

2

Actuariat - données et assurance

- La révolution numérique : nouvelles données en assurance
- Multi-compétences
- Impact du Big Data sur le secteur assurantiel
- Le contexte de l'actuariat
- Informations diverses

DEFIS STATISTIQUES

L'arrivée des Big Data a permis la découverte pour le “grand public” de méthodes statistiques fondées sur l'**apprentissage statistique** (Machine Learning quand appliqué en pratique).

Mais il faut garder en tête que

- ❶ il ne faut pas créer une usine à gaz...
- ❷ un modèle statistique est d'autant + robuste qu'il est simple,
- ❸ ces méthodes ne sont pas encore parfaitement adaptées à la gestion de tt type de données.

⇒ Beaucoup de travail préalable à faire avant un emploi judicieux !

BIG DATA, QU'ES A QUO ?

Big Data, définition simplifiée : données non traitable en une passe et dans un temps raisonnable sur une station de travail.

Deux époques :

< 2005, ordinateurs 32-bit. Taille $n > 10^7$, $p > 100 = 8\text{Go}$.

> 2005, ordinateurs 64-bit : bc + de mémoire physique, mais unités de calcul limitées.

Deux motivations principales d'utilisation : description, prévision.

Deux aspects : spatial (volume) et temporel (flux).

CARACTERISATION DES BIG DATA

On a coutume de parler de Big Data lorsqu'on dispose de données...

- en grand **volume** (énorme base de données),
- en grande **variété** (numérique, texte, images, vidéos, ...),
- en grande **vitesse** (fréquence d'arrivée de l'information, évolution des données).

Règle des 3V...qui doit déboucher sur la **création de "V"aleur** de par l'exploitation de ces données.

DEFIS PRATIQUES LIES AUX BIG DATA

- **Défi opérationnel**, essentiellement informatique :
 - système d'information, architecture, capacité de stockage...
 - calculs distribués (MapReduce) ⇒ Hadoop, Spark, ... ;
- Une réflexion sur la **donnée** :
 - qualité de la donnée et gestion de son aspect non-structuré : comment homogénéiser des formats différents à l'origine ?
 - sélection en fonction de sa pertinence, gestion,
 - visualisation : SQL (Structured Query Language), noSQL..
- Un enjeu **éthique** : anonymisation principalement (tests génétiques en assurance maladie,...) ⇒ réglementation RGPD.

D'OU VIENNENT LES NOUVELLES DONNEES ?

Essentiellement de **données externes**... Les assureurs possèdent déjà des données internes (peu exploitées, $\approx 20\%$), et accèdent maintenant à d'autres sources riches en information :

- 1 **Objets connectés** : télématique, Apple Watch, ...
- 2 Réseaux sociaux et **navigation internet** : pouvoir de nuisance des consommateurs ;
- 3 Assurance de **biens partagés** : AirBnB, AutoLib', ...
- 4 **L'Open Data** : crawling, scrapping... (Datagouv, ...).

C'est l'intégration au sein d'un même SI qui est très compliqué.

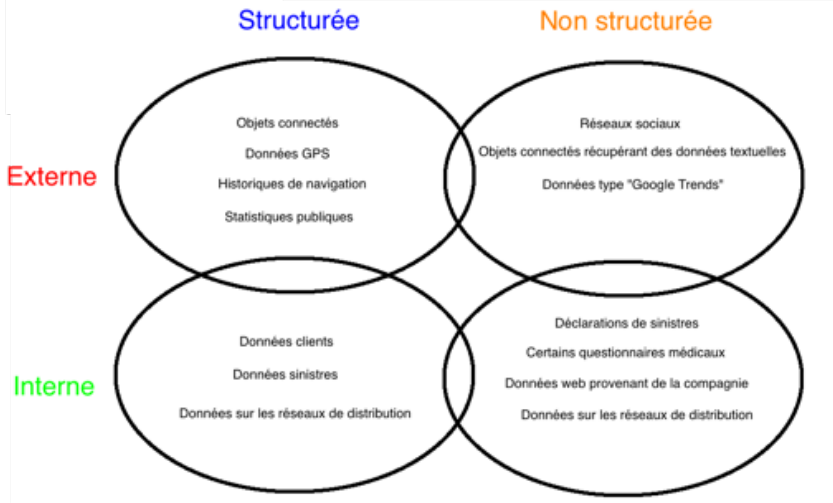
QUESTIONS ESSENTIELLES

Ces nouvelles données posent des questions fondamentales quant à leur utilisation, notamment

- **Fiabilité** des données
→ s'assurer auprès des services ayant fourni les données de leur fiabilité, de leur authenticité ;
- **Cohérence**
→ s'assurer du contenu de ces données ;
- **Sécurité**
→ cyber-risque, ...

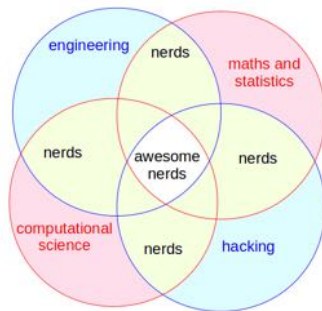
⇒ Risque opérationnel également accru !

CLASSIFICATION DES DONNEES



DATA SCIENTIST ET DATAVIZ

“statistics is the grammar of data science. It is crucial to making data speak coherently. But it takes statistics to know whether this difference is significant, or just a random fluctuation. (...) What differentiates data science from statistics is that data science is a holistic approach. We’re increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.” Mike Loukides, 2010
radar.oreilly.com



Source : blog d'Arthur Charpentier.

