

Systèmes des recommandation

Fouille de données avancées (2016-2017)

UFR MIME

Université Lille 3

7 décembre 2016

1 Motivation et données

- Netflix Prize
- Filtrage collaboratif

2 SVD

- calcul de similarité
- Imputation des valeurs

3 Évaluation d'erreurs

- Erreur moyenne quadratique

4 TP 1

[Sign In](#)

WATCH ANYWHERE. CANCEL ANYTIME

JOIN FREE FOR A MONTH



Tout pour Noël [Voir plus](#)



Ventes Flash



Toutes nos idées cadeaux

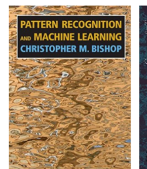
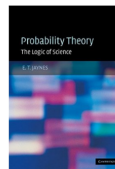
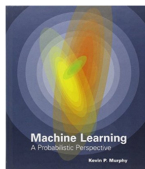
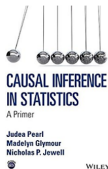


Pour les enfants



Pour les passionnés de high-tech

Recommandations pour vous en Livres





Actualités

Édition France ▼

Moderne ▼

À la une

Recommandations

Actualités près de c...

International

France

Économie

Science/High-Tech

Culture

Sport

Ligue 1

Cristiano Ronaldo

Unai Emery

FC Barcelone

Ligue Europa

Olympique lyonnais

Olympique de Marseille

Bayern Munich

Stéphane Ruffier

Lucho González

Santé

Sport

Foot - C1 - 8es - Ligue des champions (8es) : Monaco vs Manchester City, quelle fin de série ?

L'Équipe.fr - il y a 1 heure

Manchester City et Monaco seront opposés pour la première fois de leur histoire en Coupe d'Europe, en 8es de finale de la Ligue des champions. Anticipation de la confrontation en statistiques. Partager sur Facebook Tweeter Google+ 0 partages. Football ...

Monaco - Manchester City - Jardim : "On n'est pas les favoris mais..." Eurosport.fr

Ligue des champions: très compliqué pour Paris et Monaco Le Parisien

Éditorial : Ligue des champions: le Paris SG contre Barcelone, Monaco contre Manchester City en 8es de finale (tirage) La Croix

Autres

Ligue des champions de l'UEFA »

Association sportive de Monaco football club »

Le Monde

RTL.fr

Paris Match

Libération

Les Echos

Le Figaro S...

Le Parisien

Le Parisien

Les Echos

Le Huffington...

LT...

Ballon d'Or: Cristiano Ronaldo pour un quadruplé

Le Point - il y a 1 heure

C'est un tueur des surfaces et de suspense. Cristiano Ronaldo est le grand favori pour la conquête d'un quatrième Ballon d'Or après ses sacres en Ligue des champions en mai avec le Real Madrid, et à l'Euro-2016 en juillet avec le Portugal ...

Foot - C1 - PSG - Unai Emery (PSG) face au Barça, ce n'est pas trop ça

L'Équipe.fr - il y a 2 heures

Le PSG va affronter le Barça en huitièmes de finale de la Ligue des champions, pour la première fois avec Unai Emery sur son banc. Le Basque n'a jamais vraiment brillé face aux Catalans. Partager sur Facebook Tweeter Google+ 0 partages. Football ...

Foot - C3 - Ligue Europa : Les Verts joueront Manchester United en seizièmes de finale

L'Équipe.fr - il y a 2 heures

Saint-Etienne affrontera Manchester United en seizièmes de finale de la Ligue Europa, les 16 et 23 février prochains. Un match qui mettra donc aux prises les deux frères Pogba, Paul et Florentin, et marquera le grand retour de Zlatan en France dans un ...

Foot - L1 - OM - Baudot s'en va, Le Gall arrive

L'Équipe.fr - il y a 2 heures

Le médecin de l'OM Christophe Baudot va être remplacé par le médecin de Lille et de l'équipe de France, Franck Le Gall. Partager sur Facebook Tweeter Google+ 0 partages. Comme nous vous l'annoncions ce lundi, le docteur du LOSC Franck Le Gall va ...

