



Big Data et Fouille de données

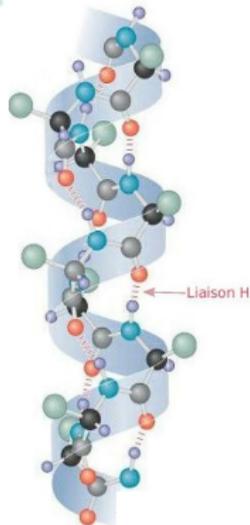
Massih-Reza Amini

Université Grenoble Alpes
Big Data HPC - 26 avril 2016



En apprentissage automatique

on cherche à trouver l'association entre les observations et leurs sorties désirées



Hélice alpha

4 → 4	2 → 2	3 → 3
4 → 4	9 → 9	0 → 0
5 → 5	7 → 7	1 → 1
9 → 9	0 → 0	3 → 3
6 → 6	7 → 7	4 → 4

From: Ms.Ella Golan <tamas.gaid@eleves.ec-nantes.fr> ☆

Subject: [SPAM MEDIUM]RE

Reply to: Ms.Ella Golan <egolan1@foxmail.com> ☆

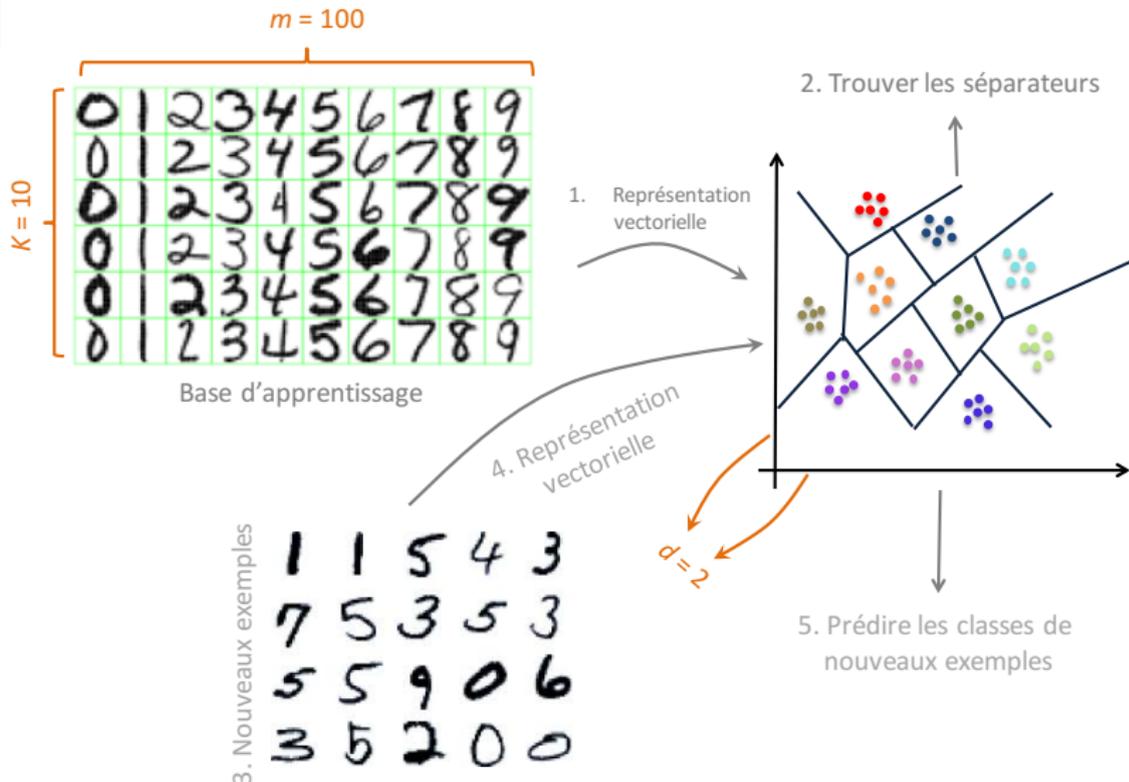
Reply Forward

I am Ms.Ella Golan, I am the Executive Vice President Banking Division with FIRST INTERNATIONAL BANK OF ISRAEL LTD (FIB). I am getting in touch with you regarding an extremely important and urgent matter. If you would oblige me the opportunity, I shall provide you with details upon your response.

