

Programme ta culture !

Jill-Jênn Vie

7 mars 2017

Hi, I'm JJ

- ▶ 2012 Président de Prologin
- ▶ 2014 Agrégé de mathématiques + Girls Can Code! + Mangaki
- ▶ 2016 Docteur en informatique + 2 livres d'algorithmique
 - ▶ *Programmation efficace* (concours de programmation) w/ Dürr
 - ▶ *Les clés pour l'info* (concours des ENS) w/ Mansuy, Belghiti
- ▶ 2017 Postdoc à RIKEN Tokyo

Recommandation de films

Problème

- ▶ Chaque utilisateur note peu de films (1 %)
- ▶ Comment inférer les notes manquantes ?

				
Bob	4	3	5	1
Ana	2	4	2	5
Elain	5	4	4	3
Sulman	2	3	1	4

Cartographie des goûts

Problème

Comment mettre un nom sur les goûts de l'utilisateur ?

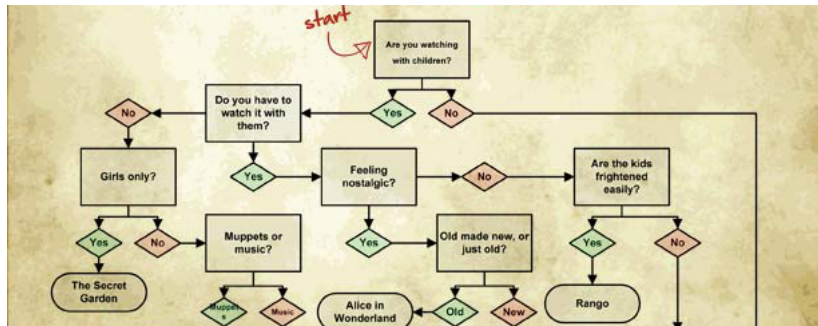
Exemple

Votre corps est composé à 80 % d'eau, à 10 % de Woody Allen, à 10 % de films noirs.

Élicitation des préférences

Problème

Comment faire un diagnostic des goûts de l'utilisateur ?









What to Watch on Netflix, Silver Oak Casino, 2013

Jeu de données 1 : Movielens

top picks [see more](#)

based on your ratings, MovieLens recommends these movies

Band of Brothers	Casablanca	One Flew Over the Cuckoo's Nest	The Lives of Others	Sunset Boulevard	The Third Man
2001 [R] 705 min ⚙	1942 [PG] 102 min ⚙	1975 [R] 133 min ⚙	2006 [R] 137 min ⚙	1950 [NR] 110 min ⚙	1949 [NR] 104 min ⚙
					
★★★★★	★★★★★	★★★★★	★★★★★	★★★★★	★★★★★

- ▶ 700 utilisateurs
- ▶ 9000 films
- ▶ 100000 notes

Jeu de données 2 : Mangaki



Death Note



Dog Days



Princesse Mononoké



The Place Promised in Our Early Days

- ▶ 2100 utilisateurs
- ▶ 15000 œuvres *anime / manga / OST*
- ▶ 310000 notes *fav / like / dislike / neutral / willsee / wontsee*
- ▶ Un utilisateur note quelques œuvres *Élicitation des préférences*
- ▶ Et reçoit des recommandations *Filtrage collaboratif*

Tout algorithme de machine learning supervisé

$\text{fit}(X, y)$

X		y
user_id	work_id	rating
24	823	like
12	823	dislike
12	25	favorite
...