Idée : le data scientist ne se limite pas à la statistique, il cherche à faire parler ses données en général... (data visualisation)

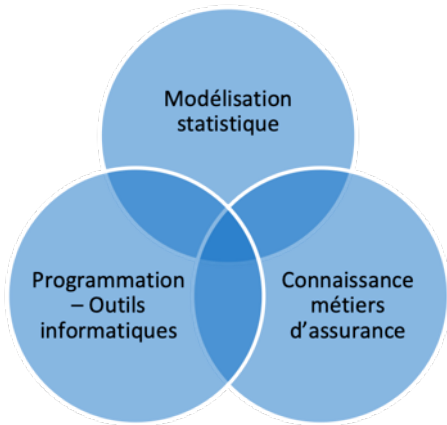
2

Actuariat - données et assurance

- La révolution numérique : nouvelles données en assurance
- **Multi-compétences**
- Impact du Big Data sur le secteur assurantiel
- Le contexte de l'actuariat
- Informations diverses

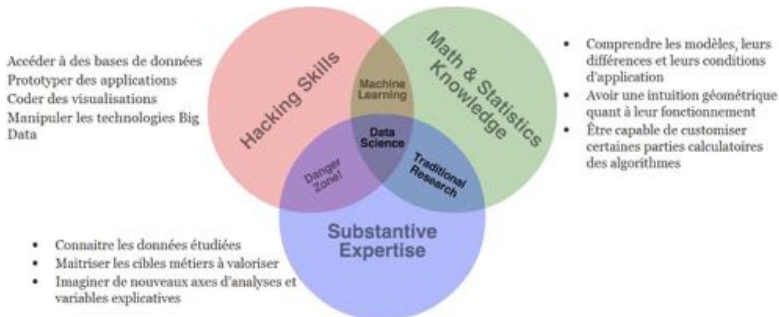
COMPETENCES DE L'ACTUAIRE TRADITIONNEL

→ Une connaissance assez approfondie de concepts variés...



COMPETENCES DU DATA SCIENTIST

→ Une connaissance assez approfondie de concepts variés...



Nécessité de bien maîtriser les outils informatiques (gestion de base de données, calcul parallèle, ...).

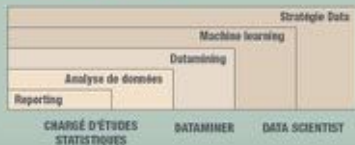
DATA SCIENTIST



UN CHEF DE PROJET... EN INTERACTION AVEC TOUS LES MÉTIERS DE L'ENTREPRISE



LES DIFFÉRENTS MÉTIERS DE LA STATISTIQUE



LE CONCEPT MAP-REDUCE POUR AGRÉGER

Un algorithme est dit **échelonnable** si le temps de calcul est divisé par le nombre de processeurs (nœuds) utilisés \Rightarrow permet aux applications de travailler avec des milliers de nœuds.

→ Le principe est de répartir les tâches parallèles (Map) puis d'intégrer (Reduce) tous les résultats obtenus.

Exemple : chaque nœud calcule la moyenne d'une variable avant de calculer la moyenne des moyennes.

Hadoop : projet destiné à faciliter la création d'applications distribuées et échelonnables (scalable).

Rq : toute méthode statistique ou d'apprentissage n'est pas scalable...

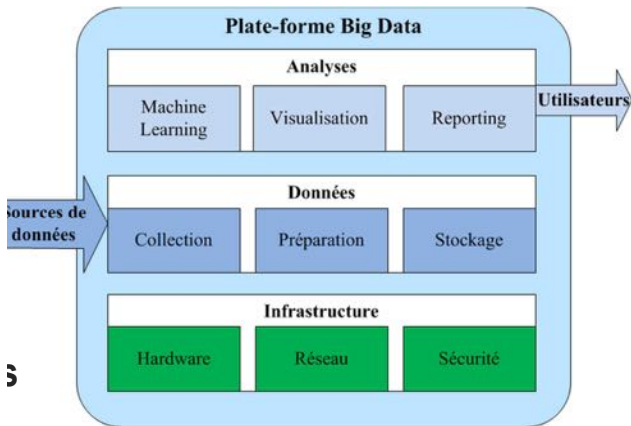
→ Les requêtes complexes comme celles de SQL (Structured Query Language...“Select From Where”) sont impossibles...

⇒ NoSQL (not only SQL) :

- Cassandra,
- MongoDB,
- Voldemort...

Le Data Scientist doit donc s'initier à une interface d'accès aux architectures Hadoop, NoSQL... Voir les outils Mahout ou RHadoop par exemple.

ELEMENTS D'UNE PLATEFORME BIG-DATA



Source : F. Soulié-Fogelman, leçon inaugurale formation DS IA.

2

Actuariat - données et assurance

- La révolution numérique : nouvelles données en assurance
- Multi-compétences
- Impact du Big Data sur le secteur assurantiel
- Le contexte de l'actuariat
- Informations diverses

APPORT PRINCIPAL DE CES NOUVELLES DONNEES EN ASSURANCE

Un des gros problèmes de l'assureur (par rapport au banquier) est la **faible fréquence de ses interactions avec l'assuré...**

En effet, ils ne se voient en général que 2 fois en tout pour tout :

→ Une fois à la souscription ;

→ Une fois lors du sinistre s'il a lieu.

⇒ Très difficile pour l'assureur de bien connaître l'assuré !

Technologies liées au Big Data vont augmenter significativement la fréquence de ces interactions...et **atténuer les particularités** de l'assurance : **antisélection et aléa moral**.

