Matrix factorization for recommender systems

Kaoutar Boulif, Eric N'guessan, Salma Elghourbal
M2 IASD

PLAN

1- Exploration des données

2- Stratégie de split en train/test

3- La méthode SVD

4- Algorithme ALS

5- Algorithme Gradient descent

6- Algorithme PMF

7- Comparaison entre différentes approches

1. Exploration des données

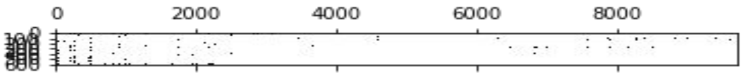
-base de données MovieLens décrit préférences utilisateurs pour des films

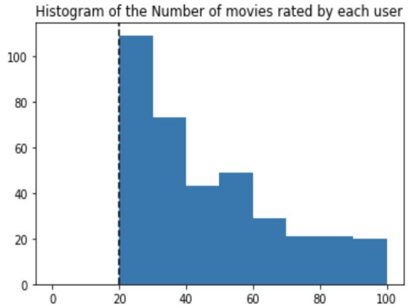
-la base de données MovieLens (version Small) contient 100,000 ratings, environ 9000 films et 600 utilisateurs.

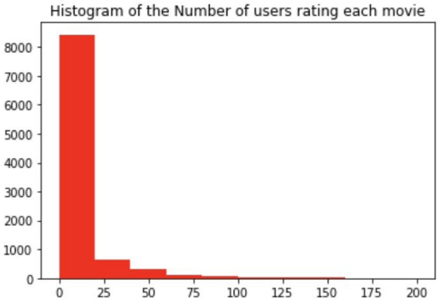
-Matrice de Ratings R small (610,9724) est creuse à 98.3%

-Chaque utilisateur a au moins évalué 20 films

-La majorité des fils ont moins de 25 notes.







2- Préparation des données et évaluation des méthodes de MF

Stratégie split train/test

Train set = la matrice R pour laquelle on enlève aléatoirement 15 ratings par user

Test set = une matrice de même dimension contenant uniquement les 15 ratings par user qu'on a enlevés

$$RMSE(R, \hat{R}, T) = \sqrt{\frac{\sum_{(i,u)\in T}(R_{iu} - \hat{R}_{iu})^2}{|T|}}$$

Where:

- R in the original rating matrix (sparse)
- $\hat{R}$ is the estimated rating matrix (dense)
- $T = \{(u, i) \mid R_{iu} \text{ is in the testing set }\}$

Objectif : Bonne prédiction des ratings du test set avec une méthode de factorisation de matrice en approchant R avec un produit de matrices de type : R ~ U* transpose(V)

2. Méthode de la SVD (Singular value decomposition)

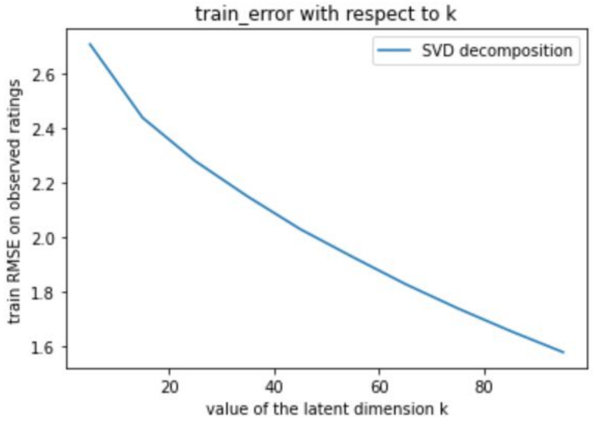
Formulation : **trouver U,V, S tels que : R ≈ USV^T**

