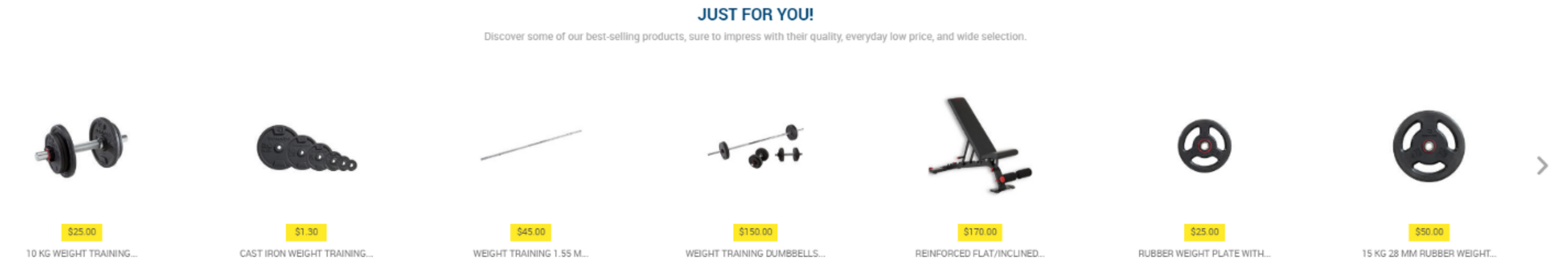
Systèmes de recommandation chez Décathlon

Apprentissage automatique I
MATH60629

Cas développé par Jérémi DeBlois-Beaucage sous la supervision de Laurent Charlin & Renaud Legoux

Présentation du cas

Q1: Quel(s) modèle(s) de système de recommandation l'équipe de science des données devrait-elle choisir, et pourquoi?



Séance en petits groupes

- <u>Question 1:</u> *Pour le prototype, il n'y a pas de contraintes sur le temps de calcul pour entraîner le modèle ou pour servir les recommandations.*

  En fonction des données disponibles et des objectifs de Décathlon, quel(s) modèle(s) de système de recommandation l'équipe de science des données devrait-elle choisir, et pourquoi?

- **15 minutes, groupe de ~2–5**

  - **Suggestion: un.e secrétaire qui présentera par équipe**

- **Ensuite: discussion en classe**

Discussions

Item-based vs. User-based

- |Users| >> |Items|
- Item-based
  - (+) Stabilité
  - (+) Meilleur performance (pas toujours)
  - (+) Explication
- User-based
  - (+) Diversité

Plan

- Modèles choisis par Décathlon
  - Modèles de base, pour référence
    - Modèle 1 : Basé sur la similarité entre images de produits
    - Modèle 2 : Filtrage collaboratif basé sur les produits
    - Modèle 3 : Factorisation matricielle
    - Modèle 4 : Réseaux de neurones récurrents
  - Présentation des métriques choisies
  - Résultats et choix final
- Limites des modèles actuels et prochaines étapes envisagées par Décathlon

Modèles de base : point de référence

- Recommandation au hasard
- Recommandation des mêmes 10 items les plus populaires à tous les usagers

Modèle 1 : Basé sur la similarité entre images de produits

Au préalable, chaque produit est lié à une image et à un vecteur qui représente l'image, créé grâce à un réseau de neurones à convolution pré-entraîné de type VGG.

1. Pour chaque utilisateur, extraire une liste de tous les produits avec lesquels il a interagi.


2. Pour chaque produit, déterminer les 10 produits les plus « similaires », basé sur la distance cosinus entre leur vecteur d'image.






3. Le produit le plus similaire reçoit 10 « points », le second 9, et ainsi de suite. Les points sont additionnés, et les 5 produits avec le plus grand score sont recommandés.

User A has interacted in the past with items **3** and **5**

For each item, we identify the 10 items in the rest of the catalog that are most similar. We give 10 pts to the most similar one, 9 pts to the second most, and so on

7	11	10 pts
12	7	9 pts
37	22	8 pts
11	30	7 pts
22	1	6 pts
6	26	5 pts
1	27	4 pts
28	12	3 pts
29	15	2 pts
30	9	1 pts

We sum all the scores, and recommend the top five items to our user

Recommendations to user **A**:
Item **7** (19 pts)
Item **11** (17 pts)
Item **22** (14 pts)
Item **12** (12 pts)
Item **1** (10 pts)

Avantages et inconvénients

- (+) Même de nouveaux produits ou des produits peu populaires peuvent être recommandés
- (+) Des explications peuvent être fournies (parce que vous avez acheté le produit X, ces produits pourraient vous intéresser)
- (+) Facile et rapide à mettre en place
- (-) Les produits recommandés seront très similaires aux produits déjà achetés
- (-) Cold-start problem pour les utilisateurs
- (-) la performance dépend de la qualité des représentations du CNN

Modèle 2 : Filtrage collaboratif basé sur les produits (*item-based CF*)

Modèle très semblable au précédent. À la place de calculer la similarité entre les vecteurs d'image, on calcule la similarité selon la matrice d'interaction entre les utilisateurs et les produits.