PSG-Nice (2-2), l'antisèche : Un PSG plombé par son inefficacité dans les deux surfaces

Eurosport.fr - il y a 4 heures

<https://news.google.com/news/section?cf=all&hl=fr&pgz=1&ned=fr&q=Lucho+Gonzalez&topics=fr-fr&itct=tny91> un break de retard, le PSG a tout de même perdu deux points face à Nice dimanche

Motivation

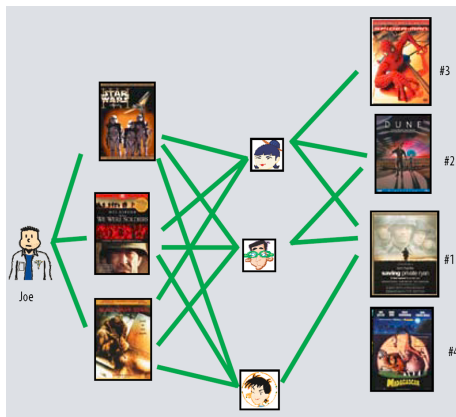


FIGURE – Recommendations ici est basé sur une similarités entres les utilisateurs.

Problème

				
	1	2	3	2
	2	3	1	?
	5	3	?	5
	2	1	4	1

FIGURE – Prédire les scores pour d'autres paires (d'utilisateurs, films) ?

Méthodologie :

- Chercher des utilisateurs qui ont les mêmes comportements avec l'utilisateur à qui l'on souhaite faire des recommandations
- Utiliser les notes des utilisateurs similaires pour calculer une liste de recommandations pour cet utilisateur.

Histoire : <https://www.youtube.com/watch?v=ImpV70uLxyw>

Création des préférences avec une grande base de données d'utilisateurs et prédiction des préférences d'utilisateurs basé sur les similarités avec d'autres utilisateurs.

- **Questions/Fouille Explicite**

- Noter les produits/items.
- Avoir un catalogue de recherche.
- Demander l'utilisateur de ordonner les produits par préférences
- De comparer deux produits
- Demander l'utilisateur de produire une liste de produits préférés

- **Questions/Fouille Implicite**

- d'observer l'activité d'utilisateur en ligne.
- d'observer durée d'observations sur des produits.
- Tenir compte des articles qu'un utilisateur achète en ligne.

Prédire les préférences d'après les “match” entre le profil d'utilisateurs et les caractéristiques de produits.

- Recommandation des films/séries (Netflix)
- Recommandation de produit relie (Amazon)
- Ordonnancements pageweb (Google)
- Recommandation sociale (Facebook)
- Recommandation nouvelles (Yahoo)
- Priority inbox et filtrage de spam (Google)
- Online dating (OK Cupid, adopte un mec)
- Computational Advertising (Yahoo)

Définition

Les modèles de factorisation matricielle mappent les utilisateurs et les éléments à un espace conjoint de dimension latente de dimension r , tel que les interactions utilisateur-produit sont modélisées par les produits scalaire dans cet espace.

Remarque

Pour chaque item/filme i est associé avec un vecteur $q_i \in \mathbb{R}^r$ et pour chaque utilisateur u il y a un vecteur $p_u \in \mathbb{R}^r$.

Décomposition SVD

$\mathbf{A} \in \mathbb{R}^{n \times d}$ est une matrice, il existe une décomposition

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$$

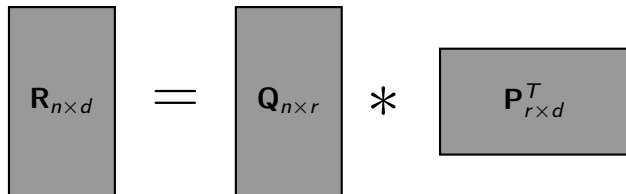
$$\boxed{\mathbf{A}_{n \times d}} = \boxed{\mathbf{U}_{n \times d}} * \boxed{\Sigma_{d \times d}} * \boxed{\mathbf{V}_{d \times n}^T}$$