$\hat{y} = \text{predict}(X)$

X		\hat{y}
user_id	work_id	rating
24	25	?disliked
12	42	?liked

Évaluation : RMSE

Si je prédis \hat{y} pour chacun des n paires à évaluer, alors que la vérité est y^* :

$$RMSE(\hat{y}, y^*) = \sqrt{\frac{1}{n} \sum_i (\hat{y}_i - y_i^*)^2}.$$

KNN → mesurer la similarité

K plus proches voisins

- ▶ Score de similarité entre utilisateurs :

$$\text{score}(u, v) = \frac{\mathcal{R}_u \cdot \mathcal{R}_v}{||\mathcal{R}_u|| \cdot ||\mathcal{R}_v||}.$$

- ▶ Identifions les k plus proches voisins d'un utilisateur
- ▶ Recommandons-lui ce que ses voisins ont aimé qu'il n'a pas vu

Astuce

Si R est la matrice $N \times M$ des $\frac{\mathcal{R}_u}{||\mathcal{R}_u||}$, alors pour obtenir la matrice des scores entre utilisateurs $N \times N$ il suffit de calculer RR^T .

PCA, SVD \rightarrow réduire la dimension

Analyse de composantes principales, déc. en valeurs singulières

$$R = \begin{pmatrix} \mathcal{R}_1 \\ \mathcal{R}_2 \\ \vdots \\ \mathcal{R}_n \end{pmatrix} = \boxed{} = \boxed{C} \boxed{P}$$

Chaque ligne \mathcal{R}_u est une combinaison linéaire des profils P .

Profils types

Si P P_1 : aventure P_2 : romance P_3 : plot twist

Et C_u 0,2 -0,5 0,6

$\Rightarrow u$ aime un peu l'aventure, déteste la romance, adore les plot twists.

$R = (U \cdot \Sigma) V^T$ où $U : N \times r$ et $V : M \times r$ sont orthogonales et $\Sigma : r \times r$ diagonale.

ALS-WR → variantes

Moindres carrés alternés avec régularisation pondérée

Rappel : R notes, C coefficients, P profils.

$$R = CP = CF^T \text{ i.e. } r_{ij} \simeq C_i \cdot F_j.$$

L'erreur de reconstruction est minimisée

SVD : $\sum_{i,j} (r_{ij} - C_i \cdot F_j)^2$ (vaut 0 si r est égal au rang)

ALS : $\sum_{i,j} \text{connus} (r_{ij} - C_i \cdot F_j)^2$

ALS-WR : $\sum_{i,j} \text{connus} (r_{ij} - C_i \cdot F_j)^2 + \lambda(\sum_i N_i \|C_i\|^2 + \sum_j M_j \|F_j\|^2)$

WALS by Tensorflow™ :

$$\sum_{i,j} w_{ij} \cdot (r_{ij} - C_i \cdot F_j)^2 + \lambda(\sum_i \|C_i\|^2 + \sum_j \|F_j\|^2)$$

À votre avis, qui gagne ?

NMF → interpréter les composantes

Non-negative matrix factorization

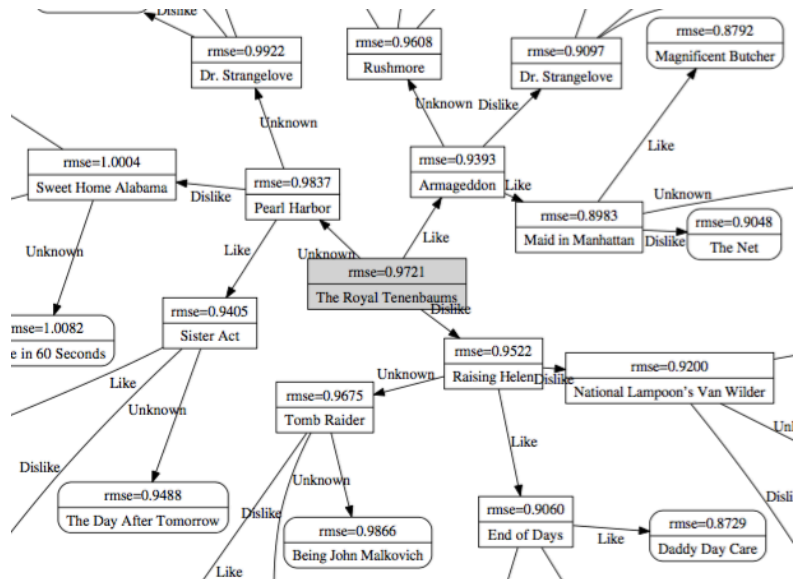
On suppose R , C et $F \geq 0$.

