Apprentissage Machine / Statistique

Résumé

Statistique, fouille ou Science des Données, les appellations changent le volume et la diversité des données explosent, les technologies se succèdent, les modèles et algorithmes se complexifient. L'estimation devient un apprentissage, la prévision remplace l'explication.

Le parcours pour devenir data scientist *est structuré en quatre parties :*

Retour *à l'introduction générale*

Saison 1 *(L3) Statistique élémentaire, descriptive* vs. *inférentielle.*

Saison 2 *(M1) Statistique Exploratoire multidimensionnelle et apprentissage non supervisé.*

Saison 3 *Apprentissage Statistique / Machine supervisé.*

Saison 4 *(M2) Technologies pour la Science des (grosses) Données.*

plus des réflexions sur : Statistique et Déontologie scientifique.

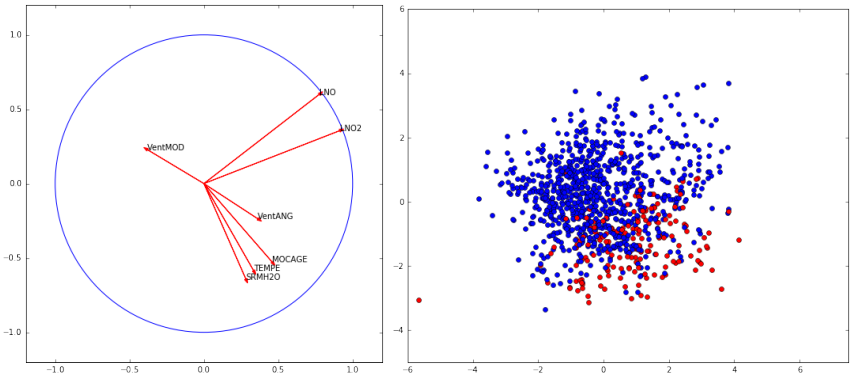

FIGURE 1 – *Ozone : Préliminaire à la prévision par adaptation statistique d'une prévision déterministe. Premier plan de l'analyse en composantes principales (47% de variance expliquée). En rouge, les jours à prévoir de dépassement du seuil critique.*

1 Introduction

1.1 Objectifs de l'apprentissage

Questions ?

Identifier les facteurs aggravants de certains types de cancer en fonction de variables cliniques et démographiques, rechercher des gènes potentiellement impliqués dans une maladie à partir de données de séquençage ou, plus généralement, des bio-marqueurs pour un diagnostic précoce, identifier des chiffres manuscrits de codes issus d'images digitalisées, prévoir un taux de pollution atmosphérique en fonction de conditions météorologiques (cf. figure 1), établir des scores d'appétence ou d'attrition en gestion de la relation client (GRC), construire des méta-modèles ou modèles de substitution à un code numérique trop complexe pour analyser la sensibilité aux paramètres, détecter ou mieux prévoir les défaillances d'un procédé... sont autant d'exemples où l'objectif est de minimiser une *erreur de prévision* ou *risque*. C'est encore la recherche d'un *modèle* plus généralement d'une *méthode optimale* au sens d'un critère à définir précisément.

Parallèlement, les méthodes et algorithmes issus de l'Intelligence Artificielle (*e.g.*réseaux de neurones) se focalisaient sur le même objectif pour devenir l'*Apprentissage Machine* incluant les méthodes et modèles de l'*apprentissage statistique*. La notion de d'apprentissage statistique (*statistical learning*) a été introduite par Vapnik (1998)[**?**] et popularisée par Hastie et al.(2001)[5].

Les choix de méthodes, de modèles, sont complexes à opérer et se déclinent en sous-objectifs qui restreignent où précisent les classes de modèles à considérer. L'objectif est-il seulement *prédictif* ? Sous-entendu, un modèle *boîte noire* suffit-il à répondre aux besoins sans interprétation détaillée ? En revanche, une compréhension du modèle, donc de l'impact des variables, attributs ou facteurs, est-elle recherchée voire indispensable ? Ou encore, plus précisément, est-ce la détermination d'un petit sous-ensemble de ces variables

(*e.g.* des biomarqueurs) qui est recherchée pour rendre opérationnelle une prévision suffisamment précise et peu coûteuse ?

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-clés.

Objectif

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.

