Statistique en Grande Dimension et Apprentissage

Fabien Panloup

Notes de Cours

Master 2 Data Science - Université d'Angers

Année universitaire 2019-2020

Introduction

Ce cours est consacré à la présentation des méthodes d'apprentissage statistique en général et à leur mise en pratique de la grande dimension, *i.e* dans le cadre où le nombre de paramètres du problème et grand. Sur le premier point, on s'intéressera dans un premier temps à quelques propriétés théoriques et aux fondements permettant d'estimer le *risque* relatif à un problème statistique donné et on tentera de présenter plusieurs familles d'algorithmes fondamentales.
Ce cours venant à la suite de celui de "Data Mining" de M1, il se focalisera principalement sur les algorithmes non traités dans ce cours, et en particulier, sur ceux qui sont spécifiques à la grande dimension.

N.B. Il s'agit pour l'instant d'une version conçue essentiellement à l'aide d'une concaténation des slides de cours. En particulier, le niveau de détail, la mise en page et la rigueur n'en font clairement pas un support suffisant pour l'instant. On n'hésitera donc pas à le compléter à l'aide des références indiquées en fin de polycopié.

Pour terminer cette introduction, voici un extrait d'article de vulgarisation écrit pour Angers Mag en 2017 (F. P.) qui fait une présentation générale des enjeux du Big Data et du point de vue mathématique sur ce sujet :

" Big Data", "Data Science", "Machine Learning", "Deep Learning",... sont des expressions qui, au coeur de ladite " Révolution Numérique ", viennent de manière de plus en plus prégnante, occuper le paysage médiatique. Mais de quelle science s'agit-il réellement et quel rôle jouent les maths dans tout ça ? La terminologie " Big Data " est généralement traduite en français par " données massives ". La " Data Science " désigne quant à elle, la (ou l'ensemble des) discipline(s) scientifique(s) permettant de " traiter " ces masses de données.

Au point zéro, on imagine un immense tableau constitué de données diverses, numériques ou non, stocké dans des serveurs conçus à cet effet. Au point un, on entrevoit la fabrication d'algorithmes permettant d'" apprendre " cet immense jeu de données. A ce stade, il est important de comprendre que le terme " Big Data " n'est pas seulement associé au nombre d'observations mais aussi et surtout au nombre de paramètres mis en jeu dans le proble?me (impliquant une " grande dimension ") ou dit plus simplement, à la complexité du phénomène considéré.

On peut penser à des situations diverses telles que la propagation d'une épidémie, le développement de certaines maladies comme le cancer par exemple, pour lesquelles l'efficacité du traitement thérapeutique peut dépendre d'un grand nombre de facteurs génétiques ou environnementaux, la tentative de prédiction de catastrophes naturelles ou de l'évolution économique, la reconnaissance d'images, la fabrication de logiciels de traduction ou encore, pour citer un exemple qui a récemment eu un retentissement médiatique important, la modélisation du jeu de Go. A des fins potentiellement plus " mercantiles ", on peut enfin citer le e-commerce et l'exploitation des données clients...

Le rôle de la " Science des Données " est alors en résumé de concevoir, d'analyser et de mettre en oeuvre des algorithmes à même de résoudre numériquement les problèmes donnés. Celle-ci implique généralement l'interaction entre plusieurs disciplines scientifiques : le domaine d'application concerné, l'informatique et les mathématiques. En ce qui concerne les deux dernières, on peut citer certains thèmes de recherche associés à la science des données tels que les intelligences artificielles, le calcul haute performance, le traitement du signal, l'algorithmique, la théorie des graphes, l'optimisation, les probabilités et bien sûr les statistiques.

Si les thèmes ci-dessus peuvent faire écho au lecteur, elles n'expliquent en revanche pas clairement le rôle du mathématicien dans la science des données. On pourrait tenter de le résumer ainsi. Le mathématicien apporte sa pierre à l'édifice en formalisant les problèmes issus de la modélisation et en tentant d'apporter des réponses théoriques générales relatives au problème posé.

D'un point de vue probabilités et statistiques, il s'agit de proposer et d'étudier des modèles aléatoires adaptés au problème donné puis d'évaluer la qualité de la prédiction obtenue en fonction du nombre de données et du nombre de variables estimées. Comme on peut l'imaginer, ces sciences ne sont pas nées avec le " Big Data ". Les fondements statistiques restent d'ailleurs les mêmes que dans un cadre " classique ". Néanmoins, l'évolution rapide de la taille des jeux de données a impliqué un certain nombre de questions nouvelles dues à la " grande dimension " du problème (ou plus simplement par la complexité de la réponse attendue).

D'un point de vue " optimisation", le mathématicien se concentre plutôt sur la proposition et l'étude d'algorithmes (souvent aléatoires) conçus pour calculer numériquement les prédictions statistiques. Plus précisément, l'approximation des objets statistiques associés à la prédiction peut être numériquement non accessible en raison de la dimension du problème et il est donc primordial de proposer des résultats théoriques permettant d'évaluer la qualité des réponses algorithmiques après un nombre fixé d'itérations. Enfin, un autre défi du mathématicien consiste à tenter de mesurer l'efficacité des modèles d'apprentissage. On pense par exemple au " Deep Learning " dont le principe issu du domaine de l'intelligence artificielle consiste à modéliser un problème par un " réseau de neurones" artificiel. Ce dernier a donné des réponses étonnantes en reconnaissance d'images et est à l'origine du logiciel Alphago, ayant vaincu un des meilleurs joueurs de Go du monde. Néanmoins, son efficacité semble encore mal comprise mathématiquement...

Table des matières

Chapitre 1

Introduction à l'Apprentissage Statistique

1.1 Introduction

1.1.1 Généralités

L'apprentissage statistique (machine learning en anglais) désigne la science dédiée à l'exploitation des données issues d'un phénomène aléatoire. Le terme "exploitation" peut avoir plusieurs sens. On peut chercher à

- Décrire un phénomène : explorer/vérifier/décrire les relations entre les différentes variables au vu des observations
- Expliquer :Tester l'influence d'une variable ou d'un ou plusieurs facteurs dans un modèle supposé connu a priori.
- Prédire : Prévoir un résultat, une réponse pour une nouvelle observation.
- Sélectionner les variables qui sont les plus influentes sur le phénomène
- Classer des individus ou des variables,...

Néanmoins, les objectifs généraux précédents et la définition elle-même ne permettent pas réellement de distinguer le terme "apprentissage" qui met en avant le caractère automatique et algorithmique de l'exploitation des données. Quelques tentatives de définitions du machine learning en vrac :

- "Field of Study that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959).
- "The goal of machine learning is to build computer systems that can adapt and learn from their experience" (Diettetiche, plus récent)

- "A computer program is said to learn from experience E with some respect to some class of tasks T and performance measure P if its performance at tasks in T as measured by P improves with experience E (Mitchell, plus récent).

Exemples

Afin d'illustrer les différents objectifs décrits ci-dessus, voici quelques situations diverses :

- Identifier les mécanismes de résistance à un traitement du cancer/Sélectionner le meilleur traitement au vu des données du patient/Prévoir sa réaction au traitement
- Identifier les gènes impliqués dans le développement d'une maladie
- Reconnaître des images (chiffres/formes/nodules,...)
- Reconnaître un spam
- Prévoir la consommation électrique d'un ensemble de foyers
- Classer les clients d'une assurance selon leurs données personnelles.
- Optimiser le choix de publicité sur un site web,...

1.1.2 Supervisé/Non Supervisé

La plupart des problèmes d'apprentissage statistique peut être classée en deux groupes : les problèmes **supervisés** et **non supervisés**:

Supervisé

: Pour chaque individu, on peut distinguer une "réponse" ou un "label" (étiquette, spécifique aux phénomènes à réponses qualitatives) que l'on notera généralement Y. Les autres variables sont appelées "prédicteurs" ou "variables explicatives" et sont souvent notées $\mathbf{X}$.

- On dispose donc d'un ensemble de données $(X_i, Y_i)_{i=1}^n$, où n est le nombre d'observations (images, patients,...), $X_i = (X_i^1, \ldots, X_i^p)$ est le vecteur (ligne ou colonne) des variables (caractéristiques) par individu.
- Dans ce cadre, l'objectif naturel est de déterminer la "meilleure fonction" permettant d'approcher au mieux la vraie réponse y étant donné d'un vecteur d'entrée $x = (x_1, \ldots, x_p)$.
- Lorsque la réponse est qualitative, on parle de classification. Dans un cadre quantitatif, on parle généralement de régression (même si ce terme est aussi utilisé dans le cadre précis de la classification : la régression logistique par exemple permet de faire de la classification). Lorsque le bruit est additif, le modèle de régression prend

la forme générale suivante : $Y = f(\mathbf{X}) + \varepsilon$ où f est une fonction appartenant à une classe de fonctions fixée au départ.

Ci-dessous, un exemple simple (2 prédicteurs) de classification binaire. Dans ce cas, le but est de déterminer une règle permettant de classer la nouvelle observation dans le bon groupe. On reviendra sur cet exemple de base dans la suite. Un exemple

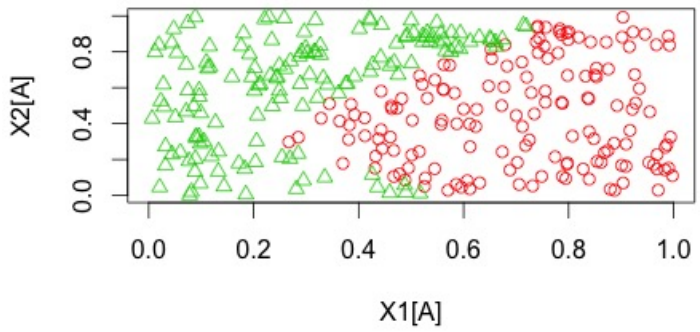

Figure 1.1: Deux groupes de couleur différente

d'algorithme : Les k-plus proches voisins : une règle naturelle. Pour un point donné, je considère les k plus proches voisins (k à déterminer) et je choisis ma réponse au vu de celles de ces k voisins les plus proches.

- Dans un cadre qualitatif, on choisit la réponse la plus fréquente.
- Dans un cadre quantitatif, on fera plutôt une moyenne de ces réponses.
- Ces règles peuvent bien sûr être raffinées (par exemple, pondérer selon la distance du voisin).

Non Supervisé

- Non Supervisé : On dispose d'un ensemble de mesures pour chaque individu mais pas de réponse naturelle. On peut alors chercher à comprendre les relations entre les variables et/ou les observations, identifier les (groupes de) variables les plus sensibles,. . . .
  Ces méthodes peuvent être vues comme un moyen de pallier l'absence de réponse. Elles peuvent être utilisées dans un but préliminaire mais sont de plus en plus présentes pour l'apprentissage de données complexes.

On peut identifier deux ou trois types importants d'algorithmes dans cette classe :

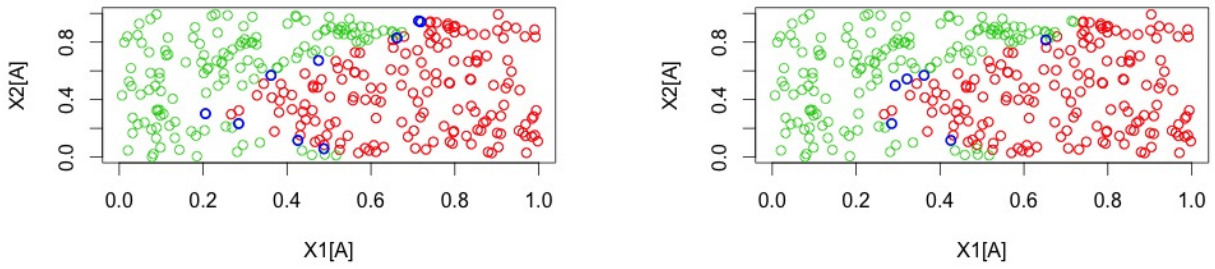

Figure 1.2: $k = 1$ à gauche, $k = 5$ à droite

- ceux qui permettent de "constituer des groupes de variables" ou plus précisément de "déterminer les variables ou combinaisons de variables les plus variantes" : c'est l'objectif de l'**analyse en composantes principales** ou de certaines de ses variantes pour le "clustering de variables".
- ceux qui sont destinés à fabriquer des groupes d'observations. On parle alors de "clustering d'individus". La méthode des **K-means** en est l'exemple le plus connu.
- ceux qui permettent d'approximer la loi de probabilité (estimation de densité/mélange...).

