Recommandation de films par modèles hybrides basés sur des embeddings de langage

Alexandre CHEN, Dineshan JOSEPH, Sidra HYDER, Egxon ZEJNULLAHI
Master 2 Machine Learning for Data Science (MLDS)

Résumé

Ce travail étudie l'impact d'embeddings textuels issus de modèles de langue modernes sur la recommandation de films. Nous utilisons le jeu MovieLens sous deux formes : une version « Small » avec des descriptions courtes (titre, genres, quelques tags) et une version « Ultra » où, pour chaque film, tous les tags des utilisateurs sont concaténés en un texte plus long. Comme ligne de base, nous entraînons une factorisation matricielle probabiliste (PMF) sur les seules interactions utilisateur–film, puis construisons des variantes hybrides intégrant des embeddings MiniLM/BERT ou Qwen gelés. Sur la version Small, ces hybrides restent proches de la PMF, tandis que sur la version Ultra, le modèle hybride Qwen améliore NDCG@10, Precision@10 et Recall@10, ce qui montre l'intérêt de descriptions textuelles plus riches.

Introduction

Les systèmes de recommandation de films reposent sur le filtrage collaboratif, qui exploite les notes passées des utilisateurs pour prédire leurs futures préférences.

Dans ce projet, l'objectif est d'enrichir une baseline de factorisation matricielle probabiliste (PMF) avec des embeddings issus de modèles de langue de grande taille, de type BERT et Qwen, afin de construire des modèles hybrides qui combinent informations issues des notes et informations textuelles. Nous évaluons ces approches sur deux configurations du jeu MovieLens, différant par la quantité d'information textuelle disponible pour chaque film.

Jeu de données et prétraitements

Nous travaillons sur le jeu de données *MovieLens*, qui contient des notes de films associées à des utilisateurs. Nous utilisons les mêmes interactions utilisateur–film mais distinguons deux configurations textuelles : une variante *Small*, où chaque film est décrit par un nombre limité de tags *(ml latest-small)*, et une variante *Ultra*, où l'on agrège pour chaque film l'ensemble des tags fournis par les utilisateurs, ce qui donne plus d'informations textuelles via les tags *(ml latest)*.

La première version, appelée *Small*, associe à chaque film un court texte constitué uniquement du titre et des genres, complété par quelques tags sélectionnés. La seconde version, notée *Ultra*, reprend les mêmes films mais concatène cette fois l'ensemble des tags disponibles, ce qui produit des descriptions beaucoup plus longues. La base Small offre donc un signal textuel limité mais relativement propre, tandis que la base Ultra fournit davantage de détails au prix d'un bruit plus important.

Pour l'évaluation, les interactions utilisateur–film sont séparées utilisateur par utilisateur : 80 % des notes servent à l'apprentissage et les 20 % restantes au test. Chaque utilisateur apparaît ainsi dans les deux ensembles et l'on prédit uniquement des éléments non vus lors de l'entraînement. Dans l'ensemble de test, les films notés au moins 4 sont considérés comme pertinents pour le calcul des métriques top-K.

Méthodologie : Baseline

Nous avons d'abord testé une baseline de type SVD en factorisant directement la matrice utilisateur–film. Cette approche s'est révélée peu adaptée à la recommandation top-K : la précision et le rappel restaient faibles et la forte sparsité du jeu de données conduisait à des listes de recommandations peu cohérentes avec les profils utilisateurs.

Nous avons donc adopté comme modèle de base une factorisation matricielle probabiliste (PMF) purement collaborative. Chaque utilisateur u et chaque film i est associé à un vecteur latent de dimension 64, notés p_u et q_i, implémentés via deux couches d'embedding distinctes initialisées aléatoirement selon une loi normale centrée de faible variance. En complément, le modèle intègre un biais par utilisateur b_u, un biais par film b_i et un biais global μ. Pour une paire (u, i), la note prédite s'écrit

$$\hat{r}_{ui} = \mu + b_u + b_i + \langle p_u, q_i \rangle,$$

où $\langle p_u, q_i \rangle$ désigne le produit scalaire entre les embeddings utilisateur et item.

L'ensemble des paramètres (facteurs latents et biais) est appris en minimisant l'erreur quadratique moyenne sur les notes observées, avec une pénalisation ℓ_2 pour limiter le surapprentissage. L'optimisation est réalisée avec l'algorithme Adam sur des mini-lots aléatoires. En pratique, cette PMF offre des performances top-K nettement plus stables et plus élevées que la SVD et sert de baseline de référence pour la comparaison avec les modèles hybrides.

