Introduction à la Data Science

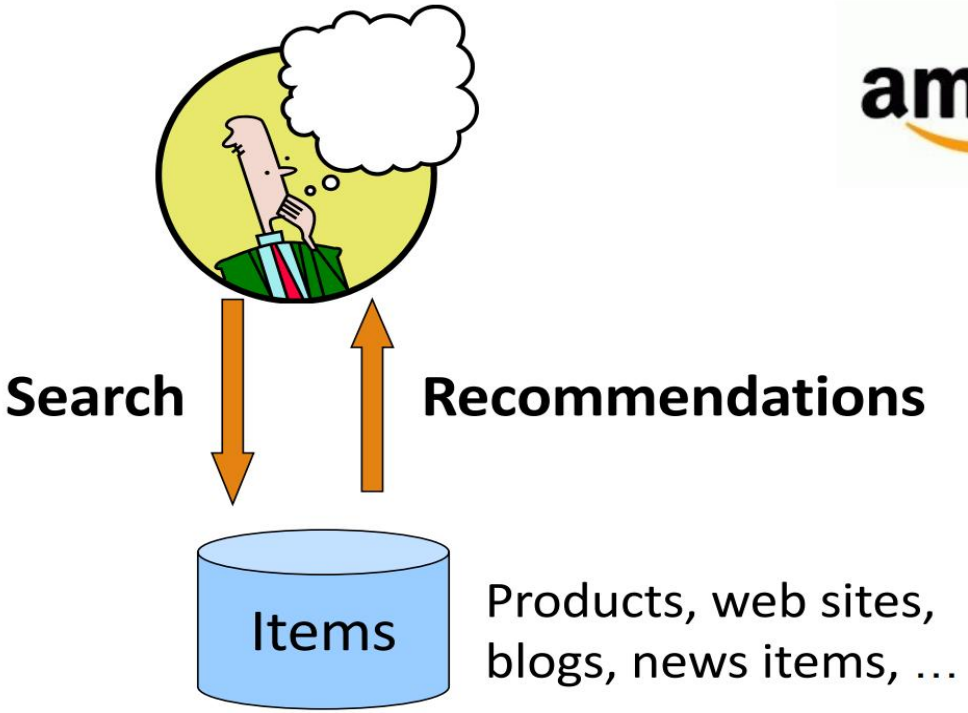
Chapitre 8.
Recommander Systems

Introduction

- Un système de recommandation est un outil conçu pour interagir, de manière automatisée et fournir des informations ou des articles susceptibles d'intéresser l'utilisateur.
  - Pour ce faire, il utilise un espace d'information complexe comme l'ensemble des articles, et ses caractéristiques, que le système recommande à l'utilisateur.
  - Ces articles peuvent être des livres, des films, des produits à acheter...

Introduction

Recommendations



Introduction

- Systèmes extrêmement courants et utilisés dans une variété d'applications tels que
  - Les films sur Netflix.
  - La musique sur Pandora ou Spotify.
  - Les produits sur Amazon.com.
  - Les actualités, les articles de recherche, les requêtes de recherche, les tags sociaux…

Quand et Pourquoi avons-nous besoin d'un système de recommandation?

- A cause de l'immensité de la quantité d'informations disponibles…
  - Les **individus ne peuvent pas être experts** dans tous les domaines où ils sont des utilisateurs.
  - Les **individus n'ont pas suffisamment de temps** pour chercher l'article parfait à acheter.
  - D'où l'utilité des systèmes de recommandation
- En particulier, les systèmes de recommandation sont intéressants lorsqu'il faut traiter les problèmes suivants:
    - Solutions pour de grandes quantités de bonnes données ;
    - Réduction de la charge cognitive sur l'utilisateur ;
    - Permettre la découverte de nouveaux éléments aux utilisateurs.

Comment fonctionne les systèmes de recommandation?

- La plupart des systèmes de recommandation suivent l'une des deux approches de base :
  - Le **filtrage basé sur le contenu** (CBF).
  - Le **filtrage collaboratif** (CF).

Filtrage basé sur le contenu

- Principe: « **Montre-moi davantage de ce que j'ai aimé** »
  - **Idée principale :** Recommander des articles au client **x** similaires aux articles précédents évalués positivement par **x**.
    - **Recommandations de films**
      - Recommander des films avec le(s) même(s) acteur(s), réalisateur, genre, ...
    - **Sites web, blogs, actualités**
      - Recommander d'autres sites avec des types similaires ou des mots similaires.

Filtrage basé sur le contenu

- Cette approche recommandera des éléments similaires à ceux que l'utilisateur a aimés auparavant.
  - Elle basera les recommandations sur les descriptions des éléments et sur un profil des préférences de l'utilisateur.
- Le calcul de la similarité entre les éléments est la partie la plus importante de ces méthodes et repose sur le contenu des éléments eux-mêmes.

