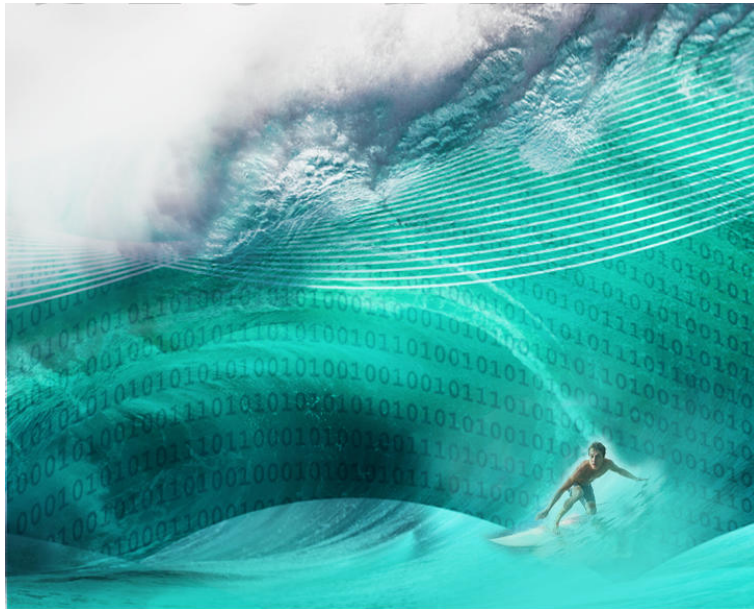


Certificat Science des Données – Module de Sensibilisation

Introduction à la Science des Données

PHILIPPE BESSE & BÉATRICE LAURENT

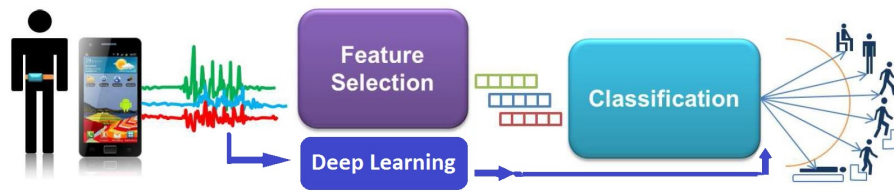
Université de Toulouse -- INSAT



Introduction

1 Comprendre Enjeux et Objectifs

1.1 Jeux de données



2 Découvrir la Science des Données

2.1 De la Statistique à la *Science des Données*

Avant d'entrer dans le vif du sujet, de savoir ce que peut la Science des Données, il est bon, en introduction, de rappeler d'où elle vient. La Science des Données est une appellation récente, datant de 2008, mais elle n'est pas née de rien. Cette pratique est la conséquence d'une longue évolution associée à la croissance des moyens de stockage des données et à celle de la puissance de calcul. Cette progression tant technologique que méthodologique et algorithmique accompagne une croissance exponentielle des masses de données analysées au sein d'une même étude : de façon indicative, un facteur 10 tous les 10 ans.

1930-70 h-Octets *Statistique inférentielle*. A partir des années 30 c'est le développement, suivant les travaux de Fisher, de la Statistique inférentielle. Une hypothèse est posée, une expérience planifiée, une statistique de test calculée, test de Student, ANOVA ; la statistique est comparée à une valeur seuil. L'hypothèse est acceptée ou rejetée avec un risque α contrôlé. Il s'agit, par exemple, de montrer qu'une molécule est significativement active, qu'une semence est significativement plus productive... De plus, la décision prise sur l'échantillon peut être inférée à la population entière. Ce cours ne s'intéresse pas à cette Statistique inférentielle et aux pratiques de test associées.

1970s kO Analyse des données et *exploratory data analysis*. Les années 70 connaissent une remise en cause de ces modèles statistiques très contraignants car basés sur des hypothèses de nature probabiliste souvent difficiles à vérifier voir même contredites. La plus grande diffusion des ordinateurs a permis une approche exploratoire multidimensionnelle de l'analyse des données, notamment avec l'Analyse en Composantes Principales (ACP), basée sur des considérations géométriques au lieu de probabilistes.