(en plus de l'inversion du cycle de production !)

IMPACT SUR LA CHAÎNE DE VALEUR

Le Big Data a un impact à plusieurs niveaux pour un assureur, parmi ses tâches “historiques” impactées :

- segmentation, tarification (Pay-As-You-Drive, HomeBox),
- provisionnement : micro-level reserving,
- détection de fraude (par géolocalisation par exemple),
- ciblage marketing (compréhension des comportements),
- scoring d'assurés : la construction d'un bon score reste issue d'une approche stat. **couplée à une connaissance métier.**

Remarque : échelle de temps de l'assurance parfois bc plus longue que dans d'autres secteurs (attention aux dérives du risque).

LE BIG DATA, JUSQU'OU ?

La base de l'assurance est la mutualisation...

...Or l'enjeu principal du Big Data est de mieux comprendre les mécanismes à l'échelle de l'individu !

“We are moving from an era of private data and public analyses to one of public data and private analyses” (Andrew Gelman)

Il y a donc un risque énorme (surtout en tarification), qui est...

...la PERTE de MUTUALISATION.

Où s'arrêtera la segmentation... ?

2

Actuariat - données et assurance

- La révolution numérique : nouvelles données en assurance
- Multi-compétences
- Impact du Big Data sur le secteur assurantiel
- **Le contexte de l'actuariat**
- Informations diverses

ETAPES D'UN PROJET BIG DATA

Voici les étapes classiques d'un projet Big Data :

- ① collecte de données,
- ② préparation des données,
- ③ **feature engineering**,
- ④ construction de modèle (via des approches ML),
- ⑤ évaluation (via un test),
- ⑥ déploiement.

LA DATA ET L'ACTUAIRE

- Matière première de l'actuaire = données ;
→ Fonction-clef responsable des données ! Responsable du rapport actuariel (annuel interne mais divulgable à l'ACPR : travaux effectués, déficiences et recommandations)
- “Feature engineering” = idée clef : construction de nouvelles covariables, + explicatives que celles d'origine, à partir de la **connaissance métier** notamment (smart data).
- Point d'attention majeur : la responsabilité de l'actuaire ne s'arrête pas lorsque l'algorithme rend son résultat.

FEATURE ENGINEERING

A quoi correspond le **feature engineering** ?

En pratique, il s'agit d'augmenter la variété des données d'entrée :

- calculer de nouvelles variables à partir de variables existantes :
 - significatives pour le métier,
 - exemples : ratios, agrégat sur fenêtre glissante temporelle, agrégat géographique...

⇒ Les nouvelles variables créées sont forcément corrélées aux anciennes : l'algorithme doit être insensible à cet effet !

- obtenir des variables de sources externes,
- alimenter avec des sources de type \neq (texte, image, ...).

C'est le facteur de succès le plus important !

CONTEXTE REGLEMENTAIRE

Le contexte réglementaire rend plus que jamais nécessaire de prendre le virage du Big Data...

- **Amendement Bourquin et Loi Hamon** : rend l'assuré plus volatile...
 - nécessite une plus grande réactivité pour comprendre ses besoins et le conserver (ou l'attirer) en portefeuille,
 - difficulté : contacts entre assureur/assuré peu nombreux.
- **Gender directive** :
 - interdiction de certains critères jugés discriminants (ici le sexe de l'assuré) pour tarifier (pas pour le provisionnement)
 - nécessité de trouver d'autres critères plus complexes qui déterminent le niveau de risque.

VIE D'UN PRODUIT D'ASSURANCE

Nouvelles technologies, nouvelles données : impact sur la vie d'un produit... Chronologiquement,

① La **souscription** :

- notamment la souscription en ligne :
- améliorer le taux de transformation...
- ... mais sans perdre de vue la “valeur-client”
- nécessite de prendre en compte les contraintes de la souscription en ligne.

② La **tarification** :

- vers une individualisation de la prime...
- ... qui nécessite de tenir compte des incertitudes
- sans porter atteinte au principe de mutualisation.

- ③ Les **résiliations**, les rachats : anticiper le comportement de l'assuré.
- ④ La **prévention** :
 - intervenir pour empêcher le sinistre de se produire
 - changement de la relation avec l'assuré
- ⑤ Le **provisionnement** :
 - amélioration de l'évaluation des engagements pris
 - atterrissage pour l'estimation des montants de sinistres graves

Deux notions-clefs à retenir :

- attention à ne pas trop segmenter pour la tarification,
- impact surtout pour la connaissance et le suivi des comportements assurés.

EXEMPLES D'EVOLUTION EN ASSURANCE

De gauche à droite un exemple et son évolution grâce aux nouvelles techno...

Taxi (G7, taxi alpha, etc.)	VTC (Uber, chauffeurs privés, the Cab, etc.) et autolib	Voitures autopilotées (ex : Google car)	Que devient l'assurance auto ?
Guichets, courtiers d'assurance	Comparateur d'assurance en ligne	Utilisation des smartphone et objets connectés	Comment acquérir les nouveaux clients et fidéliser les anciens ?
Vidéo surveillance avec de gros appareils	Capteur de présence et caméra wifi	Caméra connectée au smartphone	Assurance MRH connectée
Balance mécanique	Balance électronique	Balance connectée	Assurance santé connectée
Ordinateurs centraux	Ordinateurs personnels et portables	Ordinateurs connectés aux serveurs distants	Fonctionnement des algorithmes à distance
Appareil photo argentique	Appareil photo numérique	Téléphone mobile avec appareil photo	E-constat

REVOLUTION DU METIER D'ASSUREUR ?

