Un modèle stochastique pour les systèmes de recommandation

Gérard Biau, Benoît Cadre, Laurent Rouviere

Un modèle stochastique pour les systèmes de recommandation

Gérard BIAU [a], Benoît CADRE [b] et Laurent ROUVIÈRE [c,1]

[a] LSTA & LPMA, Université Pierre et Marie Curie – Paris VI
Boîte 158, 175 rue du Chevaleret, 75013 Paris, France
gerard.biau@upmc.fr

[b] IRMAR, ENS Cachan Bretagne, CNRS, UEB
Avenue Robert Schuman, 35170 Bruz, France
Benoit.Cadre@bretagne.ens-cachan.fr

[c] IRMAR, Université Rennes 2, CNRS, UEB
Place du Recteur Henri Le Moal, 35043 Rennes Cedex, France
laurent.rouviere@univ-rennes2.fr

Résumé

Les systèmes de recommandation établissent des suggestions personnalisées à des individus concernant des objets (livres, films, musique) susceptibles de les intéresser. Les recommandations sont généralement basées sur l'estimation de notes relatives à des objets que l'utilisateur n'a pas consommés. En dépit d'une littérature abondante, les propriétés statistiques des systèmes de recommandation ne sont pas encore clairement établies. Dans ce travail, nous proposons un modèle stochastique pour les systèmes de recommandation et nous analysons ses propriétés asymptotiques lorsque le nombre d'utilisateurs augmente. Nous établissons la convergence de la procédure sous de faibles hypothèses concernant le modèle. Les vitesses de convergence sont également présentées.

Mots clés : Systèmes de recommandation, convergence, similarité de type cosinus.

Abstract

Collaborative recommendation is an information-filtering technique that attempts to present information items (movies, music, books) that are likely of interest to a user. In its most common form, the problem is framed as trying to estimate ratings for items that have not yet been consumed by a user. Despite wide-ranging literature, very little is known about the statistical properties of recommendation systems. To provide an initial contribution to this, we propose in the present work to set out a general stochastic model for collaborative recommendation and analyze its asymptotic performance as the number of users grows. We establish consistency of the procedure under mild assumptions on the model. Rates of consistency are also provided.

Index Terms: Recommender systems, consistency, cosine-type similarity.

[1]auteur correspondant : laurent.rouviere@univ-rennes2.fr

1 Introduction

Les systèmes de recommandation émettent des suggestions à des utilisateurs concernant des objets susceptibles de les intéresser. Parmi les nombreux exemples d'applications, on peut citer la recommandation de livres, restaurants, films et remarquer que les sites Web amazon.com, match.com, movielens.com ou encore allmusic.com possèdent leur propre système de recommandation.

Le processus de recommandation débute par une série de questions posées à des utilisateurs concernant leurs préférences vis-à-vis d'un certain type d'objet. Par exemple, pour un système de recommandation concernant des films, les utilisateurs commencent par noter les films qu'ils ont déjà vus. Les notes sont alors collectées dans une matrice où chaque ligne représente un utilisateur et chaque colonne un objet (film). Un exemple est présenté dans le tableau 1 où les notes se situent entre 1 et 10, le symbole "NA" signifie que l'utilisateur n'a pas noté le film correspondant.

	Armageddon	Platoon	Rambo	Rio Bravo	Star wars	Titanic
Jim	NA	6	7	8	9	NA
James	3	NA	10	NA	5	7
Steve	7	NA	1	NA	6	NA
Mary	NA	7	1	NA	5	6
John	NA	7	NA	NA	3	1
Lucy	3	10	2	7	NA	4
Stan	NA	7	NA	NA	1	NA
Johanna	4	5	NA	8	3	9
Bob	NA	3	3	4	5	?

TAB. 1 – Un exemple de notes de 9 utilisateurs concernant 6 films. Les films sont notés entre 1 et 10 et le symbole "NA" signifie que l'utilisateur n'a pas noté le film correspondant.

Une fois les données recueillies, le système de recommandation doit dans un premier temps prédire les notes des objets non évalués, puis dans un second temps fournir une recommandation à l'utilisateur basée sur ces prévisions. De nombreuses méthodes issues de diverses communautés ont été proposées. On pourra par exemple se référer aux articles de Abernethy et al. [1], Sarwar et al. [5] ainsi qu'aux surveys de Adomavicius et Tuzhilin [3] et Adomavicius et al. [2]. Quelle que soit la méthode utilisée, le point crucial consiste à identifier des utilisateurs "proches" de l'utilisateur à qui on souhaite fournir la recommandation. La notion de proximité entre utilisateurs peut varier selon l'application, elle est néanmoins le plus souvent basée sur des notions de corrélation ou de cosinus mesurés entre les utilisateurs. Dans ce travail nous proposons un modèle stochastique permettant d'étudier les systèmes de recommandation. Ce modèle prend notamment en

compte la structure particulière des données (possibilités de non réponse ou de mise à jour des réponses de la part des utilisateurs). Les propriétés asymptotiques du modèle (convergence, vitesse de convergence) sont ensuite présentées.

