correspondant à des rapports écrits (un rapport écrit est par exemple un rapport sur le devenir des étudiants de L3). Chaque rapport est décrit par un ou plusieurs logos (associations, organismes officiels, entreprises, etc.). Autrement dit, pour chaque rapport, on sait si chaque logo est présent ou pas. Dans la suite, un rapport est aussi appelé "transaction" et un logo est aussi désigné par le terme "item".

Dans un premier temps, Alaric considère un jeu de données **D1** avec 10 rapports et 4 logos. Il souhaite extraire tous les motifs fréquents de ce jeu de données (un motif est ici un ensemble de logos).

Question 3 (0,25 point) *Quelle est la taille de l'espace de recherche ?*

Pour extraire les motifs fréquents, Alaric utilise la méthode naïve vue en TP à savoir :
— générer comme motifs candidats tous les sous-ensembles des items
— pour chaque motif candidat, tester si il est fréquent

Question 4 (0,5 point) *Est-ce que cette méthode va fonctionner sur* **D1** *? **Justifiez**.*

Alaric considère maintenant un jeu de données **D2** avec 1000 rapports et 100 logos.

Question 5 (0,25 point) *Quelle est la taille de l'espace de recherche ?*

Alaric souhaite toujours extraire les motifs fréquents, cette fois sur **D2**. Il utilise toujours la méthode naïve vue en TP.

Question 6 (0,5 point) *Est-ce que cette méthode va fonctionner sur* **D2** *? **Justifiez**.*

Eloane, qui a suivi tous les TPs de l'unité d'enseignement "Aide à la décision et intelligence artificielle" (contrairement à Alaric) fait remarquer à Alaric qu'il existe des méthodes moins naïves pour extraire les motifs fréquents.

Question 7 (1,5 point) *Quel est le principe clé pour éviter de calculer la fréquence de tous les motifs de l'espace de recherche ?* ***Expliquez en donnant un exemple concret**, par exemple avec un petit jeu de données composé de 6 à 8 transactions et de 4 à 6 items.*

Grâce aux conseils d'Eloane, Alaric arrive alors à extraire les motifs fréquents de **D2** au seuil de fréquence minimale $minfr = 10$. Alaric souhaite maintenant extraire les motifs fréquents au seuil $minfr = 15$.

Question 8 (0,5 point) *Est-ce que l'extraction au seuil* $minfr = 15$ *est plus dure, pareille ou plus facile que l'extraction au seuil* $minfr = 10$ *? **Justifiez**.*

Alaric s'intéresse maintenant aux motifs uniquement composés des logos de l'Union Européenne, de la Région Normandie, de l'Université de Caen Normandie et de la ville de Caen ou d'un sous-ensemble de ces logos.

Question 9 (0,5 point) *Le problème peut-il être simplifié ? Si oui, comment le simplifiez-vous ? **Justifiez**.*

Si vous ne trouvez pas, vous pouvez demander à Eloane.

3 Contraintes et diagnostic (14 points)

On considère un problème de configuration consistant à placer sur une grille 2×2 quatre logos, ceux de l'Union Européenne, de la Région Normandie, de l'Université de Caen Normandie et de la ville de Caen. On suppose que chaque logo est disponible en quatre formats largeur $\times$ hauteur : 10×10, 10×20, 20×10 et 20×20. Les logos doivent tous être placés sur la grille, sur des cases différentes. Par ailleurs, l'Union Européenne doit être mise au moins autant en valeur que les trois autres entités, et la région au moins autant que l'université et la ville ; on considère qu'une entité A est au moins autant mise en valeur qu'une entité B si son logo est, soit strictement plus large, soit strictement plus haut (il peut alors être plus petit sur l'autre dimension).

Enfin, pour des raisons esthétiques, on souhaite que deux logos placés sur la même ligne de la grille 2×2 aient la même hauteur, et que deux logos placés sur la même colonne aient la même largeur.

Question 10 (4 points) *Modéliser ce problème comme un problème de satisfaction de contraintes* Π*, en proposant des variables, des domaines et des contraintes, de sorte que les solutions de* Π *correspondent exactement aux configurations des logos qui respectent les règles.*

Question 11 (1 point) *Donner une solution de votre problème* Π*, en termes formels (affectation aux variables), et expliquer à quelle configuration des logos elle correspond (via un schéma, par exemple).*

Pour les deux questions suivantes, on suppose que l'utilisateur veut placer le logo de l'Union Européenne en haut à gauche de la grille au format 20×10, et celui de la Région Normandie en bas à droite avec une largeur de 10.

Question 12 (2 points) *Comment prendre en compte ces souhaits dans votre modélisation du problème comme un CSP ?*

Question 13 (2 points) *Le filtrage par arc-cohérence suffit-il à montrer que ces souhaits ne sont pas compatibles entre eux ? Justifier.*

Pour la suite, on rappelle qu'une explication E (au sens du diagnostic) du fait qu'un choix de valeurs V est impossible pour certaines variables, est un ensemble d'affectations E qui est localement cohérent pour le problème de contraintes Π sous-jacent, mais qui est tel que E et V ne peuvent pas être étendus ensemble en une solution de Π.

On suppose maintenant que l'utilisateur souhaite placer le logo de l'université en bas à gauche au format 20×10, et celui de la région en bas à droite au format 10×20, ce qu'on appelle la configuration A.

Question 14 (2 points) *Expliquer pourquoi il n'est pas possible, dans la configuration* A*, de placer le logo de l'Union Européenne en haut à gauche au format* 20×20 *(ce qu'on appelle le choix* B*). La configuration* A *est-elle une explication, au sens du diagnostic, du fait que le choix* B *est interdit ?*

Enfin, pour la question suivante, on suppose que l'utilisateur souhaite placer le logo de l'université en bas à gauche au format 10×10, et celui de la région en bas à droite au format 20×10, ce qu'on appelle la configuration C.

Question 15 (3 points) *Montrer que la configuration* C *est une explication, au sens du diagnostic, du fait que le logo de l'Union Européenne ne peut pas être au format* 10×20 *(quelle que soit sa place). Cette explication est-elle minimale au sens de l'inclusion ? Justifier.*