

© International Baccalaureate Organization 2023

All rights reserved. No part of this product may be reproduced in any form or by any electronic or mechanical means, including information storage and retrieval systems, without the prior written permission from the IB. Additionally, the license tied with this product prohibits use of any selected files or extracts from this product. Use by third parties, including but not limited to publishers, private teachers, tutoring or study services, preparatory schools, vendors operating curriculum mapping services or teacher resource digital platforms and app developers, whether fee-covered or not, is prohibited and is a criminal offense.

More information on how to request written permission in the form of a license can be obtained from <https://ibo.org/become-an-ib-school/ib-publishing/licensing/applying-for-a-license/>.

© Organisation du Baccalauréat International 2023

Tous droits réservés. Aucune partie de ce produit ne peut être reproduite sous quelque forme ni par quelque moyen que ce soit, électronique ou mécanique, y compris des systèmes de stockage et de récupération d'informations, sans l'autorisation écrite préalable de l'IB. De plus, la licence associée à ce produit interdit toute utilisation de tout fichier ou extrait sélectionné dans ce produit. L'utilisation par des tiers, y compris, sans toutefois s'y limiter, des éditeurs, des professeurs particuliers, des services de tutorat ou d'aide aux études, des établissements de préparation à l'enseignement supérieur, des fournisseurs de services de planification des programmes d'études, des gestionnaires de plateformes pédagogiques en ligne, et des développeurs d'applications, moyennant paiement ou non, est interdite et constitue une infraction pénale.

Pour plus d'informations sur la procédure à suivre pour obtenir une autorisation écrite sous la forme d'une licence, rendez-vous à l'adresse <https://ibo.org/become-an-ib-school/ib-publishing/licensing/applying-for-a-license/>.

© Organización del Bachillerato Internacional, 2023

Todos los derechos reservados. No se podrá reproducir ninguna parte de este producto de ninguna forma ni por ningún medio electrónico o mecánico, incluidos los sistemas de almacenamiento y recuperación de información, sin la previa autorización por escrito del IB. Además, la licencia vinculada a este producto prohíbe el uso de todo archivo o fragmento seleccionado de este producto. El uso por parte de terceros —lo que incluye, a título enunciativo, editoriales, profesores particulares, servicios de apoyo académico o ayuda para el estudio, colegios preparatorios, desarrolladores de aplicaciones y entidades que presten servicios de planificación curricular u ofrezcan recursos para docentes mediante plataformas digitales—, ya sea incluido en tasas o no, está prohibido y constituye un delito.

En este enlace encontrará más información sobre cómo solicitar una autorización por escrito en forma de licencia: <https://ibo.org/become-an-ib-school/ib-publishing/licensing/applying-for-a-license/>.

Informatique

Étude de cas : Puis-je vous faire une recommandation ?

A utiliser en mai et novembre 2023

Instructions destinées aux candidats

- Ce livret d'étude de cas est indispensable pour l'épreuve 3 du niveau supérieur.

Page vierge

Introduction

Vous venez de terminer une visioconférence avec vos deux amis Jungmin et Lijing. Ils ont une idée d'entreprise en ligne et ont besoin de votre aide. L'entreprise va s'appeler *NextStar*. Elle fournira une application qui permettra aux utilisateurs de visionner les œuvres d'artistes encore inconnus du grand public.

Acteurs, chanteurs, scénaristes, comiques, peintres, sculpteurs et cinéastes, ou tout autre type d'artiste voulant démontrer leurs talents, pourront télécharger des fichiers dans l'application. Tous les utilisateurs seront en mesure d'évaluer le contenu téléchargé. En fonction de leurs évaluations, l'application recommandera un nouveau contenu pour chaque utilisateur.