Assurance traditionnelle	Assurance Big Data
Gestion des risques	Gestion des données
Gestion des sinistres	Prévention
Modélisation de la sinistralité (loi et quantile)	Analyse des données (Algorithme d'apprentissage)
Assurance des stocks (ex :contrat de durée annuelle)	Assurance des flux (ex: Pay As You Drive)
Asymétrie d'information (l'assuré sait plus)	Asymétrie d'information (l'assureur sait plus)
Données clients avec la déclaration	Données clients collectées, achetées ou open data
Méthodologie en partant hypothèses de modélisation	Méthodologie en partant des données
Mutualisation	Sélection des risques à outrance

2

Actuariat - données et assurance

- La révolution numérique : nouvelles données en assurance
- Multi-compétences
- Impact du Big Data sur le secteur assurantiel
- Le contexte de l'actuariat
- Informations diverses

APARTÉ SUR LA BLOCKCHAIN

Qu'est-ce que la blockchain ?

Un protocole sécurisé et décentralisé de gestion de transaction.

Idée principale : “décentraliser le pouvoir”.

Principe de fonctionnement : chaîne de blocs pour authentifier des transactions (validité de la transaction évaluée par les utilisateurs eux-même, qui forment les maillons de la chaîne).

Toute information faisant intervenir un REGISTRE (ex : transac. bancaire) peut être mise dans une blockchain pour authentification.

Attention : la blockchain ne valide pas l'information elle-même, elle en valide la possibilité d'effectuer la transaction !

PLUS LOIN DANS LA THEORIE / PRATIQUE

Sur le tas :

- Les MOOC : FUN, COURSERA...
- Concours “Kaggle” : s’autoformer par des challenges sur de vraies données et des problèmes posés par les entreprises... Voir aussi le site datascientest.com.
- Un peu de Python (langage de référence aujourd’hui en Data Science) : voir la librairie SciKit learn par exemple...

Des formations en Data Science : Institut des Actuaire (DSA), Telecom ParisTech, ...

GROUPES DE TRAVAIL A L'IA

Parmi les forces vives qui travaillent sur ces sujets, les GT de l'Institut des Actuaires (ouverts aux non-actuaires !) :

- ① SGT1 : Norme et éthique
- ② SGT2 : Nouveaux modèles de mutualisation
- ③ SGT3 : Impact sur le métier d'actuaire
- ④ SGT4 : Algorithmes prédictifs de comportements
- ⑤ SGT5 : Risques opérationnels et pistes d'audit
- ⑥ SGT6 : Big Data à l'étranger
- ⑦ SGT7 : Assurance connectée

Plateforme : <https://actuairesbigdata.wordpress.com>

LIBRAIRIES R (non exhaustif !) ET JEU

Voici 10 packages R utilisés fréquemment pour les applications :

- pour modéliser des interactions complexes : gbm, randomForest, E1071
- pour gérer des variables explicatives avec beaucoup de modalités ou des données textuelles : glmnet, Tau
- pour accélérer le code (paralléliser, ...) : Matrix, SOAR, foreach / doMC, data.table

Et un [exemple de Machine Learning par un jeu](#) : essayez de vous connecter à Akinator... et choisissez un personnage !

SITES INTERNET INTERESSANT

cf mon navigateur internet...

3 Philosophie de l'apprentissage statistique

- Principes généraux
- Théorie de Vapnik et surapprentissage
- Agrégation d'estimateurs
- Comment analyser les résultats ?
- Rappels sur le bootstrap
- Agrégation : cas des variables catégorielles

MACHINE LEARNING, NOUVELLE APPROCHE

- Abandon d'une approche de "modélisation" pour 1 approche qui cherche à laisser parler les données ("**data-driven**"), typique du monde non-paramétrique.
- Big Data : pour rendre compte d'une réalité complexe, on s'autorise des modèles – simples, voire peu intelligibles.

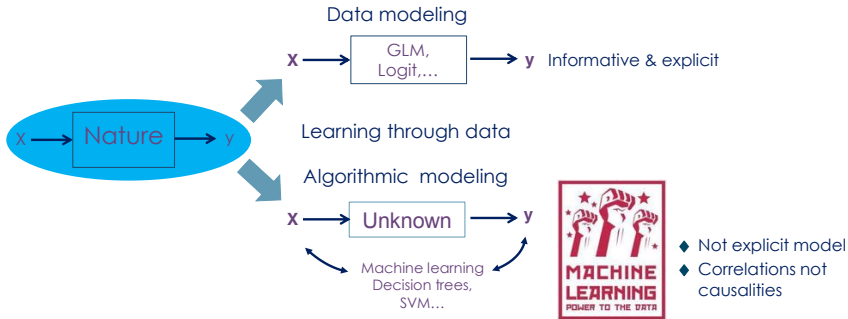
⇒ Une logique de **prévision domine**, plus qu'une logique d'analyse et d'explication des phénomènes.

RAPPEL : NON PARAMÉTRIQUE, PARAMÉTRIQUE

- Estimation **paramétrique** : on cherche m parmi une famille indexée par un paramètre de dimension finie.
→ *Exemple* : régression linéaire, $m(x) = a + bx$.
Une fonction candidate s'identifie à 2 paramètres (a, b) .
- Estimation **non paramétrique** : on ne fait plus d'hypothèse (ou bc –), on cherche $m(x)$ parmi ttes les fonctions possibles (dim. infinie).

