Formation Machine Learning

Bonnes Pratiques de Développement, Système de Recommandation

Giraud François-Marie

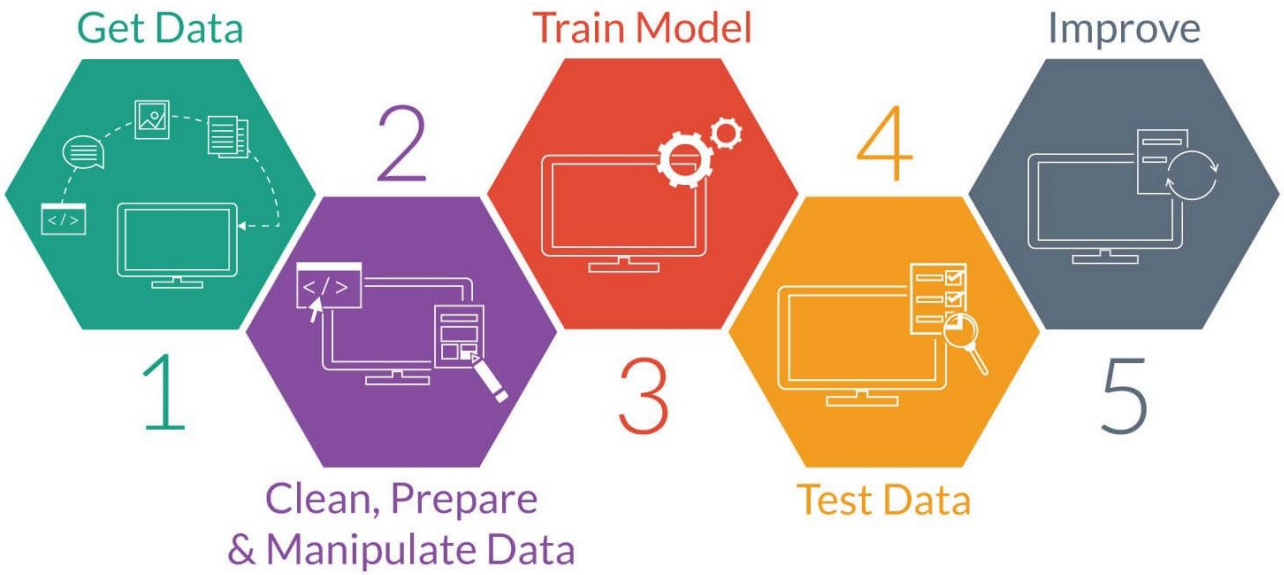
7 Juin 2019

ENI Service

Mettre en place un transition IA

Développer un projet en Machine Learning

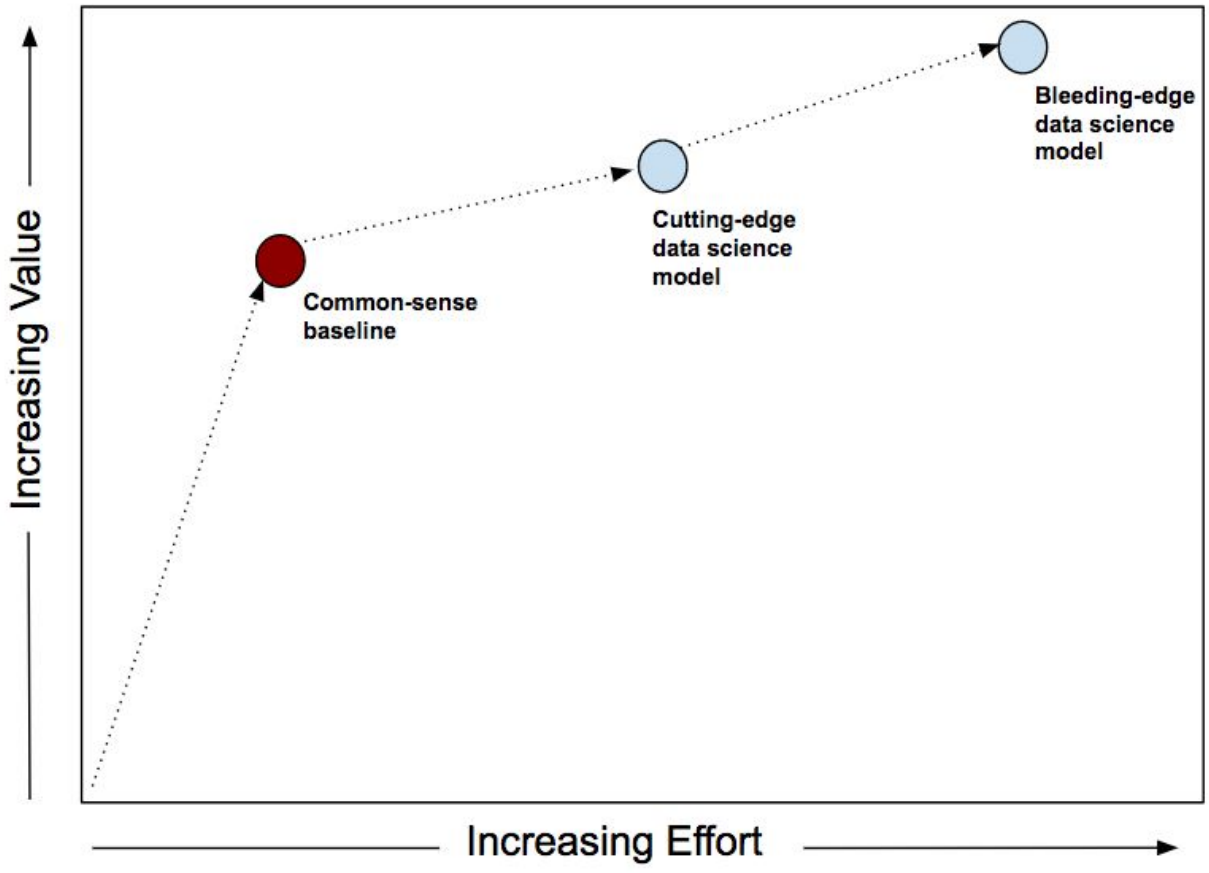
Développer un projet en Machine Learning



Développer un projet en Machine Learning

- Séparer les données en TRAIN/VALIDATION/TEST (i.e 60/20/20)
- Apprendre sur **TRAIN**
- Optimizer les hyperparamètres sur **VALIDATION**
- Observer la performance finale sur **TEST**

Développer un projet en Machine Learning



Projet académiques Vs Industriels

- $\neq$ Développement logiciel
- $\neq$ Infrastructure
- $\neq$ Performances

Développement logiciel

Académique

- Pile de scripts
- Peu de documentation
- Fonctionne le temps de l'expérience
- "Fair use"

Industriel

- Code hiérarchisé et déployable en production
- Documentation
- Code maintenable et robuste
- Galaxies de licences à respecter

Développer un projet en Machine Learning

Infrastructure

Académique	Industriel
▪ Données = un fichier	▪ Données = cloud
▪ Hardware limité	▪ Cloud computing
▪ Performance = Précision	▪ Performance = Plus-value

Développer un projet en Machine Learning

Un problème d'ingénérie avant d'être un problème de machine learning :