Rien n'empêche de construire et comparer tous types de modèles, qu'ils soient interprétatifs ou non, avec sélection de variables ou non ; les approches sont complémentaires. Compréhension préalables des données et connaissance des modèles, performances des prévisions, majoration ou contrôle des erreurs, efficacité algorithmique, sont autant de considérations à prendre en compte. Plus précisément les compétences issues des deux champs disciplinaires concernés sont nécessaires pour atteindre le but visé.

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, le sous-ensemble de l'*apprentissage machine* nommé *apprentissage statistique* (*statisticial learning*) propose un ensemble de méthodes et algorithmes pertinents car efficaces. 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. La fouille et maintenant la science des données se focalisent sur des pratiques, méthodes ou algorithmes dont Hastie et al. (2009)[5] proposent un tour d'horizon assez exhaustif.

Buts

L'objectif est bien de minimiser une erreur de prévision mais dans quel contexte ou pour quel but ? Schématiquement, s'agit-il de faire accepter un article dans une revue académique (Statistique, Apprentissage Machine, Bioinformatique...) ou de développer une solution "industrielle" (commerce électronique, détection de fraude ou de défaillance,...) ou encore de gagner un concours de prévision de type *Netflix* ou *Kaggle*. Le même objectif de minimisation d'une erreur de prévision peut alors conduire à des solutions radicalement différentes. La publication d'une nouvelle méthode d'apprentissage ou de nouvelles options de méthodes existantes nécessite de montrer qu'elle surpasse ses concurrentes sur une batterie d'exemples, généralement issus du site hébergé à l'Université de Californie Irvine (*UCI Repository*[6]). Les biais inhérents à cette démarche sont discutés dans de nombreux articles (*e.g.* Hand ; 2006)[4] et conférences (*e.g.* Donoho ; 2015)[3]. Il est notable que la pression académique de publication a provoqué une explosion du nombre de méthodes et de leurs variantes, alors que celles-ci peuvent conduire à des différences de performances peu ou pas significatives.

L'analyse du déroulement des concours de type *Kaggle* et de leurs solutions gagnantes est très instructive. La pression, donc le biais, est tout à fait différent. Il conduit à des combinaisons, voire architecture de modèles, d'une telle complexité (cf. *e.g.* figure 2) que ces solutions sont concrètement inexploitables pour des différences de performances minimes (3 ou 4ème décimale).

En effet, surtout si les données sont en plus volumineuses (cf. saison 4), les solutions opérationnelles et "industrialisées", nécessairement robustes et rapides, se contentent souvent d'outils méthodologiques assez rudimentaires et peu *glamours* dirait Donoho (2015)[3].

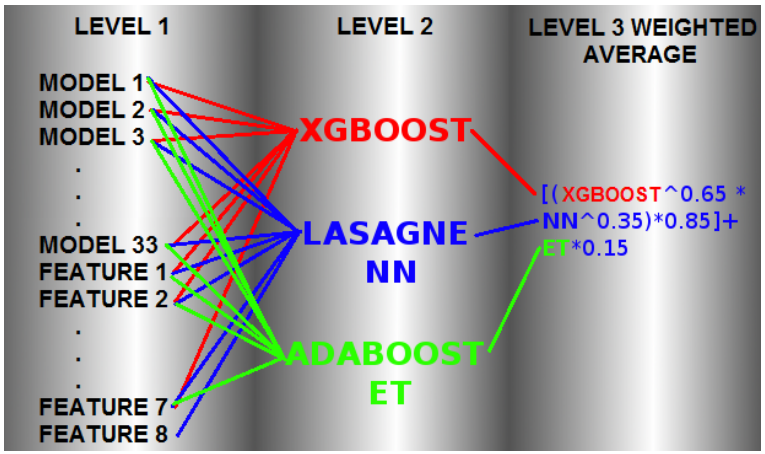

FIGURE 2 – *Solution gagnante d'un concours kaggle : Identify people who have a high degree of Psychopathy based on Twitter usage. Combinaison pondérée des combinaisons (boosting, réseaux de neurones) de trente trois modélisations (random forest, boosting, k plus proches voisins...) et 8 nouvelles variables* (features) *ad'hoc.*

Cette saison propose d'aborder la grande variété des critères et méthodes proposés, leurs conditions de mise en œuvre, les choix à opérer, notamment pour optimiser la complexité des modèles. C'est aussi l'occasion de rappeler que des méthodes robustes et linéaires ainsi que les stratégies anciennes (descendantes, ascendantes, pas-à-pas) ou plus récentes (lasso) de sélection de modèles linéaires ou polynomiaux ne doivent pas être trop rapidement évacuées des pratiques académiques ou industrielles.