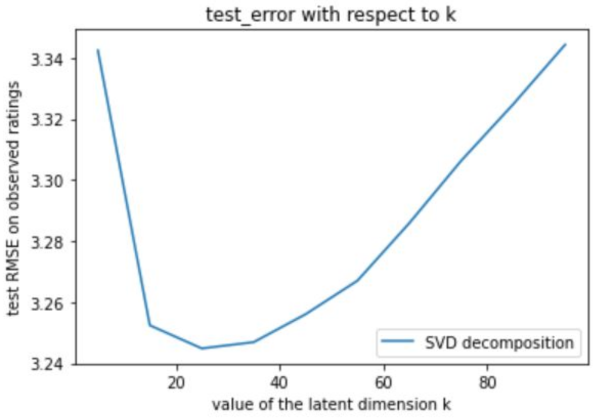
S : matrice diagonale de dimension **k*k** , U matrice des utilisateurs, V matrice des films

Implémentation : package svds de Scipy

Expériences : Impact du choix de k sur la RMSE

Valeur optimale : **k= 25**
RMSE de test minimale :
3.2448
RMSE de train correspondant :
2.2788





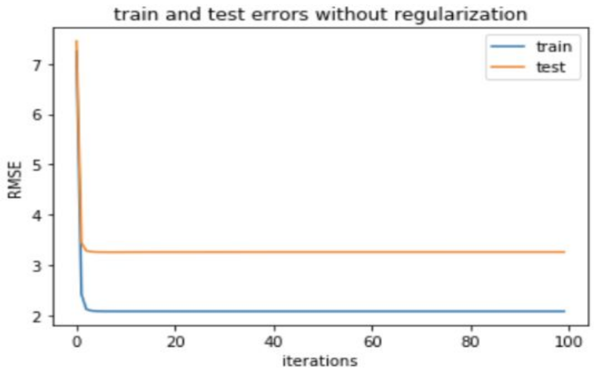
3- La méthode ALS

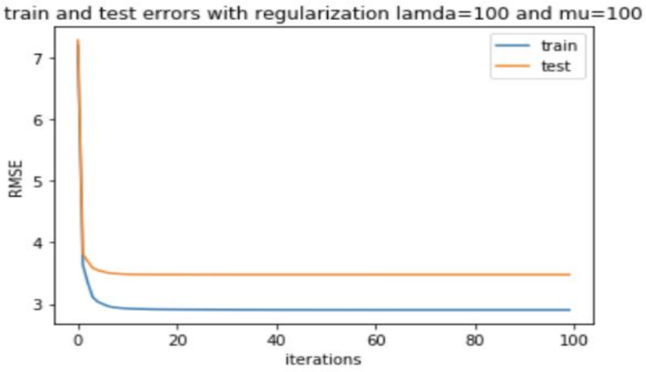
Formulation : $C(I, U) \quad = \quad \|R - IU^\top\|_{\mathcal{F}}^2 \quad + \quad \lambda\|I\|_{\mathcal{F}}^2 \quad + \quad \mu\|U\|_{\mathcal{F}}^2$

$$\frac{\partial C}{\partial I}(I, U) = -2RU + 2IU^\top U + 2\lambda I \qquad = 0 \qquad\qquad I_{t+1} = RU_t(U_t^\top U_t + \lambda\mathbb{I})^{-1}$$

$$\frac{\partial C}{\partial U}(I, U) = -2R^\top I + 2UI^\top I + 2\mu U \qquad = 0 \qquad\qquad U_{t+1} = R^\top I_t(I_t^\top I_t + \mu\mathbb{I})^{-1}$$

Expérience 1 :



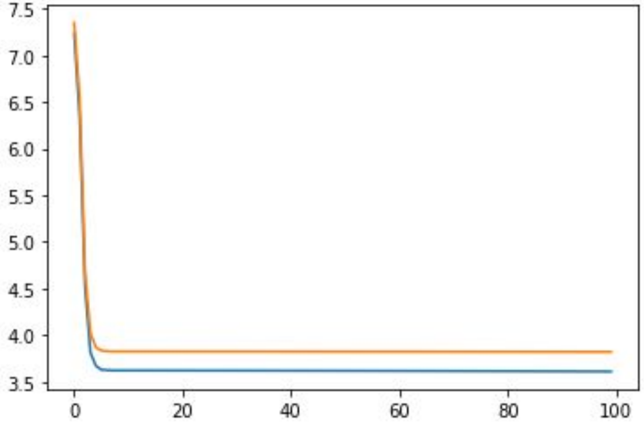


4- Descente de Gradient

Formulation : $$C(I, U) = \|R - IU^\top\|_{\mathcal{F}}^2 + \lambda\|I\|_{\mathcal{F}}^2 + \mu\|U\|_{\mathcal{F}}^2$$

