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.

1 Théorie

1. (5 points) (1,5 page maximum) Expliquez, à travers un exemple original de votre propre invention, la différence entre un clustering de type DBscan un clustering de type k-means. Mettez en évidence des cas sur lesquels l'une des méthodes est performance et l'autre non. Vous pouvez en particulier vous appuyer sur des dessins de données en deux dimensions. Vous ne pouvez pas reprendre à l'identique des examples du cours.

   J'attendais que vous preniez un exemple similaire aux cercles ou aux lunes vus en course, avec par exemple une forme allongée sur laquelle DBscan trouve une seule communauté et k-means plusieurs, car les données ne sont pas charactérisées par un centroide. Dans l'idéal, vous pouviez justifier cet exemple avec des données imaginaires quelconques.

2. (5 points) (1 page maximum) La détection de communautés a de nombreuses similarités avec le clustering. Expliquez quelles sont à votre avis les similarités et les différences entre la détection de communauté en optimisant la modularité, k-means, et DBscan. De laquelle de ces deux méthodes de clustering pensez-vous que Louvain est la plus proche ?

   Après avoir mentionné le fait que la détection de communauté pouvait être vue comme du clustering sur les graphes, vous pouviez faire certaines de ce observations: DBscan et louvain ne nécessitent pas de préciser le nombre de communautés, mais ont un paramètre qui permet de contrôler ce nombre. DBscan peut être formalisé comme un algorithme sur un graphe, comme louvain. Louvain et k-means en revanchent optimisent une fonction de qualité globale, au contraire de DBscan. Louvain et DBscan intègrent une notion de densité, mais k-means aussi explicitement. Il n'y avait pas de réponse "juste" sur lequel il est le plus proche, ce qui m'intéressait était l'argumentation.

2 Application

3. (4 points) (1 page maximum) Vous disposez de données concernant des véhicules. Pour chaque véhicule, vous disposez des informations suivantes: Date de construction, Date d'achat initial, Kilométrage, Prix, Puissance du moteur en chevaux vapeur, Consommation en ville, Consommation sur autoroute, Poids du véhicule. Nous voudrions découvrir des clusters de voiture similaire. Proposez une démarche pour faire des clusters pertinent: pré-traitement des données, transformations éventuelles, sélection de variables, méthode de clustering, éventuellement avec choix/exploration de paramètres... Vous considérerez que les données sont déjà nettoyées (pas de données aberrantes, pas de données manquantes).

J'attendais plusieurs observations: les données de type date nécessitent d'être vérifié pour être transformées dans un format numérique pertinent. Vous deviez observer aussi que les données étaient très probablement corrélées, ce qui nécessitait soit de les combiner "à la main", soit, mieux, d'utiliser une PCA. J'attendais aussi une mention de normalisation/standardisation des données. Vous pouviez proposer DBscan, gaussian mixture ou k-means, en justifiant ce choix (choisir le nombre de cluster, ou au contraire le découvrir automatiquement... Peut-être parce que vous supposez par exemple qu'il n'y a pas de clusters bien définis et que vous devez découper arbitrairement...)

4. (4 points) (1,5 page maximum) Vous disposez des données de vote des députés au cours de la dernière année. Plus précisément, pour chaque député, et pour chaque texte voté, vous disposez de l'information pour savoir s'il a voté pour, contre, ou s'il n'a pas pris part au vote. Pour chaque député, vous disposez des informations le décrivant: âge, genre, parti politique. Il y a 577 députés dans votre base de données, et 1800 votes. Pour chaque vote, vous disposez de 5 mots-clefs, qui permettent de comprendre sur quoi portait le texte. Proposez quelques applications des outils que nous avons vu en cours sur ce jeu de données: question posée, et manière d'y répondre.

   J'attendais que vous proposiez d'utiliser soit de la recommandation, soit des patterns fréquents, soit de la factorisation de matrices (NMF, SVD). Ces données s'y prêtent en effet bien, vous pouviez mentionner la création d'une matrice de taille 577x1800, parler un peu de la transformation des données en recodant vote positif comme 1 et négatif comme 0, ou tout autre choix qui permet d'appliquer ces méthodes.