Données et prétraitements de qualité $>$ algorithme de qualité

Développer un projet en Machine Learning

Une approche en 4 étapes :

- Créer un pipeline robuste de bout en bout (sans ML)
- Intégrer du ML simple
- Ajouter des caractéristiques sensées
- Conserver un pipeline robuste

Développer un projet en Machine Learning

Créer un pipeline robuste de bout en bout (sans ML) :

- Une baseline avec une heuristique
- Mettre en place des statistiques d'évaluation

Développer un projet en Machine Learning

Intégrer du ML simple :

1. Obtenir des données
2. Définir UNE métrique d'évaluation facile à observer
3. Définir des caractéristiques sensées et faciles à obtenir
4. Considérer les heuristiques comme des caractéristiques
5. Documenter TOUTES les caractéristiques utilisées

Développer un projet en Machine Learning

Intégrer du ML simple :

6. Apprendre un modèle tous les n-jours
7. Évaluer la dégradation des performances en fonction de l'âge du modèle
8. Vérifier les performances en test avant de déployer en production
9. Modèle appris sur des données jusqu'au jour N, tester sur les données après le jour N
10. Mesurer la différence entre performance en apprentissage et test
11. Plateau de performance $\Rightarrow$ trouver des nouvelles caractéristiques/augmenter la puissance du modèle
12. Supprimer des caractéristiques pas déterminantes