- $I_{t+1} = I_t - \eta_t \frac{\partial C}{\partial I}(I_t, U_t)$
- $U_{t+1} = U_t - \xi_t \frac{\partial C}{\partial U}(I_t, U_t)$

Résultats :



Pros/Cons: -temps d'exécution très lent , +erreur RMSE faible

5- Probabilistic matrix factorization

Formulation : Trouver U et V qui maximisent la distribution A posteriori

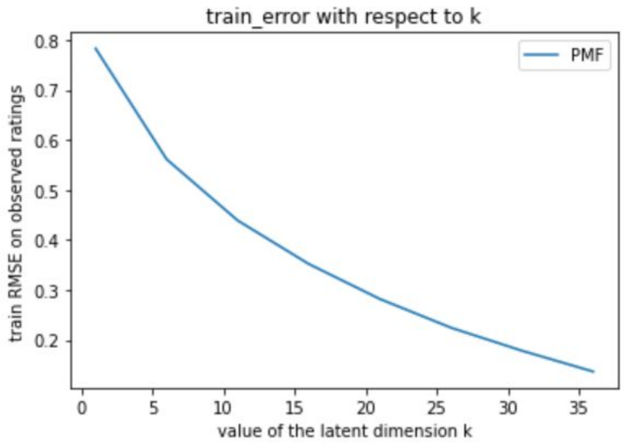
$$p(U, V | R, \sigma^2)$$

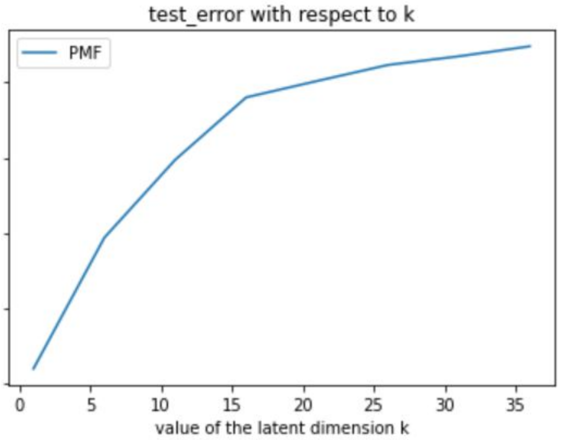
Ce qui est équivalent à maximiser L

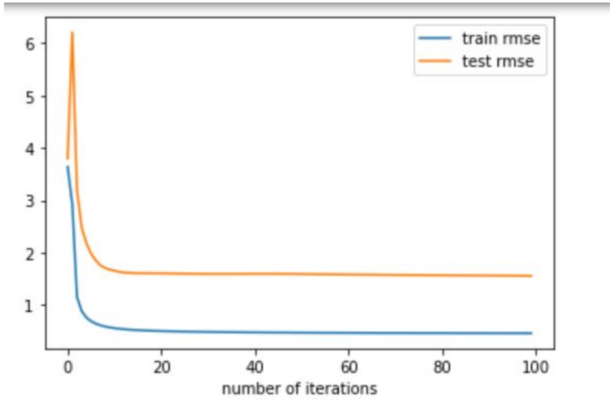
$$L = -\frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{M}(R_{ij} - U_i^T V_j)^2_{(i,j)\in\Omega_{R_{ij}}} + \lambda_U \sum_{i=1}^{N}\|U_i\|^2_{Fro} + \lambda_V \sum_{j=1}^{M}\|V_j\|^2_{Fro}\right)$$