- 10 Jungmin et Lijing envisagent une adhésion gratuite au site *NextStar* et pensent qu'ils pourront gagner de l'argent grâce à la publicité lorsque le site aura un nombre suffisant d'utilisateurs. Il sont conscients qu'à terme, l'application demandera un espace de stockage très important. Ils envisagent donc de faire appel à un hébergeur cloud. Dès qu'une quantité suffisante de contenu aura été ajoutée, l'application utilisera un système de recommandation.
- 15 Les informations qui suivent résument les recherches déjà effectuées et signalent certains défis à prendre en considération.

Cloud computing

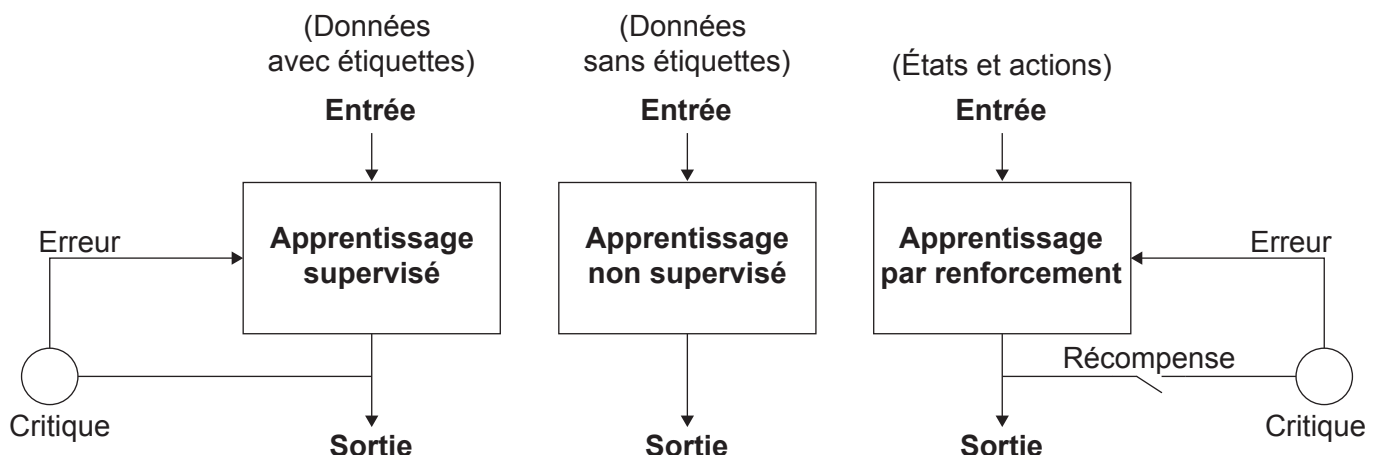
Les applications d'hébergement qui utilisent des données de niveau entreprise sont largement disponibles et bon marché grâce au cloud computing. Les utilisateurs payent uniquement pour les ressources utilisées. Ils peuvent ainsi commencer modestement puis ajouter des ressources au fur et à mesure de l'expansion de leurs entreprises. Le cloud computing est donc idéal pour une start-up comme *NextStar*.

Plusieurs *modèles de déploiement du cloud* peuvent être utilisés pour héberger les données de *NextStar*. Il existe aussi trois *modèles de services cloud* : *logiciel en tant que service (SaaS)*, *plateforme en tant que service (PaaS)* et *infrastructure en tant que service (IaaS)*. Le modèle envisagé pour *NextStar* est l'IaaS.

Apprentissage automatique

L'apprentissage automatique est une forme d'intelligence artificielle. Il en existe trois types principaux : apprentissage supervisé, apprentissage non supervisé et *apprentissage par renforcement* (voir **figure 1**).

Figure 1 : les trois types principaux d'apprentissage automatique



Un algorithme d'apprentissage supervisé utilise des *données d'entraînement* étiquetées pour apprendre une fonction qui produira une sortie appropriée lorsque des données non étiquetées seront fournies. En général, l'apprentissage supervisé sert à classifier des données ou à faire des prévisions.

