Statistique, Apprentissage, *Big–Data–Mining*

Résumé

L'objet de ce cours est d'introduire, sous une forme homogène et synthétique, les techniques de modélisation statistique ou d'apprentissage supervisé utilisées le plus couramment en fouille de données *volumineuses ou de* grande dimension *(*data mining, big data*) pour l'aide à la décision dans des champs d'applications très divers : industriels, marketing, ou encore en relation avec des thématiques de recherche en Biologie, Épidémiologie... L'objectif principal est la modélisation pour la prévision et donc la recherche de modèles optimaux (parcimonieux) pour différentes méthodes de modélisation statistique classique (modèles gaussiens et binomiaux, analyse discriminante), moins classiques (ridge, pls, lasso, arbres binaires de décision) ou encore dites d'apprentissage (réseaux de neurones, agrégation de modèles, machines à vecteurs supports) issues du* machine learning.

1 Introduction

1.1 Un peu d'histoire

1940-70 – hOctets Il était une fois la Statistique : une question, (*i.e.* biologique), associée à une hypothèse expérimentalement réfutable, une expérience *planifiée* avec $n \approx 30$ individus observés sur p (moins de 10) variables, un modèle *linéaire* supposé *vrai*, un test, une décision, une réponse.

1970s – kO Les premiers outils informatiques se généralisant, l'*analyse des données* en France, (*multivariate statistics* ailleurs : Mardia et al. (1979) [5]) explore, prétendument sans modèle, des données plus volumineuses.

1980s – MO En Intelligence Artificielle, les *systèmes experts* expirent, supplantés par l'apprentissage (*machine learning*) des *réseaux de neurones*. La Statistique aborde des modèles non-paramétriques ou fonctionnels.

1990s – GO Premier changement de paradigme. Les données ne sont plus *planifiées*, elles sont préalablement acquises et basées dans des entrepôts pour les objectifs usuels (*i.e.* comptables) de l'entreprise. L'aide à la décision les valorise : *From Data Mining to Knowledge Discovery* (Fayyad et al., 1996)[2]. Les logiciels de fouille regroupent dans un même environnement des outils de gestions de données, des techniques exploratoires et de modélisation statistique). C'est l'avènement du marketing quantitatif et de la gestion de la relation client (GRC ou CRM).

2000s –TO Deuxième changement de paradigme. Le nombre p de variables explose (de l'ordre de 10^4 à 10^6), notamment avec les biotechnologies omiques où $p >> n$. L'objectif de qualité de prévision l'emporte sur la réalité du modèle devenu "boîte noire". Face au fléau de la dimension, Apprentissage Machine et Statistique s'unissent en Apprentissage Statistique (*statistical learning*, Hastie et al. 2001-2009)[3] : sélectionner des modèles en équilibrant *biais vs. variance* ; minimiser conjointement erreurs d'*approximation* (biais) et erreur d'*estimation* (variance).

2010s – PO Troisième changement de paradigme. Dans les applications industrielles, le e-commerce, la géo-localisation... c'est le nombre n d'individus qui explose, les bases de données débordent, se structurent en nuages (*cloud*), les moyens de calculs se groupent (*cluster*), mais la puissance brute ne suffit plus à la voracité (*greed*) des algorithmes. Un troi-

sième terme d'erreur est à prendre en compte : celle d'*optimisation*, induite par la limitation du temps de calcul ou celle du volume / flux de données considéré. La décision devient adaptative ou séquentielle.

1.2 Contenu

Les données volumineuses sont au cœur des problématiques émergentes de recherche, en faisant notamment appel à des structures de données sophistiquées : graphes, fonctions, variétés. Chaque problème est porteur de sa propre originalité ; ils ne seront pas abordés dans ce cours qui se limite aux articulations : Statistique, Apprentissage Machine, fouille de données et donc au problème central de l'équilibre biais — variance.

Ainsi, le *data mining* est présenté comme la recherche d'informations pertinentes (des "pépites" d'information) pour l'aide à la décision et la prévision. Il met en œvre des techniques statistiques et d'apprentissage machine en tenant compte de la spécificité de grandes à très grandes dimensions des données.

La section 2 suivante introduit à la *fouille de données* tandis que la section 3 reprend ces objectifs dans le cadre général de la modélisation afin d'en élargir les champs d'application. La section 4 décrit la stratégie très généralement mise en place pour optimiser choix de méthodes et choix de modèles ; la section 5 décrit brièvement quelques exemples d'application et notamment ceux utilisés pour illustrer ce cours. Enfin, la section 6 liste rapidement les méthodes qui sont abordées et les raisons qui ont conduit à ces choix.

2 Motivations du *big data mining*

2.1 Origine

Le développement des moyens informatiques et de calcul permet le stockage (bases de données), le traitement et l'analyse d'ensembles de données très volumineux. Plus récemment, le perfectionnement des logiciels et de leurs interfaces offrent aux utilisateurs, statisticiens ou non, des possibilités de mise en œuvre très simples de ces méthodes. Cette évolution, ainsi que la popularisation de nouvelles techniques algorithmiques (réseaux de neurones, support vector machine...) et outils graphiques, conduit au développement et à la commercialisation de logiciels (Enterprise miner, Clementine, Insightfull miner...) intégrant un sous-ensemble de méthodes statistiques et algorithmiques utilisées sous la terminologie de *Data Mining* généralement traduit par *fouille de données* (voir Tufféry 2007 [6] pour un exposé "métier" plus détaillé). Cette approche, dont la présentation est principalement issue du marketing spécialisé dans la gestion de la relation client (GRC) (*client relation management* ou CRM), trouve également des développements et applications industrielles en contrôle de qualité ou même dans certaines disciplines scientifiques dès lors que les ingénieurs et chercheurs sont confrontés à un volume de données important. C'est même l'afflux actuel de saisies automatiques de données issues du monde industriel qui motive principalement l'émergence du *Big Data* parallèlement à l'explosion du e-commerce. Devant la complexité envisagée, lorsque les modèles physiques font défaut, un problème industriel peut changer de paradigme lorsque la modélisation déterministe atteint ses limites, les données recueillies massivement sont analysées pour l'aide à la décision comme ce fut le cas en marketing quantitatif avec la fouille de données du siècle dernier.

L'accroche publicitaire souvent citée par les éditeurs de logiciels (SAS) est :

> *Comment trouver un diamant dans un tas de charbon sans se salir les mains.*

Nous proposons d'évaluer et d'expérimenter la réalité de cette annonce qui s'adresse à un marché en pleine expansion. Les entreprises sont en effet très motivées pour tirer parti et amortir, par une aide à la décision quantifiée, les coûts de stockage des téraoctets que leur service informatique s'emploie à administrer.

