Données manquantes dans un modèle à blocs latents pour la recommandation

Gabriel Frisch, Jean-Benoist Leger, Yves Grandvalet

Données manquantes dans un modèle à blocs latents pour la recommandation

Gabriel Frisch, Jean-Benoist Leger & Yves Grandvalet

Sorbonne universités, Université de technologie de Compiègne, CNRS, Heudiasyc UMR 7253, 60203 Compiègne Cedex, France. E-mails : prénom.nom@hds.utc.fr

Résumé. Nous présentons un modèle statistique basé sur le *LBM* pour réaliser une recommandation sociale. Le modèle utilise des variables latentes pour modéliser un processus de manquement de données de type *NMAR*.

Mots-clés. Recommandation sociale, *LBM*, *NMAR*

Abstract. We present a statistical model based on the *LBM* for collaborative filtering. The model introduces some latent variables to model the *NMAR* missing data procesus.

Keywords. Collaborative filtering, *LBM*, *NMAR*

1 Introduction

Les systèmes de recommandation sont des outils d'aide à la décision réalisant une extraction de l'information afin de présenter des résultats pertinents aux utilisateurs. Il existe différents types de systèmes de recommandation. Certaines approches utilisent les caractéristiques des items ou des utilisateurs afin de les lier entre eux [7, 8]. D'autres, utilisant la recommandation sociale ou *Collaborative Filtering*, se servent de l'historique des avis pour prédire l'intérêt des utilisateurs. Les données qui y sont associées peuvent être explicites, typiquement une note indiquant l'appréciation de l'utilisateur, ou implicites ; c'est à dire basées sur le comportement des utilisateurs (achats, clics, durée sur une page). Ces systèmes de recommandation sociale présupposent l'existence de préférences qui soient communes à certains utilisateurs et à certains items.

Des modèles statistiques peuvent être développés pour découvrir ces classes de comportement commun. Un réseau de recommandation peut être modélisé par un graphe bipartite, ayant pour noeuds les utilisateurs d'une part, et les items d'autre part ; les arêtes (possiblement valuées) représentent les avis. La matrice d'adjacence de ce graphe contient ainsi l'ensemble de l'information utilisée.

Une grande variété d'algorithmes ont été proposés pour retrouver des structures dans des matrices de données. La classification double ou *co-clustering* [3] considère simultanément les lignes et les colonnes afin d'organiser les données en blocs homogènes. Le modèle à bloc latents, ou *Latent Block Model* (*LBM*) proposé par Govaert et Nadif [2] exploite ce concept en y introduisant un modèle de mélange. Les paramètres du *LBM* sont ajustés par

un algorithme itératif basé sur l'algorithme EM de Dempster *et al.* [1]. Dans nos travaux, nous étendons le modèle de base du *LBM* pour l'utiliser à des fins de recommandations.

Classiquement, les modèles de réseaux de recommandation supposent que les données manquantes sont indépendantes de leurs valeurs ; c'est l'hypothèse *Missing (completely) At Random* (*MAR*) [5]. Sous cette hypothèse, aucun processus de manquement de données n'est à considérer dans le modèle statistique et l'inférence est réalisée en utilisant uniquement la vraisemblance des données observées. Cependant, les travaux de Marlin et Zemel [6] suggèrent que le processus de manquement des jeux de données des les réseaux de recommandations sont principalement de type *Not Missing At Random* (*NMAR*). C'est-à-dire qu'il existe un lien sous-jacent dans les données, qui influe sur la présence ou non, d'avis. Ainsi, si un biais est introduit dans l'estimation du modèle, les performances du réseau de recommandation pourraient être significativement dégradées [6]. Dans nos travaux, nous faisons l'hypothèse que les données manquantes sont de type *NMAR* et nous modélisons le mécanisme de manquement. Nous supposons par exemple que la propension à donner son avis n'est pas la même pour un objet que l'on apprécie que pour un autre que l'on n'apprécie pas.

L'objectif de cet article est de proposer un modèle statistique du phénomène *NMAR* dans un *LBM*, ainsi que les grandes étapes de l'inférence. Dans un second temps, des résultats sur des données simulées sont présentés.

2 Modèle

Le modèle présenté est un modèle de graphe aléatoire bipartite, basé sur le *Latent Block Model*. La modélisation peut être faite en deux étapes. Dans la première, le graphe non totalement observé est modélisé comme un *Latent Block Model* classique. Dans la deuxième étape, le mécanisme de manquement modélise le graphe dans lequel seulement une partie des arêtes est observée.