Semi-supervisé/ Données manquantes

Bien entendu, la pratique comporte des exemples d'applications qui ne sont pas forcément parfaitement classées dans l'une ou l'autre des classes. On parle alors d'apprentissage semi-supervisé lorsque par exemple,

- seule une partie des données possède une réponse (Si la réponse est la durée de vie après traitement, les personnes encore en vie n'ont pas de réponse !!).
- les réponses ne sont pas les mêmes: en médecine, on peut penser à des patients traités dans des instituts de recherche dans plusieurs endroits du monde où l'on n'a pas considéré les mêmes réponses (parce qu'ils n'ont pas reçu le même traitement). Néanmoins, les données étant difficiles à obtenir, on souhaiterait être capable de mettre en commun ces données).
- Parmi d'autres "imperfections" (non classées dans le semi-supervisé), on peut enfin penser au problème des données manquantes. Considérons l'exemple de Netflix. Supposons que l'analyse de données se base sur les notes mises sur les films regardés auparavant. Naturellement, tous les individus n'ont pas noté les films et par ailleurs, ils n'ont pas regardé les mêmes films. Le **Système de recommandation** doit donc s'appuyer sur une matrice très creuse ce qui implique des outils d'apprentissage adéquats.

Autres exemples de cas mal posés

- les **vraies données manquantes** : comment gérer l'absence de certaines données pour une partie des individus ? Plusieurs réponses selon les situations :
  - Exclure l'individu; dépend de l'importance des variables manquantes, du nombre d'individus...
  - Remplacer la donnée absente par une valeur raisonnable : la médiane de l'ensemble des individus (pour une variable quantitative), la médiane d'une partie des individus (les plus proches par exemple)...
  - Pour plus de détails dans cette direction, voir par exemple `https://perso.univ-rennes1.fr/valerie.monbet/doc/cours/IntroDM/Chapitre4.pdf`
- Les classes **mal équilibrées** : Dans ce cas, il peut être nécessaire de
  - Adapter la *fonction de perte* en attribuant un poids plus élevé aux classes mal représentées.
  - Générer artificiellement de nouveaux individus dans les classes où l'effectif est trop faible en clonant les individus existants ou en les interpolant... (**SMOTE** par exemple).
  - Attention, cloner ou interpoler fait perdre la propriété d'indépendance et de loi identique entre les observations. Ainsi, ce type de méthode est à considérer avec prudence.

1.1.3 Adaptatif/Non adaptatif

Une autre information importante à prendre en compte dans l'étude d'un problème est la manière dont les données arrivent.

- Dispose-t-on de toutes les données à un instant fixé ?
- Les données arrivent-elles au fil du temps ("on the fly") ?

Dans la deuxième situation, le caractère adaptatif de la méthode d'apprentissage est un élément important notamment pour le calcul effectif des prédictions. Par adaptatif, on entend : la faculté de l'algorithme à être mis à jour lors de l'arrivée d'une nouvelle donnée sans devoir tout recalculer. L'algorithme des k-plus proches voisins peut être programmé de manière adaptative par exemple (cf exercice).

1.1.4 Paramétrique/Non paramétrique

Considérons le problème de base où $(Z_1, \ldots, Z_n)_{n\geq 1}$ est issu d'une loi $\mathbb{P}$ inconnue que l'on cherche à estimer. On parle de Statistique

- Paramétrique : lorsque la loi de probabilité $\mathbb{P} \in \{\mathbb{P}_\theta, \theta \in \Theta\}$ où $\Theta \subset \mathbb{R}^d$.

- Non paramétrique lorsque $\mathbb{P}$ vit dans un espace de dimension infinie (ou finie mais tendant vers $+\infty$ avec n). Deux exemples :
  - Estimation de la densité f de la loi $\mathbb{P}$ (relativement à une mesure donnée, mesure de Lebesgue par exemple). Dans ce cas, l'approche non paramétrique consiste à supposer que f vit dans un espace de fonctions fixé (par exemple, l'ensemble des fonctions $\mathcal{C}^2$ d'intégrale égale à 1).
  - Régression non paramétrique : $Y = f(\mathbf{X}) + \varepsilon$ où f vit dans un espace de fonctions de dimension infinie (alors que la régression linéaire par exemple est clairement paramétrique).

1.2 Classification supervisée : modélisation du problème

Soit $(X_1, Y_1), \dots (X_n, Y_n)$, un *échantillon d'apprentissage* issu d'une loi conjointe $\mathbb{P}$ sur $\mathcal{X} \times \mathcal{Y}$.

Définition 1.2.1. *1. Une règle de prévision/régression/décision/discrimination est une fonction $f : \mathcal{X} \to \mathcal{Y}$ qui à $\mathbf{x}$ associe la sortie $f(\mathbf{x})$.*

2. Une fonction de perte ℓ est une fonction positive définie sur $\mathcal{Y} \times \mathcal{Y}$ telle que $\ell(y, y') > 0$ dès que $y \neq y'$.

Exemples importants : En régression, $\ell(y, y') = |y - y'|^2$ ou plus généralement, $\ell(y, y') = |y - y'|^\alpha$, $\alpha > 0$, et en classification : $\ell(y, y') = \mathbf{1}_{y \neq y'}$ (voir TD pour d'autres exemples).

1.2.1 Risque

Définition 1.2.2. *Pour une fonction de perte donnée, on appelle risque ou erreur de généralisation d'une règle de prévision f la quantité :*

$$R_f = \mathbb{E}[\ell(Y, f(\mathbf{X}))].$$

On dit qu'une règle $f^\star$ est optimale si,

$$R_{f^\star} = \inf_{f \in \mathcal{F}} R_f.$$

On est capable pour certaines pertes de définir de manière formelle (mais explicite) les règles optimales.

- En régression avec $\mathcal{Y} = \mathbb{R}$, avec $\ell(y, y') = |y - y'|^2$, la fonction $f^\star(\mathbf{x}) = \mathbb{E}[Y | X = \mathbf{x}]$ est optimale (car l'espérance conditionnelle est une projection orthogonale dans L^2).
- En régression avec $\mathcal{Y} = \mathbb{R}$, avec $\ell(y, y') = |y - y'|$, la fonction $f^\star(\mathbf{x}) = \text{médiane}(\mathcal{L}(Y | X = \mathbf{x}))$ est optimale.

- En classification, le classifieur optimal est appelé "classifieur naïf de Bayes" : on a le théorème suivant.

  Plus précisément, supposons que $\mathcal{Y}$ est un ensemble fini.

Théorème 1.2.3 ((et définition)). *(i) On appelle règle de Bayes la fonction $f^\star$ telle que pour tout $\mathbf{x} \in \mathcal{X}$,*

$$f^\star(\mathbf{x}) = \text{argmax}_{y\in\mathcal{Y}}\mathbb{P}(Y = y|\mathbf{X} = \mathbf{x}).$$

Cette règle est optimale. Notons

$$\eta^\star(x) = 1 - \max_{y\in\mathcal{Y}}\{\mathbb{P}(Y = y|\mathbf{X} = x)\}.$$

A x fixé, la probabilité de se tromper avec la règle $f^\star$ est égale à $\eta^\star(x)$. On a :

$$R_{f^\star} = \mathbb{E}[\eta^\star(X)] = \inf_f R_f. \quad (R_{f^\star} \leq 1 - \frac{1}{\text{Card}(\mathcal{Y})}).$$

(ii) Pour toute autre règle de décision f, on mesure la "non-optimalité" par :

$$R_f - R_{f^\star}.$$

$R_{f^\star}$ est appelé risque de Bayes.

Preuve.

On considère seulement le cas binaire : $\mathcal{Y} = \{-1, 1\}$,

$$f^*(\mathbf{x}) = \text{Argmax}_{y\in\{-1,1\}}\,\mathbb{P}(Y = y|\mathbf{X} = \mathbf{x}).$$

Soit $f : \mathcal{X} \longrightarrow \mathcal{Y}$ une règle de prévision.

$$f(x) = \begin{cases} 1 & si \quad x \in A \\ -1 & si \quad x \in A^C \end{cases}$$

On cherche à prouver que $R_{f^*} = \min_{f:\mathcal{X}\longrightarrow\mathcal{Y}} R_f$.

$$\begin{aligned} R_f &= \mathbb{P}(Y \neq f(\mathbf{X}) = \mathbb{P}(Y = 1, f(\mathbf{X}) = -1) + \mathbb{P}(Y = 1, f(\mathbf{X}) = 1) \\ &= \mathbb{P}(\mathbf{X} \in A^C, Y = 1) + \mathbb{P}(\mathbf{X} \in A, Y = 1) \\ &= \mathbb{E}\left[\mathbb{P}(Y = 1|X)\mathbf{1}_{\{\mathbf{X}\in\mathbf{A^C}\}}\right] + \mathbb{E}\left[\mathbb{P}(Y = -1|X)\mathbf{1}_{\{\mathbf{X}\in\mathbf{A}\}}\right] \\ &= \int_A \mathbb{P}(Y = 1, \mathbf{X} = \mathbf{x})\mathbb{P}_X(\mathrm{d}x) + \int_{A^C} \mathbb{P}(Y = -1, \mathbf{X} = \mathbf{x})\mathbb{P}_X(\mathrm{d}x). \end{aligned}$$

Or, pour $j \in \{-1, 1\}$,

$$\mathbb{P}(Y = j, \mathbf{X} = \mathbf{x}) \geq \min\left\{\mathbb{P}(Y = 1, \mathbf{X} = \mathbf{x}), \mathbb{P}(Y = -1, \mathbf{X} = \mathbf{x})\right\} =: \eta^*(x)$$

Donc

$$R_f \geq \int_A \eta^*(x)\mathbb{P}_X(dx) + \int_{A^C} \eta^*(x)\mathbb{P}_X(dx) = \int_{\mathcal{X}} \eta^*(x)\mathbb{P}_X(dx) = \mathbb{E}\left[\eta^*(X)\right].$$

Par ailleurs, on peut vérifier que $f^\star$ satisfait bien $R_{f^*} = \mathbb{E}\left[\eta^*(X)\right]$, ce qui implique le résultat. ∎

Remarque : En classification binaire, la règle de Bayes est à comprendre de la manière suivante. Selon les zones de l'espace, la valeur de la réponse Y est tirée selon un jeu de pile ou face. Si la probabilité de faire "Pile" est supérieure à $1/2$, on choisit 1, si elle est plus faible que $1/2$, on choisit -1.

Néanmoins, tout ce qui précède est *formel* puisque ces règles sont construites à partir de quantités inconnues !! Elle dit simplement quel serait le choix optimal si l'on avait accès à toute l'information. En pratique le paramètre du jeu de pile ou face est inconnu, donc chercher à mimer la règle de Bayes revient à approcher cette probabilité via les observations.