$$\text{NMF} : \sum_{i,j} (r_{ij} - C_i \cdot F_j)^2 + \lambda(\sum_i \|C_i\|^2 + \sum_j \|F_j\|^2)$$

Avantage

Les composantes sont plus facilement interprétables.

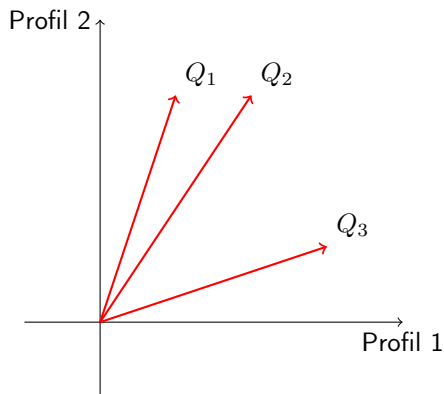
Arbre de décision



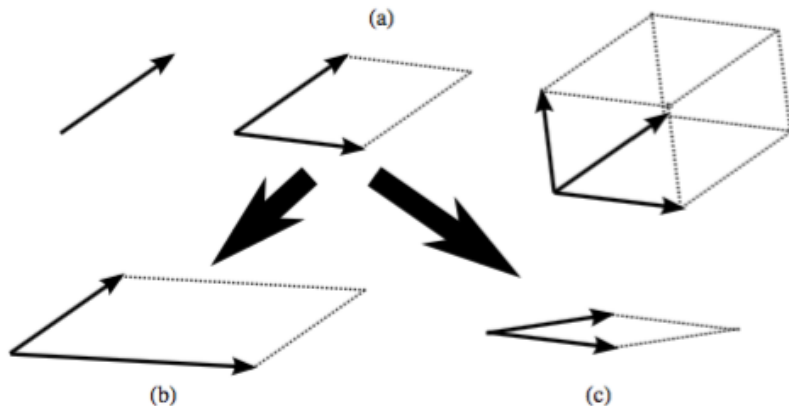
Bon équilibre entre likes, dislikes et « ne sait pas »

Comment poser des bonnes questions ?

Si on a quelques vecteurs de F :



Interprétation géométrique de la diversité



- ▶ Déterminant = carré du volume du paralléloèdre formé
- ▶ Vecteurs peu corrélés (**diversifiés**) augmentent le volume
- ▶ On souhaite échantillonner k éléments parmi n efficacement

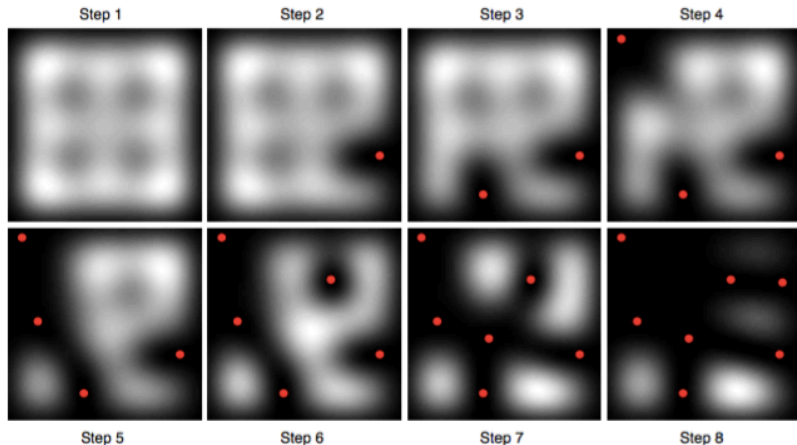
DPP : pour modéliser la diversité

Processus à point déterminantal

points éloignés les uns des autres

=

films diversifiés



DPP

On souhaite échantillonner n œuvres

$K : n \times n$ **matrice de similarité** sur les œuvres (semi-définie positive)