2.2 Environnement

Le contexte informationnel de la fouille de données est donc celui d'un système de bases de données, classique relationnel ou non, dont la mise en place est assurée par le gestionnaire de données (*data manager*) en relation avec une problématique :

- gestion des stocks (flux tendu), des ventes d'un groupe afin de prévoir et anticiper au mieux les tendances du marché,
- suivi des fichiers clients d'une banque, d'une assurance, associés à des données socio-économiques (INSEE), à l'annuaire, en vue de la constitution d'une segmentation (typologie) pour cibler des opérations de marketing ou des attributions de crédit. La *gestion de la relation client* (GRC ou CRM) vise à une individualisation ou personnalisation de la production et

de la communication afin d'évacuer la notion de *client moyen*.
- recherche, spécification puis ciblage de *niches* de marché les plus profitables (banque) ou au contraire les plus risquées (assurance) ;
- suivi en ligne des paramètres de production (traçabilité) en contrôle de qualité pour détecter au plus vite l'origine d'une défaillance ;
- prospection textuelle (*text mining*) et veille technologique ;
- *web mining*, comportement des internautes et e-commerce ;
- ...

Cet environnement se caractérise par
- une informatique hétérogène faisant intervenir des sites distants à travers le réseau de l'entreprise (intranet) ou même des accès extérieurs (internet). Des contraintes d'efficacité, de fiabilité ou de sécurité conduisent à répartir, stocker l'information à la source plutôt qu'à la dupliquer systématiquement ou à la centraliser.
- L'incompatibilité logique des informations observées sur des échantillons différents ne présentant pas les mêmes strates, les mêmes codifications.
- Des volumes et flux considérables de données issues de saisies automatisées et chiffrés en téra maintenant pétaoctets.
- Contrairement à une démarche statistique traditionnelle (planification de l'expérience), les données analysées sont stockées à d'autres fins (comptabilité, contrôle de qualité...) et sont donc *préalables* à l'analyse.
- La nécessité de ne pas exclure *a priori* un traitement *exhaustif* des données afin de ne pas laisser échapper, à travers le crible d'un *sondage*, des groupes de faibles effectifs mais à fort impact économique.

2.3 Big Data vs. Data Mining

La communication, les noms changent mais fondamentalement les méthodes restent. Le traitement des grandes masses de données, associé au "nouveau" métier de *data scientist*, occupe une grande place dans les médias notamment en relation avec les risques annoncés et réels du contrôle d'internet par *big brother*. Beaucoup d'entreprises et de formations suivent le mouvement en renommant les intitulés sans pour autant se placer dans le cadre de grandes masses de données nécessitant des traitements spécifiques. Celui-ci devient effectif à partir du moment où le volume et le flux de données imposent une parallélisation des tâches : les données sont réparties en nœuds, chacun associé à un processeur ou calculateur relié aux autres par un réseau haut débit au sein d'un *cluster*. Les mots clefs et outils de cette architecture sont *Hadoop* et *Map Reduce*, *NoSQL*. Hadoop est un projet de la fondation logicielle Apache (*open source* en java) destiné à faciliter la création d'applications distribuées et *échelonnables*. Un algorithme, une méthode est dite échelonnable (*scalable*) si le temps de calcul est divisé par le nombre de processeurs (nœuds) utilisés ce qui permet aux applications de travailler avec des milliers de nœuds et des pétaoctets de données. Le principe, initié par Google et Yahoo, est de répartir les tâches parallèles (*Map*) puis d'intégrer (*Reduce*) tous les résultats obtenus. Exemple très élémentaire : chaque nœud calcule la moyenne d'une variable avant de calculer la moyenne des moyennes. Bien entendu, toute méthode statistique ou d'apprentissage n'est pas *scalable* ou au pris d'un algorithme stochastique plus sophistiqué. D'autre part les requêtes complexes comme celle de SQL sont impossibles. D'autres systèmes dits NoSQL (*not only SQL*, Cassandra, MongoDB, Voldemort...), développés à l'origine par des grands sites comme Amazon, eBay, reposent également sur un système de fragmentation (*sharding*) des données tout en autorisant des possibilités de requêtes intermédiaires avec SQL. Bien entendu les principaux acteurs commerciaux historiques comme (Oracle) prennent position de façon complémentaire ou concurrente avec ces systèmes émergents.

Confronté à cette problématique, il appartient au statisticien *data scientist* de
- s'initier aux interfaces d'accès à une architecture Hadoop ou NoSQL, notamment par l'utilisation d'outils comme *Mahout* ou RHadoop...
- optimiser sa stratégie : sonder dans les données et se ramener à des méthodes usuelles ou traiter les données de façon exhaustives uniquement avec une technique *scalable*. Comment intervient un erreur introduite par sondage par rapport à celle de la méthode utilisée ?
- prendre en compte, ou non, les aspects temporels dus aux flux de données : estimer des modèles sur une fenêtre glissante, adopter des algorithmes adaptatifs ?
- Aborder de nouveaux (*Scala, Clojure*) anciens (basés sur *Lisp*) langages de programmation pour développer ou redévelopper des méthodes d'apprentissage directement parallélisables. C'est en effet ce que permettent ces langages fonctionnels par opposition aux langages objet (C, java...).

Nécessairement limité, ce cours, niveau M2, ne peut aborder ces dernières questions. Il ne peut non plus aborder celles liées à la complexité des données

industrielles ou issues de la recherche (biologique, médicale...) qui ouvrent souvent sur des problèmes originaux. Il peut s'agir alors de traiter non plus des tableaux ou matrices de données mais des saisies automatiques de courbes, signaux spectres, images, graphes... De telles structures posent un problème préalable de base de représentation (fourier, splines, ondelettes...) fonction de la nature des données et de l'objectif recherché. Voir par exemple le scénario d'analyse de spectres RMN décomposés sur une base d'ondelettes pour la détection de métabolites "biomarqueurs".

Il est important de noter que, s'il a une formation de base en Mathématiques et Statistique, le nouveau *d*ata scientist voit arriver avec une certaine sérénité la vague ou le tsunami du *Big Data*. Certes un travail informatique amont, perpétuellement renouvelé face à l'obsolescence rapide tes technologies, est important pour stocker les données et rendre exécutable les méthodes mais, conceptuellement, la Mathématique nécessaire prend déjà en compte des tailles et dimensions infinies pour les modèles considérés dans des espaces hilbertiens. Muni de ce bagage pérenne, il peut accompagner et suivre la recherche en développement.

