XGBoost
Origine et applications

Principe de la méthode

Mars
1er
2017

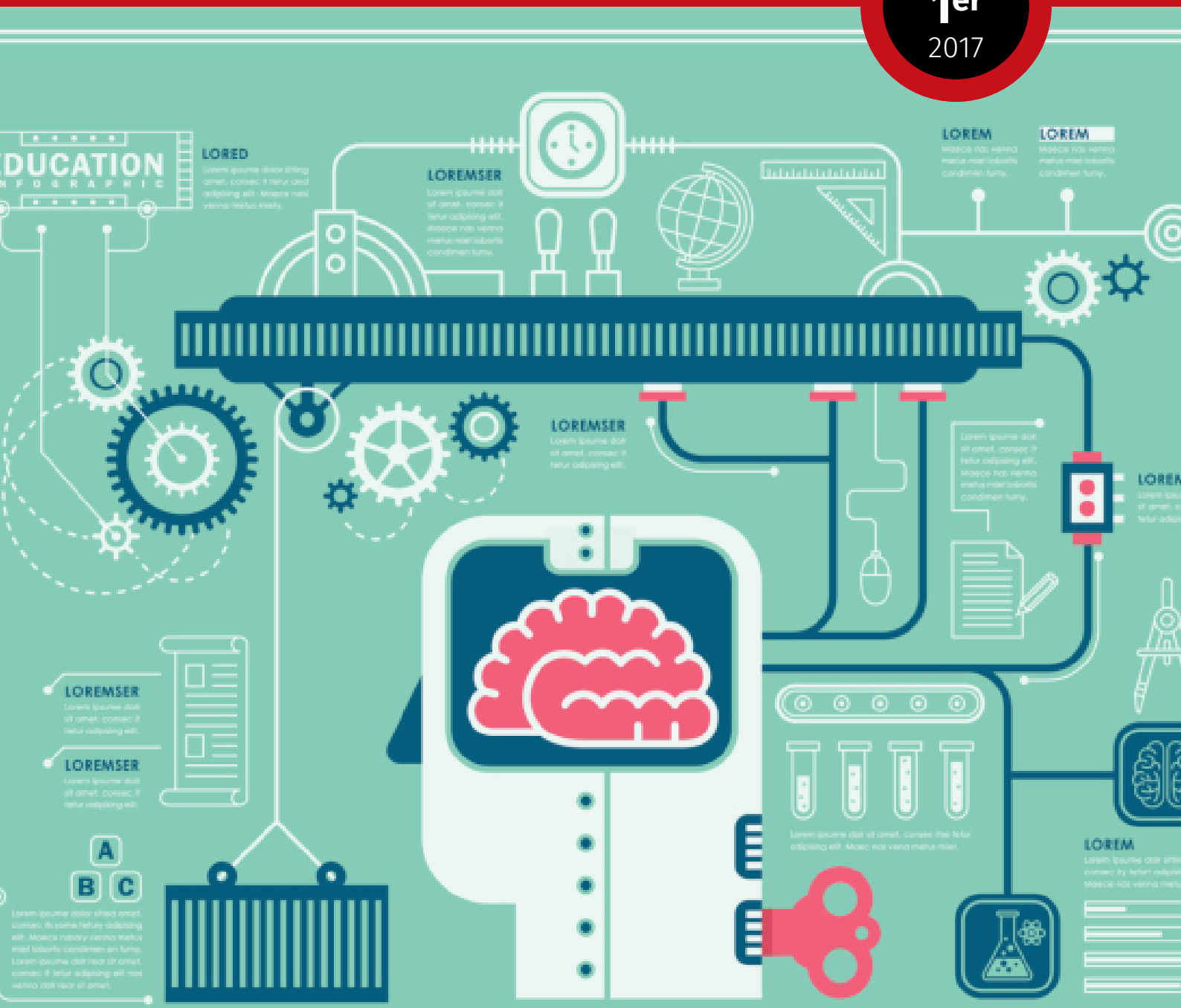

Table des matières

1. XGBoost en deux mots

L'objectif de cet article est de présenter XGBoost. Mais qu'est-ce que XGBoost ? Il s'agit d'une méthode de machine learning apparue il y a trois ans et dérivant des méthodes dites de boosting. Cette méthode est présentée par ses créateurs comme étant :

- Flexible Prise en compte de plusieurs thématiques de machine learning.
- Portable Utilisable sous toutes les plateformes (Windows, Linux, MAC)
- Multi-langages L'algorithme a été porté en Python, JAVA (Spark), C++,...
- Distribuée Depuis deux ans, l'algorithme est utilisable avec Hadoop et Spark.
- Perfomante Cet algorithme est donné pour être plus rapide que les algorithmes de sa famille et fournir de même meilleurs résultats.

Mais avant toute chose, que veut exactement dire l'acronyme « XGBoost » ?

E**X**treme **G**radiant **Boost**ing

Ainsi, la méthode XGBoost s'organise autours de trois points essentiels :

- *Boosting* Il s'agit d'une famille d'algorithme utilisés initiallement en apprentissage supervisé. Le principe du boosting sera détaillé en Section 2.
- *Gradient Boosting* Il s'agit d'une version du boosting dans laquelle l'objectif sera d'optimiser une fonctione faisant apparaître des gradients. Les détails de cette idée seront détaillées en Section **??**.
- *« Extreme »* Ce qualificatif signifie que la recherche de performances est poussée au maximum pour cette méthode, comme nous l'étudierons en Section 3.2.

Ainsi, nous allons dans un premier temps présenter les grandes lignes théoriques derrière cet algorithme avant de partir dans l'étude des implémentations et des applications de XGBoost.

2. Le boosting

2.1 • Premier algorithme

Le boosting est une méthode de Machine Learning apparue à la fin des années 1980 et ayant évoluée au fil du temps en plusieurs version, les principales étant AdaBoost ou encore les GBM (*Gradient Boosted Models*)[1].

Le succès de ces méthodes provient de leur manière originale d'utiliser des algorithmes déjà existant au sein d'une stratégie adaptative. Cette stratégie leur permet alors de convertir un ensemble de règles et modèles peu performant en les combinant pour obtenir de (très) bonnes prédictions. L'idée principale est en effet d'ajouter de nouveaux modèles au fur et à mesure, mais de réaliser ces ajouts en accord avec un critère donné. En ce sens, cette famille de méthodes se différencie des Random Forest qui vont elles miser sur l'aléatoire pour moyenner l'erreur.

[1] Historique du Boosting
1989
Premier algorithme de Boosting par R. SCHAPIRE
1996 Première implémentation d'AdaBoost par Y. FREUND et R. SCHAPIRE
1999 Apparition des modèles de boosting de gradient (GBM) par L. BREIMAN et J. FREIDMAN
2014 Implémentation et apparition de XGBoost par T. CHEN

Cet aspect fondamental du boosting permet ainsi une forte réduction du biais et de la variance de l'estimation, mais surtout garanti une convergence rapide. La contrepartie se fait au niveau de la sensibilité au bruit comme nous le verrons dans la description des algorithmes.