Ajouter des caractérstiques sensées :

- Beaucoup de caractéristiques simples $>$ peu de caractéristiques complexes
- Des caractéristiques répandues plutôt que rares
- Regarder les erreurs pour imaginer les caractéristiques qui aideraient
- Communiquer avec les experts métiers

Des questions à garder en tête :

- Ajouter des statistiques d'évaluation ?
- Revoir/Complexifier la métrique d'évaluation ?
- Les données sont-elle "stables" ?

Machine Learning

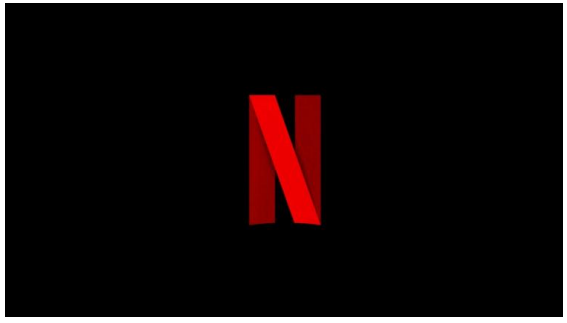
Système de recommandation

Système de recommandation

Filtrage de l'information :

Produire une information personnalisé pour un “client”

Système de recommandation



Système de recommandation

Intérêt mutuel :

- Facilitation des usages pour le “client”

Système de recommandation

Intérêt mutuel :

- Facilitation des usages pour le “client”
- Des bénéfices accrus pour le “vendeur”

Système de recommandation

3 types de recommandation :

- centrée sur l'utilisateur (user-based)

Système de recommandation

3 types de recommandation :

- centrée sur l'utilisateur (user-based)
- centrée sur le contenu (item-based)

Système de recommandation

3 types de recommandation :

- centrée sur l'utilisateur (user-based)
- centrée sur le contenu (item-based)
- colaborative (ou sociale)

Système de recommandation

2 types de collecte des données :

- active

Système de recommandation

2 types de collecte des données :

- active
- passive

Machine Learning

Filtrage Collaboratif

Filtrage Collaboratif

$\approx$ Système de recommandation

Méthode hybride :

- user-based
- item-based
- collaboratif

Filtrage Collaboratif

Données explicites et implicites :
Explicite : donnée de qualité mais contraignante pour l'utilisateur (donc potentiellement biaisé)
Implicite : donnée objective mais pauvre du point de vue de la qualité

Filtrage Collaboratif

Grosse volumétrie de données $\Rightarrow$ algorithmes simples et rapides

Machine Learning

Filtrage Collaboratif user-based

Filtrage Collaboratif user-based

Exemple de données collectées :

	film 1	film 2	film 3	film 4
u1	?	2	5	?
u2	5	1	4	3
u3	?	?	1	5
u4	3	3	?	4

Filtrage Collaboratif user-based

Principe : Une bonne recommandation est faite à partir d'individus similaires

Condition : Avoir un échantillon important d'individus représentatifs

Filtrage Collaboratif user-based

	film1	film2	film3	film4
u1	?	2	5	?
u2	5	1	4	3
u3	?	?	1	5
u4	3	3	?	4

Comment prédire les notes manquantes ? (et donc faire une recommandation)

Algorithme pour déterminer u1(film1) :

- Identifier les K plus proches utilisateurs de u1
- Agréger leurs notes pour le film1

Filtrage Collaboratif user-based

On a donc besoin :

- d'une mesure de similarité entre les utilisateur
- de trouver un K optimal
- d'une fonction d'agrégation

Filtrage Collaboratif user-based

Similarité de corrélation (Pearson) :

$$cor(a,b) = \frac{\sum_{i \in I}(a_i - \bar{a})((b_i - \bar{b})}{\sqrt{\sum_{i \in I}(a_i - \bar{a})^2 * \sum_{i \in I}(b_i - \bar{b})^2}}$$

Similarité cosinus :

$$cos(a,b) = \frac{a * b^t}{|a| * |b|}$$

Et plein d'autres dans la littérature...