1.2 Définitions

Apprentissage 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.

$$Y = \widehat{f}(X) + \varepsilon$$

où ε représente 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 algorithmes de classification ascendante hiérarchique, de réallocation dynamique ($k$means), DBSCAN ou encore des cartes auto-organisatrices (Kohonen) et bien d'autres...

Les algorithmes d'apprentissage non supervisée sont abordés dans la saison 2.

Les données

Cette saison 3 est consacrée à l'apprentissage supervisé, pour lequel on dispose d'un *ensemble d'apprentissage* constitué de données d'observations de type entrée-sortie.

Dans tous les problèmes rencontrés, des caractéristiques, attributs (*features*), facteurs ou variables $X = (X^{(1)}, \ldots, X^{(p)})$ dites explicatives ou prédictives ont été observées sur un ensemble de n objets, individus, *instances* ou unités statistiques. $d_1^n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ avec $\mathbf{x}_i \in \mathcal{X}$ quelconque (souvent égal à $\mathbb{R}^p$), $y_i \in \mathcal{Y}$ pour $i = 1 \ldots n$.

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.

Discrimination vs. *régression*

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) X. La sortie Y peut être quantitative (prix d'un stock, consommation électrique, carte de pollution ..) ou qualitative (survenue d'un cancer, reconnaissance de chiffres...) selon l'espace dans lequel elle prend ses valeurs : ensemble de cardinal fini ou réelles voire fonctionnelles. Certaines méthodes d'apprentissage ou de modélisation s'adaptent à tout type de variables explicatives tandis que d'autres sont spécialisées. 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. Certaines méthodes sont spécifiques (régression linéaire, analyse discriminante) à un type de variable à modéliser tandis que d'autres s'adaptent sans modification profonde remettant en cause leur principe (réseaux de neurones, arbres de décision...).

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

Régression *réelle* lorsque $\mathcal{Y} \subset \mathbb{R}$ et discrimination *binaire* lorsque $\mathcal{Y} = \{0, 1\}$ ou $\{-1, 1\}$.

Estimation vs. *apprentissage*

Tout au long de ce document, les termes d'*estimation* et d'*apprentissage* sont utilisés comme des synonymes mais ceci nécessite de préciser quelques nuances. 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)[8] 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 3) des Shadoks n'est pas une référence à suivre en apprentissage statistique !

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

FIGURE 3 – Deuxième devise Shadok

important, d'autre méthodes viennent concurrencer les modèles statistiques rudimentaires.

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. Le souci de calculabilité l'emporte sur la complexité mathématique du modèle et conduit à l'optimisation d'un critère d'ajustement de la fonction f sur un ensemble de solutions plus ou moins riche. Les méthodes développées dans la communauté d de l'apprentissage machine : 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.

2 Stratégies de choix

2.1 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.

2.2 Choix de modèle : équilibre biais-variance

Le point central est la construction de *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 certains algorithmes de combinaison de modèles (*e.g. bagging, random forest*) 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

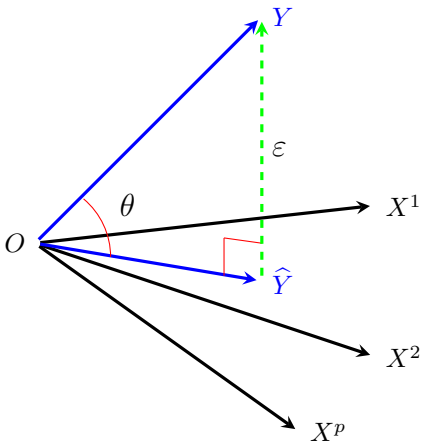

FIGURE 4 – *Régression linéaire par projection $\widehat{Y}$ de Y sur l'espace vectoriel Vect$\{\mathbf{1}, X^1, \ldots, X^p\}$*