Méthodologie : Modèles hybrides

Pour exploiter le contenu textuel, nous étendons notre modèle de factorisation matricielle probabiliste (*PMF*) en y ajoutant des informations issues d'encodeurs de langage. Côté utilisateur, nous apprenons un vecteur latent u_v de dimension fixe. Côté film, la représentation est la concaténation d'un facteur latent i_v issu de la PMF et d'une projection linéaire d'un embedding textuel c_i. L'encodeur textuel est soit un modèle de type MiniLM/BERT, soit un modèle Qwen plus volumineux capable de traiter des séquences longues. Les poids du modèle de langue restent gelés et seule la couche de projection est entraînée conjointement avec les facteurs collaboratifs. Afin d'essayer d'améliorer les modèles entraînés sur la configuration *Small*, nous réentraînons la même architecture sur le même sous-ensemble de films et d'utilisateurs, mais en remplaçant les descriptions courtes par l'ensemble des tags disponibles*(ml latest)*, de façon à disposer de davantage de contenu textuel pour l'analyse sémantique*(c'est ce qu'on appelle la version Ultra)*.

Dans ce projet, nous considérons quatre configurations hybrides : *hybride BERT (Small)*, *hybride BERT (Ultra)*, *hybride Qwen (Small)* et *hybride Qwen (Ultra)*. Dans les variantes *Small*, l'encodeur (BERT ou Qwen) reçoit uniquement le titre, les genres et quelques tags par film, ce qui fournit un texte relativement propre mais peu riche sur le plan sémantique. Dans les variantes *Ultra*, le même encodeur prend en entrée des descriptions plus longues obtenues en concaténant l'ensemble des tags disponibles pour chaque film, avec un signal textuel plus riche mais aussi plus bruité.

Métriques d'évaluation

Pour un utilisateur u, on note Rec@10(u) l'ensemble des 10 films recommandés et Rel(u) l'ensemble des films pertinents dans le test. Le *Recall@10* est défini par

$$\text{Recall@10}(u) = \frac{\big|\text{Rec@10}(u) \cap \text{Rel}(u)\big|}{\big|\text{Rel}(u)\big|}. \qquad (1)$$

il mesure la proportion de films réellement intéressants qui apparaissent dans les dix premières recommandations.

La *Precision@10* vaut

$$\text{Precision@10}(u) = \frac{\big|\text{Rec@10}(u) \cap \text{Rel}(u)\big|}{10}, \qquad (2)$$

et reflète la fraction d'items pertinents parmi les recommandations de tête.

Enfin, la *NDCG@10* (*Normalized Discounted Cumulative Gain*) tient compte de la position des films pertinents dans la liste. En notant $\text{rel}_{u,k} \in \{0, 1\}$ la pertinence du film à la position k, on définit

$$\text{DCG@10}(u) = \sum_{k=1}^{10} \frac{2^{\text{rel}_{u,k}} - 1}{\log_2(k+1)}, \qquad (3)$$

et

$$\text{NDCG@10}(u) = \frac{\text{DCG@10}(u)}{\text{IDCG@10}(u)}, \qquad (4)$$

où IDCG@10(u) correspond à la DCG pour un classement idéalement trié.

Résultats

Le tableau présente les scores de NDCG@10, Precision@10 et Recall@10 pour la PMF de base et les variantes hybrides BERT et Qwen, chacune évaluée sur les versions *Small* et *Ultra* du jeu MovieLens. On observe que, quelle que soit la variante textuelle ou l'architecture hybride considérée, les performances ne dépassent pas celles du modèle de base.

Modèle	NDCG@10	Prec@10	Rec@10
PMF (Baseline)	0.206	0.169	0.096
Hybrid BERT (Small)	0.180	0.147	0.083
Hybrid Qwen (Small)	0.178	0.147	0.081
Hybrid BERT (Ultra)	0.172	0.142	0.083
Hybrid Qwen (Ultra)	0.199	0.162	0.093

On remarque toutefois que, même pour la configuration la plus performante*(hybride Qwen Ultra)*, les scores restent relativement bons, avec un NDCG@10 qui atteint presque 0,20 et un Recall@10 proche de 0,10. Ces valeurs reflètent à la fois la taille réduite du sous-ensemble MovieLens utilisé et le fait que les modèles explorés restent volontairement simples, ce qui laisse une marge importante pour des architectures plus avancées et un réglage plus fin des hyperparamètres.

Discussion et perspectives

Les expériences montrent que la factorisation matricielle probabiliste constitue une base solide : elle exploite déjà efficacement les interactions utilisateur–film. Les variantes hybrides n'apportent un gain clair que lorsque le contenu textuel est vraiment riche. Sur la version Small, où chaque film ne possède que quelques mots-clés, les modèles BERT et Qwen restent inférieur par rapport à PMF. En revanche, sur la version Ultra, Hybrid Qwen profite mieux de la grande quantité de tags et se met presque au même niveau que la baseline, tandis que la variante BERT semble davantage pénalisée par le bruit.

Pour aller plus loin, il serait intéressant de spécialiser les embeddings textuels via un fine-tuning contrastif et de tester d'autres modèles pour la partie interactions (par exemple des architectures séquentielles). À plus long terme, il faudra aussi évaluer ces approches sur des jeux plus volumineux, avec davantage d'utilisateurs et de films, afin de vérifier si une densité d'interactions plus élevée renforce encore l'intérêt de la composante textuelle dans les modèles hybrides.