3I026 — Projet

Analyse de données de films

Ariana CARNIELLI

Introduction

Ce projet s'intéresse à l'analyse de données décrivant des films obtenues à partir des bases de données des sites MovieLens et TMDb. On s'intéresse

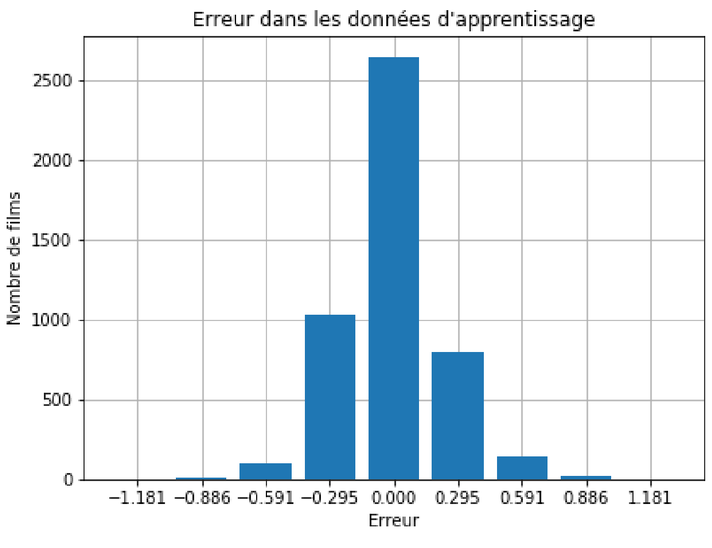
- à la régression supervisée de la note des films sur MovieLens à partir des notes d'acteurs et directeurs ;
- à la classification supervisée en genres ;
- à la catégorisation non-supervisée de films d'un même acteur.

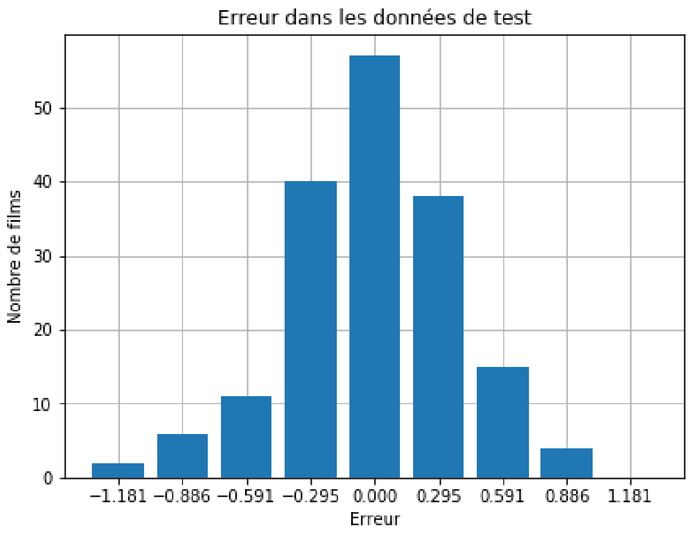
Bases de données

Les bases de données MovieLens et TMDb disponibles contiennent plusieurs informations, desquelles on a extrait, pour chaque film, ses genres, les notes obtenues sur MovieLens et TMDb et le nombre d'utilisateurs de ces sites ayant voté, la langue du film et le nombre d'acteurs et de membres de l'équipe de production séparés par chaque département. On a aussi récupéré depuis le site de TMDb les budgets et revenus des films, des données qui ne sont disponibles pour tous les films et qui n'ont donc pas été utilisées pour tous les tests.

Régression des notes

On a estimé la note d'un film à partir des informations disponibles. En particulier, on a cherché à donner une note moyenne pour chaque acteur et directeur pour les films entre 1988 et 2007 et prédire la note des films de 2008 à partir des notes des acteurs et directeurs qui y figurent. Les erreur entre la prédiction et la note réelle sont données dans les histogrammes ci-dessus.