Faithfully,
Ms.Ella Golan

↓
Spam

En apprentissage automatique



Cadre classique de l'apprentissage automatique

Nous considérons un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$ et un espace de sortie $\mathcal{Y} = \{1, \dots, K\}$.

Hypothèse: Les paires d'exemples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont *identiquement* et *indépendamment* distribuées (i.i.d) d'après une distribution de probabilité fixe \mathcal{D} .

Échantillon: Nous observons une séquence de m paires d'exemples (\mathbf{x}_i, y_i) générées i.i.d suivant \mathcal{D} .

Objectif: Construire une fonctions de prédiction $f : \mathcal{X} \rightarrow \mathcal{Y}$ qui prédit la sortie y d'un nouvel exemple \mathbf{x} avec la plus petite probabilité d'erreur.

Apprentissage supervisé

- ❑ Les modèles discriminants trouve directement la fonction de classification $f : \mathcal{X} \rightarrow \mathcal{Y}$ à partir d'une classe de fonctions \mathcal{F} ;
- ❑ La fonction trouvée devrait être celle qui minimise la probabilité d'erreur

$$R(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}} L(x, y) = \int_{\mathcal{X} \times \mathcal{Y}} L(f(x), y) d\mathcal{D}(x, y)$$

Où L est la fonction de risque définie comme

$$L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$$

La fonction de risque usuellement considérée est l'erreur de classification:

$$\forall (x, y); L(f(x), y) = \mathbb{1}_{f(x) \neq y}$$

Où $\mathbb{1}_{\pi}$ est la fonction indicatrice.

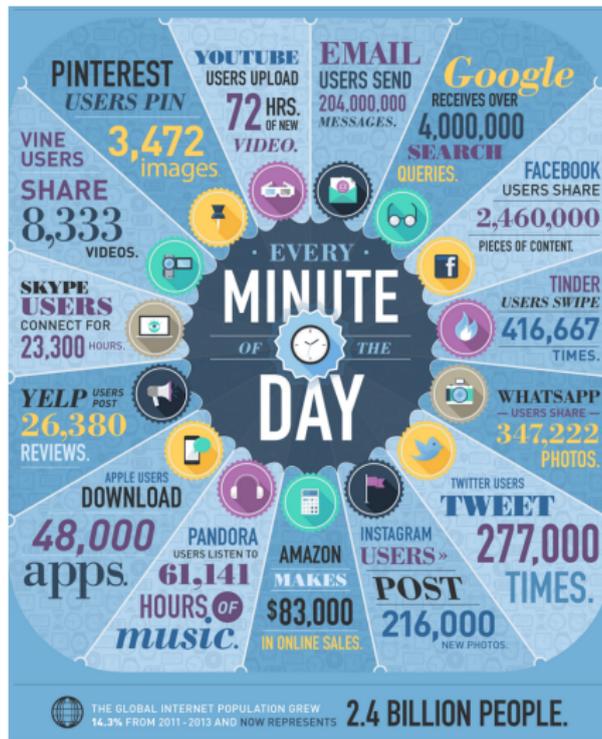
Principe de la minimisation du risque empirique

- ❑ Comme la distribution de probabilité \mathcal{D} est inconnue, la forme analytique du risque ne peut pas être calculée, la fonction de prédiction ne peut ainsi être calculée directement de $R(f)$.
- ❑ **Principe de la minimisation du risque empirique:**
Trouve f en minimisant l'estimateur non-biaisé de R sur une base d'apprentissage $S = (x_i, y_i)_{i=1}^m$:

