



Systèmes de recommandation basés sur les modèles de classification

- Problème de recommandation vu comme un problème de classification
- Recommandation discrete
 - Deux classes (aime/n'aime pas)
 - Multi-classes (1/2/3/4/5)
- Recommandation continue
 - Régression

Systemes de recommandation basés sur les modèles de classification : Naive Bayes

➤ Rappel
$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

$$P(r_{uj} = v_s | \text{Observed ratings in } I_u) = \frac{P(r_{uj} = v_s) \cdot P(\text{Observed ratings in } I_u | r_{uj} = v_s)}{P(\text{Observed ratings in } I_u)}$$

➤ I_u correspond aux ratings de l'utilisateur u

➤ On peut le simplifier par :

$$P(r_{uj} = v_s | \text{Observed ratings in } I_u) \propto P(r_{uj} = v_s) \cdot P(\text{Observed ratings in } I_u | r_{uj} = v_s)$$

➤ Avec :

$$P(\text{Observed ratings in } I_u | r_{uj} = v_s) = \prod_{k \in I_u} P(r_{uk} | r_{uj} = v_s)$$

Estimation des ratings

$$\begin{aligned}\hat{r}_{uj} &= \frac{\sum_{s=1}^l v_s \cdot P(r_{uj} = v_s | \text{Observed ratings in } I_u)}{\sum_{s=1}^l P(r_{uj} = v_s | \text{Observed ratings in } I_u)} \\ &= \frac{\sum_{s=1}^l v_s \cdot P(r_{uj} = v_s) \cdot P(\text{Observed ratings in } I_u | r_{uj} = v_s)}{\sum_{s=1}^l P(r_{uj} = v_s) \cdot P(\text{Observed ratings in } I_u | r_{uj} = v_s)} \\ &= \frac{\sum_{s=1}^l v_s \cdot P(r_{uj} = v_s) \cdot \prod_{k \in I_u} P(r_{uk} | r_{uj} = v_s)}{\sum_{s=1}^l P(r_{uj} = v_s) \cdot \prod_{k \in I_u} P(r_{uk} | r_{uj} = v_s)}\end{aligned}$$

calcul

- $q_s / \sum_{t=1}^l q_t$ estime la probabilité de $P(r_{uj}=v_s)$
- Où $q_1 \dots q_l$ est le nombre d'utilisateur ayant donné respectivement les rating $v_1 \dots v_l$
- En réalité afin que l'on obtienne de meilleurs résultats, il faut lisser les probabilités grâce au laplacian

$$P(r_{uj} = v_s) = \frac{q_s + \alpha}{\sum_{t=1}^l q_t + l \cdot \alpha}$$

- Permet d'avoir une probabilité de $1/l$ lorsqu'il n'y a pas de rating

Exemple

- On veut prédire les valeurs pour l'utilisateur 3

Item-Id ⇒	1	2	3	4	5	6
User-Id ↓						
1	1	-1	1	-1	1	-1
2	1	1	?	-1	-1	-1
3	?	1	1	-1	-1	?
4	-1	-1	-1	1	1	1
5	-1	?	-1	1	1	1

Example

$$q_s / \sum_{t=1}^l q_t$$

Item-Id \Rightarrow	1	2	3	4	5	6
User-Id \Downarrow						
1	1	-1	1	-1	1	-1
2	1	1	?	-1	-1	-1
3	?	1	1	-1	-1	?
4	-1	-1	-1	1	1	1
5	-1	?	-1	1	1	1

$$P(r_{31} = 1 | r_{32}, r_{33}, r_{34}, r_{35}) \propto P(r_{31} = 1) \cdot P(r_{32} = 1 | r_{31} = 1) \cdot P(r_{33} = 1 | r_{31} = 1) \cdot \\ \cdot P(r_{34} = -1 | r_{31} = 1) \cdot P(r_{35} = -1 | r_{31} = 1)$$

Exercice

- Prédisez les valeurs pour l'utilisateur 2 et l'utilisateur 5

Item-Id ⇒	1	2	3	4	5	6
User-Id ↓						
1	1	-1	1	-1	1	-1
2	1	1	?	-1	-1	-1
3	?	1	1	-1	-1	?
4	-1	-1	-1	1	1	1
5	-1	?	-1	1	1	1



Content-Based Models

- Non plus basé sur les avis des utilisateurs
- Basés sur le contenu des items
 - Parfois contenu trop important
 - Livres
 - Musiques
 - Films
 - Parfois pas de contenu
 - objets



Utilisation des méta-données



Méta-données



- Ex Livre

- Auteur
- Mots du titre
- Genre
- ...

