LU3IN026 — Projet

Analyse de données de Google PlayStore Apps

Ung Thierry

Introduction

Ce projet s'intéresse à l'analyse de donnéesde Google PlayStore Apps obtenues à partir de la base de donnée de Kaggle.

On s'intéresse à la :

· classification supervisée des avis laissés par les utilisateurs

. classification non-supervisée de facteur qui permette de prédire la popularité d'une d'application mobile.

Bases de données

La base de donnée Google PlayStore Apps disponibles contient plusieurs informations, qu'on a extrait :

. Le nombre de téléchargement
. La note de l'application
. Le nombre d'avis
. La taille d'une application
. Le type d'application (gratuit/payant)

Classification des avis

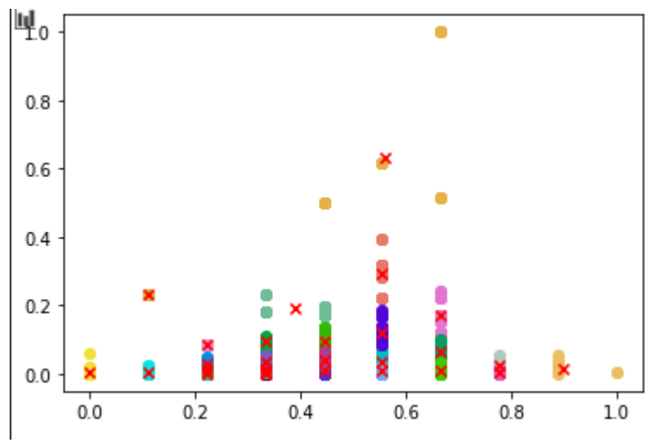
On commence les problèmes de classification par un cas simple : étant données les avis des utilisateurs, peut-on savoir s'ils ont tendances à laisser plus d'avis négatifs que positifs ? En ne gardant que les applications avec au moins 1 avis et des avis inférieurs à 50000,
afin de limiter notre modèle, les accuracies obtenues par différentes méthodes ont été :

Classifier	Minimum	Moyenne	Maximum	Ecart-type
KNN (k=3)	50%	61.5%	70%	0.057
KNN (k=10)	40%	55.4%	70%	0.099
Perceptron	38%	55.4%	66%	0.098
Perceptron biais	31%	42%	46%	0.057

Le meilleur résultat possible est eut avec le classifier KNN. Néanmoins, les fluctuations constatées sur le classifier KNN sont trop élévée, on ne peut pas utiliser ces données comme référence. Donc, avec les résultats qu'on a obtenus, on ne peut émettre de conclusion.

Classification non-supervisée

On a décidé de déterminer les facteurs qui permette de prédire la popularité d'une application mobile. On considère qu'une application est populaire quand : la note est supérieur ou égal à 4 et que le nombre de téléchargement est supérieur ou égal à 1 millions. On prédit le nombre de téléchargements en fonction des notes et des avis. On obtient le graphe suivant (chaque couleur représente un cluster et chaque croix, un centroide de clusters):



On effectue une validation croisée, pour obtenir les résultats suivants :

```
Apprentissage  1 :        |Yapp|=
Apprentissage  2 :        |Yapp|=
Apprentissage  3 :        |Yapp|=
Apprentissage  4 :        |Yapp|=
Apprentissage  5 :        |Yapp|=
Apprentissage  6 :        |Yapp|=
Apprentissage  7 :        |Yapp|=
Apprentissage  8 :        |Yapp|=
Apprentissage  9 :        |Yapp|=
Apprentissage  10 :       |Yapp|=

Résultat global avec crossval : moy
```

Il est donc possible de le nombre de téléchargement d'une application populaire 3 fois sur 4.