INF8100 - Concepts et techniques de la fouille et de l'exploitation de données

Travail Pratique 1 préparé par Nairouz Mrabah et Ange Tato

Automne 2023

1 Consignes de remise du travail

Le travail doit être remis au plus tard le **17 octobre (23:59 EDT) 2023** via *Moodle*. **Important** : Pour chaque jour de retard, vous perdrez 5% de votre note. Après 7 jours vous aurez un 0. Pas de période de grâce une fois le délai écoulé.
Votre remise doit être un fichier **.zip (-2 pts si ce n'est pas le cas)** qui contient

- Un rapport (PDF) contenant les réponses aux questions. Il doit y avoir votre nom et votre code permanent.
- deux fichiers `.ipynb` contenant le code qui vous a permis de répondre aux questions de chaque partie. Les réponses aux questions doivent être bien identifiées (numéro). Toutes les réponses doivent être justifiées par un code écrit.

2 Critères d'évaluation

- La présentation du rapport en général, le fichier ipynb : **4/30**
- Réponses aux questions : **36/40**

Ce TP est noté sur 40 et compte 20% de votre note finale.

3 Objectif

Le but de ce travail est d'appliquer les techniques vues en algèbre linéaire, statistiques et probabilités sur des jeux de données. Il nécessite un certain apprentissage individuel du langage python. Vous êtes encouragé à consulter la documentation de pandas pour trouver des fonctions ou des méthodes que vous n'avez peut-être pas encore utilisées.

4 Travail à faire

Utiliser le langage python pour trouver des réponses à toutes les questions suivantes (c'est-à-dire, ne faites aucun calcul à la main). Mettre uniquement les réponses (pas de code) dans le rapport PDF. Il est important de numéroter les questions (utiliser le markdown par exemple) dans vos notebooks.

5 Partie 1

Si vous utilisez Netflix, vous remarquerez qu'il existe une section intitulée 'Parce que vous avez regardé le film x', qui fournit des recommandations de films basées sur les films les plus récents que vous avez regardé. Dans cette première partie, nous allons générer des recommandations à l'aide d'une technique appelée filtrage collaboratif.

5.1 Description du jeu de données

Le jeu de données est formé par deux fichiers csv ('films.csv', 'notes.csv').

Le premier fichier 'films.csv' contient trois colonnes:

- *IdFilm (identifiant numérique du film)*: Discret
- *Titre* (titre du film): Chaine de caractères
- *Genre* (genre du film) : Nominal

Le deuxième fichier °notes.csv' contient quatres colonnes:

- *IdUtilisateur (identifiant numérique de l'utilisateur x)*: Discret
- *IdFilm (identifiant numérique du film y)*: Discret
- *Note* (la note donnée par un utilisateur x à un film y) : Discrète
- *Horodatage* (une date et une heure associée à un film y regardé par un utilisateur x) : Continu

5.2 Exploration des données (3pts)

1. Quel est le nombre de notes données par tous les utilisateurs ?
2. Quel est le nombre d'utilisateurs ? Quel est le nombre de films ?
3. Quel est le nombre moyen de notes par utilisateur ?
4. Quel est le nombre moyen de notes par film ?
5. Quel film a la note moyenne la plus basse?
6. Quel film a la note moyenne la plus élevée ?

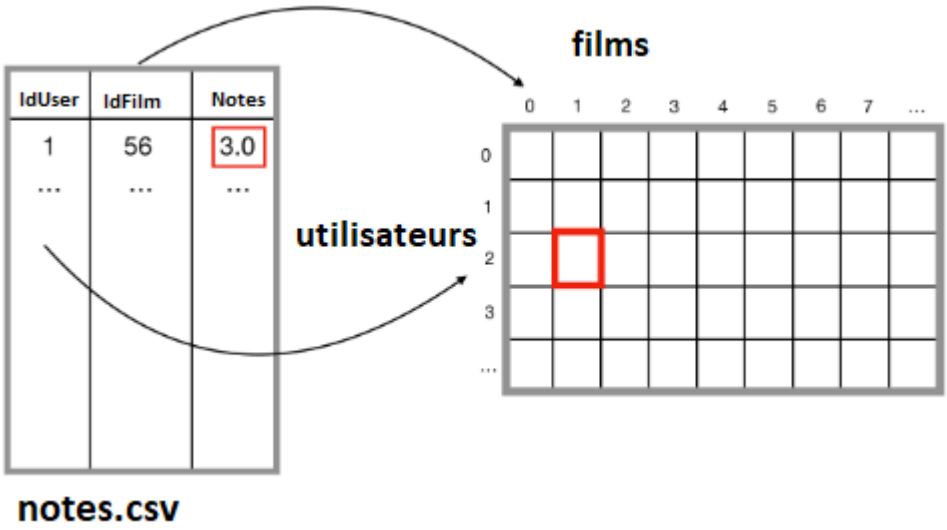

Figure 1: Illustration de la matrice utilisateur-film.