1.2.2 Algorithmes de prévision

Etant donné un échantillon de taille n, un *algorithme de prévision* (ou **prédicteur**) est une application qui à $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1, \ldots, \mathbf{X}_n, Y_n)\}$ associe une fonction notée $\hat{f}_n$ de $\mathcal{X}$ vers $\mathcal{Y}$. Par exemple, dans le cas k plus proches voisins en discrimination binaire,

$$\hat{f}_n(\mathbf{x}) = \mathrm{sgn}(\eta_n(\mathbf{x})) \text{ où } \eta_n(\mathbf{x}) = \frac{\mathrm{Card}\{ \text{ "voisins de } \mathbf{x}", Y_i = 1\}}{k} - \frac{1}{2}.$$

- En régression linéaire standard $Y = \mathbf{X}\theta + \varepsilon$, $\hat{f}_n(\mathbf{x}) = \mathbf{x}\hat{\theta}_n$ où

$$\hat{\theta}_n = \mathrm{Argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i\theta)^2.$$

- En résumé, un algorithme de prévision cherche à mimer "la" règle de prévision optimale au vu des données observées.
- **Exercice.** On pourra vérifier que dans les deux cas, $\hat{f}_n = f_n(\mathcal{D}_n)$ où f_n est une fonction déterministe.

Risque et Algorithmes de prévision

On cherche maintenant à mesurer la qualité de l'algorithme de prévision.

Définition 1.2.4. *Le risque (moyen) d'un algorithme de prévision est défini par*

$$\mathbb{E}_{(\mathbf{X},Y)}[R_{\hat{f}_n}] = \mathbb{E}_{(\mathbf{X},Y)}[\ell(Y, \hat{f}_n(\mathbf{X}))].$$

Quelques exemples :

- k-plus proches voisins, discrimination binaire, fonction de perte $1_{y \neq y'}$. On a alors

$$\mathbb{E}_{(\mathbf{X},Y)}[R_{\hat{f}_n}] = \mathbb{P}_{(\mathbf{X},Y)}(Y \neq \hat{f}_n(\mathbf{X})).$$

- N.B. Il y a ici un abus de notation : $\hat{f}_n$ est construit à partir de l'échantillon $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ et $(\mathbf{X}, Y)$ représente un échantillon indépendant de cette suite de v.a.
- Modèle linéaire, fonction de perte $\ell(y, y') = (y - y')^2$:

$$\mathbb{E}_{(\mathbf{X},Y)}[R_{\hat{f}_n}] = \mathbb{E}[(Y - \mathbf{X}\hat{\theta}_n)^2].$$

1.2.3 Consistance et algorithme par moyennisation locale

Définition 1.2.5. *L'algorithme est (faiblement)* ***consistant*** *si*

$$\mathbb{E}_{(\mathbf{X},Y)}[R_{\hat{f}_n}] \xrightarrow{n\to+\infty} \inf_{f\in\mathcal{F}} R_f.$$

Intéressons-nous à la consistance des algorithmes de prévision les plus naturels :

Définition 1.2.6. *On appelle algorithme par moyennisation locale un algorithme basé sur une moyenne "locale" des observations. Attention, le terme local ici est à comprendre comme "avec une pondération décroissant avec la distance".*

L'algorithme des k plus proches voisins en est un (pondération $1/k$ ou 0). Dans un cadre quantitatif, l'algorithme est défini par :

$$\hat{f}_n(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{n} Y_i \mathbf{1}_{\{\mathbf{X}_i \text{parmi les } k \text{ p.p.v. de } \mathbf{x}\}}.$$

Exercice. Expliquez pourquoi l'algorithme ci-dessus est une approximation de la règles optimale définie dans le slide précédent. Si l'on revient maintenant à la définition des règles optimales, on peut montrer sous des conditions assez générales (cf TD) que

Théorème 1.2.7. *Les algorithmes par moyennage local sont universellement consistants (Par exemple, pour les k-ppv, ça marche si $k_n \to +\infty$ et que $k_n/n \to 0$).*

En réalité, la consistance n'est pas une notion satisfaisante en pratique car elle est seulement asymptotique. On souhaiterait en réalité évaluer l'erreur de prévision pour un échantillon fixé. Outre l'étude de la qualité de l'algorithme, le but est souvent de pouvoir choisir dans une famille celui qui est le plus efficace. Ce choix pourra dans un premier temps de minimiser le risque (ou plutôt de son estimation).

1.2.4 Comment sélectionner un algorithme/un modèle

En apprentissage statistique, les choix du modèle et de l'algorithme sont des étapes fondamentales pour optimiser la qualité de la prévision.

- Par modèle, il s'agit de choisir la classe de probabilités $\mathbb{P}$ dans laquelle vit la loi de $(\mathbf{X}, Y)$. Par exemple, en régression linéaire (paramétrique), on fait l'hypothèse que Y et $\mathbf{X}$ sont reliés par la relation $Y = \mathbf{X}\theta + \varepsilon$.

- Par algorithme, il s'agit de choisir le type d'algorithme mais aussi et surtout le bon paramétrage. Par exemple, dans le cas des k plus proches voisins, le choix du nombre k de voisins pris en compte dans la décision est fondamental.

Pour décider, l'approche la plus naturelle consiste à minimiser le risque. Pour cela, il faut commencer par être capable de l'estimer.

1.2.5 Estimation du risque

On rappelle que le risque moyen est défini par $\mathbb{E}_{\mathbf{X},Y}[R_{\hat{f}_n}]$, ce qui donne par exemple,

- en qualitatif : l'erreur de classification (probabilité d'être mal classé)
- en quantitatif : la distance moyenne au carré de la prévision à la vraie réponse Y lorsque la fonction de perte est $\ell(y, \mathbf{x}) = (y - \hat{f}_n(x))^2$ (MSE : Mean-Squared Error)

Pour estimer le risque, on peut être tenté de simplement considérer l'erreur sur l'échantillon sur lequel on a construit $\hat{f}_n$, ce qui donnerait

- la **proportion** de mal classés dans le premier cas : $\frac{1}{n}\sum_{k=1}^{n} 1_{Y_k \neq \hat{f}_n(\mathbf{X}_k)}$.
- $\frac{1}{n}\sum_{k=1}^{n}(Y_k - \hat{f}_n(\mathbf{X}_k))^2$ dans le second cas.

Ces quantités sont appelées "**train errors**" mais ne peuvent être considérées comme des approximations du risque car elles sont mesurées sur l'échantillon. En particulier, si le modèle (ou la classe de modèles) est très **flexible**, alors cette erreur peut être beaucoup plus faible que la véritable erreur (penser aux modèles de régression polynômiale par exemple).

1.2.6 Overfitting

Ce phénomène est appelé "surapprentissage" (overfitting). Ci-dessous, deux situations simples pour bien comprendre la situation de surapprentissage. Dans la figure **??**, on considère un problème de régression. On constate que selon la richesse de la classe de fonctions considérée, on a plus ou moins de capacité à interpoler les points de l'échantillon.

On sent également que si dans le cas de gauche, la classe de fonctions envisagée est trop restreinte pour estimer le comportement de la relation entre $\mathbf{X}$ et Y, le fait de trop vouloir coller aux observations ne permet pas non plus de dégager de l'information. L'erreur d'entraînement sur la figure de droite est nulle mais pour autant, il est probable que si l'on teste ce prédicteur sur un autre échantillon, l'erreur (de test) sera bien plus importante.

Dans la figure **??**, le même type de phénomène est considéré dans un cadre de classification binaire.

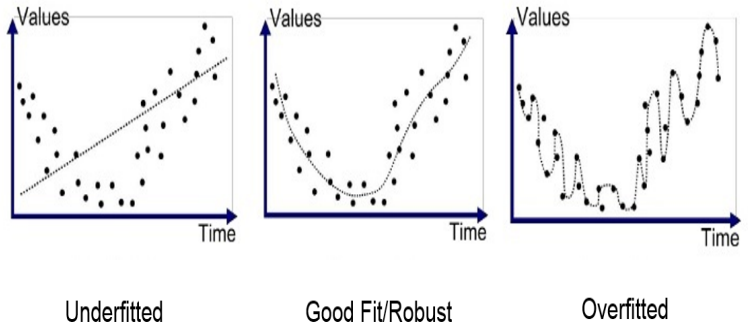

Figure 1.3: (taken from `https://medium.com/greyatom`)

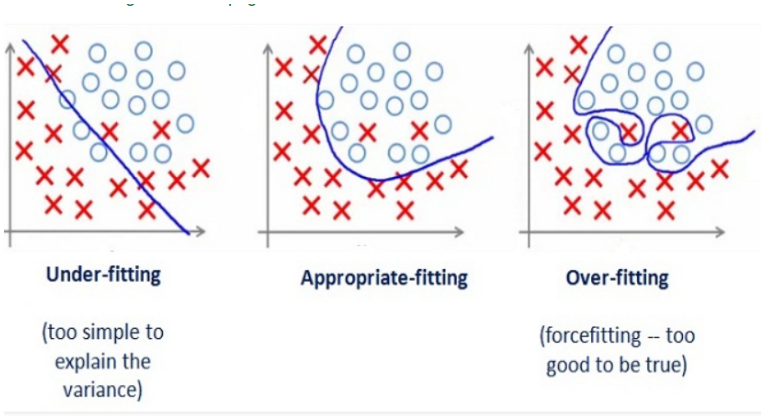

Figure 1.4: (taken from `https://medium.com/greyatom`)

1.2.7 Estimation du risque

Reprenons : pour une fonction f donnée, on a par la loi des grands nombres

$$R_f = \lim_{N\to+\infty} \frac{1}{N}\sum_{k=1}^{N} \ell(Y_k, f(X_k))$$

où $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ est une suite de v.a. i.i.d. de même loi que $(\mathbf{X}, Y)$.

- **Problème :** $\hat{f}_n$ est aléatoire puisqu'elle est construite à partir de l'échantillon.
- **Conséquence :** Deux niveaux de "moyennisation".
- **Conséquence sur l'estimation du risque :** Division de l'échantillon en deux parties (a minima), l'une pour construire $\hat{f}_n$ (échantillon d'apprentissage, *train* en anglais), l'autre pour calculer une approximation de l'espérance (échantillon **test** ou de **validation** selon la situation...).

Divisons l'échantillon en deux parties. Notons $\mathcal{D}_{n_1}$ la partie consacrée à l'apprentissage (train) et $\tilde{\mathcal{D}}_{n_2}$ la partie consacrée au test (On la note $\tilde{\mathcal{D}}$ pour éviter les confusions). L'erreur "test" (ou de validation si l'on effectue une sélection de modèle, cf suite) est alors la quantité

$$\frac{1}{n_2}\sum_{i=1}^{n_2} \ell(\tilde{Y}_i, \hat{f}_{n_1}(\tilde{X}_i))$$

A échantillon $\mathcal{D}_{n_1}$ fixé, on a alors par la LGN (sous des hypothèses appropriées) : Plaçons-nous pour simplifier dans le cas où $\hat{f}_n = f(\mathcal{D}_{n_1})$ (f déterministe pouvant dépendre de n).

$$\frac{1}{n_2}\sum_{i=1}^{n_2} \ell(\tilde{Y}_i, f(\mathcal{D}_{n_1})(\tilde{X}_i)) \to) \xrightarrow{n_2\to+\infty} \Phi(\mathcal{D}_{n_1}) = \mathbb{E}[\ell(Y, \hat{f}_{n_1}(\mathbf{X}))|\mathcal{D}_{n_1}],$$

i.e. le risque conditionnel à l'échantillon $\mathcal{D}_{n_1}$ (*i.e.* $\hat{f}_{n_1}$ vue comme fonction déterministe).

