Projet RODM - Classification associative

(à effectuer seul ou en binôme, langage de programmation libre, Julia recommandé)

Nous considérons une liste de 306 patients ayant effectués une opération pour le cancer du sein, entre 1958 et 1970, au centre hospitalier universitaire de Chicago.

Comme représenté en Table 1, pour chaque patient, nous connaissons :

— son âge ;
— son nombre de nodules (tumeurs bénigne ou maligne) ;
— l'année durant laquelle l'opération a été effectuée ;
— s'il est décédé durant les 5 années qui ont suivi l'opération.

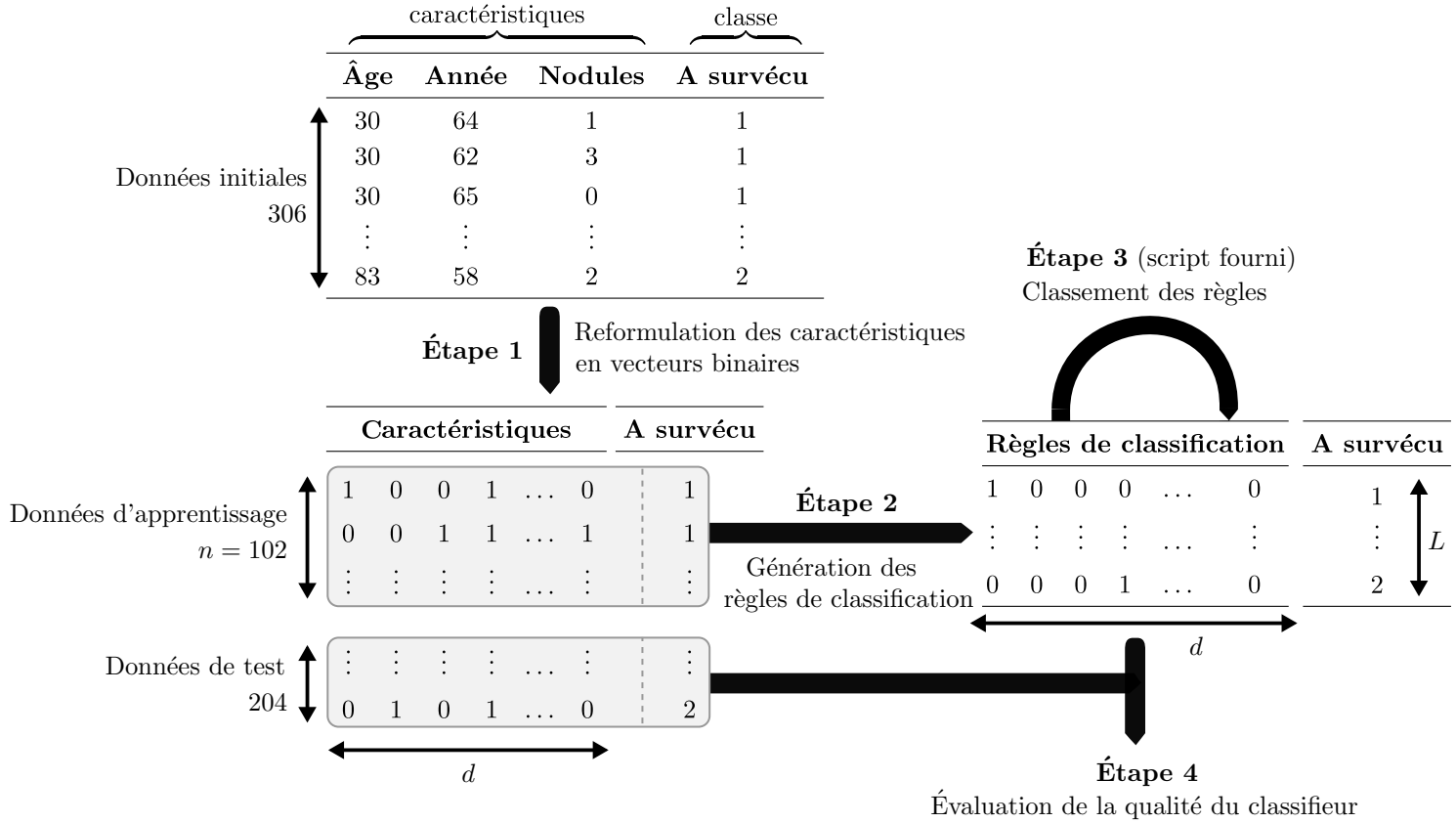
caractéristiques			classe
Âge	**Année**	**Nodules**	**A survécu**
30	64	1	1
30	62	3	1
30	65	0	1
⋮	⋮	⋮	⋮
83	58	2	2

TABLE 1 – *Extraît des données initiales. Chaque ligne correspond à un patient. La dernière colonne contient la valeur 1, si le patient a survécu pendant les 5 années suivant son opération.*

L'objectif de ce projet est de mettre en œuvre la technique de classification associative de Bertsimas, Chang et Rudin [1], présentée en cours, afin de prédire les chances de survie des patients.