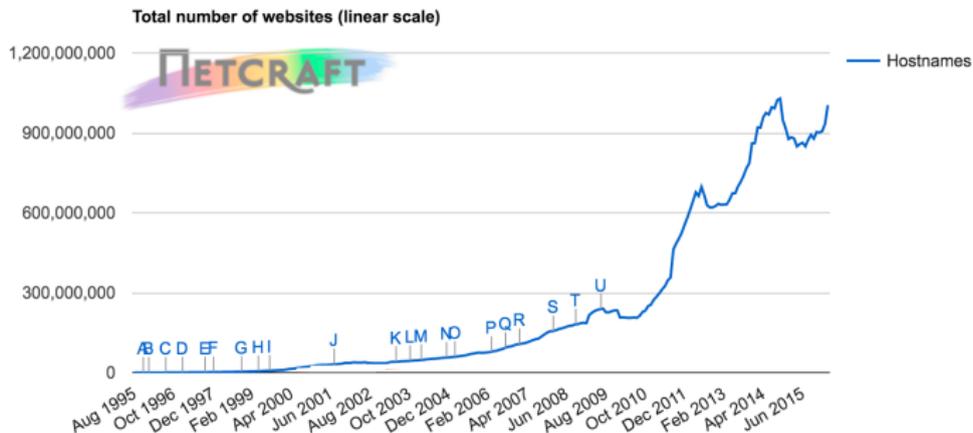
$$\hat{R}_m(f, S) = \frac{1}{m} \sum_{i=1}^m L(f(x_i), y_i)$$

- ❑ Au meilleur des cas, la complexité de la minimisation de $\hat{R}_m(f, S)$ est $O(d \times m)$, des techniques à base de gradient stochastique ont été proposées pour accélérer l'apprentissage.

Ère de Big Data dans le domaine de la RI



Ère de Big Data dans le domaine de la RI



- D'après les prévisions du projet EMC¹, 90% des données sur le web sont des données textuelles, et en 2020 il y aura 40 zetta octets (40×10^{21} octets) de données non-structurées.

¹<http://www.emc.com/leadership/digital-universe/index.htm>

Cadre classique de l'apprentissage machine

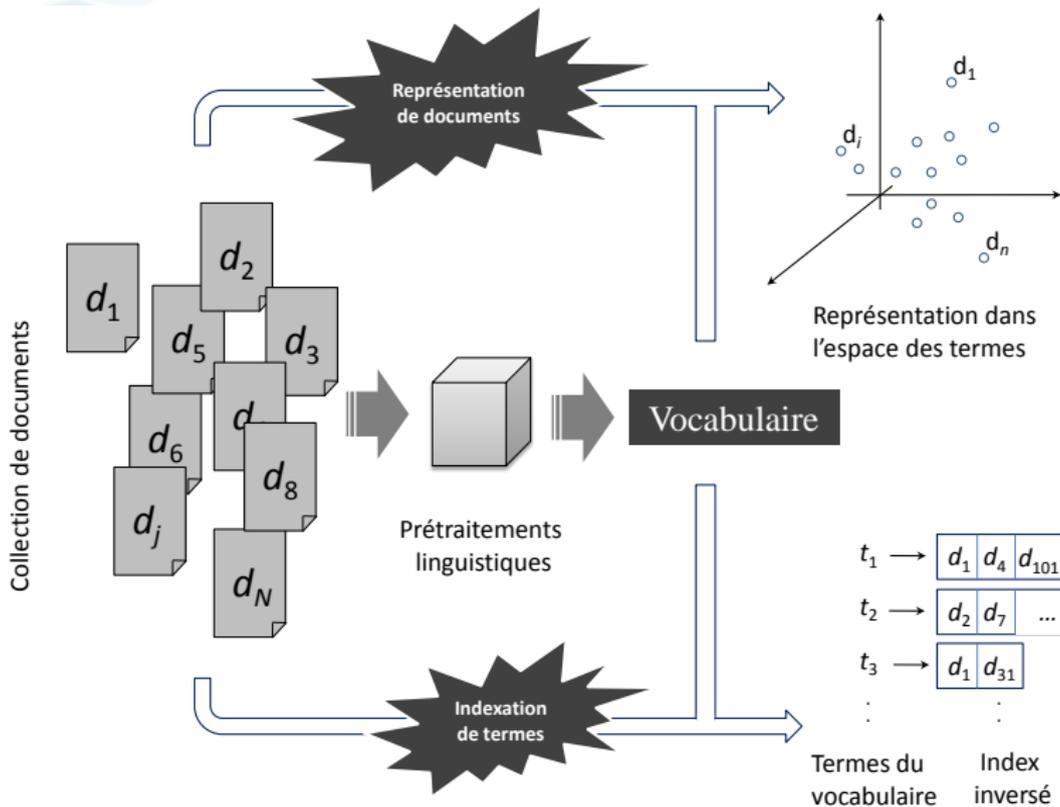
Nous considérons un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$ et un espace de sortie $\mathcal{Y} = \{1, \dots, K\}$.