Dépendance à l'échantillon d'apprentissage

- Dans ce qui précède, on constate que le risque est "conditionnel à l'échantillon d'apprentissage". Il est donc "biaisé". Une manière de réduire ce biais relatif à la base d'apprentissage sera de faire de la validation croisée.
- Néanmoins, que se passe-t-il quand n_1 est grand ? Difficile de donner une réponse générale. De façon intuitive, on peut supposer que l'échantillon devient de plus en plus représentatif de la loi de $(\mathbf{X}, Y)$ de sorte que l'espérance conditionnelle est assez peu sensible à $\mathcal{D}_{n_1}$. Plus précisément,

- Faisons par exemple l'hypothèse que $\mathbf{X}$ et Y sont liés par la relation

$$Y = f_\theta(\mathbf{X}) + \varepsilon \quad \theta \text{ inconnu},$$

et que $\hat{f}_n = f_{\hat{\theta}_n}$ (modèle linéaire par exemple). Alors, si $\hat{\theta}_n \to \theta$ (consistance de l'estimateur), on a (sous des hypothèses de continuité de $\theta \mapsto f_\theta$), $\hat{f}_{n_1} \to f$. Ainsi, par des théorèmes de convergence (type cv dominée), la quantité converge vers R_{f_θ} (déterministe).

- Ceci est une manière de traduire cette non-dépendance asymptotique à l'échantillon.

1.2.8 Validation croisée

Pour estimer le risque, on utilise de manière plus générale la validation croisée (pour limiter le "biais d'apprentissage"). Il s'agit de diviser l'échantillon en K parties (K folds) de même taille **aléatoirement** notées $I_1, \ldots, I_K$ puis

- utiliser successivement chaque échantillon $\mathcal{D}_{-I_k} = \{(\mathbf{x}_i, y_i), i \in \{1, \ldots, n\} \backslash I_k\}$, comme échantillon d'entraînement puis $\mathcal{D}_{I_k}$ comme échantillon de validation.
- Calculer le risque sur chaque sous-échantillon.
- Le risque obtenu par validation croisée notée R_{CV} est alors obtenu comme une moyenne de tous ces risques.

Remarque 1 : Par exemple, dans le cas $K = 2$, on coupe l'échantillon en 2 comme précédemment et on fait la moyenne du risque en utilisant la première puis la seconde partie de l'échantillon comme échantillon d'entraînement. Schématiquement, dans le cas $K = 5$ (K-folds cross-validation),

Plus précisément, notons $\hat{f}_{-I_k}$, l'algorithme de prévision associé à $\mathcal{D}_{-I_k}$. On a

$$\hat{R}_{CV} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \ell(y_i, \hat{f}_{-I_k}(\mathbf{x}_i)).$$

Remarques :

- Méthode générale pouvant s'appliquer dans la plupart des contextes.
- Peu de résultats théoriques sur le sujet malgré une utilisation très répandue.
- Intuitivement, l'idée est de réduire la dépendance à l'échantillon d'entraînement en moyennisant sur K échantillons d'entraînement différents.
- Très important quand l'échantillon est petit.
- $K = 5$ ou $K = 10$ sont des choix usuels (lorsque K est trop important, coût de calcul souvent élevé)
- Le cas limite est le "Leave-One Out", cas où $K = n$. (A chaque itération, l'échantillon d'entraînement est constitué de $n - 1$ individus).

Surapprentissage et nécessité de diviser en 3

Question : Supposons que l'on ait sélectionné un modèle à l'aide de l'une des méthodes précédentes. L'erreur $\hat{R}_{CV}$ est-elle une bonne approximation du vrai risque ?

Réponse : Pas si clair. On peut encore avoir sur-estimation ou sous-estimation du risque selon la flexibilité du modèle et le nombre de données. De manière plus précise, les échantillons de validation rentrent de près ou de loin dans le choix du modèle puisque l'on choisit le meilleur dans une famille de possibles. Ce phénomène est d'ailleurs amplifié si l'on empile plusieurs méthodes. Supposons par exemple que l'on *agrège* les deux meilleurs modèles de deux familles différentes et que l'on cherche la "meilleure" agrégation. Dans ce cas, les échantillons de validation servent à la fabrication de l'algorithme final.

Une autre erreur usuelle apparaît lorsque l'on fait de la sélection de variables. Imaginons que l'on fabrique un premier prédicteur et que l'on décide de réduire le nombre de variables en ne sélectionnant que les variables les plus importantes (penser par exemple à la régression linéaire). Une fois ces variables sélectionnées, on peut à nouveau recalculer le meilleur prédicteur dans la famille de modèles considérés avec pour critère de décision, l'erreur de validation croisée et dans certains cas, constater que l'erreur a diminué (si le nombre de variables p est important relativement au nombre d'observations, ce phénomène risque de se produire). On peut réitérer ce processus plusieurs fois et choisir à la fin celui qui minimise l'erreur de validation croisée. Néanmoins, dans ce cas, il est hautement probable que l'erreur de validation croisée sous-estime nettement le risque réel. Pour éviter ce phénomène, la marche à suivre est la suivante :

Règle à suivre :

- Laisser une (petite) partie de l'échantillon en dehors du processus de choix de modèle (Echantillon "Test").

- Sélectionner via une ou plusieurs méthodes le ou les meilleurs algorithmes de différentes familles sur les échantillons "Train/Validation"
- Comparer les algorithmes sur l'échantillon "vierge" à la fin du processus.

Remarque : Exemple du Kaggle (`http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/`)

Conclusions/Extensions

- Chapitre dédié aux principales notions liées à l'apprentissage statistique.
- Estimation du risque = Opération délicate.
- Peu de règles générales ; Conclusions "universelles" à envisager avec prudence.
- Présentation des méthodes de sélection de modèles via l'approche "minimisation empirique du risque". Néanmoins, pour certaines familles de modèles il existe des méthodes plus élaborées, basées sur des arguments théoriques (type AIC par exemple). Par ailleurs, dans certains problèmes, la minimisation du risque de prédiction n'est pas la seule question fondamentale d'où l'utilisation d'autres outils de mesure (Courbe ROC par exemple, voir TD).

Chapitre 2

Fléau de la dimension

Bien que ce cours soit consacré à l'apprentissage de manière générale et pas seulement à "la grande dimension", nous choisissons ici de consacrer dès maintenant un chapitre à cet aspect afin de le conserver à l'esprit dans toute la suite du cours.

2.1 Qu'est-ce que la grande dimension ?

Pour le mathématicien, le mot "grande dimension" fait surtout référence au nombre important de variables p que peut contenir le problème car la qualité de l'apprentissage en dépend fortement. Plus précisément, la dimension implique des spécificités dans la manière d'appréhender le problème. Le but de ce chapitre est d'en présenter quelques aspects.

2.1.1 Exemples

Ci-dessous, quelques situations typiques où la grande dimension apparaît de manière naturelle.

Données biotech: mesure des milliers de quantités par "individu". On peut penser en particulier aux données "omiques" où l'on cherche à appréhender une maladie par exemple via un ensemble d'information par individu de l'ordre de plusieurs dizaines à centaines de milliers de variables (cf schéma ci-dessous issu de `http://www.ipubli.inserm.fr/`)

Traitement d'images: images médicales, astrophysique, video surveillance, etc. Chaque image est constituée de milliers ou millions de pixels ou voxels.

Marketing : les sites web et les programmes de fidélité collectent de grandes quantités d'information sur les préférences et comportements des clients. Ex: syst'emes de recommandation...

Business : l'exploitation des données internes et externes de l'entreprise devient primordiale

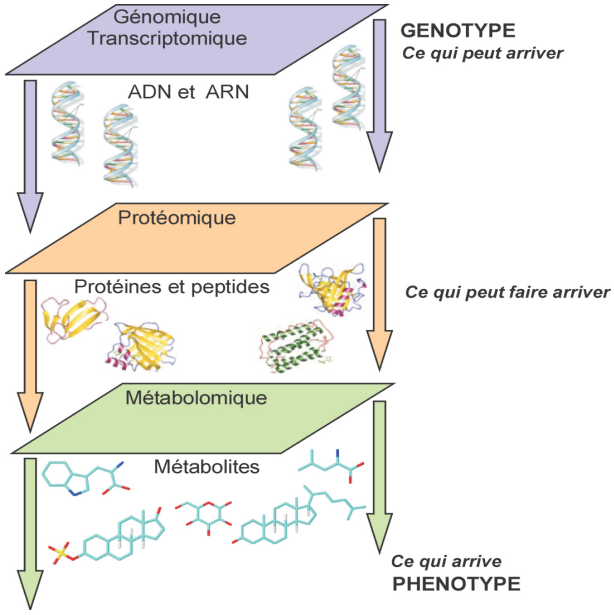

Figure 2.1: Données omiques

2.1.2 A propos de n et p

- Dans tous ces champs d'applications, l'objectif est de décrire de manière plus précise un phénomène (penser au traitement d'images ou à la médecine par exemple). Néanmoins, sous quelles conditions l'accumulation de données peut permettre d'aller dans cette direction ?

- Quantités fondamentales : n le nombre d'observations/individus (patients/images/"clics"...) et p, le nombre de variables/mesures par "individu" (éviter le terme "données" qui peut porter à confusion).

- On parle de Grande Dimension lorsque p est grand, *i.e.* lorsque l'espace sur lequel "vivent" les observations est de grande dimension.

- Du point de vue informatique, le caractère n grand est lui AUSSI un facteur de "Big Data" (puisque le temps de calcul s'en trouve impacté) mais on comprendra que du point de vue mathématique, "plus n est grand, plus on est content".

- Plusieurs situations : en médecine, $n << p$, en traitement d'images récoltées sur Internet (cf Facebook par exemple), on peut imaginer n et p grands.

- Ces deux types de situations sont totalement différents. En particulier, les objectifs d'apprentissage sont totalement dépendants du lien entre ces deux valeurs.

2.2 De la statistique classique à la statistique en grande dimension

2.2.1 n, p et statistique

1. Statistique Classique : p petit et n grand. Ainsi, on peut étudier le problème de manière asymptotique via les résultats classiques type TCL par exemple : $f : \mathbb{R}^p \to \mathbb{R}$, $X_1, \ldots, X_n$ *i.i.d.*

$$\sqrt{n}\frac{\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)]\right)}{\sqrt{\mathrm{Var}(f(X_1))}} \xrightarrow{n\to+\infty} \mathcal{N}(0,1).$$

2. Le problème est que $\mathrm{Var}(f(X_1))$ augmente avec p. Supposons que f soit Lipschitzienne. Dans ce cas, on a

$$\mathrm{Var}(f(X_1)) \leq \mathbb{E}[(f(X_1) - \mathbb{E}[f(X_1)])^2] \leq C\mathbb{E}[\|X_1\|^2] = C\sum_{i=1}^p \mathbb{E}[(X_1^j)^2] \propto p$$

   si les coordonnées ont la même loi par exemple.

3. Dans un cadre général, le ratio $\sqrt{\frac{p}{n}}$ est donc fondamental dans le calcul de l'erreur.

- Statistique en grande dimension : $n << p$ ou $n \propto p$. De manière générale, il faut repenser la statistique dans un mode non asymptotique en n (Chebyshev/Concentration) mais aussi et surtout adapter les objectifs.
- Par exemple, travailler sur un sous-espace/une sous-variété de $\mathbb{R}^p$ qui soit de dimension adaptée au nombre d'observations (LASSO).
- Utiliser des méthodes de classification qui "supportent la dimension" (arbres/SVM...)
- Réduire la dimension du problème...

2.3 Fléau de la dimension (Curse of dimensionality)

2.3.1 Quelques exemples de difficultés liées à la grande dimension

Grande Dimension et Méthodes à moyennage local :

On a vu dans le chapitre 1 que les méthodes à moyennage local (type KNN) sont des candidats non paramétriques naturels pour l'apprentissage. Qu'en est-il en grande dimension ?