P est un **processus à point déterminantal** si l'échantillon Y vérifie :

$$\forall A \subset \{1, \dots, n\}, \quad P(A \subseteq Y) \propto \det(K_A) = \text{Vol}(\{x_i\}_{i \in A})^2$$

Il existe un algo en $O(nk^3)$ pour échantillonner k parmi n !

Exemple

$$K = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 5 & 6 & 7 \\ 3 & 6 & 8 & 9 \\ 4 & 7 & 9 & 1 \end{pmatrix}$$

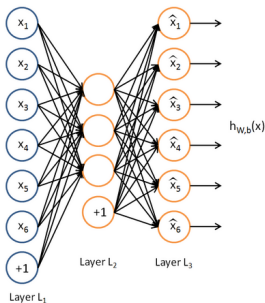
$A = \{1, 2, 4\}$ sera inclus dans la sélection avec probabilité

$$K_A = \det \begin{pmatrix} 1 & 2 & 4 \\ 2 & 5 & 7 \\ 4 & 7 & 1 \end{pmatrix}$$

SAE

Avec SVD, on passe d'une représentation des gens en dimension M films à une représentation en dimension r , qui est censée suffire à décrire leurs notes.

Ça ne vous fait pas penser à quelque chose ?



Sparse autoencoder!

Une petite anecdote

- ▶ Le 2 octobre 2006, Netflix a lancé un concours :
Le premier qui bat notre algorithme de plus de 10 % remportera 1 million de dollars.
et ont filé des données anonymisées
- ▶ La moitié de la communauté en IA s'est jetée sur le problème
- ▶ Le 8 octobre, quelqu'un a battu Cinematch
- ▶ Le 15 octobre, 3 équipes l'avaient battu, dont 1 de 1,06 %
- ▶ Le 26 juin 2009, une équipe 1 bat Cinematch de 10,05 %
→ **last call** : plus qu'un mois pour gagner
- ▶ Le 25 juillet 2009, une **équipe 2** bat Cinematch de 10,09 %
- ▶ L'équipe 1 fait 10,09 % aussi
- ▶ 20 minutes plus tard **l'équipe 2** fait 10,10 %
- ▶ ... En fait, les deux équipes étaient ex æquo sur le sous-ensemble de validation
- ▶ ... Du coup c'est la première équipe à envoyer ses résultats qui a gagné (équipe 1, 10,09 %)

Confidentialité des utilisateurs

- ▶ Août 2009, Netflix annonce une saison 2
- ▶ Entre-temps, en 2007 deux chercheurs de l'université du Texas ont été capables d'**identifier** les utilisateurs du jeu de données anonymisées en croisant les données avec IMDb
- ▶ (année approximative de naissance, code postal, films vus)
- ▶ En décembre 2009, 4 utilisateurs de Netflix ont attaqué Netflix en justice
- ▶ Mars 2010, arrangement à l'amiable, la plainte est close

Programme ta culture !

Initiation à la programmation à la création numérique
(option ICN au lycée séries L, ES et S)

Activités

- ▶ Exploration de données massives
- ▶ Création artistique numérique
- ▶ Recommandation de films

`tryalgo.org/programme-ta-culture`

Quelle série voir, finalement ?

Données



Kaiba

Chat & souris



*Lupin the
Third:
The Woman
Called Fujiko
Mine*

Merci de votre attention !



Retrouvez ces slides

Sur `research.mangaki.fr`

Retrouvez le code

Sur GitHub : `github.com/mangaki`

Suivez-nous

Sur Twitter : `@mangakifr`