2 Une modélisation séquentielle des systèmes de recommandation

2.1 Le modèle

On désigne par $d+1$ ($d \geq 1$), le nombre d'objets (films) et par n le nombre d'utilisateurs. On suppose que les notes données par les utilisateurs aux différents objets sont à valeurs dans $(\{0\} \cup [1, s])^{d+1}$ où s est un réel strictement plus grand que 1 représentant la note maximale. Par convention, la note 0 signifie que l'utilisateur n'a pas répondu à l'objet correspondant. Dans l'exemple du tableau 1, on a $n = 8$, $d = 5$ et $s = 10$. Une fois les notes des n utilisateurs collectées, un nouvel utilisateur (Bob) révèle à son tour ses préférences pour les d premiers objets mais pas pour le $(d+1)$ème (le film Titanic dans notre exemple). Le problème consiste à trouver une stratégie permettant de prédire la note de Bob pour le film Titanic en utilisant :
– les notes de Bob concernant les d premiers films ;
– les notes des autres utilisateurs.
La première étape consiste à modéliser les préférences du nouvel utilisateur par un vecteur aléatoire $(\mathbf{X}, Y)$ de dimension $d+1$ à valeurs dans $[1, s]^d \times [1, s]$. Le vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)$ représente les notes du nouvel utilisateur concernant les d premiers objets tandis que Y représente sa note pour l'objet à prédire (note du film Titanic). Le nouvel utilisateur ne révélant pas forcément ses préférences pour tous les d premiers objets, nous n'observons pas $\mathbf{X}$ mais une version *masquée* que nous noterons $\mathbf{X}^\star = (X_1^\star, \ldots, X_d^\star)$:

$$X_j^\star = \begin{cases} X_j & \text{si } j \in M \\ 0 & \text{sinon} \end{cases} \tag{1}$$

où M désigne un sous-ensemble non vide de $\{1, \ldots, d\}$ correspondant aux objets évalués par le nouvel utilisateur. Dans l'exemple du tableau 1 on a $M = \{2, 3, 4, 5\}$ et $\mathbf{X}^\star = (0, 3, 3, 4, 5)$.

Nous modélisons ensuite les préférences des autres utilisateurs (Jim, James, Steve, Mary, etc. dans le tableau 1) en utilisant une approche dynamique. Pour ce faire, on suppose que les utilisateurs entrent dans la base de données les uns après les autres et mettent à jour leurs notes de manière séquentielle. Plus précisément, à chaque temps $i = 1, 2, \ldots$, un nouvel utilisateur entre dans le processus et révèle ses préférences pour la première fois tandis que les $i-1$ premiers utilisateurs peuvent mettre à jour leurs préférences. Ainsi, au temps 1, il y a un seul utilisateur dans la base de données (Jim dans le tableau

1) et on modélise le sous-ensemble (non vide) d'objets évalués par Jim par une variable aléatoire M_1^1 à valeurs dans $\mathcal{P}^\star(\{1,\ldots,d\})$, l'ensemble des parties non vides de $\{1,\ldots,d\}$. Au temps 2, un nouvel utilisateur (James) entre ses préférences pour certains objets modélisés par une variable aléatoire M_2^1 à valeurs dans $\mathcal{P}^\star(\{1,\ldots,d\})$ et de même loi que M_1^1. Au même temps, l'utilisateur 1 (Jim) peut mettre à jour ses préférences et on désigne par M_1^2 les objets évalués par Jim au temps 2. On supposera que $M_1^1 \subset M_1^2$, c'est-à-dire que les utilisateurs ne peuvent pas enlever les notes qu'ils ont mises au préalable. En répétant ce mécanisme, on dispose au temps n d'une matrice triangulaire supérieure $(M_i^j)_{1\leq i\leq n, 1\leq j\leq n+1-i}$ de variables aléatoires (voir tableau 2).

	Temps 1	Temps 2	...	Temps i	...	Temps n
Utilisateur 1	M_1^1	M_1^2	...	M_1^i	...	M_1^n
Utilisateur 2		M_2^1	...	M_2^{i-1}	...	M_2^{n-1}
$\vdots$			$\ddots$	$\vdots$	$\vdots$	$\vdots$
Utilisateur i				M_i^1	...	M_i^{n+1-i}
$\vdots$					$\ddots$	$\vdots$
Utilisateur n						M_n^1

TAB. 2 – Modélisation séquentielle des préférences.