- **Exemple** : Supposons que l'on ait à apprendre une relation de la forme $Y = f(\mathbf{X}, \varepsilon)$ où $\mathbf{X}$ suit la loi uniforme sur l'hypercube $[0,1]^p$.

- Supposons même pour simplifier que $Y = f(\mathbf{X})$ (*i.e.* que sachant $\mathbf{X}$ la réponse est déterministe). Supposons même pour simplifier encore plus que l'on subdivise chaque dimension en 10 (sur chaque dim., on divise l'intervalle $[0,1]$ en 10 intervalles) et que f est constante sur les hypercubes de la forme $\prod_{k=1}^{p}[i_k/10, i_{k+1}/10]$. Dans ce cas, le nombre minimal d'observations pour apprendre la relation $Y = f(\mathbf{X})$ est égal à

$$10^p!! \quad \text{(Croissance exponentielle avec la dimension).}$$

- En langage plus élaboré, cette propriété peut s'interpréter comme la décroissance exponentielle du volume de la boule unité avec la dimension.

2.3.2 Boule unité en grande dimension :

On a :

$$V_p(1) = \mathrm{Vol}(B_p(0,1)) = \frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)} \overset{p\to+\infty}{\sim} \left(\frac{2\pi e}{p}\right)^{\frac{p}{2}} (p\pi)^{-\frac{1}{2}}.$$

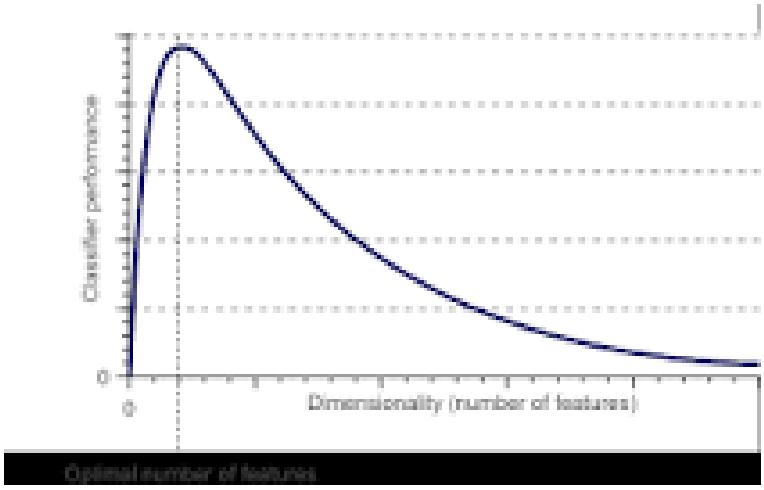

Figure 2.2: $p \to \mathrm{Vol}(B_p(0,1))$

Boule unité en grande dimension et nombre d'observations

Ainsi, si $x^{(1)}, \ldots x^{(n)}$ sont n points tels que pour tout $x \in [0,1]^p$, il existe au moins un point parmi les n tel que $\|x^{(i)} - x\| \leq 1$, on a alors

$$[0,1]^p \subset \cap_{i=1}^{n} B(x^{(i)}, p)$$

de sorte que $1 \leq nV_p(1)$. On retrouve que le nombre de points nécessaires pour remplir l'hypercube satisfait :

$$n \geq \frac{\Gamma(\frac{p}{2}+1)}{\pi^{\frac{p}{2}}} \sim \left(\frac{p}{2\pi e}\right)^{\frac{p}{2}} (p\pi)^{\frac{1}{2}}$$

En dimension 100, on trouve déjà $n \geq 42.10^{39}$.

Répartition de la masse d'une boule en grande dimension

Dans le même sens, on peut remarquer que la boule unité en grande dimension peut s'avérer peu intuitive. Considérons la boule de rayon r et la couronne $C_p(r) = \{x, 0.99r \leq \|x\| \leq r\}$. On a ‘

$$\frac{\text{Vol}(C_p(r))}{\text{Vol}(B_p(r))} = 1 - 0.99^p \to 1$$

lorsque $p \to +\infty$ (exponentiellement vite). Ainsi, très rapidement, la masse de la boule unité se trouve concentrée dans sa “croûte”. La répartition des points dans l'espace est finalement assez surprenante.

2.3.3 Moyennage local ?

1. Au vu de ce qui précède, on ne peut clairement pas envisager d'utiliser des méthodes à moyennage local en grande dimension.

2. Utiliser les k-plus proches voisins n'a donc pas de sens en général car il n'y a plus de voisins en grande dimension.

3. Ceci est confirmé par la borne théorique obtenue par Gadat *et. al.*:

$$\sup_{(\mathbf{X},Y)} |\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^\star)| \leq Cn^{-\frac{1+\alpha}{2+d}}.$$

   où $\mathcal{R}(f) = \mathbb{E}[\ell(Y, f(\mathbf{X}))]$.

4. Néanmoins, la borne ci-dessus peut être vue comme pessimiste car elle est “universelle” (pour l'ensemble des lois $(\mathbf{X}, Y)$ telles que $x \mapsto \eta(x) := \mathbb{P}(Y = 1|\mathbf{X} = x)$ varie raisonnablement par exemple).

5. *Cuisine :* Si la fonction η est très peu sensible ou si le support “réel” de X est de dimension plus faible, cela peut encore fonctionner.

2.3.4 Exemple de la régression linéaire

Supposons qu'il existe $\theta^\star \in \mathbb{R}^p$ tel que pour tout $i \in \{1, \ldots, n\}$,

$$Y_i = \langle \mathbf{x}_i, \theta^\star \rangle + \varepsilon_i$$

où $Y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$ et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Posons $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$. L'estimateur des moindres carrés $\hat{\theta}$ satisfait (lorsque $\mathbf{x}^T\mathbf{x}$ est inversible):

$$\hat{\theta} = \text{Argmin}_{\theta \in \mathbb{R}^p} \|Y - \mathbf{x}\theta\|^2 = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T Y.$$

De plus, si $\mathbf{x}^T\mathbf{x}$ est inversible (matrice $n \times p$),

$$\mathbb{E}[\|\hat{\theta} - \theta_0\|^2] = \mathbb{E}[\|(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\varepsilon\|^2] = \text{Tr}((\mathbf{x}^T\mathbf{x})^{-1})\sigma^2.$$

Supposons que les colonnes $C_1, \ldots, C_p$ de $\mathbf{x}$ soient orthonormales (ce qui implique que $p \leq n$). Dans ce cas,

$$(\mathbf{x}^T\mathbf{x})_{i,j} = \langle C_i, C_j \rangle = \delta_{i,j} \implies \mathbf{x}^T\mathbf{x} = I_p.$$

Par conséquent,

$$\mathbb{E}[\|\hat{\theta} - \theta_0\|^2] = p\sigma^2.$$

On retrouve à nouveau la croissance linéaire de l'erreur de la MSE avec la dimension.

2.3.5 Matrice de covariance

Soit X une variable aléatoire à valeurs dans $\mathbb{R}^p$ de matrice de covariance Σ (matrice $p \times p$). Pour simplifier, supposons que X est centrée. Dans ce cas, $\Sigma_{i,j} = \mathbb{E}[X_i X_j]$ et notons $(X^{(1)}, \ldots, X^{(n)})$ un échantillon *i.i.d.* issu de X.
Soit $\Sigma^{(n)}$ la matrice de covariance empirique associée :

$$\Sigma^{(n)}_{i,j} = \frac{1}{n}\sum_{k=1}^{n} X_i^{(k)} X_j^{(k)}.$$

Par la loi des grands nombres, si p est fixé et $n \to +\infty$,

$$\Sigma^{(n)} \to \Sigma.$$

Question : Si p n'est pas petit devant n, peut-on encore espérer que $\Sigma^{(n)}$ soit une bonne approximation de Σ ?

Réponse : Non, en général. Par exemple, on peut le voir sur le spectre de la matrice de covariance empirique. Dans les graphes de la figure **??**, on suppose que $X \sim \mathcal{N}(0, I_p)$ et on remarque que le spectre de la matrice de covariance ne ressemble pas du tout au spectre de la matrice cible.

2.3.6 Et l'ACP ?

On rappelle que l'Analyse en Composantes Principales consiste pour un d fixé à déterminer le meilleur sous-espace vectoriel V_d de dimension d au sens suivant :

$$V_d^{(n)} \in \text{Argmin}_{dim V \leq d} \sum_{i=1}^{n} \|X^{(i)} - P_V(X^{(i)})\|^2$$

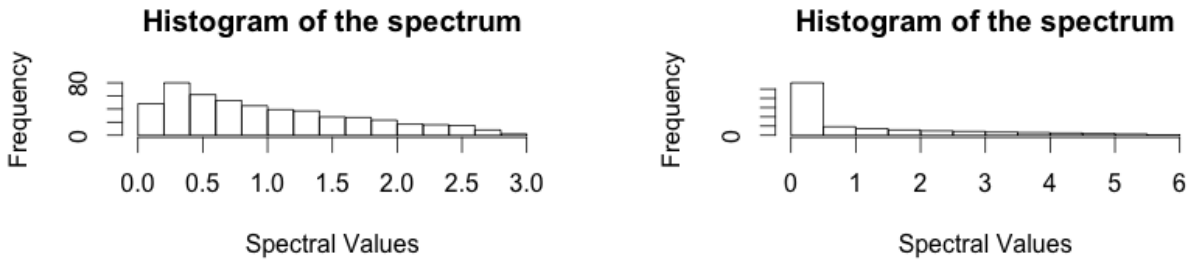

Figure 2.3: $n = 1000$, $p = 500$ (left), $p = 2000$ (right)

avec l'objectif sous-jacent de déterminer le meilleur sous-espace vectoriel au sens suivant :

$$V_d \in \text{Argmin}_{dimV \leq d} \mathbb{E}[\|X - P_V(X)\|^2].$$

Via la loi des grands nombres, on a "$V_d^{(n)} \to V_d$" lorsque $n \to +\infty$.

Question : L'intérêt de l'ACP est ainsi de réduire la dimension de l'espace "optimalement" (en perdant le moins d'information) mais est-ce que cette réduction de dimension est fiable ? Sous quelles conditions ? On rappelle le résultat suivant (lorsque les variables sont centrées, sinon il faut recentrer).

Théorème 2.3.1. *Supposons que l'échantillon soit issu d'une v.a. X centrée de matrice de covariance $\mathbb{E}[XX^T]$ de rang supérieur à d. Dans ce cas, $V_d^{(n)}$ est le sous-espace vectoriel de dimension d engendré par les d vecteurs propres associés aux plus grandes valeurs propres de $\Sigma^{(n)}$, la matrice de covariance empirique définie par :*

$$\Sigma_{i,j}^{(n)} = \frac{1}{n} \sum_{k=1}^{n} X_i^{(k)} X_j^{(k)}.$$

Pour rappel, les matrices de covariance (empirique ou non) sont symétriques positives. Ainsi, quitte à normaliser, les vecteurs propres forment une base orthonormée de V_d.
Au vu de ce qu'on a dit sur la matrice de covariance précédemment, on peut donc en conclure l'ACP ne donne pas d'information fiable en grande dimension et en toute généralité que sous la condition $n >> p$.
Néanmoins, si la matrice de covariance est de rang faible ou contient un grand nombre de petites valeurs propres, l'ACP peut s'avérer encore efficace. A voir en pratique.

2.4 Que peut-on espérer en grande dimension ?

Les messages ci-dessous (qui ne sont que des exemples de difficultés rencontrées) ne sont pas très rassurants pour aborder des problèmes en grande dimension. Néanmoins, comme

l'indique la suggestion dans le cadre de l'ACP pour les matrices de faible rang, on peut espérer tirer de l'information si :

- il existe des structures "cachées" dans les espaces de grande dimension qui sont de petite dimension: les données en grande dimension sont concentrées autour de structures de faible dimension reflétant la faible complexité des données. On peut penser à la structure géométrique des images par exemple. En médecine, ce type d'hypothèse n'est pas clair pour certaines maladies complexes mais le faible nombre de données nous astreint à une telle hypothèse.