5.3 Transformation des données (3pts)

La première étape pour construire un filtre collaboratif consiste à extraire à partir du fichier 'notes.csv' une matrice utilisateur-film. Les lignes de cette matrice représentent les utilisateurs et les colonnes repésentent les films. Pour un utilisateur x et un film y, la case située dans l'intersection de la ligne x et la colonne y désigne la note attribuée par l'utilisateur x au film y. La matrice utilisateur-film est une matrice creuse (on attribue la valeur zéro pour les cases à valeurs manquantes).

1. Filtrez les utilisateurs qui ont votés moins de 100 films et les films qui sont votés par moins de 10 utilisateurs? **Remarque**: vous pouvez utiliser pandas.DataFrame.groupby(). Dans la suite, on utilisera uniquement la matrice filtrée.

2. Construisez la matrice utilisateur-film? **Remarque**: vous pouvez utiliser la méthode pandas.DataFrame.pivot().

3. Calculez le taux de parcimonie de la matrice utilisateur-film en utilisant la formule 1?

$$S = \frac{A}{B} \tag{1}$$

A : nombre d'éléments non nuls dans la matrice utilisateur-film. B : nombre total d'éléments dans la matrice utilisateur-film.

5.4 Filtrage collaboratif (4pts)

La factorisation SVD de la matrice utilisateur-film permet de découvrir des caractéristiques latentes décrivant les interactions entre les utilisateurs et les films. Ces caractéristiques offrent une représentation plus compacte des goûts des utilisateurs et des catégories de films. Précisément, la factorisation SVD donne:

- U: matrice qui spécifie les goûts des utilisateurs (nombre d'utilisateurs, k).
- Σ: matrice carrée et diagonale qui spécifie l'importance des caractéristiques latentes (k, k).
- V: matrice qui spécifie les genres de films (nombre de films, k).

Généralemet, lorsqu'un utilisateur choisit un film, le système de recommandation lui montre quelques films similaires en fonction de la distance cosinus sur la matrice V.

1. Appliquez SVD sur la matrice utilisateur-film pour réduire la dimension de cette matrice à une dimension plus faible associée aux 'goûts des utilisateurs / genres de films" les plus pertinents? **Remarque**: vous devez conserver les k valeurs singulières les plus grandes avec k entre 5 et 20.
2. Cherchez les 5 films les plus similaires au premier film ? **Remarque**: vous devez utiliser la fonction scipy.spatial.distance.cosine sur les lignes de la matrice V (après la réduction de la dimension).
3. Affichez les films recommendés pour le premier utilisateur? **Remarque**: vous devez afficher l'identifiant de chaque film recommandé, son titre et son genre. Il ne faut pas inclure les films que l'utilisateur a déjà regardés.

6 Partie 2

6.1 Description du jeu de données

Le jeu de données 'salaires.csv' contient des informations sur les salaires annuels de personnes de 42 pays différents, mais la majorité (90 %) provient des États-Unis. Le deuxième dans cette catégorie est le Mexique à 2%, ne laissant que 8% pour les 40 autres pays. Il y a environ 32 561 entrées avec un total de 12 colonnes représentant différents attributs des personnes. Voici la liste:

- *Âge* : Discret
- *Classe de travail* (Privé, Gouvernement fédéral, etc): Nominal
- *Éducation* (le plus haut niveau d'éducation obtenu) : Ordinal (
- *Numéro d'éducation* (le nombre d'années d'études): Discret

- *État civil* : Nominal
- *Occupation* (Transport-Déménagement, Artisanat-Réparation, etc) : Nominal
- *Lien de parenté* (célibataire, hors famille, etc.) : Nominal
- *Race*: Nominal
- *Sexe* : Nominal (2 catégories)
- *Heures (travaillées) par semaine* : Discrète
- *Pays d'origine* : Nominal
- *Revenu*: booléen ($\leq$ 50 000 , $>$ 50 000)

Remarque: Les colonnes représentant des valeurs non numériques, notamment la dernière colonne représentant les salaires, peuvent être encodées en valeurs numériques lorsque nécessaire pour faciliter les analyses.