Filtrage basé sur le contenu

- Les fonctions de similarité les plus utilisées incluent :
  - La **distance euclidienne**

$$sim(a,b) = \frac{1}{1 + \sqrt{\sum_{p \in P} (r_{a,p} - r_{b,p})^2}}$$

  - La **corrélation de Pearson**

$$sim(a,b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r_a})(r_{b,p} - \bar{r_b})}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r_a})} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r_b})}}$$

  - La **cosine distance**

$$sim(a,b) = \frac{a \cdot b}{|a| \cdot |b|}$$

Filtrage collaboratif

- Principe: "**Dis-moi ce qui est populaire parmi les utilisateurs qui ont des goûts similaires aux miens**"
  - Cette approche suppose que des utilisateurs similaires ont tendance à aimer des éléments similaires.
- Deux types de filtrage collaboratif sont possibles:
  - Le **filtrage collaboratif basé sur l'utilisateur**
  - Le **filtrage collaboratif basé sur l'élément**.

Filtrage collaboratif

• Filtrage collaboratif basé sur l'utilisateur

- Au lieu d'utiliser les caractéristiques du contenu des articles pour déterminer quoi recommander**,** trouvez des **utilisateurs similaires** et **recommandez des articles qu'ils aiment !**
  - Considérons l'utilisateur **x** et l'article **i** non évalué.
    - Trouver l'ensemble **N** d'autres utilisateurs dont les évaluations sont "**similaires**" aux évaluations de **x**.
    - Estimer les évaluations de **x** pour l'article **i** en se basant sur les évaluations pour i des utilisateurs dans **N**.

Filtrage collaboratif: user to user

- Trouver des utilisateurs similaires et recommander des articles qu'ils aiment :
  - Représenter les utilisateurs par leurs lignes dans la matrice d'utilité.
  - Deux utilisateurs sont similaires si leurs vecteurs sont similaires.

	Harry Potter			Twilight	Star Wars		
	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Filtrage collaboratif: user to user

- Trouver des utilisateurs similaires
  - Supposons que r_x soit le vecteur des ratins de x

r_x = [*, _, _, *, ***]
r_y = [*, _, **, **, _]

r_x = {1, 0, 0, 1, 3}
r_y = {1, 0, 2, 2, 0}

  - La mesure de similarité cosinus est donnée par :

Cosine similarity measure

- $\text{sim}(x, y) = \cos(r_x, r_y) = \frac{r_x \cdot r_y}{||r_x|| \, ||r_y||}$

  - **Problème**: Cette représentation conduit à des résultats non intuitifs (cfr les zéros: manque de ratings ou un rating réellement égal à 0?).

Filtrage collaboratif: user to user

- Problème avec la matrice brute des utilités cosine

	Harry Potter			Twilight	Star Wars		
	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- Intuitivement, nous voulons que: $sim(A, B) > sim(A, C)$

$$\text{sim(A,B)} = \frac{4 \times 5}{\sqrt{4^2 + 5^2 + 1^2}\sqrt{5^2 + 5^2 + 4^2}} = 0.380$$

$$\text{sim(A,C)} = \frac{5 \times 2 + 1 \times 4}{\sqrt{4^2 + 5^2 + 1^2}\sqrt{2^2 + 4^2 + 5^2}} = 0.322$$

- Oui, $0.380 > 0.322$ mais cela fonctionne à peine

Filtrage collaboratif: user to user

Problem with raw cosine

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- Problem with cosine:
  - C really loves SW
  - A hates SW
  - B just hasn't seen it
- Another problem: we'd like to normalize the raters
  - D rated everything the same; not very useful

Filtrage collaboratif: user to user

Mean-Centered Utility Matrix: subtract the means of each row

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	$2/3$			$5/3$	$-7/3$		
B	$1/3$	$1/3$	$-2/3$				
C				$-5/3$	$1/3$	$4/3$	
D		0					0

- Now a 0 means no information
- And negative ratings means viewers with opposite ratings will have vectors in opposite directions!