5. (2 points) (0,5 page maximum) Regardez attentivement le réseau de la figure 1. Est-ce que vous pensez que c'est un réseau *small world*? Justifiez.

   Le réseau était clairement pas un small world: il n'y a pas de triangles, donc clustering 0. Les distances moyennes sont longues en moyenne, à cause de l'absence de "raccourcis" entre les stations longtaines.

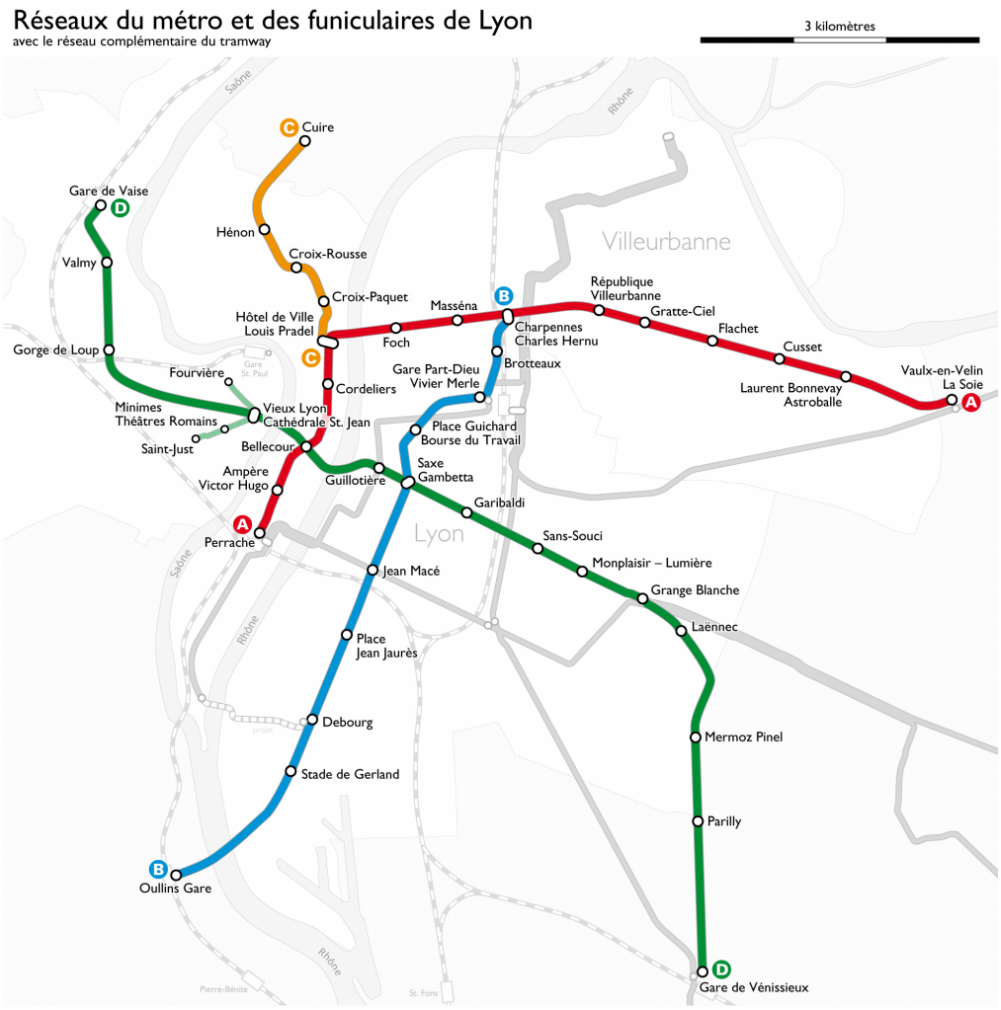

Figure 1: Réseau de métro de Lyon+funiculaires