3 Apprentissage statistique

Un peu de recul permet d'inscrire la démarche de la fouille de données dans un contexte plus large et donc potentiellement plus propice à d'autres domaines d'application.

3.1 Objectif général

Dès qu'un phénomène, qu'il soit physique, biologique ou autre, est trop complexe ou encore trop bruité pour accéder à une description analytique débouchant sur une modélisation déterministe, un ensemble d'approches ont été élaborées afin d'en décrire au mieux le comportement à partir d'une série d'observations. Voici quelques exemples de problèmes d'apprentissage :

- identifier les facteurs de risque d'un certain type de cancer, en fonction de variables cliniques et démographiques,
- rechercher des gènes potentiellement impliqués dans une maladie à partir de données de biopuces ou plus généralement des bio-marqueurs pour un diagnostic précoce,
- identifier des chiffres manuscrits sur un code postal à partir d'une image digitalisée,
- prévoir le prix d'un stock dans 6 mois à partir de mesures de performance de l'entreprise et de données économiques,
- prévoir un taux de pollution atmosphérique en fonction de conditions météorologiques ,
- prévoir une courbe de consommation électrique pour un client EDF en fonction de variables climatiques et de caractéristiques spécifiques à ce client,
- Gestion de la relation client (GRC ou CRM) et *scoring* en marketing quantitatif,
- maintenance préventive à partir de relevés d'incidents,
- construire un modèle de substitution à un code numérique complexe qui permet de prédire une carte de concentration d'un polluant dans un sol un an après un rejet accidentel en fonction de la carte initiale et des caractéristiques du sol (porosité, perméabilité...). L'objectif est de réaliser une analyse de sensibilité.

Historiquement, la Statistique s'est beaucoup développée autour de ce type de problèmes et a proposé des *modèles* incorporant d'une part des *variables explicatives ou prédictives* et, d'autre part, une composante aléatoire ou *bruit*. Il s'agit alors d'*estimer* les *paramètres* du modèle à partir des observations en contrôlant au mieux les propriétés et donc le comportement de de la partie aléatoire. Dans la même situation, la communauté informatique parle plutôt d'*apprentissage* visant le même objectif ; apprentissage machine (ou *machine learning*), reconnaissance de forme (pattern recognition) en sont les principaux mots-clefs.

L'objectif général est donc un objectif de *modélisation* qui peut se préciser en sous-objectifs à définir clairement préalablement à une étude car ceux-ci conditionnent en grande part les méthodes qui pourront être mises en œuvre :

Modéliser pour

explorer ou vérifier, représenter, décrire, les variables, leurs liaisons et positionner les observations de l'échantillon,

expliquer ou tester l'influence d'une variable ou facteur dans un modèle supposé connu a priori,

prévoir & sélectionner un meilleur ensemble de prédicteurs comme par exemple dans la recherche de bio-marqueurs,

prévoir par une éventuelle meilleure "boîte noire" sans besoin d'interprétation explicite.

Des paramètres importants du problème sont les dimensions : n nombre d'observations ou taille de l'échantillon et p nombre de variables observées sur cet échantillon. Lorsque les méthodes statistiques traditionnelles se trouvent mises en défaut pour de grandes valeurs de p, éventuellement plus grande que n, les méthodes récentes d'apprentissage sont des recours pertinents car efficaces.

Enfin, les stratégies de choix de modèle parmi un ensemble plus ou moins complexe, de choix de méthode, sont au cœur de la problématique de ce cours. L'étude de la fouille de données se focalise donc sur les pratiques ou méthodes à l'interface de l'apprentissage machine et de la Statistique. Les développements méthodologiques à cette interface ont pris depuis le début du siècle la dénomination d'*apprentissage statistique* ; Hastie et al. (2009)[3] en proposent un tour d'horizon assez exhaustif.

Attention, d'autres objectifs d'une fouille de données ou d'extensions de ces techniques, ne sont pas pris en compte dans celui d'une modélisation au sens statistique précédent et donc dans ce cours d'apprentissage statistique. Cela concerne la

- classification non-supervisée ou *clustering* traité par ailleurs et rappelé ci-dessous.
- recherche de règles d'associations ou problème du *panier de la ménagère*. Méthode qui consiste à identifié les co-occurences les plus fréquentes ou significatives par un ensemble de règles logiques associant variables et valeurs de celles-ci.
- Les *Sytèmes de recommandation* : ou modèles de *bandits manchots* pour déterminer et afficher sur un site de e-commerce les articles complémentaires susceptibles d'intéresser le visiteur.

3.2 Problématiques

Supervisé vs. *non-supervisé*

Distinguons deux types de problèmes : la présence ou non d'une variable à *expliquer* Y ou d'une *forme* à reconnaître qui a été, conjointement avec X, observée sur les mêmes objets. Dans le premier cas il s'agit bien d'un problème de modélisation ou *apprentissage supervisé* : trouver une fonction f susceptible, au mieux selon un critère à définir, de reproduire Y ayant observé X.

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

où ε symbolise le bruit ou erreur de mesure avec le parti pris le plus commun que cette erreur est additive. En cas d'erreur multiplicative, une transformation logarithmique ramène au problème précédent.

Dans le cas contraire, en l'absence d'une variable à expliquer, il s'agit alors d'apprentissage dit *non-supervisé*. L'objectif généralement poursuivi est la recherche d'une typologie ou taxinomie des observations : comment regrouper celles-ci en classes homogènes mais les plus dissemblables entre elles. C'est un problème de classification (*clustering*).

Attention, l'anglais *classification* se traduit plutôt en français par discrimination ou classement (apprentissage supervisé) tandis que la recherche de classes (*clustering*) (apprentissage non-supervisé) fait appel à des méthodes de classification ascendante hiérarchique, des algorithmes de réallocation dynamique ($k$means) ou encore des cartes auto-organisatrices (Kohonen).