Exemple connu d'estimateurs **non paramétriques** : estimateurs à noyaux,

ILLUSTRATION



PROBLEME SUPERVISÉ VS NON SUPERVISÉ

Deux types de pb : présence ou non d'une variable à expliquer Y qui a été, conjointement avec X , observée sur les mêmes objets.

Paradigme du cas supervisé : apprendre à généraliser à partir d'exemples du phénomène observé.

S'applique

- **à la régression** : cas où la réponse est continue ;
- **à la classification** : cas où la réponse est catégorielle.

Cas non supervisé : n'observe pas la valeur de la variable d'intérêt (ex. modèles mélange : classer les indiv. dans les composantes \Rightarrow on ne connaît pas leur composante d'appartenance)

EN PRATIQUE...

Dans le 1er cas (supervisé) \Rightarrow trouver une fonction f susceptible, au mieux selon un critère à définir, de reproduire Y ayant observé X :

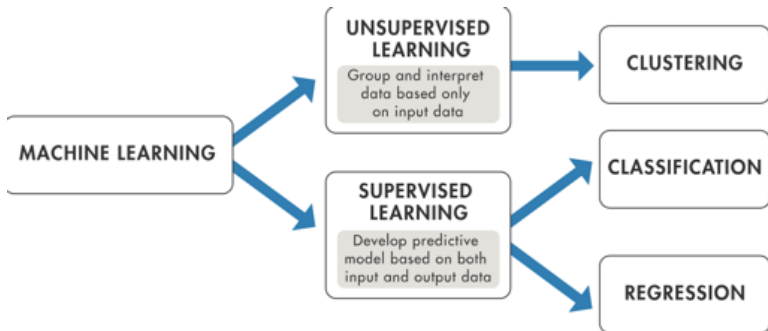
$$Y = f(X) + \epsilon$$

où ϵ symbolise le bruit ou erreur de mesure.

Dans le cas contraire (absence d' Y) \Rightarrow non-supervisé.

Objectif : recherche d'une typologie/taxinomie des observations...
Comment regrouper celles-ci en classes homogènes mais les + dissemblables entre elles \rightarrow pb de **clustering**.

SCHEMA RECAPITULATIF ET METHODES ASSOCIEES



STATISTIQUE CLASSIQUE VS APPRENTISSAGE

- **Statistique classique** : recherche le **modèle génératif** des données. Construit l'estimateur sur 1 jeu de données unique. Une **théorie asymptotique** permet de juger sa qualité (IC,...).
- **Apprentissage stat.** : recherche de **bonnes prévisions**...
 - on ne cherche pas le modèle qui génère les données !
 - les exemples du phénomène observé sont représentés par l'échantillon d'appren. : on souhaite faire apprendre à l'algo. la relation entre X et Y , puis la généraliser (prévision de Y) à des occurrences de X pour lesquelles Y inconnue.
 - la **qualité n'est plus jugée via des critères asymptotiques**, mais à l'aune d'une mesure d'adéquation à l'échantillon test.

AUTREMENT DIT...

Statistique classique : approches privilégiant la [compréhension](#) !

- Permet une compréhension du mécanisme générateur des données, avec une représentation si possible parcimonieuse ;
- Le modèle doit être “simple” et interprétable (odd-ratio, ...)

Machine learning : approches privilégiant la [prévision](#) !

- pour de nouveaux individus : pouvoir de généralisation,
- les modèles sont en fait des algorithmes.

“Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data”,
([Breiman, 2001](#))

“Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms”, ([Vapnik, 2006](#))

3 Philosophie de l'apprentissage statistique

- Principes généraux
- Théorie de Vapnik et surapprentissage
- Agrégation d'estimateurs
- Comment analyser les résultats ?
- Rappels sur le bootstrap
- Agrégation : cas des variables catégorielles

QUALITE D'ESTIMATION ET GRANDE DIMENSION MONDE NON PARAMETRIQUE

Théorème : soit $X \in \mathbb{R}^d$, et m une fonction k fois dérivable à dérivées bornées. La **vitesse optimale de convergence d'un estimateur non paramétrique** \hat{m} est

$$\hat{m}(x) - m(x) = O(n^{-k/(2k+d)}) \quad p.s.$$

- Si la fonction m est régulière (par ex. infiniment dérivable) à d fixé, la vitesse de convergence est en \sqrt{n} .
- Si d est “grand” par rapport à n , la performance d'estimation est considérablement dégradée.

ÉCHANTILLONS ET POUVOIR DE GENERALISATION

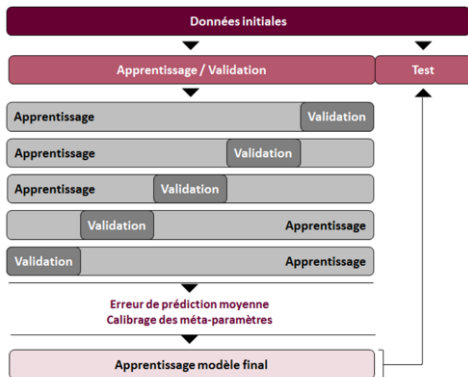
Les méthodes d'apprentissage statistique induisent le choix de **paramètres de tuning** (param. "utilisateur")... Ils jouent un rôle important dans le pouvoir de généralisation du modèle.

Pour choisir leur valeur, on peut soit

- recourir à la validation croisée, ou
- on crée plusieurs échantillons :
 - un échantillon d'apprentissage pour construire le modèle ;
 - un échantillon de validation pr optimiser les paramètres de tuning ("tuning" du modèle) ;
 - un échantillon test \perp pour évaluer la **performance** du modèle avec les paramètres de tuning choisis.