Filtrage Collaboratif user-based

Agrégation des notes des K utilisateurs retenus :

- Moyenne
- Moyenne pondéré par la similarité
- ...

Filtrage Collaboratif user-based

- Des calculs simples mais il faut parcourir toute la base dès qu'on a une nouvelle entrée ou modification, heureusement des heuristiques existent (Locality-sensitive hashing)

Filtrage Collaboratif user-based

- Des calculs simples mais il faut parcourir toute la base dès qu'on a une nouvelle entrée ou modification, heureusement des heuristiques existent (Locality-sensitive hashing)
- "Cold start problem"

Machine Learning

Filtrage Collaboratif item-based

Filtrage Collaboratif item-based

	item 1	item 2	item 3	item 4
u1	1	2	5	?
u2	5	1	4	3
u3	?	?	1	5
u4	3	3	?	4

Pour remplir les cellules manquantes, au lieu de traiter les lignes on traite les colonnes.
Une première solution est de transposer la matrice et d'appliquer l'algorithme décrit pour le user-based.

Filtrage Collaboratif item-based

Méthode couramment utilisée $\Rightarrow$ régression simplifié.

	item 1	item 2	item 3	item 4
u1	1	2	5	?
u2	5	1	4	3
u3	?	?	1	5
u4	3	3	?	4

Filtrage Collaboratif item-based

Sachant la colonne item1, on cherche uniquement une valeur de biais telle que :

$$item4 = item1 + b_1$$

Pour ce faire, on prend le biais moyen entre ces deux colonnes :

$b_1 = \frac{\sum_{i \in I}(item4_i - item1_i)}{\#I}$, dans notre exemple : $b_1 = \frac{(3-5)+(4-3)}{2} = -0.5$

	item 1	item 2	item 3	item 4
u1	1	2	5	?
u2	5	1	4	3
u3	?	?	1	5
u4	3	3	?	4

Filtrage Collaboratif item-based

On calcul alors les biais pour chaque colonne, nous donnant une prediction de la valeur u1(item4) sachant u1(item1), une autre sachant u1(item2) et une dernière sachant u1(item3).
La prédiction finale pour u1(item4) est la moyenne pondérée de toutes ces prédictions.

$b_1 = -0.5 \Rightarrow u1(item4)_1 = 0.5$

$b_2 = 1.5 \Rightarrow u1(item4)_2 = 2.5$

$b_3 = 1.5 \Rightarrow u1(item4)_3 = 2.5$

$$u1(item4) = \frac{\#I_1 * 0.5 + \#I_2 * 2.5 \#I_3 * 2.5}{\#I_1 + \#I_2 + \#I_3} = \frac{11}{6} \approx 1.83$$

	item 1	item 2	item 3	item 4
u1	1	2	5	**1.83**
u2	5	1	4	3
u3	?	?	1	5
u4	3	3	?	4

Filtrage Collaboratif item-based

- Calculs simples à mettre en oeuvre et à tenir à jour

Filtrage Collaboratif item-based

- Calculs simples à mettre en oeuvre et à tenir à jour
- Beaucoup de parmètres à maintenir quand on a beaucoup d'item : $\left(\frac{\#item(\#item-1)}{2}\right)$

Filtrage Collaboratif item-based

- Calculs simples à mettre en oeuvre et à tenir à jour
- Beaucoup de parmètres à maintenir quand on a beaucoup d'item : $\left(\frac{\#item(\#item-1)}{2}\right)$
- "Cold start problem"

Machine Learning

Filtrage Colaboratif - Évaluation

Filtrage Colaboratif - Évaluation

On utilise des techniques d'évaluation du machine learning impliquant des bases d'apprentissage et de test.

$$RMSE = \sqrt{\frac{\sum_{u,i}(u(i)_{pred} - u(i))^2}{\#Q}}$$

où $\#Q$ est le nombre de prédictions que l'on fait dans la base.

Machine Learning

Filtrage Collaboratif - Aujourd'hui

Filtrage Collaboratif - SVD

Utiilisation de SVD pour réduire le nombre de dimensions du problème.