Dans ce cours, nous allons nous intéresser essentiellement à l'apprentissage supervisé, pour lequel on dispose d'un *ensemble d'apprentissage* constitué de données d'observations de type entrée-sortie : $d_1^n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ avec $\mathbf{x}_i \in \mathcal{X}$ quelconque (souvent égal à $\mathbb{R}^p$), $y_i \in \mathcal{Y}$ pour $i = 1 \ldots n$

L'objectif est de construire, à partir de cet échantillon d'apprentissage, un modèle, qui va nous permettre de prévoir la sortie y associée à une nouvelle entrée (ou prédicteur) $\mathbf{x}$. La sortie y peut être quantitative (prix d'un stock, courbe de consommation électrique, carte de pollution ..) ou qualitative (survenue d'un cancer, reconnaissance de chiffres...).

sorties quantitatives	sorties qualitatives
$\mathcal{Y} \subset \mathbb{R}^p$	$\mathcal{Y}$ fini
$\downarrow$	$\downarrow$
régression	**discrimination, classement, reconnaissance de forme**

Nous parlerons de régression réelle lorsque $\mathcal{Y} \subset \mathbb{R}$ et de la discrimination binaire lorsque $\mathcal{Y} = \{-1, 1\}$.

Estimation vs. *apprentissage*

Tout au long de ce document, les termes de *estimation* et d'*apprentissage* sont utilisés comme des synonymes ce qui est abusif tant que les objectifs d'une étude n'ont pas été clairement explicités. Dans la tradition statistique, la notion de *modèle* est centrale surtout avec une finalité *explicative*. Il s'agit alors d'approcher la réalité, le *vrai* modèle, supposé exister, éventuellement basé sur une théorie physique, économique, biologique... sous-jacente et la forme du modèle est guidée par des indications théoriques et des critères d'*ajustement* ; les décisions de validité, de présence d'effets sont basées sur des *tests* reposant elles-mêmes sur des hypothèses probabilistes. L'interprétation du rôle de chaque variable explicative est prépondérante dans la démarche.

En revanche, si l'objectif est essentiellement la *prévision*, il apparaît que le meilleur modèle n'est pas nécessairement celui qui ajusterait le mieux le vrai modèle. La théorie de l'*apprentissage* (Vapnik, 1999) montre alors que le cadre théorique est différent et les majorations d'erreur requièrent une autre approche. Les choix sont basés sur des critères de qualité de *prévision* visant à la recherche de *modèles parcimonieux*, c'est-à-dire de complexité (nombre de paramètres ou flexibilité limitée) dont l'interprétabilité passe au deuxième plan. La deuxième devise (cf. figure 1) des Shadoks n'est pas une référence à suivre en apprentissage statistique !

Discrimination vs. *régression*

Le type des variables statistiques considérées diffèrent selon l'espace dans lequel elles prennent leurs valeur. Elles peuvent être qualitatives à valeurs dans un ensemble de cardinal fini ou quantitatives à valeurs réelles voire fonctionnelles. Certaines méthodes d'apprentissage ou de modélisation s'adaptent à tout type de variables explicatives tandis que d'autres sont spécialisées. Enfin, si Y à expliquer est qualitative, on parle de discrimination, classement ou reconnaissance de forme tandis que si Y est quantitative on parle, par habitude, d'un problème de régression. Dans ce cas encore, certaines méthodes sont spécifiques (régression linéaire, analyse discriminante) tandis que d'autres s'adaptent sans modification profonde remettant en cause leur principe (réseaux de neurones, arbres de décision...).

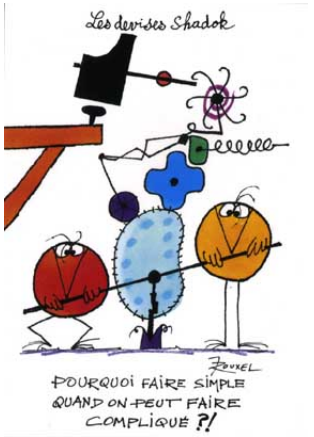

FIGURE 1 – Deuxième devise Shadok

Statistique, informatique et taille des données

Lorsque les dimensions du problèmes (n, p) sont raisonnables et que des hypothèses relatives au modèle (linéarité) et aux distributions sont vérifiées c'est-à-dire, le plus souvent, lorsque l'échantillon ou les résidus sont supposés suivre des lois se mettant sous la forme d'une famille exponentielle (gaussienne, binomiale, poisson...), les techniques statistiques de modélisation tirées du modèle linéaire général sont optimales (maximum de vraisemblance) et, surtout dans le cas d'échantillons de taille restreinte, il semble difficile de faire beaucoup mieux.

En revanche, dès que les hypothèses distributionnelles ne sont pas vérifiées, dès que les relations supposées entre les variables ou la variable à modéliser ne sont pas linéaires ou encore dès que le volume des données (*big data*) est important, d'autre méthodes viennent concurrencer l'approche statistique classique.

Prenons un exemple simple : expliquer une variable quantitative Y par un ensemble $\{X^1, \ldots, X^p\}$ de variables également quantitatives :

$$Y = f(X^1, \ldots, X^p) + \varepsilon.$$

observées sur un échantillon $(y_i, \boldsymbol{x}_i); i = 1, \ldots, n$ de taille n. Si la fonction f est supposée linéaire et p petit, de l'ordre d'une dizaine ; le problème est bien connu et largement débattu dans la littérature. Dans le cas où la fonction f n'est pas franchement linéaire et n grand, il est possible d'estimer précisément un nombre plus important de paramètres et donc d'envisager des modèles plus sophistiqués. Si on s'en tient au modèle gaussien usuel, même le cas le plus simple d'un modèle polynomial devient vite problématique. En effet, lorsque la fonction f est linéaire, prenons $p = 10$, la procédure de choix de modèle est confrontée à un ensemble de 2^{10} modèles possibles et des algorithmes astucieux permettent encore de s'en sortir. En revanche, considérer, pour estimer f, un simple polynôme du deuxième voire troisième degré avec toutes ses interactions, amène à considérer un nombre considérable de paramètres et donc, par explosion combinatoire, un nombre astronomique de modèles possibles. D'autres méthodes doivent alors être considérées en prenant en compte nécessairement la complexité algorithmique des calculs. Ceci explique l'implication d'une autre discipline, l'informatique, dans cette problématique. Le souci de calculabilité l'emporte sur la définition mathématique du problème qui se ramène à l'optimisation d'un critère d'ajustement de la fonction f sur un ensemble de solutions plus ou moins riche. Ces méthodes ont souvent été développées dans un autre environnement disciplinaire : informatique, intelligence artificielle… ; k plus proches voisins, réseaux de neurones, arbres de décisions, *support vector machine* deviennent des alternatives crédibles dès lors que le nombre d'observations est suffisant ou le nombre de variables très important.