- 35 Un algorithme d'apprentissage non supervisé apprend des modèles à partir de données non étiquetées. Ces algorithmes tirent leurs références de l'observation des données en entrée réelles. Le système peut organiser les données en sous-ensembles, ou blocs, qui n'ont pas été préalablement classifiés par les programmeurs.

- 40 Un algorithme d'apprentissage par renforcement apprend par essai-erreur dans un environnement interactif en utilisant les retours de ses propres actions et expériences. Certains systèmes de recommandation peuvent être considérés comme un type d'apprentissage par renforcement car un comportement positif, par exemple l'évaluation de contenu, est récompensé par de meilleures recommandations.

Systèmes de recommandation

- 45 Lorsqu'il existe une quantité considérable de contenu, un système de recommandation mène un utilisateur à du contenu qu'il n'a pas encore consulté mais qu'il trouvera éventuellement intéressant. Ces systèmes utilisent les données issues du contenu préalablement évalué par les utilisateurs (données réelles) pour prédire leurs préférences concernant le contenu qu'ils n'ont pas encore évalué (données prédites).

- 50 La majorité des systèmes de recommandation utilisent l'apprentissage supervisé. L'apprentissage non supervisé et l'apprentissage par renforcement sont moins courants.

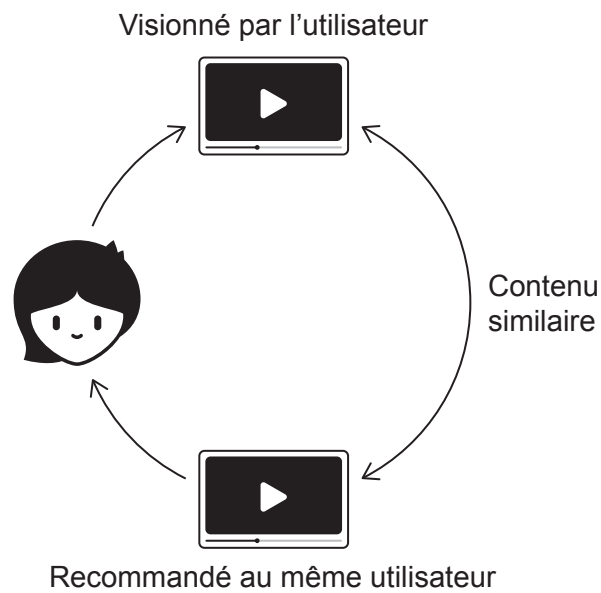
- Les systèmes de recommandation utilisent le *filtrage basé sur le contenu*, le *filtrage collaboratif* ou un mélange des deux. Les systèmes de recommandation hybrides allient plusieurs algorithmes d'apprentissage automatique. Cela a été démontré le 21 septembre 2009 lorsque
55 l'équipe BellKor's Pragmatic Chaos a gagné le prix Netflix d'un million de dollars américains pour le meilleur système de recommandation de films à filtrage collaboratif. Ce système réunit 107 algorithmes dans un modèle hybride qui a dépassé de 10,06 % le score *erreur quadratique moyenne* (RMSE en anglais) de l'algorithme de Netflix.

Filtrage basé sur le contenu

- 60 Le filtrage basé sur le contenu, parfois appelé « *item-item filtering* », se base sur les attributs d'un item plutôt que sur les interactions et commentaires des utilisateurs. Cette méthode basée sur le contenu repose sur une classification spécifique à l'utilisateur dans laquelle le classifieur apprend ce que l'utilisateur aime ou n'aime pas en se basant sur les attributs des items.

- Étant donné que le système de recommandation de *NextStar* contiendra des clips vidéo, les
65 attributs pourront par exemple inclure le genre, la date de diffusion, l'artiste, la langue, le sexe et l'âge. À titre d'illustration, si un utilisateur donne une note élevée aux clips vidéo d'humoristes, le système est susceptible de lui recommander davantage de clips d'humoristes (voir **figure 2**).