- $\Sigma \in \mathbb{R}^{d \times d}$ est diagonale avec les entrées positives $\sigma_i > 0$
- $\mathbf{U} \in \mathbb{R}^{n \times d}$ vecteurs singuliers gauches (colonnes orthogonales)
- $\mathbf{V} \in \mathbb{R}^{d \times d}$ vecteurs singuliers droits (colonnes et rangs orthogonales)
- $\mathbf{U}\mathbf{U}^T = \mathbf{I}_n$ et $\mathbf{V}\mathbf{V}^T = \mathbf{I}_d$ et $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_d)$ ou les valeurs singulières sont ordonnées : $\sigma_1 > \sigma_2, \dots, > \sigma_d$

Décomposition SVD

$\mathbf{R} \in \mathbb{R}^{n \times d}$ est une matrice d'utilité :

$$\mathbf{R} = \mathbf{Q}\mathbf{P}^T$$



The diagram shows the equation $\mathbf{R}_{n \times d} = \mathbf{Q}_{n \times r} * \mathbf{P}_{r \times d}^T$. Each matrix is enclosed in a gray rectangular box. The boxes for $\mathbf{R}_{n \times d}$ and $\mathbf{Q}_{n \times r}$ are tall and narrow, while the box for $\mathbf{P}_{r \times d}^T$ is short and wide. The symbols $=$ and $*$ are placed between the boxes.

- $r \ll d$: dimension réduite
- $\mathbf{Q}_{n \times r} = \mathbf{U}_{n \times r}$
- $\mathbf{P}_{r \times d}^T = \Sigma_{r \times r} \mathbf{V}_{r \times d}^T$
- $\mathbf{P}$ et $\mathbf{Q}$ sont des matrices mince.

Prédire les notes manquantes

Comment déterminer la note manquante pour le film i pour utilisateur u

$$\hat{r}_{ui} = \mathbf{q}_i \cdot \mathbf{p}_u^T = \sum_r q_{ir} \cdot p_{ur}$$

		USERS											
ITEMS	1		3			5			5		4		
			5	4			4			2	1	3	
	2	4		1	2		3		4	3	5		
		2	4		5			4			2		
		4	3	4	2						2	5	
	1		3		3			2			4		

r-Facteurs

ITEMS	1	-4	2
	-5	6	5
	-2	3	5
	11	21	3
	-7	21	-2
	-1	7	3

USERS

r-Facteurs	11	-2	3	5	-2	-5	8	-4	3	14	24	-9
	-8	7	5	14	3	-1	14	29	-7	12	-1	13
	21	-4	6	17	24	9	-3	4	8	7	-6	1

Prédire les notes manquantes : Calcul de similarité

Comment déterminer la note manquante pour le film i pour utilisateur u

$$\hat{r}_{ui} = \mathbf{q}_i \cdot \mathbf{p}_u^T = \sum_r q_{ir} \cdot p_{ur}$$

		USERS											
ITEMS	1		3			5			5		4		
			5	4	?		4			2	1	3	
	2	4		1	2		3		4	3	5		
		2	4		5			4			2		
		4	3	4	2						2	5	
	1		3		3			2			4		

r-Facteurs

ITEMS	1	-4	2
	-5	6	5
	-2	3	5
	11	21	3
	-7	21	-2
	-1	7	3

USERS

r-Facteurs	11	-2	3	5	-2	-5	8	-4	3	14	24	-9
	-8	7	5	14	3	-1	14	29	-7	12	-1	13
	21	-4	6	17	24	9	-3	4	8	7	-6	1

Prédire les notes manquantes

Comment déterminer la note manquante pour le film i pour utilisateur u

$$\hat{r}_{ui} = \mathbf{q}_i \cdot \mathbf{p}_u^T = \sum_r q_{ir} \cdot p_{ur}$$

		USERS										
ITEMS	1		3			5			5		4	
			5	4	2.4		4			2	1	3
	2	4		1	2		3		4	3	5	
		2	4		5			4			2	
		4	3	4	2						2	5
1			3		3			2			4	

r-Facteurs

ITEMS	1	-4	2
	-5	6	5
	-2	3	5
	11	21	3
	-7	21	-2
	-1	7	3

USERS

r-Facteurs	11	-2	3	5	-2	-5	8	-4	3	14	24	-9
	-8	7	5	14	3	-1	14	29	-7	12	-1	13
	21	-4	6	17	24	9	-3	4	8	7	-6	1