Les notes de l'utilisateur i relatives aux d premiers objets sont représentées par une variable aléatoire $\mathbf{X}_i = (X_{i1},\ldots,X_{id})$. En se basant sur la modélisation des notes du nouvel utilisateur (1), on définit la version masquée $\mathbf{X}_i^{(n)} = (X_{i1}^{(n)},\ldots,X_{id}^{(n)})$ des notes de l'utilisateur i au temps n par :

$$X_{ij}^{(n)} = \begin{cases} X_{ij} & \text{si } j \in M_i^{n+1-i} \cap M \\ 0 & \text{sinon.} \end{cases}$$

Enfin, on désigne par $Y_1,\ldots,Y_n$ les variables aléatoires à valeurs dans $[1,s]$ représentant les évaluations des utilisateurs au temps n concernant la variable d'intérêt (le film Titanic dans notre exemple). Afin de prendre en compte les possibilités de non réponse concernant cette variable, on introduit une suite $(\mathcal{R}_n)_{n\geq 1}$ de variables aléatoires à valeurs dans $\mathcal{P}^\star(\{1,\ldots,n\})$. $\mathcal{R}_n$ représente le sous-ensemble (non vide) des utilisateurs qui ont évalué la variable d'intérêt (Titanic) au temps n.

Nous disposons ainsi au temps n d'un échantillon $(\mathbf{X}_1^{(n)},Y_1),\ldots,(\mathbf{X}_n^{(n)},Y_n)$ et notre mission consiste à évaluer la note Y du nouvel utilisateur représenté par $\mathbf{X}^\star$. Le problème statistique est donc d'estimer la fonction de régression $\eta(\mathbf{x}^\star) = \mathbb{E}[Y|\mathbf{X}^\star = \mathbf{x}^\star]$.

2.2 L'estimateur

Etant donné $\mathbf{x}^\star \in (\{0\} \cup [1,s])^d - \mathbf{0}$ ($\mathbf{0}$ représente le vecteur nul de $\mathbb{R}^d$) et l'échantillon $(\mathbf{X}_1^{(n)}, Y_1), \ldots, (\mathbf{X}_n^{(n)}, Y_n)$, nous proposons d'estimer la fonction de régression $\eta(\mathbf{x}^\star)$ par un estimateur de type k_n plus proches voisins utilisant une mesure de similarité basée sur le cosinus. Plus précisément, la similarité entre le nouvel utilisateur $\mathbf{x}^\star$ et le i-ème utilisateur au temps n $\mathbf{X}_i^{(n)}$ est mesurée par

$$S(\mathbf{x}^\star, \mathbf{X}_i^{(n)}) = p_i^{(n)} \frac{\sum_{j\in\mathcal{J}} x_j^\star X_{ij}^{(n)}}{\sqrt{\sum_{j\in\mathcal{J}} x_j^{\star 2}}\sqrt{\sum_{j\in\mathcal{J}} X_{ij}^{(n)^2}}}, \quad p_i^{(n)} = \frac{|M_i^{n+1-i} \cap M|}{|M|},$$

où $\mathcal{J} = \{j : x_j^\star \neq 0 \text{ et } X_{ij}^{(n)} \neq 0\}$ et $|A|$ désigne le cardinal d'un ensemble A. On remarquera que si $M \subset M_i^{n+1-i}$ alors $p_i^{(n)} = 1$ et $S(\mathbf{x}^\star, \mathbf{X}_i^{(n)}) = \cos(\mathbf{x}^\star, \mathbf{X}_i^{(n)})$. Le terme $p_i^{(n)}$ peut être vu comme une pénalité utilisée pour ne pas trop favoriser les derniers individus entrés dans la base de données. Ainsi, au temps n, on dira que l'individu i est "plus similaire" au nouvel individu que l'individu j si $S(\mathbf{x}^\star, \mathbf{X}_i^{(n)}) > S(\mathbf{x}^\star, \mathbf{X}_j^{(n)})$. Etant donné k_n un entier vérifiant $1 \leq k_n \leq n$, la fonction $\eta(\mathbf{x}^\star)$ est alors estimée par

$$\eta_n(\mathbf{x}^\star) = \|\mathbf{x}^\star\| \sum_{i\in\mathcal{R}_n} W_{ni}(\mathbf{x}^\star) \frac{Y_i}{\|\mathbf{X}_i^{(n)}\|}$$

où $\|.\|$ désigne la norme euclidienne et

$$W_{ni}(\mathbf{x}^\star) = \begin{cases} 1/k_n & \text{si } \mathbf{X}_i^{(n)} \text{ est parmi les } k_n\text{-MS de } \mathbf{x}^\star \text{ parmi } \{\mathbf{X}_i^{(n)}, i \in \mathcal{R}_n\} \\ 0 & \text{sinon.} \end{cases}$$

