Université de Sherbrooke

IFT 599 / IFT 799 - Science de données

TP3 : Systèmes de recommandation

Automne 2024

Enseignants

	Courriel	Local	Téléphone
Shengrui Wang	shengrui.wang@usherbrooke.ca	D4-1018-1	+1 819 821-8000 x62022

Faculté des Sciences,
Département d'informatique

November 26, 2024

Sommaire

Dans le cadre de ce travail pratique est mis à la disposition des personnes étudiantes un jeu de données de films et d'évaluations par des utilisateurs. Il est question ici, à partir de ce jeu de données, de mettre en exergue les conceptes vu sur les thèmes portant sur les systèmes de recommandation. Au delà du thème portant sur les systèmes de recommandation, les notions vues dans des chapitres précédents tels que *clustering* seront également exploitées dans ce TP. À travers ce projet, vous allez acquérir une bonne capacité d'analyse et maîtriser quelques techniques nécessaires pour bâtir un système de recommandation.

Contents

1 Jeu de données et énoncé du problème

1.1 Jeu de données

Le jeu de données soumis à votre investigation est un extrait de MovieLens (*http://movielens.org*) un service de recommandation de films. Le jeu de données décrit l'activité de notation sur 5 étoiles et de balisage en texte libre provenant de MovieLens. Il contient 25 000 095 d'évaluations et 1 093 360 d'applications de balises sur 62 423 films. Ces données ont été créées par 162 541 utilisateurs entre le 9 janvier 1995 et le 21 novembre 2019. Ce jeu de données a été généré le 21 novembre 2019.

Les utilisateurs ont été sélectionnés de manière aléatoire pour inclusion. Tous les utilisateurs sélectionnés avaient évalué au moins 20 films. Aucune information démographique n'est incluse. Chaque utilisateur est représenté par un identifiant, et aucune autre information n'est fournie. Les données originales contiennent des fichiers comme genome-scores.csv, genome-tags.csv, links.csv, movies.csv, ratings.csv et tags.csv. Des détails sur le contenu et l'utilisation de tous ces fichiers, ainsi que d'autres ensembles de données GroupLens, sont disponibles publiquement en téléchargement sur *http://grouplens.org/datasets*.

Dans le cadre de ce TP, vous allez exploiter seulement les fichiers movies.csv et ratings.csv disponibles dans le repertoire des travaux pratiques (TP3).

1.2 Énoncé général du problème

À partir des fichiers movies.csv et ratings.csv on veut construire et comparer deux systèmes de recommandation basés sur le filtrage par collaboration et sur le filtrage par contenu respectivement.

2 Travail à faire

Comme on peut le constater, les fichiers soumis à votre investigation ne présentent pas clairement le contenu des films. Toutefois, à partir du genre, on pourrait bâtir le contenu associé à chaque film.

1. **Un peu de statistique de l'ensemble de données** : À partir du fichier movies.csv, construire le diagramme en bâton illustrant le nombre de films que l'on a par genre.

   NB: Vous devez ignorer les films au genre non listé: '(no genres listed)'.

2. **Nettoyage** : En ignorant les films non listés, on vous demande d'extraire le nouveau jeu de données movies1.csv et ratings1.csv.

   NB: pour cette question, il s'agit simplement de supprimer de vos fichiers toutes lignes dont l'identifiant du film est non listé. De plus, si dans votre

fichier ratings1.csv contient des ratings de valeurs $5.5, 4.5, 3.5, 2.5, 1.5, 0.5$ **changez les respectivement aux valeurs** $5, 4, 3, 2, 1, 1$.

3. **Construction d'une base de données de films en contenu** : En vous référant sur le(s) genre(s) associé(s) à chaque film, construire la matrice binaire de contenu C caractérisant chaque film. En guise d'exemple, si vous avez les films $F1$ et $F2$ ayant respectivement les genres $(G1, G3)$ et $(G1, G2)$ alors votre matrice C devrait être comme suit:

	$G1$	$G2$	$G3$
$F1$	1	0	1
$F2$	1	1	0

4. **Construire la matrice de profil des utilisateurs** P : En temps normal, le profil d'un utilisateur U_u pourrait être calculé une combinaison linéaire de l'historique de ses votes (ses *ratings*) donné comme suit,

$$Profil(U_u \,|\, \boldsymbol{I}_u) = \sum_{I_i \in \boldsymbol{I}_u} r_{u,i} C_i$$

avec $\boldsymbol{I}_u$ l'ensemble des films dont l'utilisateur U_u a donné son appréciation, $r_{u,i}$ le *rating* que l'utilisateur U_u a donné au film I_i et C_i le vecteur binaire caractérisant le film I_i (la ligne i de votre matrice C). Dans les démarches suivantes de construction et d'évaluation d'un système de recommendation basé sur le contenu, les profils d'utilisateur doivent être calculés sur les données d'apprentissage (et non sur l'ensemble de toutes les données).