1980s MO IA, *Réseaux de neurones, Statistique fonctionnelle*. Les années 80 connaissent des développements plus méthodologiques avec la possibilité d'estimer des modèles statistiques non plus paramétriques mais fonctionnels, c'est-à-dire de grande dimension, et d'apprendre des réseaux de neurones relativement complexes. C'est l'apparition de l'algorithme de rétropropagation du gradient. L'intelligence artificielle de cette période abandonne alors les systèmes experts, ou approche symbolique, au profit, temporaire, des réseaux de neurones ou approche connexionniste.

1990s GO *Data mining et données pré-acquises*. Dans les années 90, la fouille de données, ou *data mining*, fait son apparition avec principalement des applications en marketing quantitatif dans le tertiaire : banque, assurance, vente par correspondance. Ces grandes entreprises disposent déjà de bases de données clients conséquentes à des fins comptables : des milliers de clients décrits par des dizaines de variables. L'objectif est de valoriser ces données en les utilisant pour améliorer la gestion de la relation client ou GRC ; scores d'appétences pour des campagnes publicitaires, risque de crédit.

Pour ce faire, des suites logicielles associant des requêtes d'extraction dans des bases de données, des outils exploratoires et de classification non supervisée, des modèles statistiques comme la régression logistique ou des arbres de décisions, les premiers algorithmes d'apprentissage supervisé comme les réseaux de neurones...

Toutes ces capacités de gestion, traitement et analyse des données intégrés dans une même suite logicielle devient le *data mining*.

En fait, deux choses sont nouvelles : l'intégration dans une même suite logicielle et surtout le fait que les données ne soient plus issues d'une planification expérimentale comme en Statistique inférentielle. Les données sont *préalables* à l'analyse, acquises pour d'autres finalités, par exemple comptables, et il faut faire avec. C'est, pour le statisticien, devenu un prospecteur de données, un premier *changement de paradigme* : ne plus pouvoir planifier l'expérience.

2000s TO *Apprentissage statistique, bioinformatique : $p \gg n$* . Le début du siècle a connu de profondes ruptures dans les biotechnologies à la suite du premier séquençage du génome. Pour chaque échantillon biologique, ce sont maintenant des milliers voire des millions d'informations qui sont observables. Des occurrences de millions de mutations possibles sur le génome, des expressions de dizaines de milliers de gènes, des expressions de protéines, de métabolites... En conséquence, les ensembles de données à étudier présentent beaucoup plus de colonnes, variables ou *features* que de lignes, échantillons ou *instances* ; p le nombre de variables est beaucoup plus grand que n la taille de l'échantillon. Ceci crée une situation d'indétermination qui a poussé à des développements méthodologiques et algorithmiques originaux : modèles et algorithmes parcimonieux ou *sparse*.

C'est un deuxième *changement de paradigme* pour le statisticien devenu

bioinformaticien.

A cette occasion apparaissent de nouveaux algorithmes dits d'apprentissage statistique (*statistical learning*) sous-ensemble de l'apprentissage automatique lui-même sous-ensemble de l'intelligence artificielle. Ce sont ces algorithmes d'apprentissage statistique : *boosting*, *support vector machine*, *random forest*, *extrem gradient boosting* et maintenant le *deep learning*, qui connaissent un grand succès et provoquent en retour le succès ou plutôt l'énorme battage médiatique de l'intelligence artificielle.

2010s PO *Grosses Data* p et n très grands, *deep learning*. Plus récemment, avec l'avènement des réseaux sociaux, le succès planétaire de Google et des autres GAFAM, le volume des données, c'est-à-dire, le nombre de clients, d'échantillons, et la taille des bases de données explosent. La taille de ces données ainsi que la puissance de calcul autorisent l'entraînement d'algorithmes excessivement complexes avec des millions de paramètres ou poids à estimer. Ceci conduit à des succès retentissants et médiatisés pour la reconnaissance d'images, la traduction automatique ou encore le jeu de go.

Le troisième *changement de paradigme* concerne à la fois statisticiens, informaticiens, et mathématiciens, devenus *data scientists*, car le traitement de ces données massives nécessite, certes de la puissance de calcul sur des systèmes de fichiers distribués comme *Hadoop*, mais aussi de nouvelles approches pour la résolution des problèmes d'optimisation afférents.

Cet historique met en exergue trois changements de paradigme méthodologique permis par des disruptions technologiques mais, fondamentalement, les nouvelles appellations *big data*, *data science*, *deep learning*, IA, tiennent plus du battage médiatique que de la nécessité d'identifier une *nouvelle science*. La science des données n'est pas une nouvelle science. Un *data scientist* c'est d'abord une équipe pluridisciplinaire associant compétences en Statistique, Informatique et Mathématiques. Mais, mêmes assemblées, toutes ces compétences ne peuvent

rien avec des données pourries.