L'acronyme k_n-MS (k_n "most similar") signifie que l'on ne prend en compte que les k_n individus les plus similaires de $\mathbf{x}^\star$ parmi $\{\mathbf{X}_i^{(n)}, i \in \mathcal{R}_n\}$. Afin que l'estimateur soit bien défini, nous ajoutons les remarques suivantes :

- si $S(\mathbf{x}^\star, \mathbf{X}_i^{(n)}) = S(\mathbf{x}^\star, \mathbf{X}_j^{(n)})$, on dira que l'individu $\mathbf{X}_i^{(n)}$ est "plus similaire" à $\mathbf{x}^\star$ que $\mathbf{X}_j^{(n)}$ si $i < j$;
- si $|\mathcal{R}_n| < k_n$, on pose $\eta_n(\mathbf{x}^\star) = 0$;
- si $\mathbf{X}_i^{(n)} = \mathbf{0}$, on pose $W_{ni}(\mathbf{x}^\star) = 0$ avec la convention $0 \times \infty = 0$.

2.3 Propriétés asymptotiques

Sous certaines hypothèses concernant la forme de la fonction de régression η, nous obtenons les résultats suivants.

Théorème 2.1 *On suppose que $k_n \to \infty$, $|\mathcal{R}_n| \to \infty$ p.s. et $\mathbb{E}[k_n/|\mathcal{R}_n|] \to 0$ lorsque $n \to \infty$. Alors*

$$\mathbb{E}\,|\eta_n(\mathbf{X}^\star) - \eta(\mathbf{X}^\star)| \to 0 \quad \textit{quand } n \to \infty.$$

Théorème 2.2 *Soit* $\alpha_{ni} = \mathbb{P}(M^{n+1-i} \not\supset M \,|\, M)$. *On suppose que* $|M| \geq 4$. *Il existe alors une constante* $C > 0$ *telle que, pour tout* $n \geq 1$,

$$\begin{aligned}&\mathbb{E}\,|\eta_n(\mathbf{X}^\star) - \eta(\mathbf{X}^\star)| \\ &\quad \leq C\left\{k_n \mathbb{E}\left[\frac{1}{|\mathcal{R}_n|}\sum_{i\in\mathcal{R}_n}\mathbb{E}\,\alpha_{ni}\right] + \mathbb{E}\left[\left(\frac{k_n}{|\mathcal{R}_n|}\right)^{P_n}\right] + \frac{1}{k_n} + \mathbb{E}\left[\prod_{i\in\mathcal{R}_n}\alpha_{ni}\right]\right\},\end{aligned}$$

où $P_n = 2/(|M|-1)$ *si* $k_n < |\mathcal{R}_n|$, *et* $P_n = 1$ *sinon.*

On pourra notamment remarquer que dans le cas déterministe $M = \{1, \ldots, d\}$ et $\mathcal{R}_n = \{1, \ldots, n\}$ les hypothèses du Théorème 2.1 sont $k_n \to \infty$ et $k_n/n \to 0$ lorsque $n \to \infty$, c'est-à-dire les hypothèses nécessaires à la convergence de l'estimateur "classique" des k_n plus proches voisins. De même, toujours pour ce cas particulier, la borne du Théorème 2.2 devient

$$C\left\{\left(\frac{k_n}{n}\right)^{2/(d-1)} + \frac{1}{k_n}\right\}.$$

Cette borne coïncide avec la vitesse de convergence de l'estimateur des k_n plus proches voisins en dimension $d-1$ (voir [4]).

Références

[1] J. Abernethy, F.R. Bach, T. Evgeniou et J.-P. Vert : A new approach to collaborative filtering : Operator estimation with spectral regularization. *J. Mach. Learn. Res.*, 2009. in press.

[2] G. Adomavicius, R. Sankaranarayanan, S. Sen et A. Tuzhilin : Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Info. Syst.*, 2005.

[3] G. Adomavicius et A. Tuzhilin : Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, 17:734–749, 2005.

[4] L. Györfi, M. Kohler, A. Krzyżak et H. Walk : *A Distribution Free Theory of Nonparametric Regression*. Springer-Verlag, 2002.

[5] B. Sarwar, G. Karypis, J. Konstan et J. Riedl : Item-based collaborative filtering recommendation algorithms. *In Proceedings of the 10th International WWW Conference*, pages 285–295, 2001.