Implémentation une ALS sur les lignes Ui et Vj

Expériences : Impact du choix de k sur la RMSE

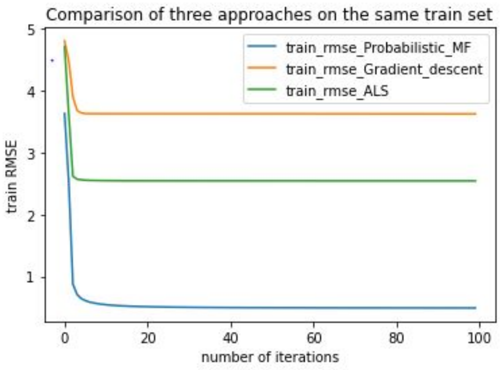


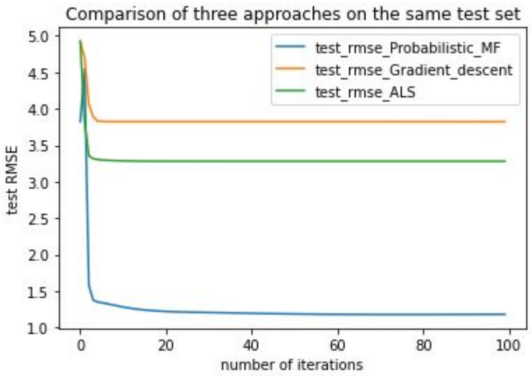




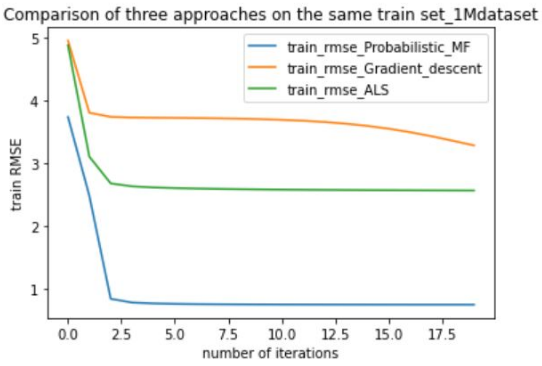
6- Comparaison des méthodes

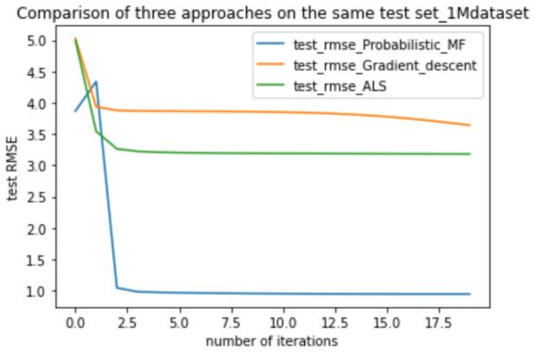
a- Dataset 1 (610 users, 9742 movies)



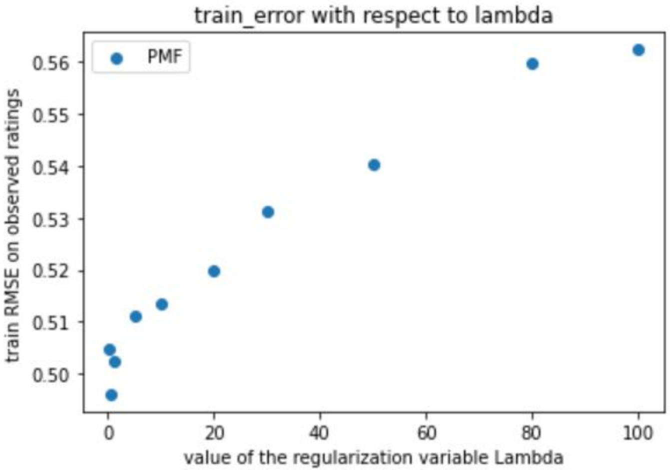


b- Dataset 2 (6040 users, 3706 movies)