Hypothèse: Les paires d'exemples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont *identiquement et indépendamment* distribuées (i.i.d) d'après une distribution de probabilité fixe \mathcal{D} .

Échantillon: Nous observons une séquence de m , **$m \gg 1$** , paires d'exemples (\mathbf{x}_i, y_i) générées i.i.d suivant \mathcal{D} .

Objectif: Construire une fonctions de prédiction $f : \mathcal{X} \rightarrow \mathcal{Y}$ qui prédit la sortie y d'un nouvel exemple \mathbf{x} avec la plus petite probabilité d'erreur.

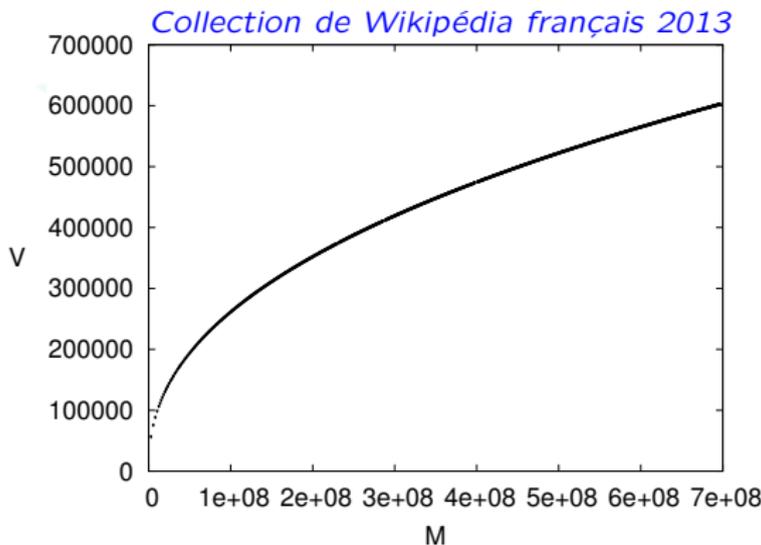
Indexation et représentation



Loi de Heaps (1978)

La loi de Heaps stipule que la taille du vocabulaire (V) a une croissance sous-linéaire en fonction du nombre d'occurrence des mots (M) ou de la taille du corpus.

$$V = K \times M^\beta$$



Cadre classique de l'apprentissage machine

Nous considérons un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$, $d \gg 1$,
et un espace de sortie $\mathcal{Y} = \{1, \dots, K\}$.

Hypothèse: Les paires d'exemples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont
identiquement et indépendamment distribuées (i.i.d)
d'après une distribution de probabilité fixe \mathcal{D} .

Échantillon: Nous observons une séquence de m ,
 $m \gg 1$, paires d'exemples (\mathbf{x}_i, y_i) générées i.i.d
suivant \mathcal{D} .

Objectif: Construire une fonctions de prédiction
 $f : \mathcal{X} \rightarrow \mathcal{Y}$ qui prédit la sortie y d'un nouvel exemple \mathbf{x}
avec la plus petite probabilité d'erreur.

Classification à large échelle

dmoz open directory project

In partnership with
Aol Search.

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

Search *advanced*

Arts

[Movies, Television, Music...](#)

Games

[Video Games, RPGs, Gambling...](#)

Kids and Teens

[Arts, School Time, Teen Life...](#)

Reference

[Maps, Education, Libraries...](#)

Shopping

[Clothing, Food, Gifts...](#)

World

[Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...](#)

Business

[Jobs, Real Estate, Investing...](#)

Health

[Fitness, Medicine, Alternative...](#)

News

[Media, Newspapers, Weather...](#)

Regional

[US, Canada, UK, Europe...](#)

Society

[People, Religion, Issues...](#)

Computers

[Internet, Software, Hardware...](#)

Home

[Family, Consumers, Cooking...](#)

Recreation

[Travel, Food, Outdoors, Humor...](#)

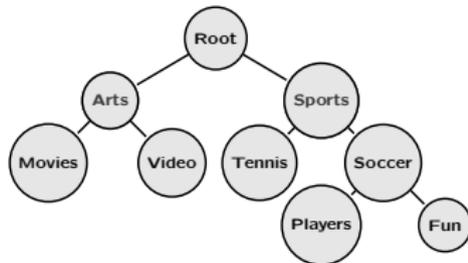
Science

[Biology, Psychology, Physics...](#)