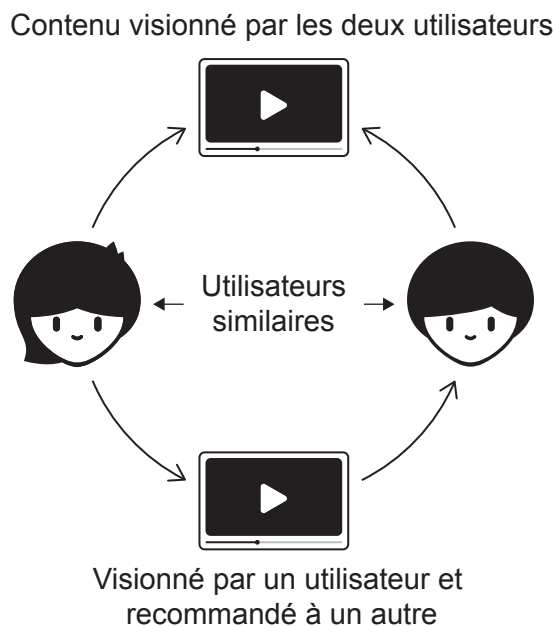
Figure 2 : exemple de filtrage basé sur le contenu



Filtrage collaboratif

Dans le filtrage collaboratif, les recommandations à chaque utilisateur sont générées en utilisant des données d'évaluation fournies par d'autres utilisateurs et items. L'hypothèse de base est que les utilisateurs qui étaient d'accord dans le passé seront en général d'accord à l'avenir. Donc, si deux utilisateurs ont noté du contenu de la même façon, un autre contenu bien noté par un des utilisateurs est susceptible de plaire à l'autre. Ce contenu lui est alors recommandé (voir **figure 3**).
 L'une des faiblesses des systèmes de recommandation à filtrage collaboratif est le *biais de popularité* : le contenu populaire est recommandé trop fréquemment.

Figure 3 : exemple de filtrage collaboratif



Le filtrage collaboratif peut utiliser différents algorithmes pour recommander du nouveau contenu. Deux types d'algorithmes utilisables sont la méthode des *k plus proches voisins* (*k-NN*) et la *factorisation de matrices*.

K plus proches voisins (k-NN)

80 L'algorithme k-NN utilise la similarité des caractéristiques pour prédire les valeurs de nouvelles données ou de données manquantes. Cela signifie que l'algorithme attribue des valeurs aux nouveaux points de données en fonction de leur degré de ressemblance aux autres points de jeu de données d'entraînement.

85 L'algorithme k-NN fait ses prévisions en fonction des voisins les plus proches. « K » représente le nombre de voisins. Il s'agit d'un *hyperparamètre* qui peut être modulé par essai-erreur.

Factorisation de matrices


90 La factorisation de matrices est un autre algorithme possible. La méthode standard de factorisation de matrices est difficile à utiliser dans les systèmes de recommandation car le jeu de données n'est pas complet. Pour surmonter cette contrainte, les valeurs sont estimées en utilisant des matrices plus petites à l'aide d'un algorithme itératif.

Dans la **figure 4**, la matrice d'interaction utilisateur/item représente la note de chaque utilisateur (lignes) donnée à chaque item de contenu (colonnes). L'utilisateur 1, représenté par US1, a noté les trois premiers items mais pas les items 4 et 5, représentés par IT4 et IT5.

Figure 4 : matrice d'interaction utilisateur/item

		Item				
		IT1	IT2	IT3	IT4	IT5
Utilisateur	US1	3	2	3		
	US2		2	2	2	
	US3	3	4		4	2
	US4	2	2	4		1
	US5	2		3	2	1
	US6			3	1	1

Légende :

 Contenu inconnu

95 La factorisation de matrices décompose la matrice d'interaction utilisateur/item en deux matrices plus petites, une matrice item/caractéristique et une matrice utilisateur/caractéristique, pour saisir les caractéristiques les plus importantes nécessaires à l'apprentissage. Si les valeurs des matrices item/caractéristique et utilisateur/caractéristique sont modifiées, les valeurs correspondantes seront également changées dans la matrice d'interaction utilisateur/item (voir **figure 5**).