1. Pour chaque utilisateur, extraire une liste de tous les produits avec lesquels il a interagi.

2. Pour chaque produit, déterminer les 10 produits les plus « similaires », basé sur la distance cosinus entre leur ligne respective dans la matrice d'interaction utilisateurs-produits.

3. Le calcul du score final est similaire à celui du modèle 1. Cependant, à la place d'attribuer des points arbitrairement (10 pts pour le 1er, 9 pour le 2e, etc.), les scores de similarité sont directement utilisés. Les points sont additionnés, et les 5 produits avec le plus grand score sont recommandés.

Exemple : similarité entre deux produits

Produits

```
[0, 0, 1, 0, 1, 0]
[0, 1, 0, 0, 0, 0]
[0, 0, 1, 0, 1, 0]
[1, 0, 0, 0, 1, 0]
```

Usagers

- Dans la matrice ci-contre, les lignes représentent des utilisateurs et les colonnes, des produits. Une valeur de 1 indique un intérêt, et 0 indique une absence d'interaction.
  - P.ex., la première colonne indique que seulement l'utilisateur 4 a interagi avec le produit 1.
- La similarité cosinus entre le premier (a =[0 0 0 1]) et le second produit (b=[0 1 0 0]) serait de 0. Aucun utilisateur n'a interagi avec les 2 produits.
- La similarité entre le troisième (c=[1 0 1 0]) et le cinquième produit (e=[1 0 1 1]) est de 0.82. Une valeur près de 1 indique une grande similarité entre les produits.

Avantages et inconvénients

- (+) Des explications peuvent être fournies (parce que vous avez acheté le produit X, ces produits pourraient vous intéresser)
- (+) Facile et rapide à mettre en place
- (+) Le filtrage collaboratif donne habituellement de bons résultats
- (-) Cold-start problem pour les utilisateurs et les produits
- (-) Coût computationnel important

Modèle 3 : Factorisation matricielle

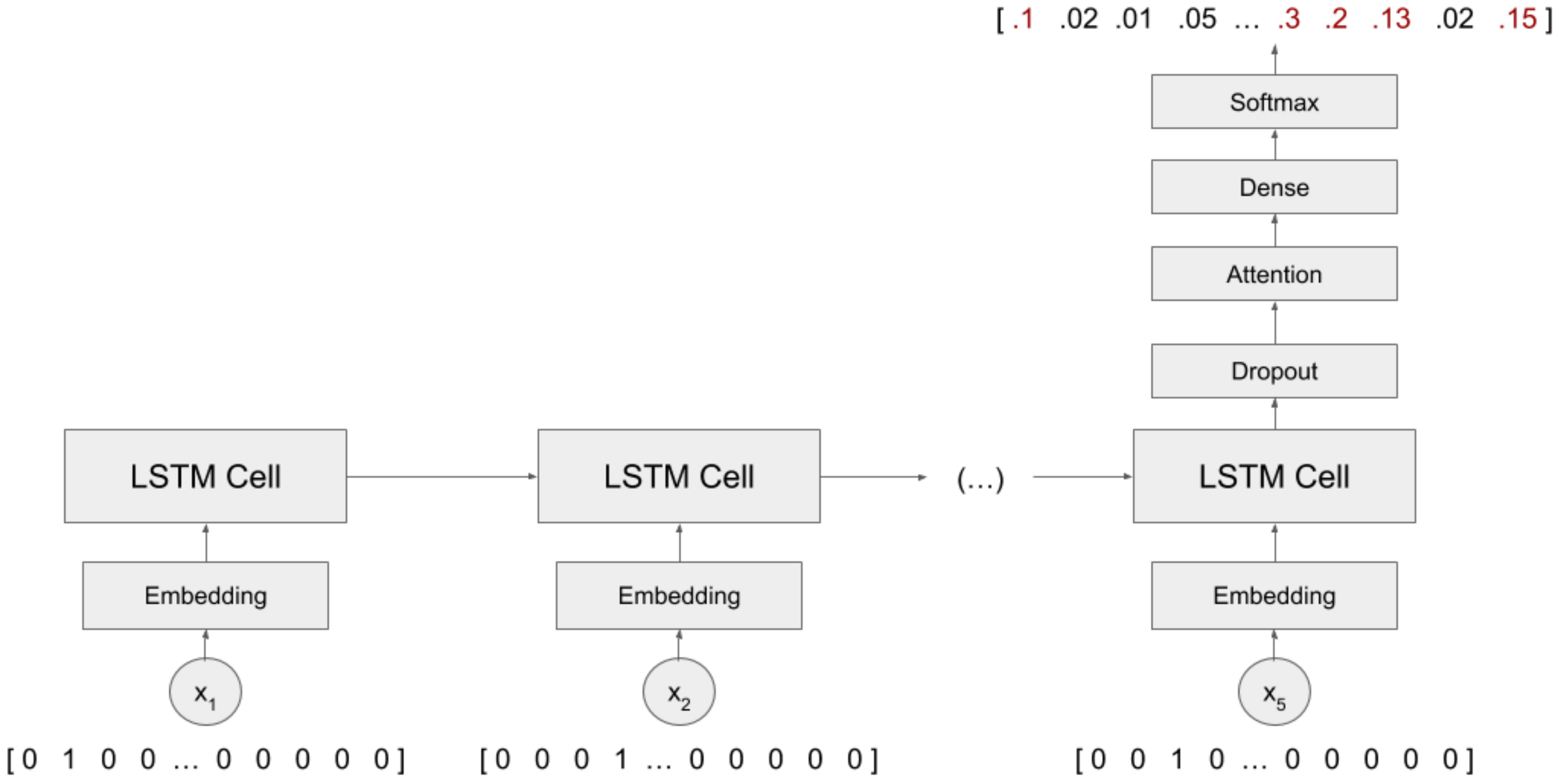
- Complétion de la matrice d’interaction utilisateurs-produits, à travers la factorisation matricielle
  - Plusieurs méthodologies ont été essayées, dont Singular Value Decomposition (SVD) et Non-negative Matrix Factorization (NMF)
- Prédit la probabilité d’interagir avec chacun des produits. Les produits recommandés sont ceux avec la plus grande probabilité d’interaction.