Les méthodes de facteur latent

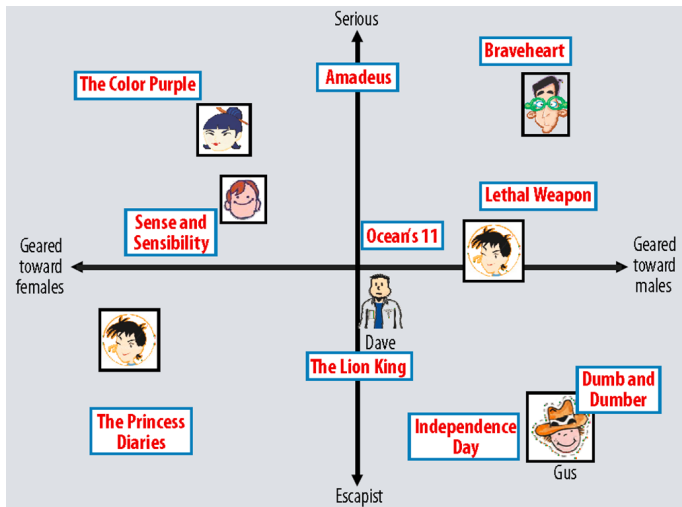


FIGURE – Une illustrations simple qui montre l'approche en facteur latent avec deux axes : 1. Sérieux/Surreal 2. Pour les Male/Female.

Comment calculer le SVD sans avoir tous les valeurs de matrice de utilité ?

- Remplir avec zéros ?
- Remplir avec moyennes de colonnes ?
- Remplir avec moyennes de rangs ?
- Quelle est la différences ?

Souvent les matrices **sparse** ont des problèmes avec imputations et il est mieux de retrouver des valeurs de p_u, q_i qui minimise :

$$\min_{q,p} \sum_{(u,i) \in X} (r_{ui} - q_i^T p_u)^2 + \lambda \cdot (\|q_i\|^2 + \|p_u\|^2) \quad (1)$$

résolu avec Descente en gradient stochastique ou *Alternating least squares*

Évaluation de performance

Users	Filmes				
	1	4	5	-	-
		3	4	-	5
			3		
			4		
			5		
	2		2	2	
		2	1	1	2
1					

Évaluation de performance

		Filmes			
Users	1	4	5	-	-
		3	4	-	5
			3		
			sous-matrice T pour evaluation d'erreur ou sous-matrice Test		
	2				
		2			
	1				

On cache un certain sous-matrice de donnée M sur laquelle on vas évaluer une erreur de prédiction.

Erreur moyenne quadratique

On va comparer les valeurs prédit par SVD (ou d'autres méthodes) et les vraies valeurs en sous-matrice T

$$\text{RMSE} = \frac{\sum_{(u,i) \in T} (r_{ui} - r_{ui}^*)^2}{|T|}$$

ou

- r_{ui} sont les valeurs prédites par l'algorithme
- r_{ui}^* sont les vraies valeurs dans la sous-matrice T
- $|T|$ est le nombre d'entrées dans la sous-matrice T

- Sur-Spécialisation

- Diversités des prédictions : Les films de **Harry potter's** ne sont pas les autres facteurs/similarités
- La contexte des prédictions : La même utilisateurs peut préférer des recommandations différents pendant des périodes différents
- Ordres de prédictions : Prendre en comptes les tomes, trilogies

- En pratique on veut mieux prédire les meilleurs (4-5) notes que les mauvais notes (1-2)

- Le RMSE pénalise les deux et il est indifférent
- En réalité pour un utilisateur on veut bien prédire son top-K films
- On peut du coup calculer par exemples les erreurs dans les top-K meilleurs films pour un utilisateurs, mais en globale il devient difficile a donner une bonne erreur.

Les fonctions nécessaires

Jeu de données

Enjeux TP