Le graphe étant bipartite, nous aurons deux types de nœuds. Typiquement dans le cas d'un réseau de recommandation, les nœuds de type (1) représentent les personnes et les nœuds de type (2) représentent les objets. Nous considérons un graphe composé de n_1 nœuds de type (1), et n_2 nœuds de type (2). Le modèle sera présenté pour k_1 classes sur les nœuds de type (1) et k_2 classes de type (2). $\boldsymbol{X}^{(\mathrm{c})}$ représente la matrice d'adjacence du graphe valué complet. Les données associées à notre réseau de recommandation sont des avis binaires et explicites de l'utilisateur : nous notons $X_{ij}^{(\mathrm{c})} = +1$ un avis positif et $X_{ij}^{(\mathrm{c})} = -1$ un avis négatif d'un utilisateur i pour un objet j. Nous soulignons le fait que notre modèle peut être facilement étendu pour utiliser un système de notation à valeurs entières.

Nous notons $\mathcal{B}_{a,b}(p)$ la loi de transformation de la variable de Bernoulli suivante : une variable aléatoire qui prend b avec probabilité p et a sinon. Conceptuellement, $\mathcal{B}_{a,b}(p) = a + (b-a)\mathcal{B}(p)$, où $\mathcal{B}$ est la loi de Bernouilli classique.

Latent Block Model L'appartenance de chaque nœud à une classe est ici modélisée par une variable latente. Toutes ces variables latentes sont considérées comme indépendantes.

$$\begin{cases} \forall i,\ Y_i^{(1)} \overset{\text{iid}}{\sim} \mathcal{M}\big(1; \boldsymbol{\alpha}^{(1)}\big) & ,\ \boldsymbol{\alpha}^{(1)} \in \mathbb{R}^{k_1}, \sum_{q=1}^{k_1} \alpha_q^{(1)} = 1 \\ \forall j,\ Y_j^{(2)} \overset{\text{iid}}{\sim} \mathcal{M}\big(1; \boldsymbol{\alpha}^{(2)}\big) & ,\ \boldsymbol{\alpha}^{(2)} \in \mathbb{R}^{k_2}, \sum_{l=1}^{k_2} \alpha_l^{(2)} = 1 \end{cases}$$

Puis, l'arête entre le nœud i de type (1) et le nœud j de type (2), notée $X_{ij}^{(\text{c})}$ est considérée comme dépendant uniquement des classes des nœuds i et j.

$$\forall i, j,\ \left(X_{ij}^{(\text{c})} \middle| Y_i^{(1)} = q, Y_j^{(2)} = l \right) \overset{\text{ind}}{\sim} \mathcal{B}_{-1,+1}(\pi_{ql}),\ \pi_{ql} \in [0,1]$$

Loi de l'observation Nous considérons que $\boldsymbol{X}^{(\text{c})}$ n'est que partiellement observé. Nous notons $\boldsymbol{X}^{(\text{o})}$ la matrice d'adjacence du graphe observé, c'est-à-dire soumis au mécanisme de manquement. La valeur observée de l'arête entre le nœud i de type (1) et le nœud j de type (2) noté $X_{ij}^{(\text{o})}$ dépend de la valeur de $X_{ij}^{(\text{c})}$ et des caractéristiques latentes des nœuds i et j. Le fait que $X_{ij}^{(\text{o})}$ dépende de $X_{ij}^{(\text{c})}$ rend le modèle *Not Missing At Random*, c'est-à-dire que la probabilité d'observer une variable dépend de la valeur de la variable. Une donnée manquante entre l'utilisateur i et l'objet j sera notée $X_{ij}^{(\text{o})} = 0$

Le mécanisme de manquement d'un lien est dépendant de la nature du réseau considéré. Nous considérons que le comportement des personnes peut être modélisé par deux variables latentes continues. Pour chaque personne i, nous considérons son affinité A_i à évaluer des objets, et B_i la sur-affinité à évaluer ces objets qu'elle apprécie. Ainsi, pour la personne i, son affinité à évaluer un objet apprécié sera $A_i + B_i$ et son affinité à évaluer un objet non apprécié sera $A_i - B_i$.

$$\begin{cases} \forall i,\ A_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_A^2) & ,\ \sigma_A^2 \in \mathbb{R}_+^* \\ \forall i,\ B_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_B^2) & ,\ \sigma_B^2 \in \mathbb{R}_+^* \end{cases}$$

De façon similaire, nous introduisons une variable latente continue P_j modélisant la popularité d'un objet j.

$$\forall j,\ P_j \overset{\text{iid}}{\sim} \mathcal{N}\big(0, \sigma_P^2\big),\ \sigma_P^2 \in \mathbb{R}_+^*$$