Sports

[Baseball, Soccer, Basketball...](#)

- ❑ 5×10^6 sites
- ❑ 10^6 catégories
- ❑ 10^5 editeurs



Become an Editor Help build the largest human-edited directory of the web



Copyright © 2013 Netscape

5,292,731 sites - 99,941 editors - over 1,020,828 categories

Cadre classique de l'apprentissage machine

Nous considérons un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$, $d \gg 1$,
et un espace de sortie $\mathcal{Y} = \{1, \dots, K\}$, $K \gg 1$.

Hypothèse: Les paires d'exemples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont
identiquement et indépendamment distribuées (i.i.d)
d'après une distribution de probabilité fixe \mathcal{D} .

Échantillon: Nous observons une séquence de m ,
 $m \gg 1$, paires d'exemples (\mathbf{x}_i, y_i) générées i.i.d
suivant \mathcal{D} .

Objectif: Construire une fonctions de prédiction
 $f : \mathcal{X} \rightarrow \mathcal{Y}$ qui prédit la sortie y d'un nouvel exemple \mathbf{x}
avec la plus petite probabilité d'erreur.

Challenges

- ❑ Face à ces challenges de nouveaux paradigmes d'apprentissage sont en train de voir le jour.
- ❑ Les nouveaux algorithmes travaillent sur des bases distribuées à travers différentes machines et en prenant en compte les caractéristiques du réseau,
- ❑ Exemple: Collaboratif Filtering

$R = \text{User}$

	Item			
User/ Item	A	B	C	D
Ted		4		3
Carol	3		2	
Bob		5		2
Alice			4	

Exemple: Filtrage Collaboratif

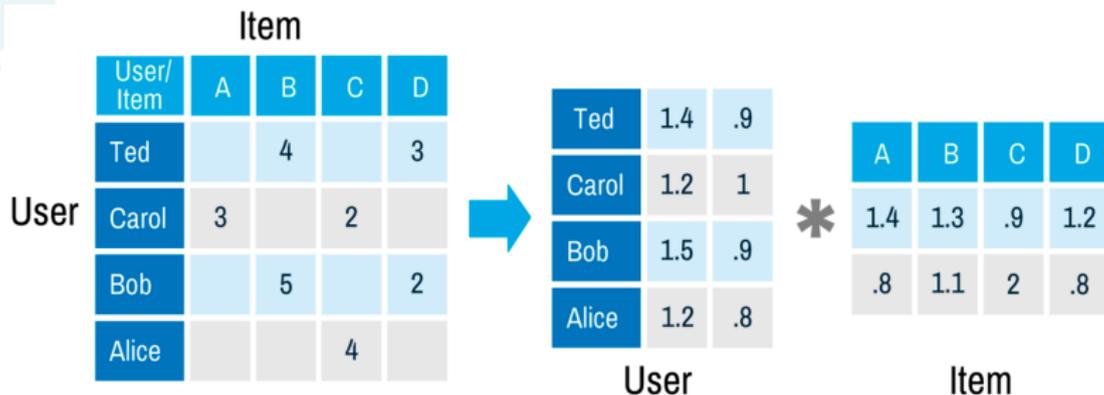
- Les corpus publics

Collection	$ \mathcal{U} $	$ \mathcal{I} $	sparsité
ML-100K	943	1682	93.7 %
ML-1M	6040	3952	95.8 %
ML-10M	71567	10681	98.7 %
NetFlix	480189	17770	98.8 %
Yahoo! Music	1000990	624961	99.6%