3.3 Stratégies de choix

Choix de méthode

Avec le développement du *data mining*, de très nombreux articles comparent et opposent les techniques sur des jeux de données publics et proposent des améliorations incrémentales de certains algorithmes. Après une période fiévreuse où chacun tentait d'afficher la suprématie de sa méthode, un consensus s'est établi autour de l'idée qu'il n'y a pas de "meilleure méthode". Chacune est plus ou moins bien adaptée au problème posé, à la nature des données ou encore aux propriétés de la fonction f à approcher ou estimer. Sur le plan méthodologique, il est alors important de savoir comparer des méthodes afin de choisir la plus pertinente. Cette comparaison repose sur une estimation d'erreur (de régression ou de classement) qu'il est nécessaire de conduire avec soin.

Choix de modèle : équilibre biais-variance

Tous les auteurs s'accordent pour souligner l'importance qu'il y a à construire des modèles *parcimonieux* quelque soit la méthode utilisée. Toutes les méthodes sont concernées : nombre de variables explicatives, de feuilles dans un arbre ou de neurones dans une couche cachée…. Seuls les algorithmes de combinaison de modèles (bagging, boosting) contournent cette étape au prix d'un accroissement sensible du volume des calculs et surtout de l'interprétabilité des résultats obtenus.

L'alternative est claire, plus un modèle est complexe et donc plus il intègre de paramètres et plus il est flexible donc capable de s'ajuster aux données engendrant ainsi une erreur faible d'ajustement. En revanche, un tel modèle peut s'avérer défaillant lorsqu'il s'agira de prévoir ou généraliser, c'est-à-dire de s'appliquer à des données qui n'ont pas participé à son estimation.

L'exemple élémentaire de la figure 2 illustre ce point fondamental dans le cas d'un problème de discrimination dans $\mathbb{R}^2$. Une frontière dont le modèle "vrai" est quadratique est, à cause d'"erreurs de mesure" sous-ajustée par une régression linéaire mais sur-ajustée par un polynôme de degré plus élevé ou l'algorithme local des k plus proches voisins.

Ce problème s'illustre aussi facilement en régression classique. Ajouter des variables explicatives dans un modèle ne peut que réduire l'erreur d'ajustement (le R^2) et réduit le biais si le "vrai" modèle est un modèle plus complet. Mais, ajouter des variables fait rédhibitoirement croître la variance des estimateurs et donc celle des prévisions qui se dégradent, voire explosent, avec la multicolinéarité des variables explicatives. Un risque pour le modèle, ou erreur quadratique de prévision, s'exprimant comme le carré du biais plus la variance, il est important d'optimiser le dosage entre biais et variance en contrôlant le nombre de variables dans le modèle (sa complexité) afin de minimiser le risque. Ces remarques conduisent à la définition de critères de choix de modèle dont le C_p de Mallows fut un précurseur en régression suivi par d'autres propositions : Akaïke (AIC), Schwartz (BIC)…

Parfois plus que celui de la méthode, le choix du bon modèle dans une classe ou ensemble de modèles pour une méthode donnée est primordial. En consé-

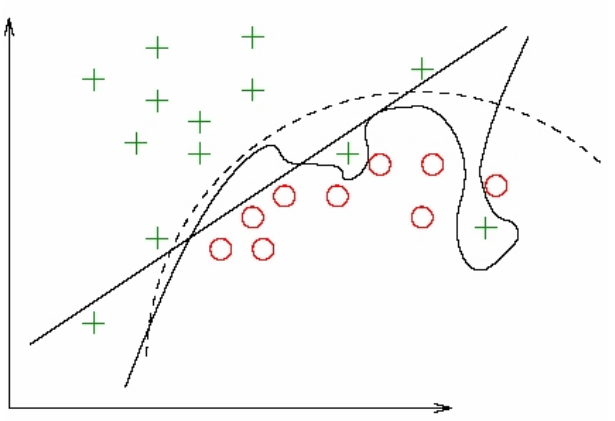

FIGURE 2 – *Sous-ajustement linéaire et sur-ajustement local (proches voisins) d'un modèle quadratique.*

quence, les problèmes d'optimisation considérés doivent mettre en œuvre un critère qui prend en compte la *complexité du modèle*, c'est-à-dire la complexité de l'espace ou de la classe dans lequel la solution est recherchée.

Choix de modèle : sélection vs. *régularisation*

Selon la méthode considérée, la complexité du modèle s'exprime de différentes façons. Simple lors d'une sélection de variable en régression linéaire, la complexité est directement liée à la dimension de l'espace engendré et donc au nombre de variables. Les choses se compliquent pour les modèles non-linéaires lorsque, à dimension fixée, c'est la plus ou moins grande flexibilité des solutions qui doit être pénalisée.

C'est typiquement le cas en régression non-paramétrique ou fonctionnelle. Une pénalisation faisant intervenir la norme carrée de la dérivée seconde contrôle la flexibilité d'un lissage spline. La "largeur de fenêtre" du noyau contrôle également la régularité de la solution. En régression linéaire, si le nombre et les variables sont déterminés, la version "ridge" de la régression pénalise la norme carrée du vecteur des paramètres et restreint ainsi, par *régularisation*, l'espace des solutions pour limiter l'effet de la multicolinéarité.

Enfin, pour aborder en toute généralité les situations les plus compliquées, Vapnik (1999) a formalisé la théorie de l'apprentissage en introduisant une notion particulière de dimension pour toute famille de modèles.

4 Stratégie de l'apprentissage statistique

4.1 Les données

Dans la majorité des problèmes rencontrés, des caractéristiques ou variables $X = (X^1, \ldots, X^p)$ dites explicatives ou prédictives ont été observées sur un ensemble de n objets, individus ou unités statistiques. Un premier travail, souvent fastidieux mais incontournable, consiste à mener une exploration statistique de ces données : allure des distributions, présence de données atypiques, corrélations et cohérence, transformations éventuelles des données, description multidimensionnelle, réduction de dimension, classification. C'est l'objet d'un cours distinct d'exploration statistique tandis que ce cours décrit les outils de modélisation statistique ou encore d'apprentissage utilisables pour la modélisation à fin de prévision d'une variable *cible* Y par les variables explicatives ou prédictives X^j.

L'enchaînement, éventuellement itératif, de ces étapes (exploration puis apprentissage) constitue le fondement de la fouille de données.

Pour comprendre la structure et bien appréhender le contenu de ce cours, il est important d'intégrer rapidement ce qu'est la stratégie à mettre en œuvre pour aboutir au bon *apprentissage* ou encore au bon *modèle prédictif* recherché à partir des données observées.

