Factorisation de matrices pour les systèmes de recommandation

Bibliographie Scientifique

Projet Transversal

Orange – Polytech'Nantes

Fabien Richard Valentin Proust

Introduction

Les systèmes de recommandation s'appuient sur des techniques de recueil de l'information variés et ont des fonctions et des objectifs divers. Mais la plupart sont basées sur le même outils mathématique : la factorisation de matrices.

Pour tenter de comprendre l'intérêt de ces factorisations et leur principe, nous commencerons dans une première partie par rappeler les deux principales stratégies des systèmes de recommandation. Puis nous distinguerons deux grands domaines dans la stratégie des systèmes de recommandation collaboratifs. Enfin nous détaillerons une méthode utilisée pour la factorisation : la décomposition en valeurs singulières.

Sommaire

I. Les deux stratégies des systèmes de recommandation

Le filtrage de contenu consiste à dresser un profil des produits en leur affectant différents attributs (genre, acteurs, popularité, etc) et des utilisateurs (âge, sexe, origine démographique, etc). A partir de ces profils, les algorithmes tentent de faire correspondre utilisateurs et produits.

La deuxième stratégie est le filtrage collaboratif. Il est basé sur l'historique comportemental des utilisateurs (ce qu'ils ont acheté ou consulté) et sur leur avis sur les produits. L'avis sur un produit peut être un simple "pouce vers le haut" ou "pouce vers le bas", une note ou un commentaire entier. La stratégie du filtrage collaboratif est confrontée à un important problème : le "démarrage à froid". En effet, sans information sur un nouvel utilisateur, il est difficile de lui proposer des produits correspondants à ses goûts.

II. La Méthode des plus proches voisins ou des facteurs latents

Dans le filtrage collaboratif on peut distinguer deux grands domaines.

Le premier est celui des plus proches voisins. La méthode des voisins peut être centrée sur les produits ou sur les utilisateurs. On cherche à rapprocher les produits qui obtiennent des notes similaires par des utilisateurs qui ont le même profile.

Le deuxième domaine est celui des facteurs latents. On va chercher à partir d'un grand nombre de facteurs (entre 20 et 100 par exemple) à opposer et rassembler des utilisateurs et des produits. Si les produits sont des films on va par exemple pouvoir trouver des oppositions entre les films d'action et les comédies, les films lents ou rapides, etc.

III. La factorisation de matrices

Pour pouvoir exploiter la méthode des facteurs latents, la factorisation de matrices est très efficace. En effet, elle permet d'utiliser plusieurs types de données en entrée : les données explicites comme une note ou un commentaire sur un produit et les données implicites comme l'historique de navigation, le mouvement de la souris etc. Ce système s'apparente au SVD (Singular Value Decomposition ou Décomposition en valeurs singulière) qui est une technique très utilisée pour réduire l'information et récupérer les facteurs latents. Cette technique est utile ici à cause du caractère creux des matrices due au très grand nombre de facteurs.

IV. La décomposition en valeurs singulières (SVD)

La décomposition en valeurs singulières s'appuie sur un théorème d'algèbre linéaire qui dit qu'une matrice rectangulaire A peut être décomposée en le produit de trois matrices (une matrice orthogonale U, une matrice diagonale S et la transposée d'une matrice orthogonale V. On peut présenter ce théorème comme :

$$A_{mn} = U_{mm} \times S_{mn} \times V^{T}_{nn}$$

où $U^TU = V^TV = I$; les colonnes de U sont les vecteurs propres de AA^T; les colonnes de V sont les vecteurs propres orthogonaux de A^TA; et S est une matrice diagonale qui contient les racines carrées des valeurs propres de U ou V dans l'ordre décroissant.

1. Calculer la transposée de A
2. Multiplier A par sa transposée
3. Trouver les valeurs propres de AA^T en résolvant $\det(AA^T - \lambda Id) = 0$
4. Pour chaque valeur propre, trouver le vecteur propre correspondant
5. Ces vecteurs propres deviennent les vecteurs en colonne d'une matrice. Ils sont classés dans l'ordre décroissant de leur valeur propres correspondantes
6. Pour obtenir la matrice orthogonale U, il suffit enfin d'appliquer le processus d'orthonormalisation de Gram-Schmidt
7. Le calcul de V^T est identique en se basant sur A^TA au lieu de AA^T comme pour et en prenant la transposée
8. Pour trouver S, on prend simplement la racine carrée des valeurs propres non nulles et on les place dans l'ordre décroissant en diagonale : s_{11} est la plus grande valeur, puis s_{22}, etc.

Finalement on a réussi à décomposer A en un produit de trois matrices. Les données dans la diagonale de A sont les valeurs singulières de A, les colonnes de U sont les vecteurs singuliers gauches de A et les colonnes de V les vecteurs singuliers droits de A.

Conclusion

Nous avons pu voir que l’intérêt pour les grandes matrices creuses, telles qu’utilisent les systèmes de recommandation, est de réduire leur dimension pour pouvoir extraire des tendances et fournir les meilleures recommandations possibles.

Références

BAKER Kirk, Singular Value Decomposition Tutorial, 2005 révisé en 2013

KOREN Yehunda, BELL Robert, VOLINSKY Chris, Matrix factorization techniques for recommender systems, IEE Computer society 2009