Classification selon des genres

On commence les problèmes de classification par un cas simple : étant données les évaluations qu'un film a eu sur MovieLens et TMDb, peut-on savoir s'il appartient à un genre en particulier ? On choisit ici le genre « drame » car il est très bien représenté dans le dataset movies, avec environ 50% des films classifiés dans ce genre. En ne gardant que les 8434 films de la base de données ayant plus de 100 votes d'utilisateurs, les *accuracies* obtenues par différentes méthodes ont été :

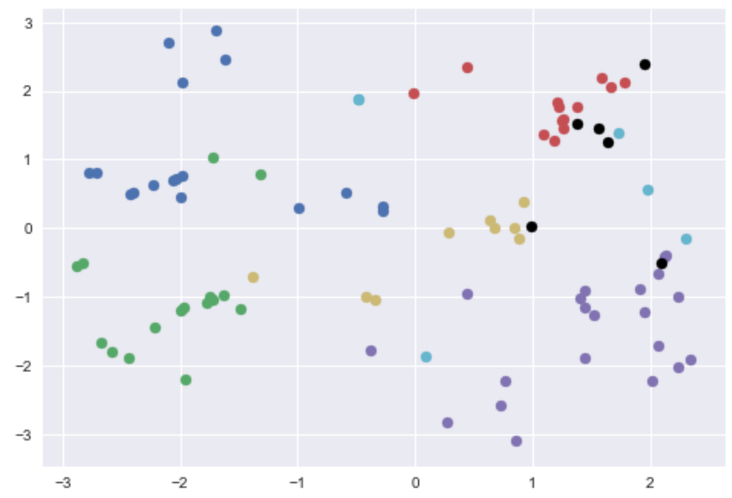
Classifieur	Apprent.	Test
Perceptron par gradient stochastique	64%	65%
Perceptron par gradient stochastique kernelisé	64%	64%
k plus proches voisins ($k = 5$)	73%	60%
Arbre de décision simple	64%	65%
Arbre de décision par bagging	65%	64%
Évaluation par OOB	65%	

Ces résultats sont meilleurs que ceux obtenus avec une classification par classe majoritaire. Cela, cependant, dépend du genre considéré. On a aussi testé la classification multi-genres à l'aide d'un perceptron par gradient stochastique pour chaque genre et une combinaison des résultats par la méthode un-contre-tous. Cela a donné des *accuracies* de 63% en apprentissage et 61% pour le test.

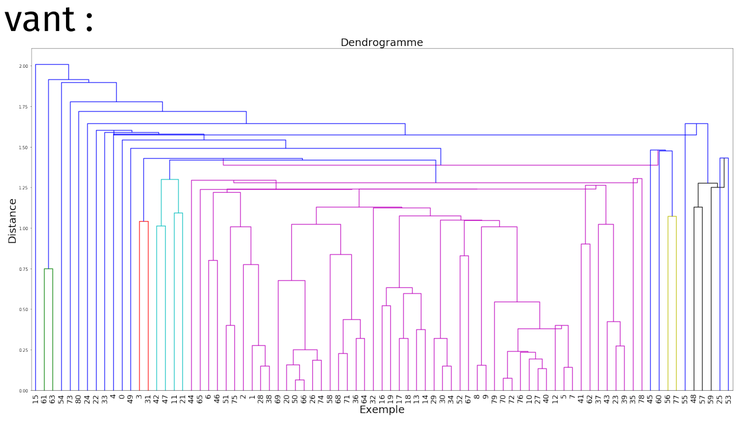
Catégorisation non-supervisée

On a décidé de catégoriser les films d'un même acteur. Cela a été fait dans un premier moment pour les films de Gérard Depardieu en utilisant la méthode des k-moyennes. On a utilisé plusieurs valeurs de k et évalué chacun d'entre eux par la méthode de Dunn pour choisir la meilleure valeur, qui est $k = 6$.

Avec une projection 2D par la méthode Isomap, on obtient le graphe suivant représentant les classes avec $k = 6$ (les points en noir représentent les centroïdes) :



On a aussi catégorisé les films de Jackie Chan à l'aide de la méthode de clustering hiérarchique. Le dendrogramme résultat est le suivant :



Le cluster le plus grand, représenté en magenta sur le dendrogramme, contient 53 films, tous d'action, genre par lequel Jackie Chan est connu. Ce sont les films où Jackie Chan est l'un des acteurs principaux. La sortie de ces films a été dans une plage de 20 ans autour de 1987, ce qui correspond à la plage d'années où il était très actif. La majorité de ces films n'est pas en anglais mais il y a un grand mélange avec d'autres langues.