Attention, contrairement à une démarche statistique traditionnelle dans laquelle l'observation des données est intégrée à la méthodologie (planification de l'expérience), les données sont généralement *préalables* à l'analyse. Néanmoins il est clair que les préoccupations liées à leur analyse et à son objectif doivent intervenir le plus en amont possible pour s'assurer quelques chances de succès.

4.2 Les étapes de l'apprentissage

Les traitements s'enchaînent de façon assez systématique selon le schéma suivant et quelque soit le domaine d'application :

1. Extraction des données avec ou sans échantillonnage faisant référence à des techniques de sondage appliquées ou applicables à des bases de données.
2. Exploration des données pour la détection de valeurs aberrantes ou seulement atypiques, d'incohérences, pour l'étude des distributions des structures de corrélation, recherche de typologies, pour des transformations des données...
3. Partition aléatoire de l'échantillon (apprentissage, validation, test) en fonction de sa taille et des techniques qui seront utilisées pour estimer une erreur de prévision en vue des étapes de choix de modèle, puis de choix et certification de méthode.
4. Pour chacune des méthodes considérées : modèle linéaire général (gaussien, binomial ou poissonien), discrimination paramétrique (linéaire ou quadratique) ou non paramétrique, k plus proches voisins, arbre, réseau de neurones (perceptron), support vecteur machine, combinaison de modèles (bagging, boosting)...
   - estimer le modèle pour une valeur donnée d'un paramètre (ou plusieurs) de *complexité* : nombre de variables, de voisins, de feuilles, de neurones, durée de l'apprentissage, largeur de fenêtre... ;
   - optimiser ce paramètre (ou ces paramètres) en fonction de la technique d'estimation de l'erreur retenue : échantillon de validation, validation croisée, approximation par pénalisation de l'erreur d'ajustement (critères C_p, AIC).
5. Comparaison des modèles optimaux obtenus (un par méthode) par estimation de l'erreur de prévision sur l'échantillon test ou, si la présence d'un échantillon test est impossible, sur le critère de pénalisation de l'erreur (AIC d'Akaîke par exemple) s'il en existe une version pour chacune des méthodes considérées.
6. Itération éventuelle de la démarche précédente (validation croisée), si l'échantillon test est trop réduit, depuis (*iii*). Partitions aléatoires successives de l'échantillon pour moyenner sur plusieurs cas l'estimation finale de l'erreur de prévision et s'assurer de la robustesse du modèle obtenu.
7. Choix de la méthode retenue en fonction de ses capacités de prévision, de sa robustesse mais aussi, éventuellement, de l'interprétabilité du modèle obtenu.
8. Ré-estimation du modèle avec la méthode, le modèle et sa complexité optimisée à l'étape précédente sur l'ensemble des données.
9. Exploitation du modèle sur la base complète et de nouvelles données.

La conclusion de cette stratégie peut éventuellement être modifiée par la construction d'un *meilleur compromis* entre les différentes méthodes testées plutôt que de sélectionner la meilleure. Deux approches proposent cette démarche conduisant à une *collaboration* entre modèles : COBRA de Biau et al. (2013)[1] et *SuperLearner* de van der Laan et al. (2007) [7]. La première revient à exécuter une forme d'algorithme des k plus proches voisins avec une définition très particulière de la distance tandis que la deuxième cherche, par minimisation d'une estimateur d'erreur par validation croisée, une meilleure combinaison convexe des prévisions. Ces deux approches sont développées dans la vignette consacrée à l'agrégation de modèles et testé dans un exemple de données de criblage vituel de molécules.

5 Exemples

En plus des exemples "pédagogiques" illustrant simplement les différentes méthodes étudiées, d'autres exemples en "vraie grandeur" permettent d'en évaluer réellement l'efficacité mais aussi toute la complexité de mise en œuvre. D'autres exemples sont encore plus concrètement proposés en travaux dirigés ou sous formes de scénarios avec leur traitement informatique explicite.

5.1 Banque, finance, assurance : Marketing

L'objectif est une communication personnalisée et adaptée au mieux à chaque client. L'application la plus courante est la recherche d'un *score* estimé sur un échantillon de clientèle pour l'apprentissage puis extrapolé à l'ensemble en vue d'un objectif commercial :

- *Appétence* pour un nouveau produit financier : modélisation de la probabilité de posséder un bien (contrat d'assurance...) puis application à l'ensemble de la base. Les clients, pour lesquels le modèle prédit la possession de ce bien alors que ce n'est pas le cas, sont démarchés (télé marketing, publipostage ou mailing, phoning,...) prioritairement.
- *Attrition* ; même chose pour évaluer les risques de départ ou d'attrition (churn) des clients par exemple chez un opérateur de téléphonie. Les clients pour lesquels le risque prédit est le plus important reçoivent des

incitations à rester.
- *Risque* pour l'attribution d'un crédit bancaire ou l'ouverture de certains contrats d'assurance ; risque de faillite d'entreprises.
- ...

L'exemple traité suit un schéma classique d'analyse de données bancaires. Après la phase exploratoire, il s'agit de construire un score d'appétence de la carte Visa Premier dans l'idée de fidéliser les meilleurs clients. La variable à prévoir est binaire : possession ou non de cette carte en fonction des avoirs et comportements bancaires décrits par $p = 32$ variables sur $n = 825$ clients.

5.2 Environnement : pic d'ozone

L'objectif est de prévoir pour le lendemain les risques de dépassement de seuils de concentration d'ozone dans les agglomérations à partir de données observées : concentrations en 03, NO3, NO2... du jour, et d'autres prédites par Météo-France : température, vent... Encore une fois, le modèle apprend sur les dépassements observés afin de prévoir ceux à venir.

Il s'agit d'un problème de régression : la variable à prévoir est une concentration mais elle peut aussi être considérée comme binaire : dépassement ou non d'un seuil. Il y a 8 variables explicatives dont une est déjà une prévision de concentration d'ozone mais obtenue par un modèle déterministe de mécanique des fluides (équation de Navier et Stockes). L'approche statistique vient améliorer cette prévision en modélisant les erreurs et en tenant compte d'observations de concentration d'oxyde et dioxyde d'azote, de vapeur d'eau, de la prévision de la température ainsi que de la force du vent.

Cette étude est proposée en exemple ou en travaux dirigés mais pas sous la forme d'un scénario car les données propriétés de MétéoFrance ne sont pas publiques.