6.2 Exploration des données (3pts)

1. Quel est la taille du jeu de données ?
2. Combien de valeurs manquantes y a-t-il pour chaque colonne ? Quelle est la colonne (*feature*) qui a le plus de valeurs manquantes ?
3. Combien d'hommes y a-t-il dans ce jeu de données ?
4. Quel est l'âge le plus élevé et le plus bas de ce jeu de données ?
5. Combien de femmes travaillent dans le secteur privé?
6. Quelle est la liste des différents pays représentés dans la colonne représentant le pays natal/d'origine de chaque personne de notre jeu de données?

6.3 Algèbre linéaire (6pts)

Pour les questions de cette partie, nous allons uniquement travailler avec une petite partie des données : les femmes du jeu de données. Vous allez garder le 1/10 des données pour les tests.

1. Peut-on prédire l'âge d'une personne (femme) en fonction de son niveau d'éducation et son nombre d'heures travaillées par semaine ? Vous devez passer par la technique de résolution d'équations linéaires par la factorisation QR. Vous devez tester votre modèle sur les données de tests.
2. Si le niveau d'éducation et le nombre d'heures travaillées étaient réduits à 1 seule dimension qui capture les informations importantes des 2 (SVD) ? est-ce que la prédiction serait améliorée ou pas ? Vous devez tester votre modèle sur les données de tests.

3. Peut-on prédire si le salaire d'une femme sera $\leq$ 50K en fonction de son niveau d'éducation, son âge et son nombre d'heures travaillées par semaine ? Vous devez passer par la technique de résolution d'équations linéaires. **Remarque** : puisque le Y à prédire est binaire alors ce n'est normalement pas un problème de régression linéaire. Vous devez donc adapter votre solution en conséquence. Par exemple, si la prédiction est inférieure à 0,5 alors le salaire est $\leq$ 50k sinon c'est $>$ 50k. Vous devez tester votre modèle sur les données de tests.

6.4 Analyses statistiques et probabilistes (12pts)

1. Quelle est la moyenne des âges ? Quelle est la moyenne du nombre d'heures travaillées par semaine ? et les valeurs médianes ?

2. Quelle sont les âges (prendre les 5 premiers) les plus représentatifs de ce jeu de données ? Représentent-ils une tranche d'âge en particulier (ex: 30-35 ans) ?

3. Comment est la variance (à quel point les données sont éparpillés/dispersées) des heures travaillées par semaine ? Et les âges ?

4. Est ce que le niveau d'éducation, comparé à l'âge influe le plus sur la valeur du salaire ? En d'autres termes est que le niveau d'éducation a plus de poids sur le salaire que l'âge ?

5. Est ce que le niveau d'éducation, comparé au nombre d'heures travaillé par semaine influe le plus sur la valeur du salaire ? En d'autres termes est que le niveau d'éducation a plus de poids sur le salaire que le nombre d'heures travaillées par semaine ?

6. À quel point l'âge est un facteur qui influe le salaire? Pour répondre à cette question vous devez vérifier que l'âge moyen de ceux qui gagne moins de 50k est plus bas que ceux qui gagnent plus de 50k.

7. À quel point le nombre d'heures travaillé par semaine est un facteur qui influe le salaire? Pour répondre à cette question vous devez vérifier que le nombre moyen d'heure travaillé par semaine pour ceux qui gagnent moins de 50k est plus bas que ceux qui gagnent plus de 50k.

8. Est ce qu'il y a une différence de salaire entre les personnes mariées et les personnes célibataires ? Quel groupe gagne le mieux leurs vies ?

9. Quelle est la proportion des hommes qui ont un doctorat et gagnent plus de 50k ? et chez les femmes ?

10. Quelle est la proportion des femmes mariées qui n'ont pas fait de masters ni de doctorat ? et chez les hommes ?

11. Les ressortissants de quel pays ont les salaires les plus élevés uniquement dans notre échantillon?

12. Quelle race a les salaires les plus élevés uniquement dans notre échantillon?

6.5 Tests d'hypothèses (5pts)

Pour ces questions, vous devez utiliser les t-tests.

1. Est-ce qu'on peut dire que la moyenne d'heures travaillées par semaine est autour de 40h/semaine ? Est-ce que l'affirmer serait significatif ?

2. Est ce que les hommes ont en moyenne un salaire supérieur à 50k comparé aux femmes ? est-ce significatif ?

3. Est ce qu'il y a un lien quelconque entre l'âge et le nombre d'heures travaillé par semaine ? est-ce significatif ?