$M = U^p * V^p * I^p$, où :

M est la matrice user/item de $\mathbf{R}^{n*m}$

U^p la matrice des vecteurs de base orthonormée de $\mathbf{R}^n$ (input)

V^p la matrice diagonale des valeurs singulières de U

I^p la matrice des vecteurs de base orthonormée de $\mathbf{R}^m$ (output)

$$\mathrm{M} = \begin{array}{|c|c|c|}\hline 2 & 5 & ? \\ \hline 1 & 4 & 3 \\ \hline ? & 1 & 5 \\ \hline 3 & ? & 4 \\ \hline \end{array} =$$

$$\begin{array}{|c|c|}\hline U^p_{(1,1)} & U^p_{(1,2)} \\ \hline U^p_{(2,1)} & U^p_{(2,2)} \\ \hline U^p_{(3,1)} & U^p_{(3,2)} \\ \hline U^p_{(4,1)} & U^p_{(4,2)} \\ \hline \end{array} * \begin{array}{|c|c|}\hline V^p_{(1)} & 0 \\ \hline 0 & V^p_{(2)} \\ \hline \end{array} * \begin{array}{|c|c|c|}\hline I^p_{(1,1)} & I^p_{(1,2)} & I^p_{(1,3)} \\ \hline I^p_{(2,1)} & I^p_{(2,2)} & I^p_{(2,3)} \\ \hline \end{array}$$

Filtrage Collaboratif - SVD

Avantages :

On choisit le nombre de valeurs singulières de notre transformation (PCA).

Filtrage collaboratif user-based $\Rightarrow U^p$

Filtrage collaboratif item-based $\Rightarrow I^p$

Filtrage Collaboratif - SVD

La SVD est extrèmement coûteuse sur des grandes matrices :

$$O(min(m * n^2, m^2 * n))$$

Heureusement des algorithmes approchés existent. Ces approches tirent parti du fait que les matrices sont généralement sparses et que l'on ne s'intéresse qu'a un nombre limité de valeurs singulières.
Par exemple, on peut calculer les 20 premières valeurs singulières d'une matrice 100k x 100k avec 1M valeurs non-nulles en moins d'une seconde (redsvd).

Filtrage Collaboratif - Embeddings

Utilisation d'Embeddings sur les utilisateurs et les items afin de pouvoir exprimer la matrice des ratings ainsi :
$M = U^{embed} * I^{embed}$

Pour Faire une recommandation :
$argmax(u_{query} * I^t)$

Filtrage Collaboratif - Recommandation par le contenu

Jusqu'à présent, la nature des items était sans importance.
Que les items soient des livres ou des voitures, l'algorithme ne faisait aucune différence.
On peut utiliser les détails textuels, les images (, etc...) sur les items afin d'améliorer les recommandations. (Embeddings, Catégorisation des items)

Machine Learning

Filtrage Collaboratif - Conclusions

Filtrage Collaboratif - Conclusions

- Défis technologique :

Filtrage Collaboratif - Conclusions

- Défis technologique :
  - Grande volumétrie $\Rightarrow$ algorithmes simples et rapides

Filtrage Collaboratif - Conclusions

- Défis technologique :
  - Grande volumétrie $\Rightarrow$ algorithmes simples et rapides
  - Quel est le minimum nécessaire d'information sur un utilisateur ?

Filtrage Collaboratif - Conclusions

- Défis technologique :
  - Grande volumétrie $\Rightarrow$ algorithmes simples et rapides
  - Quel est le minimum nécessaire d'information sur un utilisateur ?
- Défis sociologique :

Filtrage Collaboratif - Conclusions

- Défis technologique :
  - Grande volumétrie $\Rightarrow$ algorithmes simples et rapides
  - Quel est le minimum nécessaire d'information sur un utilisateur ?
- Défis sociologique :
  - aimer $\neq$ acheter

Filtrage Collaboratif - Conclusions

- Défis technologique :
  - Grande volumétrie $\Rightarrow$ algorithmes simples et rapides
  - Quel est le minimum nécessaire d'information sur un utilisateur ?
- Défis sociologique :
  - aimer $\neq$ acheter
  - prédiction par agrégation $\Rightarrow$ peu propice à l'exploration

TP 5

Filtrage Colaboratif

TP 5 : Filtrage Colaboratif

www.filtrage-colaboratif.ipynb