PRINCIPE DE LA VALIDATION CROISÉE (5-fold)

Utilisée pr la **sélection de modèle** ! Permet de choisir le param. et/ou modèle optimal.

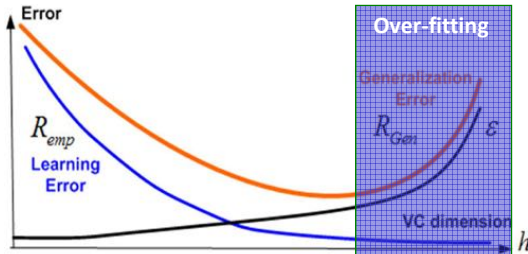


THÉORIE DE VAPNIK

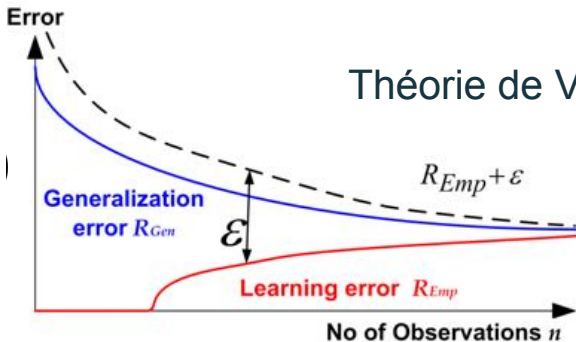
ERREURS EN FONCTION DE LA VC DIMENSION (h)

$$R_{Gen}(\theta) \leq R_{emp}(\theta) + \varepsilon(n, h)$$

$$\varepsilon(n, h) = \sqrt{\frac{1 + \ln(2n/h)}{n/h} - \frac{\ln \eta}{n}}$$



ET EN FONCTION DE n ?



QUELQUES PREMIERES REMARQUES

On voit très bien à travers l'inégalité de Vapnik que :

- l'erreur de généralisation croît quand la dimension augmente :
⇒ les modèles de grande dim. ont un faible biais au prix d'une grande variance (et inversement).
- l'erreur est dépendante du rapport n/h (rapport du nombre de données sur complexité du modèle),
- on \nearrow la capacité prédictive si $h \nearrow$ mais moins vite que n ,
- on peut \nearrow la complexité du modèle si on \nearrow aussi n .

3 Philosophie de l'apprentissage statistique

- Principes généraux
- Théorie de Vapnik et surapprentissage
- **Agrégation d'estimateurs**
- Comment analyser les résultats ?
- Rappels sur le bootstrap
- Agrégation : cas des variables catégorielles

L'AGREGATION

- Approche “modèle” VS **agrégation** :
 - **modèle** : déterminer une distribution de probabilité “simple” et unique qui rende compte des données ;
 - **agrégation** : faire la synthèse de plusieurs approches, ne plus se reposer sur un modèle unique.
- Les 2 approches ne sont **pas totalement antagonistes**.
- Les approches d'estimation basées sur l'agrégation sont + précises mais + difficilement interprétables (ex : agréger 3 modèles de régression paramétriques, comment ?).

Rq : on dit que les modèles simples (ex : logit) sont interprétables.
Loin d'être vrai car les covariables sont svt corrélées, donc la valeur des param. ne reflète pas exactement leur impact !

META-MODELES ou METHODES D'ENSEMBLE

Soit $\hat{m}_j(x)$ l'estimateur obtenu en utilisant le modèle j . Pour agréger B modèles et obtenir l'estimateur ensembliste

$$\hat{m}_a(x) = \sum w_j \hat{m}_j(x),$$

on peut mener :

- construction parallèle, \perp de +sieurs estimateurs individuels, puis combinaison \Rightarrow **bagging**
- construct. séquentielle, puis combinaison \Rightarrow **boosting** !
- construct. parall., puis **imbrication** (meta-modèle) \Rightarrow **stacking**

Rq : $\sum_{j=1}^B w_j = 1$ avec w_j poids affecté à l'estimateur j (version fréquentiste du Bayesian Model Averaging).

	Avantages	Inconvénients
"Modèle" unique	<p>Interprétation des paramètres</p> <p>Analyse de l'impact des variables</p> <p>Communication plus aisée</p>	<p>Biais important (erreur de modèle)</p> <p>Choix entre deux modèles ?</p>
Agrégation	<p>Moins de biais car hypothèses plus faibles</p> <p>Plus de difficulté liée au choix de modèle (à nuancer)</p>	<p>Interprétation complexe</p> <p>Analyse de l'impact des variables plus compliquée (mais possible)</p> <p>Communication sur le modèle hardue</p>

SYNTHÈSE DES PRINCIPALES DIFFÉRENCES

A travers ce que nous venons de voir, les différences essentielles de l'apprentissage statistique par rapport à une approche classique de modélisation résident dans les points suivants :

- **les hypothèses** : beaucoup moins d'hypothèses (\perp entre observations, entre facteurs de risque, hypothèses de distribution paramétrique, ...)
- **l'agrégation potentielle de modèles** : on construit plusieurs modèles et on synthétise,
- **l'interprétabilité des résultats** : on perd en interprétabilité à cause de l'agrégation.

3 Philosophie de l'apprentissage statistique

- Principes généraux
- Théorie de Vapnik et surapprentissage
- Agrégation d'estimateurs
- **Comment analyser les résultats ?**
- Rappels sur le bootstrap
- Agrégation : cas des variables catégorielles

MESURE DE L'INCERTITUDE - CAS D'UN MODÈLE

On dispose de résultats asymptotiques...