- La structure est plus complexe mais le nombre de données est lui aussi élevé ce qui permet via des algorithmes performants d'aller plus loin dans la flexibilité et donc d'augmenter la prédiction, l'analyse...

Chapitre 3

Régression linéaire en Grande Dimension

3.1 Régression Linéaire Classique : Rappels

3.1.1 Modèle

On considère une variable à expliquer Y (supposée réelle pour l'instant) et p covariables (ou variables explicatives) $X_1, \dots, X_p$. L'objectif de la *régression linéaire* est de modéliser Y comme une combinaison linéaire bruitée de $X_1, \dots, X_p$ de la forme :

$$Y = \theta_0 + \sum_{i=1}^{p} \theta_i X_i + \varepsilon$$

où $\theta = (\theta_0, \theta_1, \dots, \theta_p)^T$ appartient à $\mathbb{R}^{p+1}$ et ε est une variable aléatoire centrée (bruit) indépendante de la variable aléatoire X. Quitte à considérer le vecteur $(1, X_1, \dots, X_p)$, on remarque que l'on peut mettre ce modèle sous la forme plus synthétique :

$$Y = X\theta + \varepsilon.$$

Pour un couple (X, Y) donné, le (un des) but(s) est alors de déterminer :

$$\theta^\star = \mathrm{Argmin}_{\theta \in \mathbb{R}^{p+1}} \mathbb{E}[(Y - X\theta)^2].$$

Remarque : Dans la suite, on note pour simplifier $\theta = (\theta_1, \dots, \theta_p)^T$ et $X = (X_1, \dots, X_p)$ (quitte à supposer que la première composante de X est égale à 1).

3.1.2 Régression et Apprentissage

Si l'on fait l'hypothèse que le "vrai" modèle est de la forme $Y = X\theta^\star + \varepsilon$ (ce que l'on fera dans la suite), alors le prédicteur de Bayes (pour la fonction de perte $\ell(y, y') = (y - y')^2$ est donné par

$$f_{\mathrm{Bayes}}(x) = \mathbb{E}[Y|X = x] = x\theta^\star.$$

Dans ce cas, si $\mathrm{Var}(\varepsilon) = \sigma^2$, alors

$$\inf_{f:\mathbb{R}^p \to \mathbb{R}} \mathcal{R}_f = \mathbb{E}[(Y - X\theta^\star)^2] = \mathbb{E}[\varepsilon^2] = \sigma^2.$$

Soit $\mathcal{D}_n = \{(X^{(1)}, Y^{(1)}), \dots, (X^{(n)}, Y^{(n)})\}$ un échantillon d'apprentissage (en particulier *i.i.d*). On suppose dans la suite que pour $k \in \{1, \dots, n\}$,

$$Y^{(k)} = X^{(k)}\theta^\star + \varepsilon_k.$$

où $\theta^\star$ est une quantité "à apprendre". Sauf mention contraire, on fera aussi l'hypothèse que $\varepsilon_k \sim \mathcal{N}(0, \sigma^2)$.

3.1.3 Estimateur de $\theta^\star$

Pour estimer $\theta^\star$, on utilise l'estimateur

$$\hat{\theta} = \mathrm{Argmin}_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y^{(i)} - X^{(i)}\theta)^2 = \mathrm{Argmin}_{\theta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\theta\|^2$$

où $\|\,,\,\|$ désigne la norme euclidienne,

$$\mathbf{Y} = \begin{pmatrix} Y^{(1)} \\ \dots \\ Y^{(n)}) \end{pmatrix} \text{ et } \mathbf{X} = \begin{pmatrix} X_1^{(1)} & \dots & X_p^{(1)} \\ & \dots & \\ X_1^{(n)} & \dots & X_p^{(n)} \end{pmatrix}$$

Attention aux notations : X et Y désignent des variables aléatoire. $\mathbf{X}$ et $\mathbf{Y}$ désignent la matrice et la réponse associées à l'échantillon. Dans la suite, ε désigne le vecteur $(\varepsilon_1, \dots, \varepsilon_n)$.

3.1.4 Loi du vecteur Y et premières propriétés

Avec ces notations, on a donc l'égalité dans $\mathbb{R}^n$

$$\mathbf{Y} = \mathbf{X}\theta^\star + \varepsilon.$$

En particulier, comme $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, on en déduit que conditionnellement à $\mathbf{X}$ (*i.e.* connaissant $\mathbf{X}$)

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\theta^\star, \sigma^2 I_n).$$

Dans la suite, on raisonnera toujours "conditionnellement à $\mathbf{X}$" de sorte que $\mathbf{X}$ comme une matrice déterministe.

- Dans le modèle linéaire classique, on fait maintenant l'hypothèse que $(\mathbf{X}^T\mathbf{X})$ est inversible, ce qui revient à dire que $\mathrm{rg}(\mathbf{X}) = p$ et donc que $p \leq n$.
- Dans ce cas, si l'on note X_i la i-ème colonne de $\mathbf{X}$, $V_{\mathbf{X}} = \mathrm{Vect}(X_1, \dots, X_p)$, alors la projection orthogonale sur $V_{\mathbf{X}}$ est donnée par : $\forall U \in \mathbb{R}^n$,

$$\mathrm{Proj}_{V_{\mathbf{X}}}(U) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T U.$$

Calcul de $\hat{\theta}$

Proposition 3.1.1. *Si $\mathbf{X}^T\mathbf{X}$ est inversible, on a :*

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

De plus,

$$\hat{\theta} \sim \mathcal{N}(\theta^\star, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

Pour n observations, le prédicteur est alors défini pour tout $x \in \mathbb{R}^p$ par :

$$\hat{f}_n(x) = x^T\hat{\theta} = \langle x, \hat{\theta}\rangle.$$

Proposition 3.1.2. *Si $\mathbf{X}^T\mathbf{X}$ est inversible, alors (théorème de Cochan),*

$$x^T\hat{\theta} \sim \mathcal{N}(x^T\theta^\star, \sigma^2 x^T(\mathbf{X}^T\mathbf{X})^{-1}x).$$

Propriétés de $\hat{Y}$

On note $\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta}$. $\hat{Y}$ est donc le vecteur des prédictions relatives à l'échantillon d'apprentissage. D'après ce qui précède,

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathrm{Proj}_{V_\mathbf{X}}(Y).$$

On a :

Proposition 3.1.3.

$$\|\hat{\mathbf{Y}} - \mathbf{X}\theta^\star\|^2 = \|\mathbf{X}(\hat{\theta} - \theta^\star)\|^2 \sim \sigma^2\chi^2_p$$

Par conséquent,

$$\mathbb{E}[\frac{\|\hat{\mathbf{Y}} - \mathbf{X}\theta^\star\|^2}{n}] = \sigma^2\frac{p}{n}.$$

Preuve : On a vu que

$$\hat{\theta} - \theta^\star \sim \mathcal{N}(0, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

de sorte que

$$\mathbf{X}(\hat{\theta} - \theta^\star) \sim \mathcal{N}(0, \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \mathcal{N}(0, \sigma^2 P_{V_\mathbf{X}}).$$

Le théorème de Cochran nous permet alors d'en déduire le résultat (On peut par exemple le vérifier facilement dans le cas où $\mathbf{X}^T\mathbf{X}$ est la matrice I_p).

Commentaires

Lorsque p augmente, on constate que l'erreur d'apprentissage ou plutôt l'erreur d'estimation du "signal non bruité" (sur l'échantillon d'apprentissage) se comporte mal avec p. Par ailleurs, ce qu'on a fait jusqu'ici ne prend pas en compte les phénomènes d'overfitting qui peuvent apparaître lorsque le modèle est trop flexible (puisqu'on n'a pas encore considéré d'échantillon test). Se posent alors plusieurs questions :

- Dans un cadre classique, comment estimer la qualité du modèle et si besoin comment peut-on sélectionner des variables ?
- Lorsque $p > n$ (et donc en particulier lorsque $p >> n$), ce qui précède n'a pas de sens car $\mathbf{X}^T\mathbf{X}$ ne peut être inversible car de rang inférieur ou égal à $\min(n, p)$.

3.1.5 Qualité du modèle linéaire classique

Dans le modèle linéaire classique, une quantité classique est le R^2 basée sur la décomposition suivante de la somme des carrés (SC) :

$$SCT = SCM + SCR$$

où

$$SCT = \sum_{i=1}^{n}(\mathbf{Y}_k - \bar{\mathbf{Y}}_n)^2, \quad SCM = \sum_{i=1}^{n}(\hat{\mathbf{Y}}_k - \bar{\mathbf{Y}}_n)^2, \quad SCR = \sum_{i=1}^{n}(\mathbf{Y}_k - \hat{\mathbf{Y}}_k)^2.$$

Remarque : Cette décomposition de la variance empirique s'écrit sous forme condensée de la manière suivante :

$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2.$$

On remarque en particulier que le "double-produit" a disparu. Ceci est dû à la propriété suivante :

Coefficient de détermination

Proposition 3.1.4. $\bar{\mathbf{Y}} = \frac{1}{n}\sum_{k=1}^{n}\hat{\mathbf{Y}}_k$. *En particulier,* $\bar{\mathbf{Y}} \in V_{\mathbf{X}}$ *et* $\hat{\mathbf{Y}} - \bar{\mathbf{Y}} \in V_{\mathbf{X}}$. *Par conséquent,*

$$\langle \hat{\mathbf{Y}} - \bar{\mathbf{Y}}, \mathbf{Y} - \hat{\mathbf{Y}} \rangle = 0.$$

Pour prouver ce résultat, il faut se rappeler que $\mathbf{1}$ appartient à $V_{\mathbf{X}}$... Le coefficient de détermination est alors défini par :

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SST} \quad \text{(en anglais SCR=RSS)}.$$

C'est en fait le rapport de carrés de deux longueurs : si l'on "oublie" la moyenne $\bar{\mathbf{Y}}$, c'est le cosinus entre $\mathbf{Y}$ et sa projection $\hat{\mathbf{Y}}$.

Remarque: De cette interprétation géométrique, on comprend facilement que $0 \leq R^2 \leq 1$ et que l'ajout de variables explicatives augmente le R^2. Plus R^2 est proche de 1 plus l'ajustement sur les données observées est bon.

Coefficient de détermination ajusté

Proposition 3.1.5. *Soient $\mathcal{M}_1$ et $\mathcal{M}_2$ deux modèles de régression linéaire de la forme :*

$$\mathcal{M}_1 : Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$$
$$\mathcal{M}_2 : Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \beta_{p+1} X_{p+1} + \varepsilon.$$

Leurs coefficients de détermination satisfont $R_1^2 \leq R_2^2$.

Cette propriété implique que le R^2 ne détecte pas l'overfitting. On lui préfère souvent le R^2 ajusté défini par :

$$R^2_{adj} = 1 - \frac{\frac{SCR}{n-p-1}}{\frac{SST}{n-1}}.$$

Le R^2 ajusté est en fait un coefficient de détermination normalisé en fonction des paramètres p et n. Pour rappel, si $\text{Vect}(1, \mathbf{X}_1, \ldots, \mathbf{X}_p) = p + 1$, alors, on obtient via le théorème de Cochran ($Y - \hat{Y} = P_{V_{\mathbf{X}}^\perp} Y = P_{V_{\mathbf{X}}^\perp} \varepsilon$),

$$\frac{SCR}{\sigma^2} \sim \chi^2(n - p - 1).$$

Sous l'hypothèse $(H_0) := \theta_1 = \ldots = \theta_p = 0$, on a également que

$$\frac{SCT}{\sigma^2} \sim \chi^2(n - 1).$$

Plus p augmente, plus A ce stade, on aborde la "sélection de modèle" au sens où l'on prend en compte la flexibilité dans la comparaison des modèles.