On considère que les affinités à noter du nœud i de type (1) et la popularité du nœud j de type (2) influent linéairement sur la log-cote de la probabilité d'observation, telle que classiquement utilisée dans une régression logistique. On a donc :

$$\forall i, j,\ \left(X_{ij}^{(\text{o})} \middle| X_{ij}^{(\text{c})}, A_i, B_i, P_j \right) \overset{\text{ind}}{\sim} \begin{cases} \mathcal{B}_{0,+1}\big(\text{logit}^{-1}\left(\mu_{+1} + A_i + B_i + P_j\right)\big) & \text{si } X_{ij}^{(\text{c})} = +1 \\ \mathcal{B}_{0,-1}\big(\text{logit}^{-1}\left(\mu_{-1} + A_i - B_i + P_j\right)\big) & \text{si } X_{ij}^{(\text{c})} = -1 \end{cases}$$

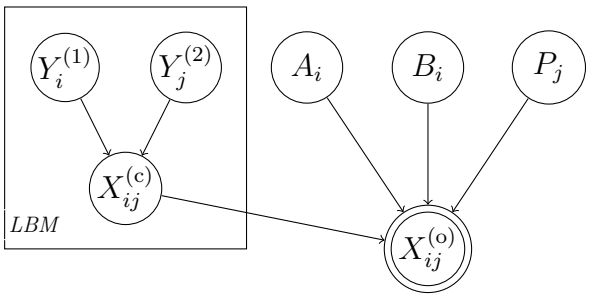

FIGURE 1 – Modèle graphique des variables aléatoires pour un individu i et un item j donnés. Le double cercle représente la variable observée.

avec logit^{-1} la fonction sigmoïde, μ_{+1} le paramètre de moyenne d'avis positifs et μ_{-1} le paramètre de moyenne d'avis négatifs de l'ensemble des individus.

Le modèle graphique représentant les variables aléatoires latentes et observées est présenté figure 1.

3 Inférence

Durant l'inférence, le modèle équivalent suivant liant $\boldsymbol{X}^{(o)}$ à $\boldsymbol{Y}^{(1)}$ et $\boldsymbol{Y}^{(2)}$ sera utilisé :

— La modélisation des variables latentes (excepté $\boldsymbol{X}^{(c)}$ est la même) :

$$\begin{cases} \forall i,\ Y_i^{(1)} \overset{\text{iid}}{\sim} \mathcal{M}\left(1; \boldsymbol{\alpha}^{(1)}\right) \\ \forall j,\ Y_j^{(2)} \overset{\text{iid}}{\sim} \mathcal{M}\left(1; \boldsymbol{\alpha}^{(2)}\right) \\ \forall i,\ A_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_A^2) \\ \forall i,\ B_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_B^2) \\ \forall j,\ P_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_P^2) \end{cases}$$

— Le graphe observé dépend maintenant directement de toutes les variables latentes.

$$\left(X_{ij}^{(o)} \middle| Y_i^{(1)} = q, Y_j^{(2)} = l, A_i, B_i, P_j\right) \overset{\text{ind}}{\sim} \text{Cat}\left(\begin{bmatrix} +1 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} \pi_{ql} \,\text{logit}^{-1}\left(\mu_{+1} + A_i + B_i + P_j\right) \\ (1 - \pi_{ql}) \,\text{logit}^{-1}\left(\mu_{-1} + A_i - B_i + P_j\right) \\ (\cdots) \end{bmatrix}\right)$$

avec $(\cdots)$ le terme calculé tel que le vecteur somme à 1, et où $\text{Cat}\left(\mathbb{X}, \boldsymbol{p}\right)$ représente la distribution catégorielle à valeurs dans $\mathbb{X}$ avec les probabilités associées $\boldsymbol{p} \in \mathbb{R}_+^{\text{card}\,\mathbb{X}}$ et $\sum_k p_k = 1$. Les paramètres du modèle sont donc $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \sigma_A^2, \sigma_B^2, \sigma_P^2, \mu_{-1}, \mu_{+1}, \boldsymbol{\pi}$.

Afin d'inférer les paramètres et de prédire les variables latentes, nous maximisons la vraisemblance. La vraisemblance incomplète, c'est-à-dire celle de $\boldsymbol{X}^{(o)}$ n'est pas calculable, pour maximiser celle-ci, l'algorithme EM est utilisé. Toutefois, le calcul de l'espérance de la log-vraisemblance complète conditionnellement à $\boldsymbol{X}^{(o)}$ à paramètres fixés est impossible. Une formulation variationnelle de l'EM est utilisée [4], et une restriction sur l'espace de recherche est faite, ce qui constitue une approximation du critère variationnel et des

éléments qui le composent. Pour le calcul de certaines espérances sous la distribution variationnelle, une approximation au moyen d'un développement de Taylor à l'ordre trois est utilisée. L'erreur est contrôlée en utilisant la variance variationnelle.