- En paramétrique, théorie du max. de vraisemblance. On a en général des IC sur le paramètre estimé...

Exemple : modèle linéaire,

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \approx \mathcal{N}(0, \sigma^2)$$

Donc $\mathbb{P}(|\hat{\beta}_1 - \beta_1| \geq \epsilon) \approx \mathbb{P}(|Z| \geq \epsilon) \quad \text{où } Z \sim \mathcal{N}(0, \sigma^2/n).$

→ σ^2 indique la précision de l'estimation : à estimer ! → D'où la possibilité d'évaluer $\mathbb{P}(|\hat{m}_1(x) - m_1(x)| \geq \epsilon).$

- En non paramétrique, théorie de Vapnik-Chervonenkis.

CAS DE L'AGREGATION D'ESTIMATEURS

C'est différent...(notons $\hat{m}_a(x)$ l'estimateur agrégé)

- En général, pas de résultat du type $\hat{m}_a(x) - m(x) \sim \mathcal{N}(0, \sigma^2)$.
- La qualité se mesure en premier lieu par rapport à un échantillon de validation.

Vocabulaire : soit un échantillon de $n + m$ observations, avec

- un **échantillon d'apprentissage** : sous-échantillon de n observations à partir desquelles on construit \hat{m}_a .
- un **échantillon de validation** : le reste (m observations) sur lequel on juge de la qualité de l'estimateur.

EXEMPLE EN RÉGRESSION

On dispose d'un échantillon de $(n + m)$ observations i.i.d., de même loi qu'un vecteur aléatoire (Y, X) .

But : estimer $m(x) = E[Y|X = x]$.

- (1) On tire au sort n observations, d'où l'échantillon $(Y_i, X_i)_{1 \leq i \leq n}$.
- (1bis) Les m autres observations $(Y_i, X_i)_{n+1 \leq i \leq n+m}$ constituent l'échantillon de validation.
- (2) Construction de $\hat{m}_a(x)$ à partir de $(Y_i, X_i)_{1 \leq i \leq n}$.
- (3) Calcul de l'erreur de prédiction sur l'échantillon de validation :

$$e(\hat{m}_a) = \sum_{i=n+1}^{n+m} (Y_i - \hat{m}_a(x_i))^2.$$

Plus cette quantité est petite, plus l'estimateur est jugé bon.

Questions :

- 1 Pourquoi ce critère ?
- 2 Pourquoi ne pas directement regarder l'erreur sur l'échantillon d'apprentissage ?
- 3 Choix de n et m ?

Q1 : POURQUOI CE CRITÈRE ?

- $m(x) = E[Y|X = x]$: “meilleure façon d'approcher Y par une fonction de X , au sens de l'écart quadratique” ;
- Si \hat{m}_a est bon estimateur, alors $\hat{m}_a(X_i)$ est proche de $m(X_i) \forall i$.
Or $m(x)$ étant la fonction la + proche de Y sachant $X = x$, + $e(\hat{m}_a)$ est petit, + \hat{m}_a devrait être proche de m (inconnue!).
- Si on calcule d'autres quantités, le coût quadratique ne sera pas forcément utilisé pour l'erreur.
→ Ex. : pour estimer la médiane de $Y|X = x$, on minimisera

$$\sum_{i=n+1}^{n+m} |Y_i - \hat{m}_a(x_i)|.$$

Q2 : POURQUOI NE PAS REGARDER L'ERREUR SUR L'ÉCHANTILLON D'APPRENTISSAGE ?

En d'autres termes, pourquoi ne pas prendre $m = 0$?

- Risque = **surapprentissage** (on capte le bruit au lieu du signal), le signal étant l'information principale...
- Exemple d'estimateurs faisant de l'overfitting : arbre maximal dans les estimateurs CART possibles...

Q3 : CHOIX DE n ET m

Pas de règle gravée dans le marbre (choix classique, et très arbitraire : $m = n/2$), mais

- en général, $n > m$, et
 - la proportion de l'échantillon d'apprentissage tend vers 50% quand la taille globale des données est grande ;
 - elle tend vers 80 voire 90% le cas contraire.
- pourquoi a-t-on besoin d'un n grand ?
 - besoin de + de données pour calculer un estimateur \hat{m}_a précis (sa CV est, en général, en $n^{-\alpha}$ pour un certain $\alpha > 0$).
- pourquoi m ne doit pas être trop petit ?
 - Pour que la validation ait un sens...

AGREGATION : ESTIMATEURS SUR ÉCHANTILLONS $\perp\!\!\!\perp$

→ **Limite** : on rappelle que $\hat{m}_a(x) = \frac{1}{B} \sum_{j=1}^B \hat{m}_j(x)$, où les $\hat{m}_j(x)$ sont **corrélés si calculés sur le même échantillon...**

Si les \hat{m}_j sont $\perp\!\!\!\perp$ car calculés sur \neq échantillons $\perp\!\!\!\perp$ (en notant $\sigma_j^2(x)$ la variance de $\hat{m}_j(x)$) :

$$\text{Var}(\hat{m}_a(x)) = \frac{1}{B^2} \sum_{j=1}^B \sigma_j^2(x).$$

En somme, si $\sigma^2(x) = \sup_{j=1,\dots,B} \sigma_j^2(x)$, $\text{Var}(\hat{m}_a(x)) \leq \frac{\sigma^2(x)}{B}$:

⇒ **Variance estimateur agrégé \ll variance estimateur unique.**

LIMITE PRATIQUE

Néanmoins, il est difficile de calculer des estimateurs sur des échantillons \neq , car la taille des données n'est évidemment pas infinie en pratique... D'où :

- prendre B sous-échantillons pour calculer B estimateurs \neq est une solution de riche (n doit être très grand pour l' \perp) ;
- la solution de couper l'échantillon en sous-échantillons atteint vite ses limites.