- Approche gagnante le prix NetFlix à base de la factorisation matricielle

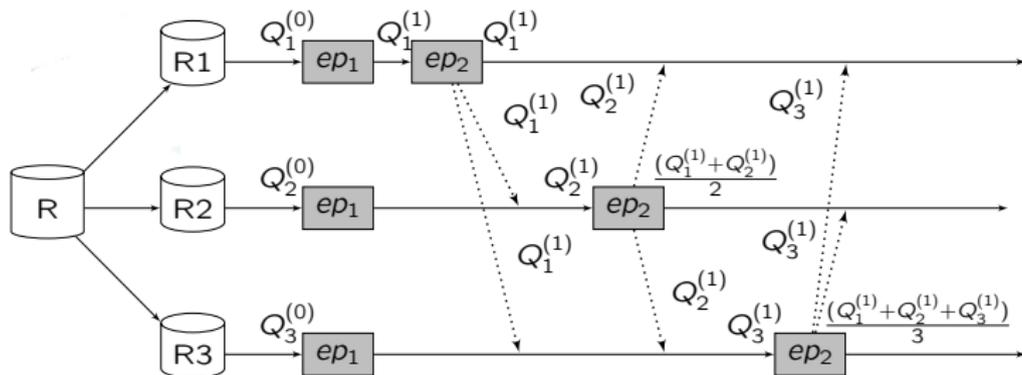
$$\min_{P,Q} \sum_{i,j:r_{ij}\text{existe}} \left(r_{ij} - q_j^\top p_i \right)^2 + \lambda (\|p_i\|^2 + \|q_j\|^2)$$

Filtrage Collaboratif

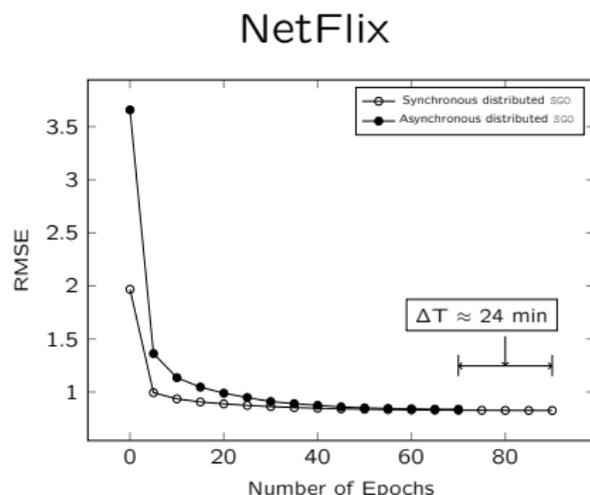
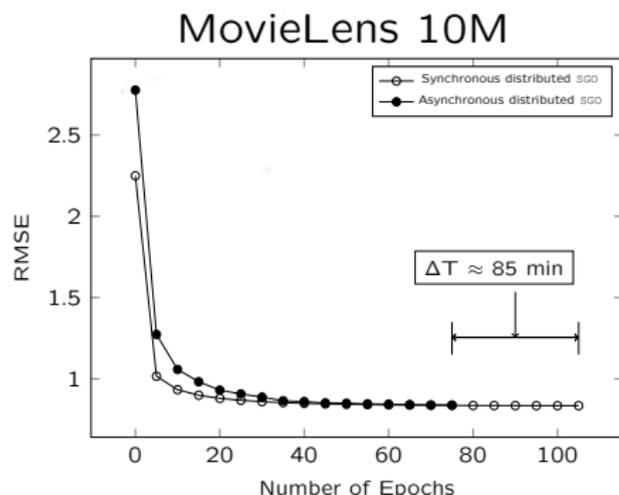


- Une façon de faire est de distribuer la matrice des scores R à travers les machines et de minimiser itérativement la fonction objectif avec la descente de gradient stochastique et puis moyenner les matrices P et Q une fois que les machines ont fait leurs mises à jour.

Factorisation Matricielle avec une approche asynchrone



Factorisation Matricielle avec une approche asynchrone: temps de convergence



Conclusion

- ❑ Le boom des données numériques se traduit par un accroissement fort des trois dimensions sous-jacents au cadre de l'apprentissage automatique
- ❑ Cela a amené de nouvelles problématiques en apprentissage automatique
- ❑ Les nouveaux algorithmes d'apprentissage sont de plus conçus en mode distribués et prennent en compte les caractéristiques du réseau pour atteindre de meilleures performances computationnelles.

Références

-  M.-R. Amini. **Apprentissage Machine: de la théorie à la pratique.** *Éditions Eyrolles*, 2015.
-  M.-R. Amini and É. Gaussier. **Recherche d'Information: applications, modèles et algorithmes.** *Éditions Eyrolles*, 2013.
-  I.H. Witten, A. Moffat and T.C. Bell. **Managing Gigabytes,** *Morgn Kaufmann Publishing*, 1999.