Autres mesures de la qualité du modèle

Le R^2 ajusté prend donc plus en compte la fiabilité du modèle et pose la question : vaut-il mieux un modèle complet moins fiable statistiquement qu'un modèle réduit biaisé mais d'estimation plus précise (question du compromis Biais-Variance) ? Il existe d'autres moyens de comparer les modèles (pénalisant la montée en dimension). On en présente ici succinctement quelques-uns (dans le cadre du modèle linéaire gaussien uniquement) :

- Le C_p de Mallows défini par

$$\frac{1}{n}(SCR + 2p\hat{\sigma}^2) \quad \text{que l'on cherche à minimiser.}$$

- Le critère AIC (Akaike Information Criterion), proportionnel au C_p dans le cas du modèle linéaire (uniquement)

$$AIC = \frac{1}{n\hat{\sigma}^2}(SCR + 2p\hat{\sigma}^2)$$

- Le critère BIC (Bayesian) :

$$BIC = \frac{1}{n}(SCR + \log(n)\hat{\sigma}^2)$$

  Là encore, on cherche à minimiser ce critère.

Et la validation ?? Et dans la pratique ??

- Validation : on a pu constater que dans ce chapitre les aspects "erreur test" et "erreur de validation". Les outils présentés plus hauts peuvent être vus comme des alternatives à la validation. Ces outils ont en particulier été développés pour pallier la difficulté computationnelle générée par la validation croisée. Néanmoins, aujourd'hui, ce problème est moins prégnant avec l'accélération des calculateurs.

- Pratique : qu'on utilise validation croisée, R^2_{adj}, C_p, AIC ou BIC, une question se pose, comment comparer l'ensemble des modèles que l'on peut fabriquer avec p variables ? En effet, il en a 2^p !!

- Réponse : Algorithmes de sélections de variables "Pas à Pas", "Mixtes", "Globaux" selon les situations.

Algorithmes de sélection de variables

Présentons quelques méthodes (qui possèdent toutes leurs variantes) :

- Forward : A chaque pas, une variable est ajoutée au modèle. C'est celle dont le R^2 est maximal (ou de manière équivalente celle dont le SCR est minimal). On obtient ainsi p modèles contenant 1, 2, 3, ..., p variables. On choisit ensuite celui qui minimise un critère pénalisé ($R^2_{adj}, \ldots$).

- Backward : A chaque pas, une variable est retirée au modèle : celle qui augmente le moins le SCR (ou qui diminue le plus moins le R^2.

- On a aussi des approches hybrides (qui mélangent ces deux approches).

3.2 LASSO/RIDGE/ELASTIC NET

3.2.1 La pénalisation par la norme de θ

Les méthodes présentées ci-dessus souffrent de deux problèmes :

- Lorsque p est grand, leur mise en oeuvre est coûteuse et l'exploration des modèles reste nécessairement très partielle.
- Lorsque $p > n$, elles ne sont a priori pas applicables puisque l'estimateur $\hat{\theta}$ n'est pas bien défini. En effet, comme $rg(X) < p$, Ker$X$ est non réduit à 0. Ainsi,

  $$\text{Argmin}_\theta \|\mathbf{Y} - \mathbf{X}\theta\|^2$$

  n'est pas unique.
- Une théorie assez récente permet de pallier le second problème en introduisant une pénalisation dans l'estimateur. Il permet aussi de donner une alternative à la sélection de modèles dans le cadre classique.

Principe général de la régression pénalisée

- Pour un $\lambda > 0$, on cherche à minimiser la fonction

  $$\Phi_{\mathcal{P}}(\theta) = \frac{\|\mathbf{Y} - \mathbf{X}\theta\|^2}{2n} + \lambda \mathcal{P}(\theta).$$

  $\mathcal{P}$ désigne la *pénalité*.
- Quelques exemples de pénalités :
  - $\mathcal{P}(\theta) = \|\theta\|_0 := \text{Card}\{i, \theta_i \neq 0\}$.
  - $\mathcal{P}(\theta) = \|\theta\|_1 = \sum_{i=1}^p |\theta_i|$ (LASSO)
  - $\mathcal{P}(\theta) = \frac{\|\theta\|_2^2}{2} = \frac{1}{2}\sum_{i=1}^p |\theta_i|^2$ (Ridge).
  - $\mathcal{P}(\theta) = (1-\alpha)\frac{\|\theta\|_2^2}{2} + \alpha\|\theta\|_1$, $\alpha \in [0,1]$ (Elastic Net).

Rôle de la pénalisation

- On favorise les solutions pour lesquelles $\mathcal{P}(\theta)$ est petit.
- Ainsi, par exemple, en norme $\|\ ,\ \|_0$, on pénalise le nombre de variables. On favorise les solutions qui ont peu de variables "allumées".

- **Remarque** : les C_p, AIC et BIC ressemblent à des pénalisations par cette norme sauf que la pénalisation est "uniforme".
- Interprétation Lagrangienne : la fonction $\Phi_{\mathcal{P}}(\theta)$ est à comprendre comme un *Lagrangien*, *i.e.* une fonction fabriquée pour *régulariser* le problème d'optimisation sous contrainte
$$\min_{\mathcal{P}(\theta)\leq s(\lambda)} \frac{\|\mathbf{Y}-\mathbf{X}\theta\|^2}{2n}.$$
(voir plus loin pour quelques rappels)

Représentation graphique

Représentons ici le problème sous contrainte (en dimension 2) :

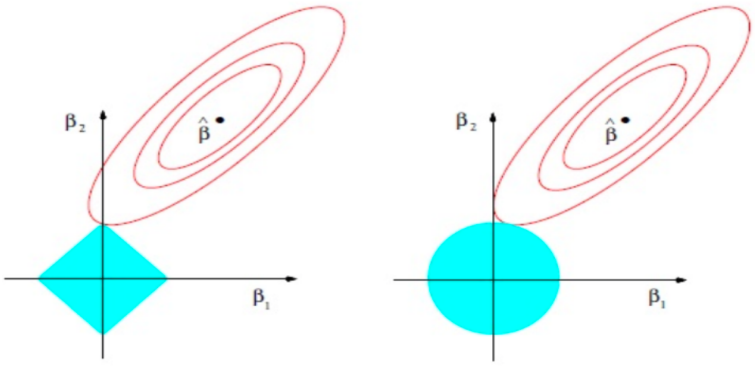

Figure 3.1: Gauche : en norme 1, Droite, en norme 2 ($\hat{\beta} = \text{Argmin}_{\beta\in\mathbb{R}^2}\|Y - X\beta\|^2$, en rouge, les lignes de niveau de $\|Y - X\beta\|^2$ autour du minimum)

3.2.2 Rappels succincts sur le Lagrangien

On considère le problème suivant : minimiser la fonction J sur un sous-ensemble G défini par $G = \{\theta \in \mathbb{R}^p, f(\theta) = 0\}$ (f supposée régulière ici).

- Exemple : $J(\theta) = \|Y - X\theta\|^2$ et $f(\theta) = \|\theta\|_2^2 - R$.
- Dans ce cas, le théorème des multiplicateurs de Lagrange donne la condition nécessaire suivante :

Théorème 3.2.1. *Si $\theta^\star$ est un point de G où J atteint son minimum, alors il existe $\lambda^\star$ tel que*
$$\nabla J(\theta^\star) + \lambda^\star \nabla f(\theta^\star) = 0.$$
Ainsi, si l'on note L la fonction définie par
$$L(\lambda, \theta) = J(\theta) + \lambda f(\theta)$$

alors, $(\lambda^\star, \theta^\star)$ *est solution de*

$$\partial_\theta L(\lambda^\star, \theta^\star) = \partial_\lambda L(\lambda^\star, \theta^\star) = 0.$$

Admettons maintenant que J et f sont convexes. Dans ce cas,

- $\mathcal{C} = \{\theta, f(\theta) \leq 0\}$ est convexe.
- Si $\min_{\theta \in \mathbb{R}^p} J(\theta)$ est en dehors de $\mathcal{C}$, on peut alors vérifier que $\min_{\theta \in \mathcal{C}} J(\theta)$ est nécessairement sur le bord du convexe.
- Les points critiques de $\theta \mapsto L(\lambda, \theta)$ sont des minimums.
- **Pour aller plus loin** :Pour rendre la condition nécessaire du théorème précédent suffisante, on s'appuie souvent sur les conditions "KKT".
- **Important :** Considérer la fonction $\theta \mapsto L(\lambda, \theta)$ permet de considérer un problème d'optimisation non contraint. Lorsque la fonction L est convexe, alors, on peut mettre en oeuvre des algorithmes d'optimisation (type descente de gradient) pour déterminer le minimum de L
- **En pratique** : on cherche le minimum $\theta^\star_\lambda$ de $\theta \mapsto L(\lambda, \theta)$ (à λ fixé) puis on fait varier λ en étudiant l'évolution de $J(\theta^\star)$.
- **Cas limites** :Lorsque $\lambda = 0$, on regarde le problème non pénalisé, lorsque $\lambda \to +\infty$, on cherche à déterminer la meilleure approximation sur une boule de rayon tendant vers 0.

3.2.3 A propos de la norme $\|\,,\,\|_0$

- Si l'on veut pénaliser le nombre de variables pour réduire la variance du modèle, le choix naturel est la norme $\|\,,\,\|_0$.
- **Parcimonie** : on parle dans ce cas d'hypothèse de parcimonie ("sparsity") : on fait l'hypothèse que peu de variables sont réellement explicatives. Si tel est le cas, *i.e.* que le vrai modèle est de la forme

  $$Y = X\theta^\star + \varepsilon$$

  avec $\|\theta^\star\|_0 = \{j, \theta^\star_j \neq 0\} = s_0$ petit devant p (et si possible devant n également).
- Notons $\mathbf{X}_{s_0}$ la sous-matrice de $\mathbf{X}$ de taille $n \times s_0$ constituée des colonnes "allumées" de $\mathbf{X}$. Si l'on suppose que $\mathbf{X}^T_{s_0}\mathbf{X}_{s_0}$ est inversible, alors, une fois "déterminé ce support", l'application du modèle linéaire standard permet d'obtenir une erreur en $\frac{s_0}{n}$.
- **Problème** : La norme $\|\,,\,\|_0$ n'est pas une fonction convexe !! Chercher le minimum de la fonction $\Phi_{\|\,\|_0}$ n'est pas un problème soluble numériquement car il faut explorer toutes les configurations de manière exhaustive (pour un support de taille s_0, on a s_0^p possibilités).

De la norme $\|\,,\,\|_0$ à la norme $\|\,,\,\|_1$

- **Idée** : Remplacer la norme $\|\,,\,\|_0$ par la norme convexe "minimale" : la norme $\|\,,\,\|_1$ (pour comprendre la notion de minimalité de convexité, penser à l'application $x \mapsto |x|$ en dimension 1).

- **LASSO** : Acronyme de Least Absolute Shrinkage and Selection Operator (introduit par Tibshirani en 1996).

- **Objectifs du LASSO** : Obtenir avec une norme moins "sélective" que la norme $\|\,,\,\|_0$ des résultats qui sont sensiblement similaires à ceux que l'on obtiendrait avec la norme "idéale. Notons

$$\hat{\theta}_\lambda = \text{Argmin}_{\theta \in \mathbb{R}^p} \frac{\|\mathbf{Y} - \mathbf{X}\theta\|^2}{2n} + \lambda \|\theta\|_1.$$

  - Estimation : $\hat{\theta}_\lambda$ proche de $\theta^\star$?
  - Sélection : si J_0 est le "vrai" support, obtenir avec grande probabilité $J(\theta^\star_\lambda)$ proche de J_0 ?
  - Prédiction : $\frac{1}{n}\|\mathbf{X}(\hat{\theta}_\lambda - \theta^\star)\|_2^2 = O(\frac{s_0}{n})$ avec grande probabilité ?