Figure 5 : Factorisation de matrices

Matrice item/caractéristique		Item				
		IT1	IT2	IT3	IT4	IT5
Caractéristique	C1	1,2	3,1	2,1	4,5	0,7
	C2	2,6	1,5	4,4	0,4	1,1

Matrice utilisateur/ caractéristique		Caractéristique	
		C1	C2
Utilisateur	US1	0,3	0,7
	US2	0,3	0,4
	US3	0,7	0,8
	US4	0,4	0,6
	US5	0,7	0,4
	US6	0,1	0,6

Matrice d'interaction utilisateur/item		Item				
		IT1	IT2	IT3	IT4	IT5
Utilisateur	US1	2,18	1,98	3,71	1,63	0,98
	US2	1,4	1,53	2,39	1,51	0,65
	US3	2,92	3,37	4,99	3,47	1,37
	US4	2,04	2,14	3,48	2,04	0,94
	US5	1,88	2,77	3,23	3,31	0,93
	US6	1,68	1,21	2,85	0,69	0,73

100 L'ajustement des matrices se réalise en générant des prévisions de préférences pour le contenu déjà doté de données de préférence. Lorsque les valeurs prédites se rapprochent des notes réelles, l'hypothèse est que les matrices seront capables de prédire efficacement les préférences pour le contenu sans données.

105 Une *algorithme de gradient stochastique* utilise une *fonction objectif* pour moduler chaque cellule en modifiant légèrement les matrices item/caractéristique et utilisateur/caractéristique. Par exemple, dans la **figure 4**, la valeur de la cellule correspondant à l'intersection US1/IT1 est de 3, mais dans la **figure 5** cette valeur est de 2,18. L'erreur pour cette cellule est donc $(3 - 2,18)^2$, soit 0,6724.

Entraînement des systèmes de recommandation

110 On peut évaluer un système de recommandation en divisant les données contenant les notes en un jeu de données d'entraînement et un jeu de données de test. Une division courante consiste à attribuer 80 % des données au jeu d'entraînement et les 20 % restants au jeu de test.

Un système de recommandation apprend les relations entre les items ainsi que celles entre les utilisateurs. Une fois entraîné, il prédit la note qu'un utilisateur attribuera à un item qui n'a pas encore été évalué par ce dernier.

115 Un problème courant associé à l'entraînement d'un algorithme d'apprentissage automatique est le *surapprentissage* (ou *overfitting*), dans lequel le modèle est trop proche du jeu de données d'entraînement. Lorsque le modèle est entraîné pendant trop longtemps sur les données d'entraînement, ou qu'il est trop complexe, il peut commencer à apprendre des caractéristiques non pertinentes. Par conséquent, le modèle ne généralise plus efficacement lorsqu'il est confronté à de nouvelles données.

120 Évaluation des systèmes de recommandation

On peut évaluer l'exactitude des systèmes de recommandation en employant deux mesures différentes : l'*erreur absolue moyenne* (EAM) et l'erreur quadratique moyenne (RMSE). Ces mesures donnent une idée des performances des systèmes pour les données d'entraînement et celles de test.

- 125 Cependant, l'efficacité d'un système de recommandation n'est pas entièrement connue avant que le public ne s'en serve. Un tel système n'est pas performant s'il ne recommande pas du contenu que l'utilisateur va aimer ou qu'il recommande du contenu qu'il n'aime pas.

La *précision* et le *rappel* sont les indicateurs de performances utilisés sur les données réelles.

- 130 La précision est une mesure d'exactitude, qui est la proportion de cas pertinents parmi les cas extraits. Le rappel est une mesure d'exhaustivité. La *F-mesure* donne un score unique dans lequel la précision et le rappel sont pondérés de façon égale.

La manière dont les recommandations sont affichées aux utilisateurs est également importante. Une liste peut suffire, mais on peut aussi sélectionner le contenu recommandé par groupe ou sous-groupe. Les groupes peuvent être organisés par genre, sexe, âge ou tout autre catégorie.

