Examen Data Mining (20 points)
15 Novembre 2022

Tous documents papier autorisé. Lisez les questions attentivement et répondez de manière argumentée. Essayez d'être concis mais précis. Vous pouvez rédiger en français ou en anglais. En général, il n'y a pas **une bonne réponse unique**. Vous pouvez utiliser des schémas ou des équations si vous le souhaitez.

Vous vous voyez confier une mission de Data Mining, pour laquelle vous aurez du temps, des moyens et une petite équipe. L'objectif est de chercher à mieux comprendre ce qui favorise le succès ou l'échec des étudiants du département informatique d'une université. En particulier, nous voudrions répondre à certains questions bien précises. Pour tous les étudiants, nous disposons déjà d'une valeur binaire Succès/Échec, obtenu l'année précédente, qui nous dit si l'étudiant a validé son année ou non.

Lisez bien toutes les questions d'abord avant de commencer à répondre à la première.

1. (5 points) (1,5 page maximum) Nous voudrions comprendre comment le réseau social de l'étudiant, en particulier ses connexions avec les autres étudiants, joue un rôle dans sa réussite. Nous pouvons faire passer un questionnaire aux étudiants, leur demandant de lister les noms des autres étudiants avec lesquels ils ont parlé récemment (A vous de définir le récemment). Vous pouvez poser jusqu'à 3 questions de ce type, formulées de la manière que vous voulez. Quelles seraient ces questions et quelle analyse feriez-vous à partir des réponses des étudiants ?

   J'attendais que vous proposiez d'utiliser une approche orientée graphe. Vos questions devaient donc permettre de récupérer le nom/identifiant des contacts. Idéalement, vous deviez mentionner si votre réseau était pondéré ou dirigé, ou dynamique. Vous pouviez proposer de regarder la relation entre les propriétés graphe des nœuds (degré, centralité, communautés) et la réussite par exemple, ou que vous proposiez votre propre approche (corrélation entre la réussite d'un nœud et de ses voisins)

2. (4 points) (1 page maximum) Nous voudrions créer 3 cours de support pour aider les étudiants dans les matières dans lesquelles ils ont des difficultés. Mais tous les étudiants n'ont pas des difficulté dans les mêmes matières. Il nous faudrait donc créer 3 groupes d'étudiants ayant des profils similaires en terme de résultat aux examens. Nous avons accès toutes les notes obtenues dans toutes les matières. Que proposez-vous?

   J'attendais la proposition d'utiliser du clustering, en justifiant le choix. Typiquement, k-means était un bon choix car il permet de choisir le nombre de cluster comme demandé, et son "biais" tend à créer des clusters de taille comparable. J'attendais un minimum d'explication sur comment passer des clusters aux groupes d'étudiants, éventuellement en mentionnant les problèmes possibles.

3. (5 points) (1,5 page maximum) Nous voudrions savoir comment le profil de l'étudiant et son milieu socio-économique (revenu, profession des parents, travail pour financer ses études, lycée d'origine, etc.) affecte sa réussite. En considérant que nous pouvons faire passer une enquête aux étudiants, en leur posant un maximum de 5 questions, quelles questions proposez-vous ? Quelles analyses feriez-vous ?

J'attendais que les questions permettent de récupérer des données exploitables, c'est à dire soit catégorielles, soit numériques, soit en expliquant comment passer de données complexes (nom du lycée) à des données exploitables (récupération de la moyenne des notes du lycées, etc.) J'attendais ensuite une proposition d'analyse de ces données ayant du sens. Idéalement, ça pouvait être des patterns fréquents, bien adapté en raison de la présence d'une donnée cible (réussite oui/non). Mais une autre approche bien justifiée marchait aussi.

4. (3 points) (1 page maximum) Nous avons accès à l'ensemble des absences des étudiants dans les différentes matières. Nous savons que certains étudiants sont plus absent que d'autres pour des raisons diverses. Mais ce que nous souhaitons savoir, c'est si les absences sont plus fréquentes dans certaines matières que dans d'autres. En particulier, nous voudrions savoir si ces variations sont aléatoires ou si certaines matières ont un plus haut taux d'absentéisme. Quelles méthodes proposez-vous pour étudier cela?

   Ici, il s'agissait au moins de calculer des taux d'absentéisme moyen par matière (moitié des points), puis pour avoir plus, mentionner soit l'usage de la variance, soit un test statistique. Des méthodes plus avancées était possible, comme l'usage de la normalisation utilisateur/item utilisée en recommendation, ou certains d'entre vous ont également décrit une méthode consistant à calculer des moyennes d'absentéisme par étudiant puis par classe ou le contraire, c'était satisfaisant également.

5. (3 points) (1 page maximum) Nous voudrions faire une cartographie des étudiants, sous la forme d'un nuage de points en 2 dimensions, telle que l'une des dimensions synthétise sa situation socio-économique et l'autre son absentéisme. Nous voudrions utiliser cette cartographie pour voir si le succès des étudiants est corrélé avec ces facteurs. Que proposez-vous ?

   Ici, il s'agissait de faire attention à ne pas appliquer directement une PCA en 2D (ou autre méthode en 2D), car cela ne permettait pas d'avoir ce qui était demandé. Il fallait donc proposer une méthode, n'importe laquelle, pour transformer les absences en 1D et la situation socio-eco en 1D. Beaucoup ont proposé d'utiliser la réduction de dimension, d'autres simplement de faire des moyennes ou des sommes. Tout cela était correct. Il fallait ensuite mentionner la façon d'interpréter ce plot, on mettant une couleur aux nœuds et en regardant si les nœuds ayant des couleurs similaires étaient proche par exemple, ou dans un endroit spécifique du graphe, selon votre choix des dimensions.