Filtrage collaboratif: user to user

Modified Utility Matrix: subtract the means of each row

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	$2/3$			$5/3$	$-7/3$		
B	$1/3$	$1/3$	$-2/3$				
C				$-5/3$	$1/3$	$4/3$	
D		0					0

$$\text{Cos(A,B)} = \frac{(2/3) \times (1/3)}{\sqrt{(2/3)^2 + (5/3)^2 + (-7/3)^2}\sqrt{(1/3)^2 + (1/3)^2 + (-2/3)^2}} = 0.092$$

$$\text{Cos(A,C)} = \frac{(5/3) \times (-5/3) + (-7/3) \times (1/3)}{\sqrt{(2/3)^2 + (5/3)^2 + (-7/3)^2}\sqrt{(-5/3)^2 + (1/3)^2 + (4/3)^2}} = -0.559$$

Now A and C are (correctly) way further apart than A,B

Filtrage collaboratif: user to user

Terminological Note: subtracting the mean is **mean-centering**, not **normalizing**

(normalizing is dividing by a norm to turn something into a probability), but the textbook (and common usage) sometimes overloads the term “normalize”

Filtrage collaboratif: user to user

Finding similar users with overlapping-item mean-centering

Let r_x be the vector of user x's ratings

r_x = {1, 0, 0, 1, 3}

r_y = {1, 0, 2, 2, 0}

r_x = [*, _, _, *, ***]

r_y = [*, _, **, **, _]

Mean-centering:

- For each user x, let $\bar{r_x}$ be mean of r_x (ignoring missing values)
- $\bar{r_x}$ = (1 + 1 + 3)/3 = 5/3 $\bar{r_y}$ = (1 + 2 + 2)/3 = 5/3
- Subtract this average from each of their ratings
  - (but do nothing to the "missing values"; they stay "null").
  - mean centered r_x = {-2/3, 0, 0, -2/3, 4/3}

One new idea: Keep only items they both rate (unlike 2 slides ago)

r_x = {-2/3, [redacted], -2/3, [redacted]}

r_x = {-2/3, -2/3}

r_y = {-2/3, [redacted], 1/3, [redacted]}

r_y = {-2/3, 1/3}

Now take cosine:

- Now compute cosine between user vectors
  - cos([-2/3, -2/3], [-2/3, 1/3])

Filtrage collaboratif: user to user

Mean-centered overlapping-item cosine similarity

Let r_x be the vector of user x's ratings,and $\overline{r_x}$ be its mean (ignoring missing values)

Instead of basic cosine similarity measure

- $\text{sim}(x, y) = \cos(r_x, r_y) = \frac{r_x \cdot r_y}{||r_x||\ ||r_y||}$

Mean-centered overlapping-item cosine similarity

(Variant of Pearson correlation)

- S_{xy} = items rated by both users x and y

$$sim(x,y) = \frac{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x})(r_{ys} - \overline{r_y})}{\sqrt{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x})^2} \sqrt{\sum_{s \in S_{xy}} (r_{ys} - \overline{r_y})^2}}$$

Filtrage collaboratif: user to user

Rating Predictions

From similarity metric to recommendations for an unrated item *i*:

Let r_x be the vector of user x's ratings

Let N be the set of k users most similar to x who have rated item i

Prediction for item *i* of user *x*:

- Rate i as the mean of what k-people-like-me rated i

$$r_{xi} = \frac{1}{k} \sum_{y \in N} r_{yi}$$

- Even better: Rate i as the mean weighted by their similarity to me …

$$r_{xi} = \frac{\sum_{y \in N} s_{xy} \, r_{yi}}{\sum_{y \in N} s_{xy}}$$

- **Many other tricks possible…**

Shorthand:

$$s_{xy} = sim(x, y)$$

Filtrage collaboratif

- **Filtrage collaboratif basé sur l'élément**
  - Trouve des éléments similaires à ceux que j'ai aimés précédemment.
  - Dans ce type, nous construisons d'abord une matrice élément-élément qui détermine les relations entre les paires d'éléments ;
  - Puis en utilisant cette matrice et les données sur l'utilisateur U actuel, nous déduisons les goûts de l'utilisateur.
  - Typiquement, cette approche est utilisée dans le domaine où les gens qui achètent x achètent aussi y (Utilisé par Amazon).

Filtrage collaboratif: item item

So far: User-user collaborative filtering

Alternate view that often works better: Item-item

- For item ***i***, find other similar items
- Estimate rating for item ***i*** based on ratings for those similar items
- Can use same similarity metrics and prediction functions as in user-user model
- "Rate i as the mean of my ratings for other items, weighted by their similarity to i"

$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \; r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

N(i;x)…set of items rated by ***x*** *and* similar to ***i***
s_{ij}… similarity of items ***i*** and ***j***
r_{xj}…rating of user ***x*** on item ***j***

Filtrage collaboratif: item item

- **In practice, item-item often works better than user-user**
- **Why?** Items are simpler, users have multiple tastes
  - (People are more complex than objects)

Système de recommandation hybride

- Les **approches hybrides** peuvent être mises en œuvre de plusieurs manières :
  - en faisant des prédictions basées sur le contenu et collaboratives séparément, puis en les combinant ;
  - en ajoutant des capacités basées sur le contenu à une approche collaborative (et vice versa) ;
  - en unifiant les approches en un seul modèle.