NB: Pensez à normaliser les profils d'utilisateur surtout si vous utilisez des fonctions de distances impliquant les profils.

5. Pour evaluer (et comparer) vos systèmes de recommandation construits aux points 6 et 7 ci-dessous, préparez les données selon les instructions suivantes : soit subdiviser votre fichier ratings1.csv en deux fichiers, ratings_apprendissage.csv (contenant 80% des données) et ratings_evaluation.csv (contenant 20% des données), pour une évaluation "simple"; soit subdiviser votre fichier ratings1.csv en cinq fichiers (de 20% chacun) comme ratings_partie1.csv, ratings_partie2.csv, ratings_partie3.csv, ratings_partie4.csv et ratings_partie5.csv, pour une évaluation-croisée à 5-blocs. Pour ce TP, vous choisissez l'une de ces deux options d'évaluation (pas les deux). Bien entendu, les divisions du fichier ratings1.csv doivent êtes faites aléatoirement.

NB: Dans les discussions suivantes, on suppose que vous ayez choisi l'évaluation "simple" avec un fichier pour l'apprentissage et un fichier pour l'évaluation.

6. **Filtrage collaboratif basé sur utilisateurs** : Construire votre système sur le fichier .csv d'apprentissage en utilisant la formule de corrélation de Pearson (page 8, partie 2 du Thème 6) comme la fonction de similitude et le moyennage de scores

des utilisateurs similaires pour estimer un nouveau score. Pour cette évalution, c'est vous qui détermine votre définition de voisinage et la formule pour le moyennage.

NB: Pour l'évaluation de votre système, il faut ESTIMER les scores $r_{u,i}$ qui sont donnés dans le fichier ratings_evaluation.csv et comparer les scores estimés avec les scores connus. Pour estimer la performance globale de votre système, vous pouvez utiliser l'erreur moyenne calculée entre l'ensemble des scores estimés et ceux fournis dans le fichier ratings_evaluation.csv.

7. **Filtrage basé sur contenu en utilisant le clustering** : Pour ce système, vous appliquez un algorithme de clustering sur les données de profils d'utilisateur et proposer et implémenter une façon d'utiliser les résultats du clustering pour faire de la recommandation par contenu. Utilisez le fichier .csv d'apprentissage pour construire les profils d'utilisateur (point 4) et utilisez la combinaison "K-moyenne + silhouette" ou "clustering spectral + silhouette" pour déterminer le bon nombre de clusters et les résultats de clustering.

   Si vous choisissez de faire "clustering spectral + silhouette" avec votre matrice de profil P, vous pouvez employer l'impantation suivante *spectral clustering*, *https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html*).

   NB: Pour l'évaluation de votre système, suivez essentiellement les même consignes que le point 6.

8. **Filtrage basé sur contenu, une approche simple** : Ceci est une question de réflxion. Proposer une idée d'algorithme (en quelques phrases ou en un pseudo code pas trop long) pour réaliser une filtrage basé sur le contenu soit par l'approche Baysienne (page 18 partie 1 du Thème 6), soit par une approche s'inspirant des k plus proches voisins.

3 Remise du TP

- Pour le point 7, les résultats à présenter incluent aussi ceux vous permettant de choisir le nombre optimal de clusters;
- La date officielle de remise du TP est le lundi 9 décembre 2023 à 23h59. Aucun TP ne sera accepté après cette date;
- Soignez votre rapport, une pénalité de 5% pourrait être appliquée pour un rapport mal rédigé;
- Les fichiers à soumettre sont le rapport (en Word ou pdf) et l'ensemble de vos programmes. Ne soumettez pas des fichiers de données car il risque ne pas pouvoir passer à cause de la limitation imposée par turnin. De ce fait, on vous demande d'ajouter, dans votre rapport, quelques informations sur vos données utilisées. Les

informations essentielles à ajouter sont le nombre d'utilisateurs et le nombre de films dans les fichiers que vous avez utilisés pour l'apprentissage et pour l'évaluation.

- N'oubliez pas d'identifier les membres du groupe de travail. Indiquez les noms et cips (ou matricules) des membres du groupe dans chacun des fichiers que vous soumettez. La remise doit être faite par http://turnin.dinf.usherbrooke.ca