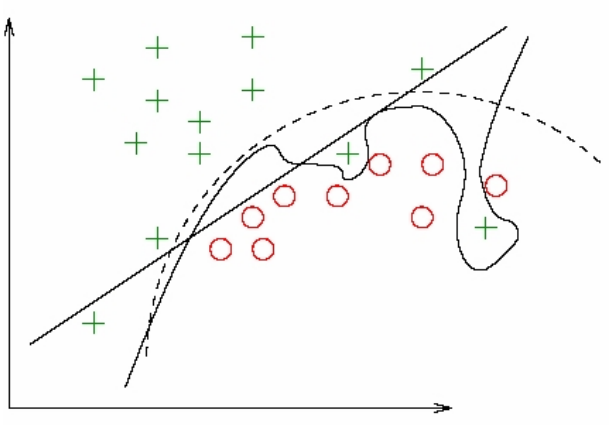

FIGURE 5 – *Sous-ajustement linéaire et sur-ajustement local (proches voisins) d'un modèle quadratique.*

de s'appliquer à des données qui n'ont pas participé à son estimation.

Un exemple élémentaire illustre ce point fondamental dans le cas d'un problème de régression (figure 4) : la qualité d'ajustement du modèle (R^2) augmente nécessairement avec le nombre de variables mais la variance des estimateurs peut exploser dès que la matrice à inverser ($\mathbf{X}'\mathbf{X}$) devient mal conditionnée. Dans un problème de discrimination dans $\mathbb{R}^2$ (figure 4) : une frontière dont le modèle "vrai" est quadratique est, par exemple 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 avec k petit.

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 conséquence, 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.

2.3 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. Ce principe de régularisation par pénalisation a été transposée à la plupart des méthodes connues.

3 Stratégie de l'apprentissage statistique

3.1 Les étapes d'une analyse

En situation réelle, la préparation initiale des données, simple nettoyage (*cleaning*, points 1 et 2 ci-dessous) ou trafic (*munging, wrangling*) plus complexe : extraction, nettoyage, vérification, imputation éventuelle de données manquantes, transformations... est la phase la plus ingrate, celle qui demande le plus de temps, de ressources humaines et des compétences très variées : informatique, statistique et métier. Ne nécessitant pas de développements théoriques majeures mais plutôt beaucoup de bon sens, d'expérience et une bonne connaissance des données du métier, cette préparation est négligée dans les travaux académiques généralement illustrés, par souci de concision, sur des données déjà pré-traitées donc nettoyées. Une fois bien menée à terme, la phase de modélisation ou apprentissage en découle finalement assez automatiquement même pour des méthodes et algorithmes sophistiqués.

De façon systématique et aussi très schématique, l'analyse, maintenant la *Science, des Données* enchaînent les étapes suivantes pour la plupart des domaines d'application. On retrouve l'essentiel des items des 6 divisions de Donoho (2015)[3] avec un découpage différent.

1. Extraction des données avec ou sans échantillonnage faisant référence à des techniques de sondage appliquées ou applicables à des bases de données structurées (SQL) ou pas (NOSQL).
2. Visualisation, exploration des données pour la détection de valeurs atypiques, incohérences, erreurs ou anomalies ; étude des distributions et des structures de corrélation et recherche de transformations des variables, construction de nouvelles variables et/ou représentation dans des bases (Fourier, spline, ondelettes...) adaptées, recherche de typologies des observations...
3. Prise en compte, par simple suppression, par imputation ou non, des données manquantes.
4. Partition aléatoire de l'échantillon (apprentissage, validation, test) en fonction de sa taille et choix d'une fonction perte ou critère qui seront utilisées pour estimer une erreur de prévision en vue des étapes d'optimisation de modèle, puis de choix de méthode.
5. Pour chacune des méthodes considérées : modèle linéaire général (gaussien, binomial ou poissonien), discrimination paramétrique (linéaire ou quadratique) ou non paramétrique (k plus proches voisins), réseau de neurones (perceptron), arbre binaire de décision, machine à vecteurs supports, combinaison de modèles (*bagging, boosting, random forest*...
   - estimer le modèle pour une valeur donnée d'un paramètre (ou plusieurs) de *complexité* : nombre de variables, de voisins, de feuilles, de neurones, pénalisation ou régularisation... ;
   - optimiser ce paramètre (ou ces paramètres) en fonction de la technique d'estimation de l'erreur retenue : critère de pénalisation, échantillon de validation, validation croisée...
6. Comparaison des modèles optimaux obtenus (un par méthode) par estimation de l'erreur de prévision sur l'échantillon test.
7. Itération éventuelle de la démarche précédente ou validation croisée *Monte Carlo*, si l'échantillon test est trop réduit, depuis (*4*). Partitions aléatoires successives de l'échantillon (apprentissage et test) pour étudier la distribution des erreurs et moyenner sur plusieurs cas l'estimation finale de l'erreur de prévision et s'assurer de la robustesse du modèle obtenu.
8. Choix de la méthode retenue en fonction de ses capacités de prévision, de sa robustesse mais aussi, éventuellement, de l'interprétabilité recherchée du modèle.