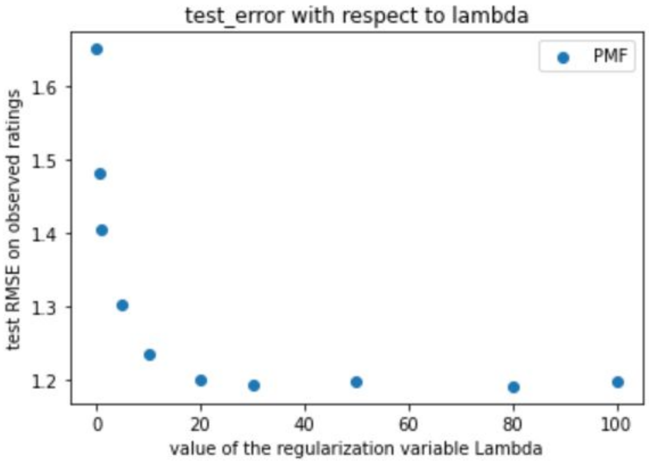




ANNEXES - PMF

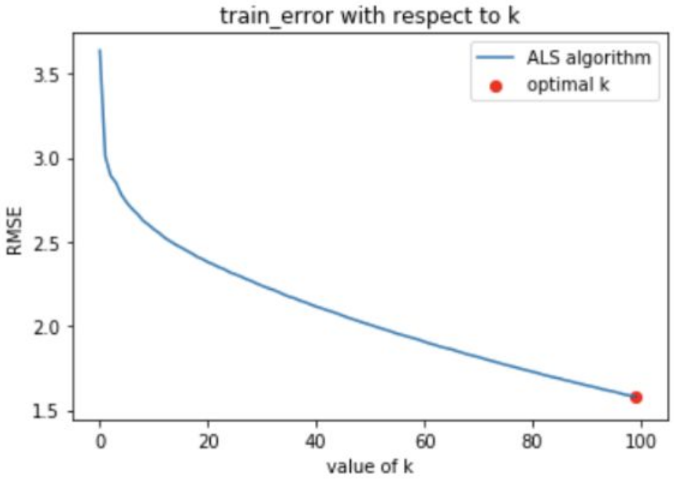


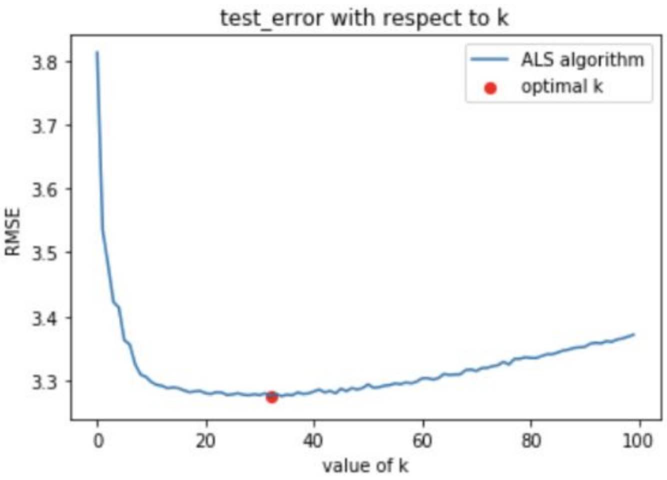
train_error with respect to lambda
PMF
0.56
0.55
0.54
0.53
0.52
0.51
0.50
train RMSE on observed ratings
0
20
40
60
80
100
value of the regularization variable Lambda



test_error with respect to lambda
PMF
1.6
1.5
1.4
1.3
1.2
test RMSE on observed ratings
0
20
40
60
80
100
value of the regularization variable Lambda

ANNEXES - ALS


train_error with respect to k
ALS algorithm
optimal k
RMSE
3.5
3.0
2.5
2.0
1.5
0
20
40
60
80
100
value of k


test_error with respect to k
ALS algorithm
optimal k
RMSE
3.8
3.7
3.6
3.5
3.4
3.3
0
20
40
60
80
100
value of k