Avantages et inconvénients

- (+) La factorisation matricielle donne habituellement de très bons résultats
- (+) Peut révéler des caractéristiques latentes intéressantes
- (-) Cold-start problem pour les utilisateurs et les produits
- (-) Importants coûts de calculs
- (-) Nécessite une infrastructure virtuelle plus complexe pour déployer les modèles*

Modèle 4 : Réseau de neurones récurrent

- Envisager le problème comme étant une séquence de produits avec lesquels l'utilisateur interagit, et prédire le prochain.
  - Chaque produit est représenté par un vecteur ***one-hot.***
  - Les produits entrent dans le réseau séquentiellement
    - Le réseau prédit le prochain produit
  - La sortie du réseau est une distribution des probabilités pour chacun des produits dans le catalogue.
  - Le réseau utilise des neurones de type ***Long Short-Term Memory (LSTM)***, de la régularisation de type ***dropout*** et des principes d'attention


[.1 .02 .01 .05 … .3 .2 .13 .02 .15]
Softmax
Dense
Attention
Dropout
LSTM Cell
LSTM Cell
(…)
LSTM Cell
Embedding
Embedding
Embedding
x1
x2
x5
[0 1 0 0 … 0 0 0 0 0]
[0 0 0 1 … 0 0 0 0 0]
[0 0 1 0 … 0 0 0 0 0]

Avantages et inconvénients

- (+) Facile d’ajouter de nouveaux utilisateurs
- (+) Les réseaux de neurones récurrents donnent habituellement de très bons résultats
- (-) Cold-start problem pour les produits
- (-) Fonctionne surtout lorsqu’un utilisateur a interagi avec plusieurs produits

Métriques

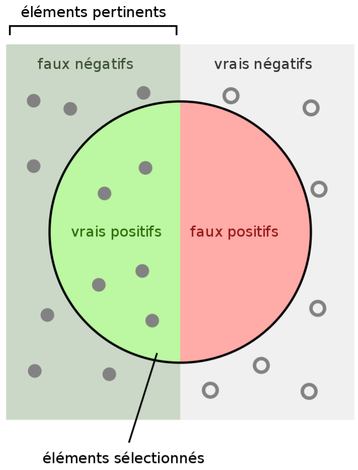
Uniquement des métriques hors ligne : l'équipe n'a pas accès à de l'évaluation en ligne ou à des études utilisateurs

