Filtrage Collaboratif

Mohamed EL KHMISSI
Mohamed Abdelmalek BOUARROUDJ
Rapport de projet d'Apprentissage Statistique
2022/2023

1 Introduction

De nos jours tous les géants mondiaux de la tech mettent en place des systèmes de recommandation pour déterminer les produits les plus susceptibles d'intéresser les clients à partir d'un certain nombre d'informations, c'est ce qu'on appelle filtrage collaboratif. les systèmes les plus utilisés sont basés sur la recherche de quelques facteurs latents susceptibles d'expliquer en faible dimension les interactions entre clients et produits.
Dans ce rapport nous allons essayer de prédire les notes qu'un lecteur pourrait donner à un livre qui n'a pas lu encore par la méthode de filtrage collaboratif appeler NMF (factorisation matricielle non négative).

2 Présentation des données

Tout d'abord BookCrossing est un site de réseau social intelligent qui met les gens en contact par les livres, le but est de faire passer un livre d'un lecteur à un autre sans frontière.
Nous avons récolté notre base de données "Book-Crossing Dataset" de ce site,[1] elle comprend 278.860 BookCrosseurs (lecteurs) et 271.360 livres, et on dispose pour chaque lecteur les évaluations qu'il a donné à des livres lus (plus que 1.149.781 évaluations entre 1 le pire et 10 le meilleur).
Pour notre cas pratique nous avons pris tous les livres qui ont été noté par au moins 100 lecteurs et tous les lecteurs qui ont notés au moins 300 livres soit au total 553 lecteurs et 731 livres.

	User-ID	ISBN	Book-Rating	Book-Title	Book-Author
0	276725	034545104X	0	Flesh Tones: A Novel	M. J. Rose
1	276726	0155061224	5	Rites of Passage	Judith Rae
2	276727	0446520802	0	The Notebook	Nicholas Sparks
3	276729	052165615X	3	Help!: Level 1	Philip Prowse
4	276729	0521795028	6	The Amsterdam Connection : Level 4 (Cambridge ...	Sue Leather

FIGURE 1 – Extrait de la base de données

A partir de cette base de données nous avons crées une matrice $X \in \mathbb{R}^{n\times p}$, contenant pour chaque lecteur i (ligne) une note d'appréciation de 1 à 10 d'un livre j (colonne), la matrice X est donc trés creuse et contient que des valeurs non négatives.

$$X = \begin{array}{cc} & p \\ & \begin{pmatrix} 8 & 0 & .. & 0 & 1 \\ 1 & 3 & .. & 6 & 0 \\ : & : & . & : & : \\ 0 & 0 & .. & 0 & 0 \\ 0 & 0 & .. & 0 & 5 \end{pmatrix} n \end{array}$$

Remarque la valeur "0" signifie une donnée manquante (le livre n'a pas encore été lu ou évalué), c'est d'ailleurs la valeur qu'on veut prédire pour recommander au lecteur (i) le livre (j) si la note prédite dépasse un certain seuil.

3 Principe de la Factorisation Matricielle Non-Négative

L'idée principale derrière un problème de factorisation matricielle non-négative est d'apprendre un modèle latent de lecteurs $W \in \mathbb{R}^{n\times r}$ et de livres $H \in \mathbb{R}^{r\times p}$ de sorte que la reconstruction $\widehat{X}_{ij} = W_i H_j$ entre un lecteur i et un livre j estime la note X_{ij}. Autrement dit la recherche de deux

matrices $W_{n\times r}$ et $H_{r\times p}$ avec un faible rang (r) et ne contenant que des valeurs positives ou nulles et dont le produit se rapproche de X.

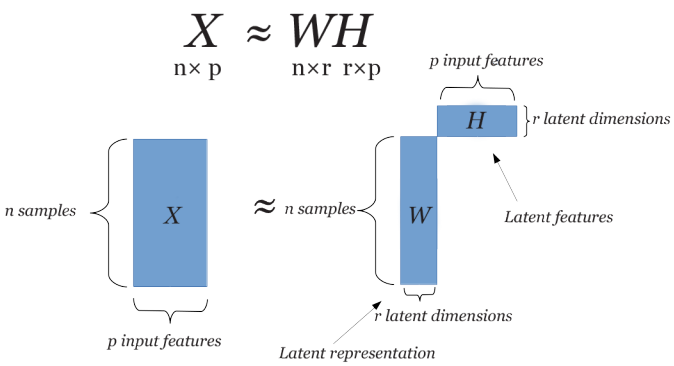

FIGURE 2 – Factorisation Matricielle

Schématiquement, W_{ik} dénote l'appétence du i-ème utilisateur pour le k-ème facteur latent, tandis que H_{kj} décrit quelle part du j-ème lecteur intervient dans le k-ème facteur latent ; le modèle suppose que la note X_{ij} est la somme, sur tous les facteurs latents k, des produits $W_{ik} \times H_{kj}$.