Méfiez vous du battage médiatique et surtout des faux espoirs qu'il fait miroiter.

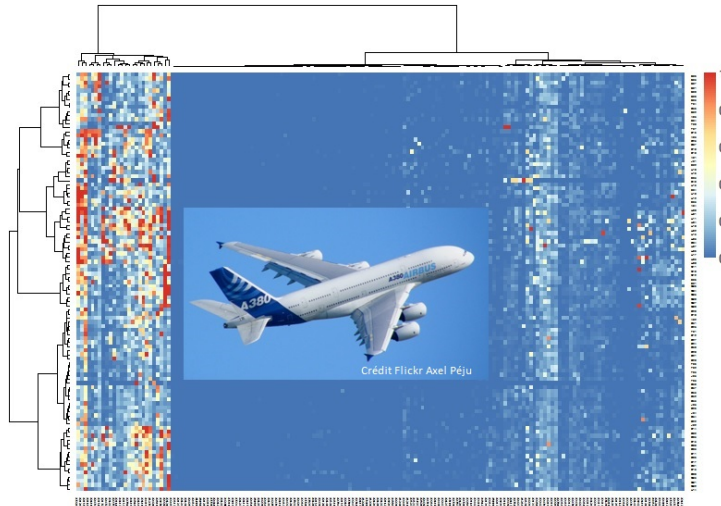
Data Scientist (n) : Person who is better at statistics than any software engineer and better at software than any statistician (J. Wills, Cloudera).

2.2 Algorithme et décision automatique

Voici quelques exemples illustratifs d'applications au quotidien de la science des données ou, plus précisément, d'algorithmes d'apprentissage automatique.

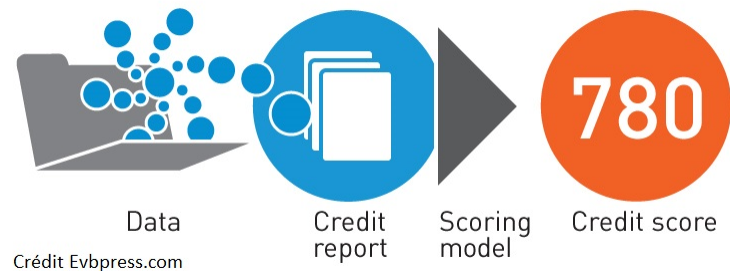
- Le choix d'un traitement médical, d'une action commerciale, d'une action de maintenance préventive, d'accorder ou non un crédit, de surveiller un individu... toutes les décisions qui en découlent sont la conséquence d'une prévision.
- La prévision du risque ou de la probabilité de diagnostic d'une maladie, le risque de rupture d'un contrat par un client, qui est le score d'attrition ou *churn*, le risque de défaillance d'un système mécanique, de défaut de paiement d'un client ou encore de radicalisation d'un individu... Les

exemples sont très nombreux et envahissent notre quotidien.



Le graphique ci-dessus est une visualisation avant analyse prédictive d'un recueil durant 6 mois de 700 000 messages d'incidents sur une flotte de 139 appareils. En ligne les appareils ordonnés par une classification ascendante hiérarchique, en colonne les types d'incidents également ordonnés. La fausse couleur illustre un taux d'incidents.

- Ces prévisions de risques ou scores, par exemple de crédit, sont produits par des algorithmes d'apprentissage supervisé, après entraînement sur une base de données où sont connus le comportement bancaire des clients et l'observation de bon remboursement ou non d'un emprunt. Le graphique ci-dessous illustre la démarche d'estimation puis prévision d'un risque de crédit à partir de l'enregistrement de l'historique des comportements des clients d'une banque.



3 Identifier les Étapes de la Science des Données

Pour introduire plus précisément la science des données, décrivons schématiquement, les deux principales phases d'une démarche d'apprentissage automatique supervisé.

3.1 Préparation des données ou *data munging*

La première phase constitue le prétraitement des données ou *data munging*, de l'extraction à l'exploration des données.