- Ex Film

- Genre
- Titre du film
- Acteurs
- Réalisateurs

- Ex Musique

- Instruments
- Auteurs
- Type

- Produit

- Nom
- Caractéristiques

Méta-données : encodage

- Idée : représenter l'ensemble des caractéristiques par un vecteur
 - Présence => 1
 - Absence => 0
- Vecteur très grand
 - Taille du vocabulaire
 - Très sparse (beaucoup de 0)
- Peu d'information stockée
 - Seul présence/absence

	d1	d2	d3	...
<i>cinéma</i>	0	1	0	...
<i>football</i>	1	1	0	...
<i>rugby</i>	0	0	1	...
...



Calcul de similarité

- Lors de l'arrivée de nouveaux items
 - Calcul de la similarité par rapport à des items d'un utilisateur
 - Ceux qu'il a déjà acheté, vu ...
 - Ceux qu'il a noté positivement
- Utilisation d'une mesure de similarité
 - Cosinus
 - Distance de manathan

Exemple similarité : cosinus

$$\text{C}_{\text{osine}}(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^d x_i y_i}{\sqrt{\sum_{i=1}^d x_i^2} \sqrt{\sum_{i=1}^d y_i^2}}$$

Exemple

Genre ⇒ Movie-Id ↓	Comedy	Drama	Romance	Thriller	Action	Horror	Like or Dislike
1	1	0	1	0	0	0	Dislike
2	1	1	1	0	1	0	Dislike
3	1	1	0	0	0	0	Dislike
4	0	0	0	1	1	0	Like
5	0	1	0	1	1	1	Like
6	0	0	0	0	1	1	Like
<i>Test-1</i>	0	0	0	1	0	1	?
<i>Test-2</i>	0	1	1	0	0	0	?

Exercice : Quel est le film à recommander ?

$$C_{\text{psine}}(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^d x_i y_i}{\sqrt{\sum_{i=1}^d x_i^2} \sqrt{\sum_{i=1}^d y_i^2}}$$

Naive bayes

- Même principe de pour le collaborative filtering

$$P(c(\bar{X}) = 1 | x_1 \dots x_d) \propto P(c(\bar{X}) = 1) \cdot \prod_{i=1} P(x_i | c(\bar{X}) = 1)$$

$$P(c(\bar{X}) = -1 | x_1 \dots x_d) \propto P(c(\bar{X}) = -1) \cdot \prod_{i=1}^d P(x_i | c(\bar{X}) = -1)$$

Naive Bayes

Table 4.1: Illustration of the Bayes method for a content-based system

Keyword \Rightarrow	Drums	Guitar	Beat	Classical	Symphony	Orchestra	Like or Dislike
Song-Id \Downarrow							
1	1	1	1	0	0	0	Dislike
2	1	1	0	0	0	1	Dislike
3	0	1	1	0	0	0	Dislike
4	0	0	0	1	1	1	Like
5	0	1	0	1	0	1	Like
6	0	0	0	1	• 1	0	Like
<i>Test-1</i>	0	0	0	1	0	0	?
<i>Test-2</i>	1	0	1	0	0	0	?

$$P(c(\bar{X}) = 1 | x_1 \dots x_d) \propto P(c(\bar{X}) = 1) \cdot \prod_{i=1}^d P(x_i | c(\bar{X}) = 1)$$

$$P(c(\bar{X}) = -1 | x_1 \dots x_d) \propto P(c(\bar{X}) = -1) \cdot \prod_{i=1}^d P(x_i | c(\bar{X}) = -1)$$

Evaluation

- Mesure l'écart des erreurs (MSE):

$$MSE = \frac{\sum_{(u,j) \in E} e_{uj}^2}{|E|}$$

$$e_{uj} = \hat{r}_{uj} - r_{uj}$$

- Ou RMSE

$$RMSE = \sqrt{\frac{\sum_{(u,j) \in E} e_{uj}^2}{|E|}}$$

- C'est la mesure la plus couramment utilisé



Evaluation



- Diversité
 - Ne donne pas toujours des recommandations très proches
- Sérendipité
 - Notion de surprise. Doit être nouveau mais surprenant.
- Nouveauté
 - Qui n'a pas été vu avant par l'utilisateur