Le travail minimum demandé pour ce projet est :

— la réalisation des 4 étapes représentées dans le schéma ci-dessous ;
— le traitement d'au moins une des questions d'ouvertures de l'étape 5 (totalement ou partiellement suivant la difficulté).



Étape 1 Reformulation des caractéristiques sous la forme de vecteurs binaires

La méthode considérée nécessite que chaque patient soit représenté par un vecteur de caractéristiques binaires. Les caractéristiques sont, initialement, 3 entiers : l'âge, l'année de l'opération et nombre de nodules.

1. Récupérer l'archive contenant les fichiers du projet à l'adresse : ? ? ?
2. Proposer une représentation binaire pertinente des trois caractéristiques.
3. Reformuler les données initiales, situées dans le fichier "data/haberman.dat", en utilisant cette représentation.
4. Créer le fichier de données "data/haberman_train.dat" contenant un tiers des patients reformulés, choisi aléatoirement, ainsi qu'un fichier "data/haberman_test.dat" contenant le reste des patients reformulés.

Étape 2 Génération des règles de classification

1. Implanter l'algorithme `RuleGen` pour obtenir l'ensemble des règles de classification dans le fichier "res/haberman_rules.dat". Les valeurs des différents paramètres de la méthode sont fixées dans le fichier "src/02_createRules.jl" (que vous pourrez utiliser comme base si vous programmez en Julia).
2. Ajouter automatiquement (*i.e.*, pas manuellement) les deux règles nulles à la liste obtenue.

Étape 3 Classement des règles

Utiliser le fichier "src/03_sortRules.jl" afin de déterminer un classement optimal des règles. Pour que ce programme fonctionne il faut que le fichier :

— "data/haberman_train.dat" contienne les variables :

1. d : nombre de caractéristiques de votre représentation ;
2. n : nombre de patients dans le fichier de train ;
3. t : un tableau dont chaque ligne correspond au vecteur de caractéristiques d'un patient ;
4. transactionClass : un vecteur contenant la classe de chaque transaction (tel que la classe de t[i, :] soit transactionClass[i]).

— "data/haberman_rules.dat" contienne les variables :

1. rules : tableau dont chaque ligne correspond à une règle générée ;
2. ruleClass : la classe de chaque règle.

Après exécution, les règles ordonnées se trouvent dans le fichier "res/haberman_ordered_rules.dat".

Étape 4 Évaluation de la qualité du classifieur

1. Calculer la précision du classifieur obtenu en l'utilisant sur les patients figurant dans le fichier "data/haberman_test.dat"
2. Calculer le rappel du classifieur en utilisant le classifieur sur le fichier "data/haberman_train.dat".

Étape 5 Questions d'ouverture

1. Comparer les performances du classifieur lorsque d'autres vecteurs de caractéristiques sont considérés.
2. Tester l'influence des divers paramètres de la méthode.
3. Appliquer la méthode à un autre corpus (par exemple un de ceux mentionnés en Section 5 de [1]) ;
4. Les données d'apprentissage, que vous avez utilisées, sont constituées de deux tiers des patients (notés tiers A et tiers B) tandis que le tiers restant (noté tiers C) constitue les données de test. Afin de limiter les risques de biais dans les résultats, calculer les performances des deux classifieurs obtenues lorsque l'apprentissage est effectué en utilisant les tiers A et C, puis B et C.

5. L'étape 2 consiste à générer les règles qui maximisent le support tout en minimisant la couverture. L'algorithme `RuleGen` fais le choix de fixer une borne supérieure sur la couverture afin de ne pas se trouver face à un problème bi-objectif. Dans cette question nous allons résoudre directement le problème bi-objectif. Pour ce faire, utilser les packages vOptGeneric et MultiJuMP de Julia qui permettent de résoudre facilement des problèmes de ce type en retournant une liste de solutions (chaque solution correspondant, ici, à une règle de classification). Étudier l'intérêt de cette approche par rapport à `RuleGen` (nombre de règles obtenues, temps de calcul, performance du classifieur, ...).

Étape 6 Rendu

A l'issue de ce projet (data limite de rendu le 04/03), vous rendrez vos fichiers sources, les fichiers de données générés ainsi qu'un rapport qui contiendra notamment :

- Une description argumentée de la représentation binaire que vous avez choisie.
- Le nombre de règles générées, le temps de calcul associé, le solveur choisi, ainsi que sa version, et les caractéristiques de la machine utilisée (type et nombre de processeurs et mémoire RAM)
- Le nombre de règles du classifieur et le temps de calcul pour les ordonner.
- Les performances de votre classifieur (précision, rappel).
- Un tableau représentant les règles ordonnées de votre classifieurs ainsi que leur classe.
- Le travail effectué sur les questions d'ouvertures.
- La liste des opérations à effectuer pour reproduire vos expériences (quels programmes exécuter, avec quels paramètres, dans quel ordre, ...).

Références

[1] Allison Chang, Dimitris Bertsimas, and Cynthia Rudin. An integer optimization approach to associative classification. In *Advances in neural information processing systems*, pages 269–277, 2012.