5.3 Santé : aide au diagnostic

Les outils statistiques sont largement utilisés dans le domaine de la santé. Ils le sont systématiquement lors des essais cliniques dans un cadre législatif stricte mais aussi lors d'études épidémiologiques pour la recherche de facteurs de risques dans des grandes bases de données ou encore pour l'aide au diagnostic. L'exemple étudié illustre ce dernier point : il s'agit de prévoir un diagnostic à partir de tests biologiques et d'examens élémentaires. Bien entendu, la variable à prédire, dont l'évaluation nécessite souvent une analyse très coûteuse voire une intervention chirurgicale, est connue sur l'échantillon nécessaire à l'estimation des modèles.

Dans l'exemple étudié (breast cancer), il s'agit de prévoir le type de la tumeur (bénigne, maligne) lors d'un cancer du sein à l'aide de $p = 9$ variables explicatives biologiques observées sur $n = 700$ patientes.

5.4 Biologie : sélection de gènes

Les techniques de microbiologie permettent de mesurer simultanément l'expression (la quantité d'ARN messager produite) de milliers de gènes dans des situations expérimentales différentes, par exemple entre des tissus sains et d'autres cancéreux. L'objectif est donc de déterminer quels gènes sont les plus susceptibles de participer aux réseaux de régulation mis en cause dans la pathologie ou autre phénomène étudié. Le problème s'énonce simplement mais révèle un redoutable niveau de complexité et pose de nouveaux défis au statisticien. En effet, contrairement aux cas précédents pour lesquels des centaines voire des milliers d'individus peuvent être observés et participer à l'apprentissage, dans le cas des biopuces, seuls quelques dizaines de tissus son analysés à cause essentiellement du prix et de la complexité d'une telle expérience. Compte tenu du nombre de gènes ou variables, le problème de discrimination est sévèrement indéterminé. D'autres approches, d'autres techniques sont nécessaires pour pallier à l'insuffisance des méthodes classiques de discrimination.

L'exemple concerne les expressions de gènes dans une expérience croisant deux facteurs le régime alimentaire (5 niveaux) chez $n = 40$ souris de 2 génotypes. Il s'agit de mettre en évidence l'impact des facteurs sur les expressions de $p = 120$ gènes puis d'expliquer un ensemble de $q = 21$ variables phénotypiques (concentrations d'acides gras) par ces mêmes expressions.

5.5 Exemples industriels

Données de spectrométrie

depuis de très nombreuses années, l'industrie agroalimentaire est confrontée à des problèmes de grande dimension pour l'analyse de données de spectrométrie comme par exemple dans le proche infra-rouge (NIR). Sous l'ap-

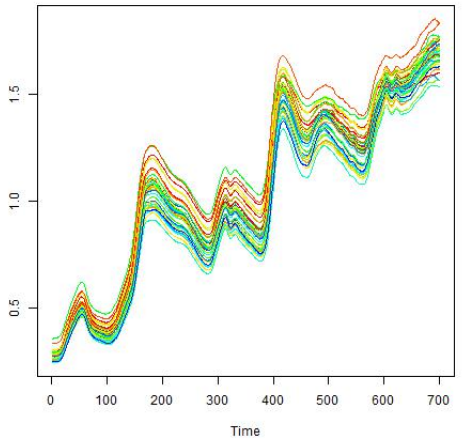

FIGURE 3 – Cookies : Spectres proche infrarouge (NIR) d'échantillons de pâtes à gâteaux. La couleur dépend du taux de sucre.

pellation de *Chimiométrie* de très nombreuses méthodes et stratégies ont été développées ou enrichies (*i.e.* la régression PLS) afin de prendre en compte la spécificité des problèmes rencontrés par la discrétisation de spectres conduisant très généralement à un nombre de variables $p > n$. Dans un premier exemples, il s'agit de modéliser, la teneur en sucre d'une pâte à gâteau (cookies où $n = 72, p = 700$)) à partir des spectres (cf. figure 3) tandis que dans un deuxième (Tecator ou $n = 215, p = 100$), c'est la teneur en matière grasse qui est recherchée. Ces questions sont considérées comme des problèmes de *calibration* d'un appareil de mesure (le spectromètre) pour arriver à la quantification d'une mesure chimique dont l'évaluation classique est beaucoup plus coûteuse ou encore destructive.

Criblage virtuel de molécules

Une stratégie classique de l'industrie pharmaceutique consiste à tester *in silico* un nombre considérable de molécules avant de ne synthétiser que celles jugées intéressantes pour passer aux étapes de recherche clinique *in vitro* puis *in vivo*. Une propriété thérapeutique d'un ensemble de molécules d'apprentissage (perméabilité de la paroi intestinale ou à la barrière sanguine du cerveau, adéquation à une cible donnée...) étant connue, un grand ensemble de caractéristiques physico-chimiques sont évaluées, calculées par un logiciel spécifique : ce sont des données dites QSAR *Quantitative structure-activity relationship*. S'il est possible de raisonnablement prévoir la propriété thérapeutique à partir des caractéristiques physico-chimiques, ce modèle est systématiquement appliqué à un grand ensemble de molécules virtuelles ; c'est le criblage ou *sreening* virtuel de molécule. Deux jeux de données sont étudiés l'un illustrant un problème de régression (blood brain barrier data) avec $n = 208, p = 134$ tandis que l'autre est un problème de discrimination à deux classes (multidrig resistance reversal) avec $n = 528, p = 342$.

D'autres exemples sont cités à titre illustratif mais leur complexité, inhérente à beaucoup de problèmes industriels, ne permet pas de les détailler à des fins pédagogiques.

Industrie des semi-conducteurs : Détection de défaillance

Un procédé de fabrication de microprocesseurs comporte des centaines d'étapes (photogravures, dépôts, cuissons, polissages, lavages...) dont tous les paramètres, équipement et mesures physiques (températures, pressions...), sont enregistrés dans une grande base de données permettant la traçabilité des produits manufacturés. Le test électrique de chaque microprocesseur ne peut se faire qu'en fin de fabrication lorsque ceux-ci sont achevés. Il est évidement important de pouvoir déterminer, lors de l'apparition d'une baisse du rendement et en utilisant les données de la base, l'équipement ou la fourniture responsable de la défaillance afin d'y remédier le plus rapidement possible.