⇒ Une solution : le **rééchantillonnage** (par exemple bootstrap).

Rq : l'indépendance entre les estimateurs unitaires n'est pas garantie car certains échantillons bootstrap peuvent fortement se ressembler...

3 Philosophie de l'apprentissage statistique

- Principes généraux
- Théorie de Vapnik et surapprentissage
- Agrégation d'estimateurs
- Comment analyser les résultats ?
- **Rappels sur le bootstrap**
- Agrégation : cas des variables catégorielles

APPLICATION : REECHANTILLONNAGE BOOTSTRAP

- Si on souhaite agréger B estimateurs, on génère B échantillons bootstrap suivant la méthode ci-dessous.
- On utilise l'échantillon j pour calculer l'estimateur \hat{m}_j .

Bootstrap : pour $j = 1, \dots, B$ et $i = 1, \dots, n$, on tire $Z_i^{(j)}$ i.i.d. de loi unif. sur $\{1, \dots, n\}$. Le $j^{\text{ème}}$ échantillon bootstrap est $(Y_i^{(j)}, X_i^{(j)})_{1 \leq i \leq n}$ où

$$Y_i^{(j)} = Y_{Z_i^{(j)}} \quad X_i^{(j)} = X_{Z_i^{(j)}}.$$

En moyenne, $e^{-1} = 36,7\%$ des observations initiales ne sont pas tirées dans un échantillon bootstrap donné.

ILLUSTRATION

Bootstrap: $x = (x_1, x_2, x_3)$

$$\bar{x}_n = \frac{1}{3}(x_1 + x_2 + x_3)$$

$\bar{X}_n = \frac{1}{3}(X_1 + X_2 + X_3)$: variable aléatoire, estimateur de la moyenne.

Q: cet estimateur est-il fiable?

Créons des échantillons bootstrap, au nombre de 3^3 . Par exemple le 1^{er}:

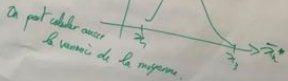


$$x_1^* = (x_1, x_1, x_1)$$

On dérive le raisonnement... $\bar{x}_{n,1}^* = x_1$

\Rightarrow B moyennes:

répartition moyenne



POURQUOI LE BOOTSTRAP ?

Idées derrière le bootstrap :

- On va créer artificiellement des échantillons semblables à celui d'origine en simulant des données, ce qui permettra de construire des modèles cohérents entre eux.
- Problème : les échantillons étant corrélés, les estimateurs seront corrélés (même si différents !)...

On aura besoin d'introduire des éléments supplémentaires pour décorréliser au mieux les estimateurs !

3 Philosophie de l'apprentissage statistique

- Principes généraux
- Théorie de Vapnik et surapprentissage
- Agrégation d'estimateurs
- Comment analyser les résultats ?
- Rappels sur le bootstrap
- Agrégation : cas des variables catégorielles

PRÉVISIONS AGRÉGÉES SUR VARIABLE BINAIRE

Soit le contexte suivant :

- une v.a. Y qui vaut 0 ou 1, avec X = caractéristiques.
 - ex. 1 : $Y = 1$ si accident dans l'année, 0 sinon.
 - ex. 2 : $Y = 1$ si défaut de paiement dans l'année, 0 sinon.
 - ex. 3 : $Y = 1$ si le client souscrit un contrat, 0 sinon.
- 1^{ère} solution : déterminer $E[Y|X] = P(Y = 1|X)$ pour chaque modèle. Puis approche similaire à précédemment : moyenner.
→ Cette solution fournit un estimateur $\hat{m}_a(x)$ qui prend des valeurs entre 0 et 1.

⇒ Si $X = x$ et $\hat{m}_a(x) > 0.5$, on prédit $Y = 1$. Sinon $Y = 0$.

DEUXIÈME SOLUTION : LE VOTE MAJORITAIRE

- Au lieu d'agréger les estimations des espérances conditionnelles \hat{m}_j , on agrège les **prédictions** associées.
- i.e. on définit, pour $j = 1, \dots, B$, $\hat{p}_j(x) = 1_{\hat{m}_j(x) > 0,5}$.
- Pour $X = x$, on prédit Y par

$$\hat{p}_a(x) = \begin{cases} 1 & \text{si majorité de } \hat{p}_j(x) \text{ égaux à } 1 \\ 0 & \text{sinon.} \end{cases}$$

- Rq : c'est ce que fait `randomForest(.)` de `rpart`.

GÉNÉRALISATION AUX VARIABLES CATÉGORIELLES

On s'intéresse à une variable Y prenant un nombre fini de modalités, $\{1, \dots, k\}$, avec X = caractéristiques.

- Exemple : Y = gravité sinistre, classé sur échelle de 1 à k .
- Stratégie : transformation en un problème binaire.

$$Z_l = 1_{Y=l}$$

pour $l = 1, \dots, k$, on estime $E[Z_l | X] = P(Y = l | X)$ pour tout l .

- Si on note $\hat{m}_{j,l}(x)$ l'estimateur de $P(Y = l | X = x)$ basé sur la méthode j , la prédiction $\hat{p}_j(x)$ associée est

$$\hat{p}_j(x) = \arg \max_{l=1, \dots, k} \hat{m}_{j,l}(x).$$