9. Ré-estimation du modèle avec la méthode, le modèle et sa complexité optimisée à l'étape précédente sur l'ensemble des données.
10. Mise en exploitation sur la base complète ou de nouvelles données.

La fin de cette démarche peut être modifiée par la construction d'un *meilleur compromis* ou d'une combinaison des différentes méthodes testées plutôt que de sélectionner la meilleure. C'est souvent le cas des solutions gagnantes "usines à gaz" des concours de type *Kaggle*. Cela a aussi été théorisé en deux approches conduisant à une *collaboration* entre modèles : COBRA de Biau et al. (2016)[2] 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 spécifique de la distance tandis que la deuxième cherche, par minimisation d'une estimation de d'erreur par validation croisée, une meilleure combinaison convexe des prévisions.

L'une des principales difficultés pratiques est d'arriver à déterminer où faire porter l'effort ou les efforts :

- la saisie, la gestion, la sélection des données et variables ; la préparation des données est de toute façon essentielle ;
- la sélection des méthodes à comparer ;
- l'optimisation des choix de modèles ;

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

3.2 Les méthodes

Chaque méthode ou famille de méthodes de modélisation et d'apprentissage parmi les plus répandues, est présentée de façon plus ou moins succincte dans une vignette distincte 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 lequel reposent les choix opérationnels décisifs : de modèle, de méthode mais aussi l'évaluation de la précision des résultats escomptés.
- La régression linéaire ou modèle gaussien, classique en statistique, donne lieu a une bibliographie abondante. Conceptuellement plus simple, elle permet d'introduire plus facilement les questions rencontrées comme celle du choix d'un modèle selon les deux approches types : *sélection de variable* ou *pénalisation* (*ridge, Lasso*).
- La régression PLS propose une autre solution de choix de modèle par projection sur une base de facteurs orthogonaux avec une versions *sparse* ou parcimonieuse pour simplifier l'interprétation.
- Toujours dans le cadre du modèle linéaire général, la régression logistique ou modèle binomial reste toujours très utilisée même et surtout lorsque les données sont massives.
- La présentation de l'analyse discriminante décisionnelle, paramétrique ou non paramétrique (dont les k plus proches voisins), permet d'introduire également des notions de théorie bayésienne de la décision.
- La vignette suivante est consacrée aux arbres binaires de décision (*classification and regression trees* ou CART)
- puis à des algorithmes plus directement issues de la théorie de l'apprentissage machine : réseau de neurones (perceptron).
- Viennent ensuite les méthodes d'agrégation de modèles (*boosting, random forest*),
- de support vector machine (SVM).
- imputation de données manquantes.
- Détection d'anomalies ou observation atypiques applicable à la détection de défaillances, fraudes...
- avant de conclure par une présentation synthétique des différentes méthodes exposées ainsi que des considérations éthiques sur la loyauté des décisions algorithmiques.

3.3 Tutoriels : cas d'usage et "fil rouge"

Cette saison accorde beaucoup d'importance aux exemples pour illustrer le comportement des critères généralement utilisés et analyser les performances des méthodes étudiées tout en se formant à leur usage. 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.

L'analyse de ces différents cas d'usage se présente sous la forme de tutoriels contenus dans des calepins (*jupyter notebook*) en R ou Python. Ils sont ou seront disponibles dans le dépôt `github.com/wikistat`

Fil Rouge

Exécuter le tutoriel (calepin) : Prévision du pic d'ozone en R et Python en se référent aux descriptifs des méthodes autant que de nécesaire. L'exécution des tutoriels est découpé en épisodes. Cet exemple étudié par Besse et al. (2007)[1] est une situation réelle dont l'objectif est de prévoir, pour le lendemain, les risques de dépassement du seuil légal de concentration d'ozone dans des agglomérations. Le problème peut être considéré comme un cas de régression : la variable à prévoir est une concentration en ozone, mais également comme une discrimination binaire : dépassement ou non du seuil légal. Il n'y a que 8 variables explicatives dont une est déjà une prévision de concentration d'ozone mais obtenue par un modèle déterministe de mécanique des fluides (équations de Navier et Stockes). Il s'agit d'un exemple d'*adaptation statistique*. La prévision déterministe sur la base d'un maillage global (30 km) est améliorée localement, à l'échelle d'une ville, par un modèle statistique incluant cette prévision ainsi que des informations connues sur la base d'une grille locale, spatiale et temporelle plus fine : concentration d'oxyde et dioxyde d'azote, de vapeur d'eau, température, vitesse et direction du vent.