- Acquérir, archiver, extraire mettre en forme des données sont des tâches de *data management* en lien avec un gestionnaire de base de données incluant un langage de requête comme SQL ou *not only SQL* associé au gestionnaire de données distribuées *Hadoop*.
- Il est important ensuite de s'assurer de l'intégrité des données, de leur cohérence. Cette étape exploratoire avec des outils élémentaires est un préalable important pour détecter des valeurs atypiques, éventuellement des erreurs, gérer les données manquantes par suppression des observations ou imputation de ces valeurs.
Cette étape permet également d'analyser la structure des données ; leurs sources de variabilité et si possible de la représenter (ACP). Ces premiers résultats permettent d'apporter des réponses aux questions : faut-il transformer les données, calculer de nouvelles variables, caractéristiques ou *features* ?
- *Garbage in garbage out*. Même rudimentaires d'un point de vue méthodologique, ces étapes occupent la majeure partie du temps de l'analyse ; environ 80%. Enfin insistons lourdement : la qualité des données recueillies, leur représentativité par rapport à la question posée, sont fondamentales afin d'obtenir des prévisions robustes, généralisables à d'autres données, c'est-à-dire, d'un point de vue statistique, non biaisées et de faible variance. La qualité des données est essentielle ; aussi performants que soient les algorithmes d'apprentissage, ils ne peuvent rien avec des données pourries.
Cette étape nécessite, pour être efficace et pertinente, des compétences et savoir faire en Statistique exploratoire multidimensionnelle.

3.2 Apprentissage supervisé

Une fois qu'une base de données d'entraînement fiable et représentative a été constituée, la phase d'apprentissage proprement dite est généralement bien définie et suit schématiquement la structure suivante.

1. *Tirage aléatoire des échantillons d'apprentissage et de test*.
Commencer par construire, de façon aléatoire, deux sous échantillons, le premier d'apprentissage, le second de test. Le premier sert à estimer

les modèles ou faire apprendre les algorithmes, le deuxième n'est utilisé que pour en *prédire la variable cible* Y afin, comme les vraies valeurs ou les vraies classes de Y sont connues, de pouvoir estimer sans biais l'erreur de prévision. Il faut en effet distinguer l'erreur d'ajustement du modèle calculée sur l'échantillon d'apprentissage, erreur généralement biaisée car optimiste, de l'estimation de l'erreur de prévision calculée sur l'échantillon test, pas biaisée car ces observations n'ont pas participé à l'estimation des paramètres du modèle ou à l'entraînement de l'algorithme.

2. **Pour** chaque modèle ou algorithme considéré :
 Itérer ensuite les étapes suivantes pour chaque famille d'algorithme. Citons quelques possibilités : régression, SVM, réseaux de neurones, *boosting*, *random forest*, des plus utilisés parmi une grande farandole d'algorithmes disponibles.
 - (a) Première étape : estimation du modèle en fonction des valeurs de certains hyper-paramètres qui contrôlent la complexité du modèle, à savoir sa flexibilité ou capacité à s'ajuster finement aux données ;
 - (b) Deuxième étape : optimiser, généralement en minimisant une estimation de l'erreur de prévision par validation croisée, ce ou ces hyper-paramètres : le nombre de variables dans un modèle, le nombre de feuilles dans un arbre, le nombre de neurones d'un réseau, la pénalisation des SVM...
 - (c) Troisième étape : calculer la prévision de l'échantillon test pour le modèle "optimal" courant afin,
 - (d) Quatrième étape, d'estimer l'erreur de prévision.
3. On obtient ainsi une erreur par méthode d'apprentissage considérée ; il suffit alors de choisir la meilleure méthode ou le meilleur algorithme : celui qui minimise l'erreur de prévision mais éventuellement en tenant compte de la facilité d'interpréter le modèle. C'est généralement le cas des modèles de régression ou d'arbres de décision, cela ne l'est plus pour les SVM, neurones et autres *boosting* ou *random forest*.
Attention, ce cours consacré à un cas d'usage ne propose pas une étude exhaustive de l'ensemble des algorithmes d'apprentissage ; seuls ceux nécessaires à la réalisation des objectifs sont brièvement décrits. Une initiation systématique aux autres algorithmes est proposée dans les tutoriels du dépôt github.com/wikistat.

Contenu du cours

Référence

Anguita D., Ghio A., Oneto L., Parra X., Reyes-Ortiz J.L. (2013). A Public Domain Dataset for Human Activity Recognition Using Smartphones,

*ESANN 2013 proceedings, European Symposium on Artificial Neural Networks,
Computational Intelligence and Machine Learning, 437-442.*