Fouille de données et medias sociaux
TP4 : Filtrage collaboratif

Olivier Schwander

8 octobre 2018

Tous les modèles seront évalués sur le jeu de données MovieLens http://files.grouplens.org/datasets/movielens/ :

- Version 100k http://files.grouplens.org/datasets/movielens/ml-100k.zip
- Version 1M http://files.grouplens.org/datasets/movielens/ml-1m.zip

Aide pour le chargement :

```
def loadMovieLens(path='/data/movielens'):

  # Get movie titles
  movies={}
  for line in open(path+'/u.item'):
    (id,title)=line.split('|')[0:2]
    movies[id]=title

  # Load data
  prefs={}
  for line in open(path+'/u.data'):
    (user,movieid,rating,ts)=line.split('\t')
    prefs.setdefault(user,{})
    prefs[user][movies[movieid]]=float(rating)
  return prefs
```

Il faudra mettre les données sous la forme de matrice sparse (voir https://www.scipy-lectures.org/advanced/scipy_sparse/index.html).

Question 1

En utilisant une SVD, construisez et évaluez un modèle de filtrage collaboratif.

Suggestion : http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

Question 2

En utilisant une NMF, construisez et évaluez un modèle de filtrage collaboratif.

Suggestion : http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html

Question 3

En utilisant une descente de gradient stochastique, construisez et évaluez un modèle de filtrage collaboratif (avec une pénalité L2).

Suggestion : https://pytorch.org

Question 4

Étudiez l'impact d'un terme de régularisation L2.

Question 5

Rajoutez la gestion des biais.

Question 6

Rédigez un rapport synthétique présentant vos résultats et comparant les différentes méthodes.