Cet exemple, à la fois de régression et de discrimination, présente des vertus pédagogiques certaines qui permettent de l'utiliser comme *fil rouge* de comparaison entre toutes méthodes. L'étude préliminaire rudimentaire a conduit à la transformation (log) de certaines variables de concentration. Les données sont résumées par leur représentation dans le premier plan de l'analyse en composantes principales réduite (cf. figure 1). Ce graphique résume la structure de corrélation assez intuitives des variables et met en évidence les difficultés à venir pour discriminer les deux classes : présence ou non d'un pic de concentration avec dépassement du seuil légal.

Exemples jouet

Jeux de données élémentaires dans dans $\mathbb{R}^2$. Exécuter les calepins en faisant varier les paramètres de complexité des modèles. Discrimination en R de mélanges gaussiens. Discrimination en Python de mélanges de points ou *blobs*.

Diagnostic d'une maladie coronarienne

Cas d'usage : Diagnostic d'une maladie coronarienne en R avec initiation à l'utilisation du package xgboost.

Spectrographie en proche infa-rouge (NIR)

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'appellation 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 6) 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.

QSAR : criblage virtuel de molécule

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 $n = 528, p = 342$.

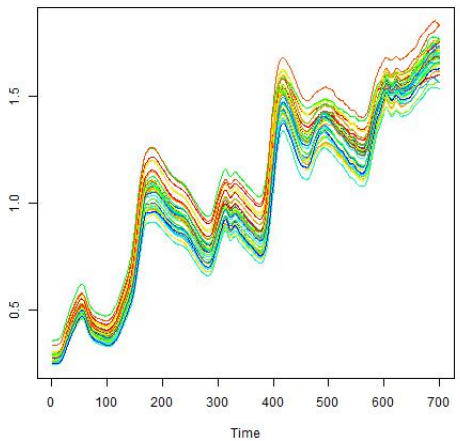

FIGURE 6 – Cookies : Spectres proche infrarouge (NIR) d'échantillons de pâtes à gâteaux. La couleur dépend du taux de sucre.

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.

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.
- ...

Des jeux de données sont abordés, le premier construit un score d'appétence de la carte visa premier. Le deuxième analyse une enquête de l'INSEE sur les patrimoines des français pour l'estimation d'un score d'appétence des produits d'assurance vie.

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.

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.

Adult census

Étude d'une enquête aux USA et prévision du dépassement d'un seuil de revenu.

Détection de pourriel

Les données sont composées d'un ensemble de messages ou courriels dont certains sont identifiés comme pourriels ou *spams*. Les données ont déjà été préparées ou simpifiées, les variables ou *features* sont les effectifs ou présence / absence de certains mots ou caractères spécifiques (`$, !` ...). L'objectif est d'apprendre à détecter un pourriel.

Références

[1] P. Besse, H. Milhem, O. Mestre, A. Dufour et V. H. Peuch, *Comparaison de techniques de Data Mining pour l'adaptation statistique des prévisions d'ozone du modèle de chimie-transport MOCAGE*, Pollution Atmosphérique **195** (2007), 285–292.

[2] G. Biau, A. Ficher, B. Guedj et J. D. Malley, *COBRA : A Nonlinear Aggregation Strategy*, Journal of Multivariate Analysis **146** (2016), 18–28.

[3] David Donoho, *50 years of Data Science*, Princeton NJ, Tukey Centennial Workshop, 2015.

[4] David J. Hand, *Classifier Technology and the Illusion of Progress*, Statist. Sci. **21** (2006), n° 1, 1–14.

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

[6] M. Lichman, *UCI Machine Learning Repository*, 2013, http://archive.ics.uci.edu/ml.

[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).

[8] V.N. Vapnik, *Statistical learning theory*, Wiley Inter science, 1999.