4 Problème de la factorisation

La factorisation est résolue par la recherche d'un optimum local du problème d'optimisation :

$$\min_{W,H\geq 0} [L(X, WH), P(W, H)]$$

L est une fonction perte mesurant la qualité d'approximation (critère de Frobenius). Comme on connaît les valeurs (les évaluations) x_{ij} pour les couples (i, j) appartenant à un ensemble χ, et les autres valeurs sont inconnues. On s'intéresse donc qu'aux termes connus de la matrice. Ainsi, on cherche à minimiser :

$$L(X, WH) = \frac{1}{2} \|X - WH\|_F^2 = \frac{1}{2} \sum_{(i,j)\in\chi} (x_{ij} - w_i h_j)^2$$

Et P une fonction de pénalisation optionnelle.

$$\begin{aligned} P(W, H) = {} & \alpha_W \times l1_ratio \times p \times ||vec(W)||_1 \\ & + \alpha_H \times l1_ratio \times n \times ||vec(H)||_1 \\ & + 0.5 \times \alpha_W \times (1 - l1_ratio) \times p \times ||W||_F^2 \\ & + 0.5 \times \alpha_H \times (1 - l1_ratio) \times n \times ||H||_F^2 \end{aligned}$$

- α_W (resp α_H) est une constante qui multiplie les termes de régularisation de W (resp H).
- $l1_ratio$ est un paramètre de mélange de régularisation, la pénalité est une pénalité :

$$\begin{cases} L1 & si \quad l1_ratio = 1 \\ L2 & si \quad l1_ratio = 0 \\ Elastic\ net & si \quad 0 < l1_ratio < 1 \end{cases}$$

Remarque Les algorithmes NMF convergent au mieux vers des optimums locaux (à cause de la contrainte de positivité de W et H), alors que la SVD bénéficie d'une convergence "globale" néanmoins la SVD est moins adaptée au contexte car les solutions ne sont pas cohérentes avec l'objectif recherché : des évaluations (notes) positives.

5 Exemple

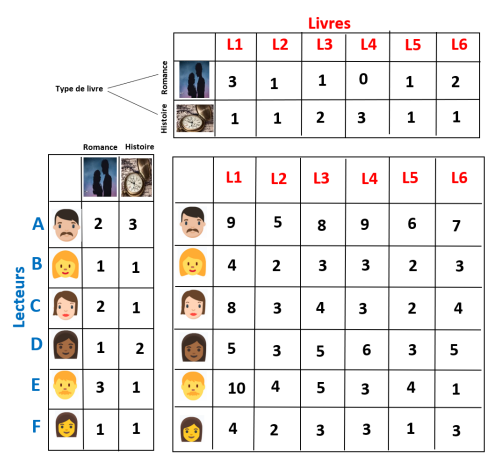

FIGURE 3 – Exemple de factorisation

Si on prend l'exemple ci-dessus, la factorisation matricielle non-négative a permis de trouver deux facteurs latents (r=2) qui sont le type de livres (juste pour l'exemple, car en réalité, nous

ne pouvons pas les décrire, ça serait plutôt f1 et f2).
Les coordonnées de la matrice W représentent l'appétence qu'un lecteur à pour le type de livre k, et les coordonnées de la matrice H représentes le poids du facteur k pour un livre donné.

Le lecteur E il a une appétence de 3 pour les livres de romance et 1 pour les livres d'histoire, et le livre L1 a un poids de 3 pour les livres de romance et 1 pour les livres d'histoire, en faisant le produit des facteurs latents du lecteur E et le livre L1 nous retrouvons la note que ce lecteur est successible de donnée à ce livre.

6 Cas pratique

Pour répondre à notre problématique qui est de prédire les notes qu'un lecteur pourrait donner à un livre nous allons utiliser la méthode NMF de la librairie sickit-learn qui se base sur la résolution du problème que nous avons présentés dans la section 4.
Nous avons pris les paramètres suivants :

- La méthode "mu" (Multiplicative Update solver) pour la factorisation de la matrice X en WH.
- $l1_ratio = 1$ et donc pénalité L1 ce qui va forcer certaines valeurs de H et W a être égale à 0.
- Rang de factorisation r=(8, 9, 10, 11, 12, 13, 14)
- $\alpha = \alpha_W = \alpha_H$ = (0, 0.025, 0.05, 0.075, 0.1)

Nous avons pris la matrice X de taille (553, 731) qui contient 404.243 coordonnées dont a peu prés 9.000 valeurs remplis et 395.243 valeurs manquantes que nous allons essayer de prédire.

Pour calibrer le meilleur couple de paramètre (r, α) nous allons procéder comme suit :

- Prendre un couple (r,α)
- Prendre 100 coordonnées non nulles de la matrices X les vider puis calculer $\widehat{X}$
- Calculer le RMSE entre X et $\widehat{X}$ (pour les 100 coordonnées dont on connaît les vraies valeurs)
- Répéter le procédure 20 fois
- Calculer l'erreur le RMSE moyen de chaque couple (r,α) puis prendre l'optimal.

À la fin, nous avons créé une fonction qui, à partir d'un seuil donné elle renvoie les titres des livres à recommander à des groupes de lecteurs. (voir le script pour plus détails[2])

Références

[1] http ://www2.informatik.uni-freiburg.de/ cziegler/BX/

[2] https ://github.com/ELKHMISSI/NMF