Une implémentation de cette inférence a été réalisée en python, permettant d'analyser des réseaux de quelques centaines de nœuds. Cette implémentation utilise la differentiation automatique afin de calculer de manière exacte les gradients, jacobiens, et hessiens impliqués dans le développement de Taylor, et de calculer exactement les gradients des critères. Un algorithme de quasi-Newton est utilisé pour la maximisation. Le nombre de paramètres variationnel étant grand, le L-BFGS est utilisé afin de ne pas calculer ni stocker explicitement le hessien approché, et de le manipuler seulement de manière implicite.

4 Expérimentation sur données simulées

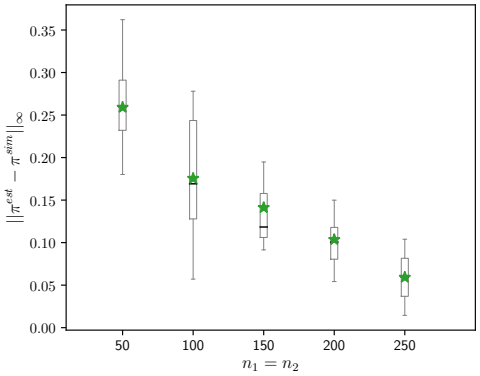

FIGURE 2 – Erreur en norme ∞ sur $\boldsymbol{\pi}$ en fonction de la taille n du graphe ($n_1 = n_2$). $\star$ représente la moyenne.

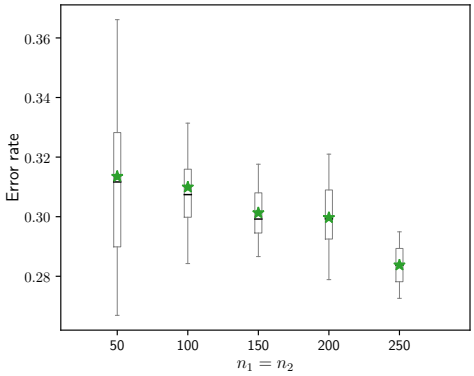

FIGURE 3 – Erreur de prédiction des avis manquants en fonction de la taille du graphe ($n_1 = n_2$) en utilisant un prédicteur construit à partir du MAP sous hypothèse variationnelle. $\star$ représente la moyenne.

L'objectif principal de l'étude expérimentale est de vérifier si l'algorithme proposé peut inférer les paramètres du modèle et prédire de façon cohérente les données manquantes. Les données d'expérimentation sont générées suivant notre modèle. Les résultats sont présentés à nombre de classes fixées ($k_1 = k_2 = 3$), et en faisant varier le nombre d'individus et d'items ($n_1 = n_2$). Les paramètres μ_{+1}, μ_{-1}, σ_A^2, σ_B^2 et σ_P^2 sont fixés à 1 donnant ainsi un taux de manquement moyen de 0.34.

Notre algorithme rencontrant encore des difficultés d'initialisation, similaires à celles rencontrées dans le *LBM* standard, l'algorithme est initialisé avec des valeurs proches des

variables latentes simulées afin de minimiser le risque de se retrouver dans un minimum local.

Les résultats sur 10 répétitions, présentés figure 2, montrent qu'avec l'augmentation de la taille du graphe, la distance infinie entre le paramètre $\boldsymbol{\pi}$ simulé et $\boldsymbol{\pi}$ inféré tend à diminuer. Il en est de même avec l'erreur de prédiction des données manquantes, présentée figure 3, calculée en utilisant un prédicteur construit à partir des maximums a posteriori sous hypothèse variationnelle. Ces premiers résultats permettent de penser que l'algorithme peut donc inférer de façon cohérente les paramètres du modèle.

Références

[1] A. P. Dempster, N. M. Laird et D. B. Rubin : Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1):1–22, 1977.

[2] G. Govaert et M. Nadif : Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473, 2003.

[3] G. Govaert et M. Nadif : *Co-clustering : models, algorithms and applications*. ISTE Ltd, 2013.

[4] T. S. Jaakkola : Tutorial on variational approximation methods. *In Advanced mean field methods : theory and practice*, p. 129–159. MIT Press, 2000.

[5] R. J. Little et Rubin : *Statistical analysis with missing data*. John Wiley & Sons, 1987.

[6] B. M. Marlin et R. S. Zemel : Collaborative prediction and ranking with non-random missing data. *In Proceedings of the third ACM conference on Recommender systems*, p. 5–12. ACM, 2009.

[7] B. Sarwar, G. Karypis, J. Konstan et J. Riedl : Item-based collaborative filtering recommendation algorithms. *In Proceedings of the 10th International Conference on World Wide Web, WWW*, p. 285–295. ACM, 2001.

[8] G. Takács, I. Pilászy, B. Németh et D. Tikk : Matrix factorization and neighbor based algorithms for the netflix prize problem. *In RecSys'08 : Proceedings of the 2008 ACM Conference on Recommender Systems*, p. 267–274, 2008.