Airbus : Aide au pilotage

Les graphes de la figure 4 tracent les enregistrements des commandes et positions d'un avion en vol. Ceux-ci mettent en évidence un phénomène de résonance entre l'appareil et le comportement du pilote qui est très dangereux

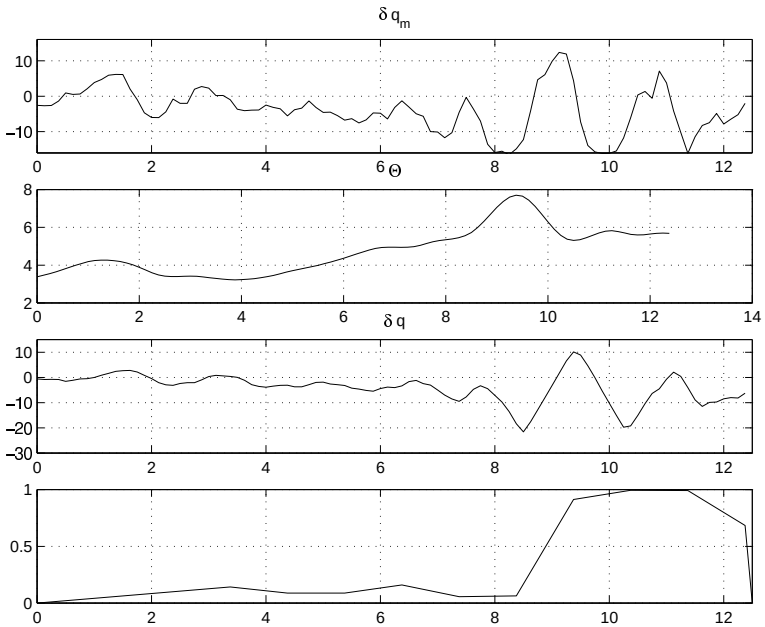

FIGURE 4 – Airbus : Pompage piloté révélé par l'observation des paramètres en temps réel. De (haut en bas) : manche, assiette, gouverne comparer avec la prévision qu'en fait un réseau de neurones.

pour la sécurité. L'objectif est de construire un modèle susceptible, en temps réel, de détecter une telle situation afin d'y remédier par exemple en durcissant les commandes de vol électriques. Le problème est très spécifique car les données, ou signaux, sont mesurées en temps réel et constituent des discrétisations de courbes.

6 Contenu

Il a fallu faire des choix dans l'ensemble des techniques proposées et leurs nombreux avatars. La forme et le contenu sont guidés par les besoins exprimés lors des stages réalisées par les étudiants du département Génie Mathématique de l'INSA ou par les thèmes des collaborations industrielles et scientifiques de l'équipe Statistique et Probabilités de l'Institut de Mathématiques de Toulouse. Le lecteur peut se faire une idée du nombre très important de méthodes et variantes concernées par l'apprentissage supervisée en consultant une aide en ligne de la librairie caret (Kuhn, 2008)[4] du logiciel R. Remarquons que les principaux logiciels commerciaux (SAS, Splus, SPSS, Matlab, KXEN, SPAD, Statsoft. . .) ou gratuits (R, Weka, Tanagra), performants et s'imposant par des interfaces très conviviales (Enterprise Miner, Insightfull Miner, Clementine, Statistica Data Miner), contribuent largement à la diffusion, voire la pénétration, de méthodes très sophistiquées dans des milieux qui seraient imperméables à une conceptualisation mathématique trop abstraite.

Chaque méthode ou famille de méthodes de modélisation et d'apprentissage parmi les plus répandues, est présentée de façon plus ou moins succincte dans un chapitre distinct avec un objectif de prévision. Une première vignette incontournable est consacrée aux techniques d'estimation d'une erreur de prévision ou d'un *risque* sur lesquelles reposent les choix opérationnels décisifs : de modèle, de méthode mais aussi l'évaluation de la précision des résultats escomptés. La régression linéaire classique en statistique prend une place particulière à titre pédagogique. Très antérieure aux autres, elle donne lieu a une bibliographie abondante. Conceptuellement plus simple, elle permet d'introduire plus facilement les problématiques rencontrées comme celle du choix d'un modèle par ses deux approches types : la sélection de variable ou la régularisation (*ridge*, Lasso). Le modèle linéaire général fournit le cadre théorique nécessaire à l'unification des régressions linéaire, loglinéaire et logistique ; cette dernière reste toujours très utilisée en scoring. La présentation de l'analyse dis-

criminante décisionnelle, paramétrique ou non paramétrique (dont les k plus proches voisins), permet d'introduire également des notions de théorie bayésienne de la décision. Les vignettes suivantes sont consacrées aux techniques algorithmiques : arbres binaires de décision (*classification and regression trees* ou CART) et à celles plus directement issues de la théorie de l'apprentissage machine (*machine learning*) : réseau de neurones et perceptron, agrégation de modèles (*boosting, random forest*), support vector machine (SVM). Enfin une vignette de conclusion tâche de synthétiser le panorama et propose une comparaison systématique des méthodes sur les différents jeux de données.

Le choix a été fait de conserver et expliciter, dans la mesure du possible, les concepts originaux de chaque méthode dans son cadre disciplinaire tout en tâchant d'homogénéiser notations et terminologies. L'objectif principal est de faciliter la compréhension et l'interprétation des techniques des principaux logiciels pour en faciliter une *utilisation pertinente et réfléchie*. Ce cours ne peut être dissocié de séances de travaux dirigés sur ordinateur à l'aide de logiciels (SAS, R...) pour traiter des données en vraie grandeur dans toute leur complexité. La principale difficulté pratique est d'arriver à déterminer où faire porter l'effort ou les efforts :

- la saisie, la gestion, la sélection des données et variables,
- la sélection des méthodes à comparer,
- l'optimisation des choix de modèles,

et ceci en fonction des méthodes considérées, de la structure des données, des propriétés des variables notamment celle à modéliser.

Références

[1] G. Biau, A. Ficher, B. Guedj et J. D. Malley, *COBRA : A Nonlinear Aggregation Strategy*, (2013), http://arxiv.org/abs/1303.2236.

[2] U. M. Fayyad, G. Piatetsky-Shapiro et P. Smyth, *From data mining to knowledge discovery : an overview*, Advances in Knowledge Discovery and Data Mining (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth et R. Uthurusamy, réds.), AAAI Press/MIT Press, 1996, p. 1–34.

[3] T. Hastie, R. Tibshirani et J Friedman, *The elements of statistical learning : data mining, inference, and prediction*, Springer, 2009, Second edition.

[4] Max Kuhn, *Building Predictive Models in R Using the caret Package*, Journal of Statistical Software **28** (2008), n° 5.

[5] K.V. Mardia, J.T. Kent et J.M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

[6] S. Tufféry, *Data Mining et Statistique décisionnelle : l'intelligence des données*, Technip, 2007.

[7] M. J. van der Laan, E. C. Polley et A. E. Hubbard, *Super learner*, Statistical Applications in Genetics and Molecular Biology **6 :1** (2007).