135 Considérations sociales et éthiques

Lorsqu'un modèle est élaboré à partir des comportements des utilisateurs, on peut utiliser deux types de *données de comportement* : les données explicites et les données implicites.

Les données de comportement explicites sont extraites des données fournies par les utilisateurs, par exemple, lorsque l'utilisateur évalue un clip vidéo, saisit ses préférences ou recherche un item.

- 140 Les utilisateurs pourraient penser qu'il s'agit des seules données utilisées pour l'élaboration des recommandations.

Les données de comportement implicites sont des données dont l'utilisateur n'a pas connaissance du recueil. Par exemple, il peut s'agir de ses clics, de ses achats, voire de ses frappes recueillies à l'aide d'un enregistreur.

- 145 La qualité des données utilisateur est essentielle au succès du projet *NextStar*, mais des considérations éthiques sont à prendre en compte concernant le recueil, le stockage et l'utilisation des données de comportement. Le *droit à l'anonymat* ainsi que le *droit à la vie privée* des utilisateurs sont également à prendre en considération.

Défis rencontrés

- 150 Pour aider vos amis à monter leur entreprise commerciale, vous devez vous pencher sur une série de défis :

- comprendre les similarités des apprentissages supervisé, non supervisé et par renforcement ainsi que leurs différences ;
- comprendre comment utiliser l'algorithme k-NN et la factorisation de matrices dans les

- 155 systèmes de recommandation ;

- comprendre comment entraîner, tester et évaluer un système de recommandation ;
- comparer les systèmes de recommandation à filtrage basé sur le contenu et à filtrage collaboratif ;
- comprendre les considérations éthiques associées au recueil, au stockage et à l'utilisation

- 160 des données de comportement des utilisateurs.

Les candidats n'ont pas besoin de connaître les équations mathématiques relatives aux systèmes de recommandation.

Terminologie supplémentaire

Algorithme des k plus proches voisins (k-NN, *k-nearest neighbour algorithm*)

Algorithme du gradient stochastique (*stochastic gradient descent*)

Apprentissage par renforcement (*reinforcement learning*)

Biais de popularité (*popularity bias*)

Données de comportement (*behavioural data*)

Données d'entraînement (*training data*)

Droit à l'anonymat (*right to anonymity*)

Droit à la vie privée (*right to privacy*)

Erreur absolue moyenne (EAM, *mean absolute error*)

Erreur quadratique moyenne (RMSE, *root-mean-square error*)

Factorisation de matrices (*matrix factorization*)

Filtrage basé sur le contenu (*content-based filtering*)

Filtrage collaboratif (*collaborative filtering*)

Fonction objectif (*cost function*)

F-mesure (*F-measure*)

Hyperparamètre (*hyperparameter*)

Modèles de déploiement du cloud (*cloud deployment models*)

Modèles de services cloud (*cloud delivery models*) :

infrastructure en tant que service (IaaS)

logiciel en tant que service (SaaS)

plateforme en tant que service (PaaS)

Précision (*precision*)

Rappel (*recall*)

Surapprentissage (*overfitting*)

Certains produits, sociétés et individus mentionnés dans cette étude de cas sont fictifs. Toute ressemblance avec des entités réelles ne saurait être que fortuite.

Avertissement :

Le contenu utilisé dans les évaluations de l'IB est extrait de sources authentiques issues de tierces parties. Les avis qui y sont exprimés appartiennent à leurs auteurs et/ou éditeurs, et ne reflètent pas nécessairement ceux de l'IB.

Références :

Figure 1 Jones, M. Tim, 2017. Models for machine learning. [en ligne] Disponible sur Internet : <https://developer.ibm.com/articles/cc-models-machine-learning/> [Référence du 15 octobre 2021]. Source adaptée.

Tous les autres textes, graphiques et illustrations : © Organisation du Baccalauréat International 2023