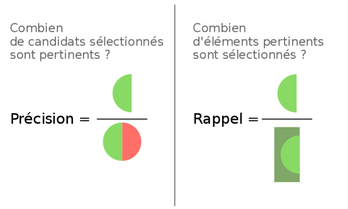
A. **Précision :** proportion des produits recommandés qui ont réellement été achetés par les utilisateurs

B. **Rappel :** proportion des produits réellement achetés qui ont été recommandés

C. **Couverture** : proportion des produits recommandés à au moins 1 utilisateur

Choix de ne pas considérer la diversité et la sérendipité

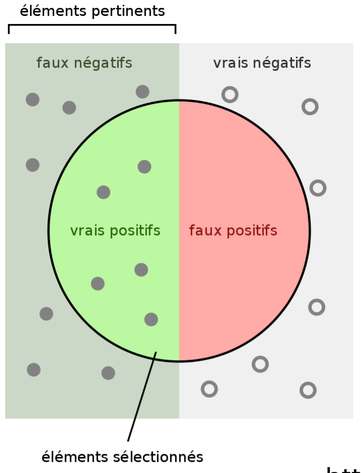


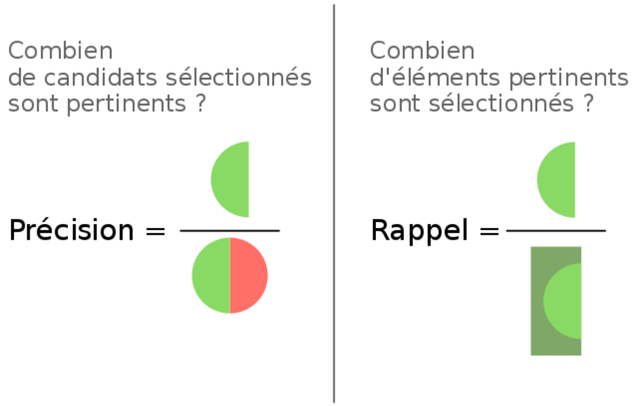


https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel

Résultats

Modèle	Aléatoire	Plus populaires	1. Similarité visuelle	2.Collaboratif, basé sur les produits	3. Factorisation matricielle	4. Réseaux de neurones récurrents
Précision	0,06%	1,5%	1,9%	3,4%	3,9%	4%
Rappel	0,07%	1,8%	2,3%	4,1%	5,3%	5,7%
Couverture	91%	0,07%	37,1%	74%	69%	57%





https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel

Choix final

D’abord, les modèles plus simples ont été implémentés :

1. Produits les plus populaires

2. Collaboratif basé sur les produits (modèle 2)

- Le modèle de similarité visuelle (modèle 1) ne présentait pas une performance suffisante pour justifier son implémentation.

- La factorisation matricielle (modèle 3) a été écartée: le gain de performance ne valait pas la peine d’investir dans des architectures logistiques plus complexes.

- Aujourd’hui, le modèle à réseaux de neurones récurrents est en place (modèle 4).

Comment choisir le bon modèle?

- Tâche plus subjective que les tâches habituelles d’apprentissage automatique
- Impératif : le modèle doit pouvoir traiter de grandes quantités de données
- Commencer par des modèles plus simples, jusqu’à plus complexes
- Choix final selon la performance sur les métriques choisies et des considérations logistiques
- 4 modèles ont été retenus

Limites

- Intuition qu’un autre modèle performerait mieux sur les métriques choisies
  - L’équipe cherche un modèle qui utilise à la fois les données d’interaction entre membres et produits, et les caractéristiques des produits.
- Le modèle actuel se concentre sur la performance à court terme
  - L’équipe aimerait un modèle qui puisse explorer davantage la diversité des préférences des utilisateurs, et éventuellement leur proposer des produits agréablement surprenants.

Pour aller plus loin

Quelles améliorations pourraient être proposées, ou quels modèles plus avancés devraient être testés en priorité?

Séance en petits groupes

Si le temps le permet…

Améliorations envisagées

1. Curiosité dans les réseaux de neurones récurrents
   - Ajouter des fonctionnalités aux réseaux utilisés : ajouter des techniques de curiosité
   - Permet d’explorer davantage la diversité des préférences des utilisateurs
2. Graph Neural Networks
   - Structurer différemment les données : graphe hétérogène, qui permet d’utiliser les données d’interaction et les caractéristiques des produits
   - Prédiction de lien entre noeuds d’un graphe
3. Apprentissage par renforcement
   - Modéliser la tâche différemment : processus séquentiel et interactif, considéré comme une boucle de propositions de produits et de rétroactions par l’utilisateur
   - Permet d’explorer davantage la diversité des préférences des utilisateurs