Université de Sherbrooke

IFT 599 / IFT 799 - Science de données

TP3 : Systèmes de recommandation

Automne 2023

Enseignants

	Courriel	Local	Téléphone
Shengrui Wang	shengrui.wang@usherbrooke.ca	D4-1018-1	+1 819 821-8000 x62022
Etienne G. Tajeuna	etienne.gael.tajeuna@usherbrooke.ca		+1 819 821-8000 x

Faculté des Sciences,
Département d'informatique

December 3, 2023

Sommaire

Dans le cadre de ce travail pratique (TP) est mis à la disposition des personnes étudiantes un (01) jeu de données. 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. Pas seulement limité aux thèmes portant sur les systèmes de recommandation, les notions vu dans les 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 se trouvent dans les fichiers genome-scores.csv, genome-tags.csv, links.csv, movies.csv, ratings.csv et tags.csv. Plus de détails sur le contenu et l'utilisation de tous ces fichiers suivent. Ce jeu de données, 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 exclusivement exploiter les fichiers movies.csv et ratings.csv disponibles dans le repertoire des travaux pratiques (TP3).

1.2 Énoncé du problème

À partir des fichiers movies.csv et ratings.csv on voudrait construire un système de recommandation mixte (basé sur le contenu et la collaboration).

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. 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. 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, si vous avez 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. 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. En supposant que le profil d'un utilisateur U_u est 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).

Construire la matrice de profil des utilisateurs P.

5. Afin de limiter la recherche d'items et d'utilisateurs, on voudrait trouver les groupes d'utilisateurs et d'items présentant les même caractéristiques. En utilisant votre matrice de profil P, en se basant sur le regroupement spectral (*spectral clustering*, *https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html*), trouver les différents groupes d'utilisateurs.

NB: Pour cette question 5., vous devez tester plusieurs valeurs de $K = \{2, 3, 4, 5\}$ **et retenir celle dont le score de silhouette est le plus élevé. Dans votre rapport, on devrait visualiser les nuages de points coloriés par clustering; on devrait voir la courbe des scores associée aux différents regroupements.**

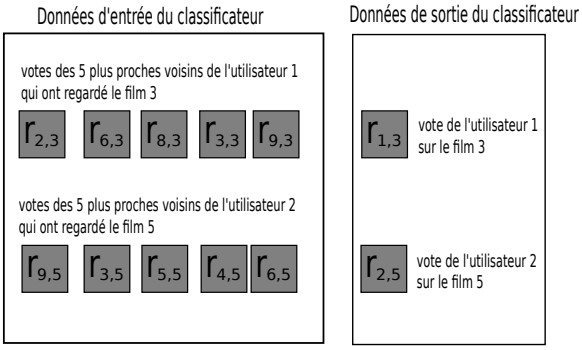
6. On voudrait à présent préparer les données qui permettront de valider votre future système de recommandation. Sachant que les votes varient de 1 à 5, on voudrait s'assurer que tous les cas de figures soient pris en compte lors de l'apprentissage. Pour ce faire, suivant votre fichier ratings1.csv, on vous demande de le subdiviser en trois fichiers distincts, ratings_train.csv, ratings_evaluation.csv et ratings_test.csv. Le fichier ratings_train.csv doit contenir 60% des données de ratings1.csv tandis que ratings_evaluation.csv et ratings_test.csv doivent contenir 20% des données de ratings1.csv chacun.

NB: La subdivision doit se faire de manière aléatoire. Assurez-vous d'avoir le même pourcentage de votes par fichier.

(a) On suppose que le vote $r_{u,i}$ donné par un utilisateur U_u sur un film I_i peut en tout temps être déterminé par une fonction de classification $\mathcal{D}()$ qui prend en entrée les votes des cinq (05) utilisateurs qui sont le plus proches de U_u (en terme de profils) et qui ont déjà regardé le film I_i.

Comme fonction de classification, on voudrait utiliser un arbre de décision *https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html*. À partir de vos données d'entrainement et d'évaluation construire votre modèle permettant de prédire le vote des utilisateurs.

En guise d'illustration, votre classificateur doit avoir en entrée et sortie les données comme suit,



NB: Il faudrait tester plusieurs hyper-paramètres pour votre abre de décision et retenir celui qui vous donnerait les meilleurs résultats suivant le critère $F1$ score sur les données d'évaluation. Vous devez également exploiter les clusters pour retrouver les plus proches voisins d'un utilisateur.

(b) À partir de votre prédicteur $\mathcal{D}()$ on vous demande de prédire les votes des utilisateurs qui se trouvent dans le fichier ratings_test.csv. Suivant le critère de $F1$ évaluer la performance de votre modèle.

3 Remise du TP

- Vous devez respecter vos groupes initiaux. Pour ceux qui ont travaillé seul dans le dernier TP seront mis le même groupe;
- La date de remise du TP est le 22 décembre 2023 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), movies1.csv, ratings1.csv, ratings_train.csv, ratings_evaluation.csv, rating_test.csv et l'ensemble de vos programmes. Vous devez vous assurez que l'on puisse reproduire les résultats reportés dans vos rapports à partir des fichiers que vous allez rendre.

- 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