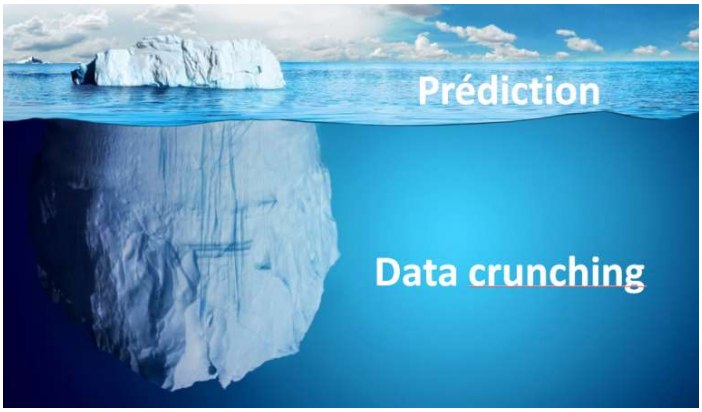
La préparation des données, enjeu majeur du machine learning

octobre 18, 2016

La préparation de données, enjeu trop souvent sous-estimé



Quand on se confronte pour la première fois à un projet de machine learning, on peut être tenté de croire que la problématique principale s'articule autour de l'algorithmie. S'il est vrai que cette partie doit être parfaitement maîtrisée pour permettre au prédictions d'atteindre des résultats d'une fiabilité suffisamment élevée pour être exploitables, le choix et le paramétrage des algorithmes ne représentent que peu souvent la problématique la plus complexe, ni la plus chronophage du projet (en considérant ici surtout les sujets de prédiction « classiques » qui nous sont soumis les plus fréquemment par nos clients et qui peuvent être traités en très large partie par les algorithmes existants, dans leur version originale ou adaptée par nos soins).

Dans la plupart des projets prédictifs que nous réalisons pour nos clients en nous appuyant sur les capacités sur la plateforme Verteego, la préparation des données, au sens large, constitue l'enjeu majeur et trop souvent négligé.

Principales étapes de la préparation des données dans un projet de machine learning

Voici les principales étapes de la préparation des données et les questions auxquelles il faudra répondre au cours d'un projet de machine learning :

- Identification des différentes sources de données accessibles et exploitables
  Où sont stockées les données que je peux utiliser ? Sont-elles dans un format immédiatement utilisable ? Sont-elles clusterisées (ex. architecture de type Hadoop) ? Peuvent-elles quitter le firewall ou l'architecture de machine learning doit-elle respecter le lieu de stockage ? Sont-elles cryptées, si oui par quelle technologie ?
- Définition d'un lieu d'entreposage
  Les données transformées et enrichies doivent-elle être entreposées ? Si oui, quel endroit servira d'entrepôt ? Stockage en base de données ou système de fichiers ? Stockage cryptée ou non ? Les données transformées doivent-elles alimenter des tableaux de bord opérationnels ?
- Nettoyage des données
  Quel est le degré de propreté de mes données (doublons, identification des types, données manquantes, format des dates et séries temporelles,…) ? Quelle technique utiliser pour nettoyer (fill down/up, dédoublonnage, clustering, supression des outliers, etc.) ?
- Feature generation
  Quelles sont les dimensions de données qui auront une utilité *a priori* dans mon analyse prédictive ? Comment

améliorer la qualité et la pertinence des différentes features (ex. suppression des éléments textuels à faible valeur ajoutée dans des messages SMS, ex. mots de liaison et modalisateurs) ?

- Feature extraction
  Quelles sont les dimensions de données susceptibles de renforcer l'expressivité de mon dataset à l'égard du degré de corrélation calculé par les algorithmes ?
- Construction du pipeline de données
  Quelles étapes sont nécessaires pour couvrir les différents besoins de transformation et enrichissement identifiés ? A quelle fréquence le flux de données sera-t-il actif (streaming, planifié, déclenché par des événements (ex. arrivée de nouvelles données dans un répertoire, réception d'un message POP3, étapes précédent du pipeline terminée,...)) ? Le pipeline doit-il être modifiable en cours d'utilisation (*on run*) ?

La plateforme Verteego met à disposition l'ensemble des outils nécessaires à la gestion efficace du flux de données, de manière à permettre aux équipes de paramétrage et analystes de passer un maximum de temps sur les tâches liées à la résolution des problématiques métiers.

Recent Posts from Rupert Schiessl

Remise du rapport sur la stratégie d'IA de la France au Président Hollande (http://www.verteego.com/fr/france-intelligence-artificielle/)
March 28, 2017

Verteego au lancement de #FranceIA (http://www.verteego.com/fr/franceia/)
January 24, 2017

Use Case HR – Reduce employee attrition and make talents stay longer (Part 2: Prediction) (http://www.verteego.com/hr-predict-employee-attrition-retain-talents/)
December 6, 2016

See more posts (http://www.verteego.com/fr/author/rupert/)

Categories: Big Data (http://www.verteego.com/fr/category/big-data-fr/), Innovation (http://www.verteego.com/fr/category/innovation-fr/), Technologie (http://www.verteego.com/fr/category/technologie-fr/)
Tags: Apache (http://www.verteego.com/tag/apache/), data flow (http://www.verteego.com/fr/tag/data-flow/), data pipeline (http://www.verteego.com/fr/tag/data-pipeline/), data preparation (http://www.verteego.com/fr/tag/data-preparation/), feature extraction (http://www.verteego.com/fr/tag/feature-extraction/), feature generation (http://www.verteego.com/fr/tag/feature-generation/), hadoop (http://www.verteego.com/fr/tag/hadoop/), hdfs (http://www.verteego.com/fr/tag/hdfs/), machine learning (http://www.verteego.com/fr/tag/machine-learning/), nifi (http://www.verteego.com/fr/tag/nifi/)

Leave a Reply

Votre adresse de messagerie ne sera pas publiée. Les champs obligatoires sont indiqués avec *

Start typing...

Nom *

Adresse de messagerie *

Site web

Post Comment