3.2.4 Lasso vs Ridge et Elastic Net

Ridge et Elastic Net sont aussi basés sur des normes convexes (même strictement convexes) (Ridge a même une solution explicite). Quel est leur rôle statistique ?

- LASSO : tente de mimer la norme $\|, \|_0$ en tentant d'extraire des variables d'intérêt et en annulant les autres.

- Ridge est une méthode plus "douce" qui réduit ("shrinks") la taille des variables, notamment celles qui sont corrélées.

- Elastic Net tente donc de mélanger ces deux approches.

- Dans la suite, on parlera principalement du LASSO et du Ridge.

3.2.5 Evolution avec λ

- La fonction $\lambda \to \Phi_{\mathcal{P}}$ est croissante avec λ.

- Comment choisir λ ? Il s'agit d'un compromis biais-variance. Plus λ est petit, plus la variance est élevée. Plus λ est grand, moins le modèle est flexible et donc plus le biais est élevé.
- On choisit donc λ en minimisant (l'estimation de) l'erreur de prédiction.
- Ce choix est fait par validation (simple ou croisée) (voir figure slide suivant).
- On trace aussi généralement les coefficients sélectionnés en fonction de λ (voir figure slide suivant).

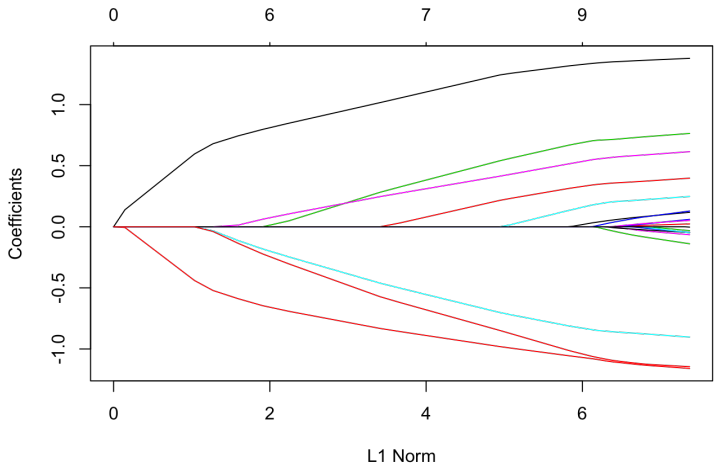

Figure 3.2: Apparition des coefficients dans le LASSO en fonction de $\|\hat{\theta}_\lambda\|_1$ (remarque : à l'origine $\lambda = +\infty$)

3.2.6 Un résultat simple pour le Ridge

- si $\mathcal{P}(\theta) = \frac{\|\theta\|_2^2}{2}$, alors la fonction $\Phi_{\mathcal{P}}$ définie par

$$\Phi_{\mathcal{P}}(\theta) = \frac{\|\mathbf{Y} - \mathbf{X}\theta\|^2}{2n} + \lambda \mathcal{P}(\theta)$$

est strictement convexe de classe $\mathcal{C}^2$ et coercive (tend vers $+\infty$ lorsque $\|\theta\|_2 \to +\infty$). On a alors le résultat suivant :

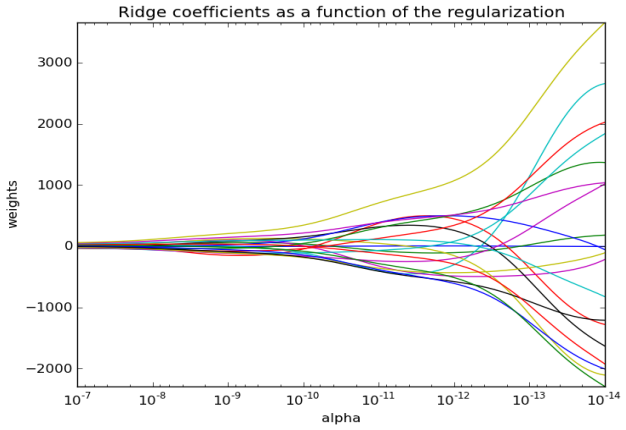

Figure 3.3: Apparition des coefficients dans le RIDGE en fonction de λ (plus "smooth")/"Chemin de régularisation"

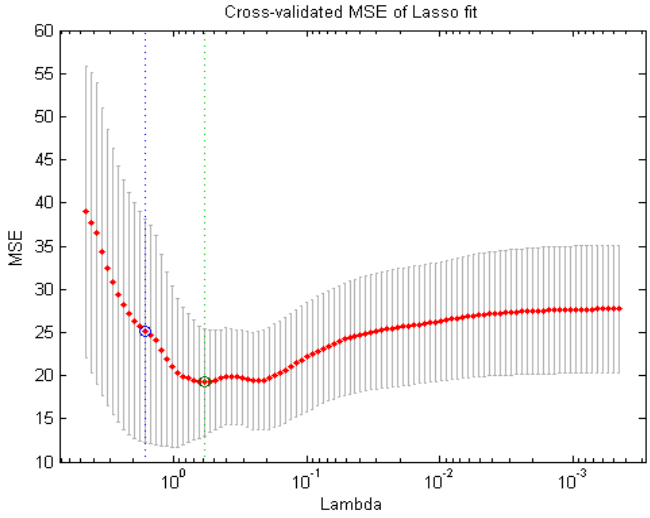

Figure 3.4: Estimation de la MSE en fonction de λ

Proposition 3.2.2. *Si* $\mathcal{P}(\theta) = \frac{\|\theta\|_2^2}{2}$, *alors* $\Phi_{\mathcal{P}}$ *admet un unique minimum en*

$$\hat{\theta}_\lambda = (X^T X + \lambda n I_p)^{-1} X^T Y.$$

Remarque : 1 -Inverse bien définie car $X^T X + \lambda n I_p$ est une matrice définie positive.
2 - On voit l'effet de la pénalisation qui apparaît au "dénominateur" et réduit la taille des coefficients. On voit également que l'on ne "tue" pas les coefficients.

3.2.7 "Calcul" de $\hat{\theta}_\lambda$ pour le LASSO

Le problème est plus difficile car $\|\theta\|_1$ est non strictement convexe et non dérivable sur les axes.

- **Un cas particulier** : $\mathbf{X}^T\mathbf{X} = I_p$ (ce qui implique que $p \leq n$. Dans ce cas, la première forme quadratique est définie positive donc $\Phi_{\mathcal{P}}$ est strictement convexe et coercive. Elle admet donc un unique minimum. On a de plus

Proposition 3.2.3 (Seuillage doux (Soft-Thresholding))**.** *Si* $\mathbf{X}^T\mathbf{X} = I_p$, *alors si l'on note* $\hat{\theta} = X^T Y$ *la solution du modèle linéaire non pénalisé, on a :*

$$\hat{\theta}_\lambda(j) = \text{sgn}(\hat{\theta}(j))(|\hat{\theta}(j)| - \lambda)_+.$$

Remarque : On voit donc sur cet exemple que le LASSO sélectionne les coordonnées de la solution classique qui sont supérieures à λ mais les "shrinke" (les réduit) aussi.

Dans le cas général, il n'y a pas nécessairement unicité à la solution du LASSO. De plus, "la" solution n'est pas explicite. On doit faire recours à un algorithme d'optimisation pour déterminer $\hat{\theta}_\lambda$. On peut néanmoins affirmer que :

Proposition 3.2.4. $\theta \in \mathbb{R}^p$ *est solution du LASSO de paramètre* λ *si et seulement si les conditions de "stationnarité" suivantes sont satisfaites : pour tout* $j \in \{1, \ldots, p\}$

$$\begin{cases} \mathbf{X}_j^T(Y - \mathbf{X}\theta) = \lambda \text{sgn}(\theta_j) & \text{si } \theta_j \neq 0 \\ |\mathbf{X}_j^T(Y - \mathbf{X}\theta)| \leq \lambda & \text{sinon.} \end{cases}$$

3.2.8 En pratique

2 types d'algorithmes pour calculer $\hat{\theta}_{LASSO}$ (voir TD pour plus de détails)

- **L'algorithme LARS** (Least Angle Regression). Calcul par palier des variables allumées (cf graphe d'apparition des coefficients). On fait décroître λ jusqu'à l'apparition d'une première variable (la plus corrélée avec y). Ensuite, on cherche le seuil pour lequel la deuxième variable apparaît (toujours par un principe de corrélation ou d'angle)

- **La descente coordonnées par coordonnées.** Il s'agit d'un algorithme d'optimisation où l'on minimise coordonnées par coordonnées. Etape 1 : on fige $\theta_2, \dots, \theta_p$ et on regarde la fonction

$$\theta_1 \mapsto \frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{k\neq j} x_{ik}\theta_k - x_{i1}\theta_1)^2 + \lambda\sum_{k\neq 1}\theta_k + \lambda|\theta_1|.$$

  On en tire un minimum (explicite dans le cas du LASSO) $\hat{\theta}_1$ puis on réitère la minimisation sur la deuxième coordonnée en conservant la valeur de $\hat{\theta}_1 \dots$. Par construction, il s'agit d'une suite décroissante. Sous conditions standard, l'algorithme converge vers un minimiseur (même principe que EM).

3.3 Prédiction/Estimation : Résultats

3.3.1 La condition RE

Définition 3.3.1. *Soit $\alpha > 0$ et J_0 le support de $\beta^\star$. On note*

$$C_\alpha(J_0) = \{\Delta \in \mathbb{R}^p, \|\Delta^c_{J_0}\|_1 \leq \alpha\|\Delta_{J_0}\|_1\}.$$

La matrice de design $\mathbf{X}$ vérifie la Restricted Eigenvalue (RE) condition sur J_0 avec les paramètres (κ, α) si

$$\|\mathbf{X}\Delta\|_2 \geq \kappa\|\Delta\|_2 \quad \forall \Delta \in C_\alpha(J_0).$$

3.3.2 Estimation

Théorème 3.3.2. *Supposons que la condition RE vérifiée avec $\alpha = 3$ et $\kappa > 0$. (i) (Résultat algébrique) Alors, si $\lambda > \|\frac{X^T\varepsilon}{n}\|_\infty$,*

$$\|\hat{\theta} - \theta^\star\|_2 \leq \frac{3\lambda\sqrt{s_0}}{\kappa}.$$

(ii) (Résultat probabiliste) Supposons que les ϵ_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et que les colonnes de $\mathbf{X}$ ont été standardisés de sorte pour tout $j \in \{1, \dots, p\}$, $n^{-\frac{1}{2}}\|\mathbf{X}_j\| \leq C$. Alors, si $\delta > 0$ et $\lambda = 2C\sigma\sqrt{\frac{2\log p + \delta^2}{n}}$, le point (i) est vérifié avec probabilité $1 - 2e^{-\frac{\delta^2}{2}}$. Ainsi, avec grande probabilité,

$$\|\hat{\theta} - \theta^\star\|_2^2 \lesssim \frac{s_0 \log p}{n}$$

où $s_0 = \text{Card}\{i, \beta_i^\star \neq 0\}$.

Idée de preuve de (ii)

$\|\frac{X^T\varepsilon}{n}\|_\infty$ correspond au maximum de p variables gaussiennes. Sous les hypothèses sur les colonnes, la variance de chacune de ces variables est bornée par $C\frac{\sigma^2}{n}$. Ainsi, via un résultat similaire à celui obtenu en TD sur le maximum de variables gaussiennes, on a :

$$\mathbb{P}(\|\frac{X^T\varepsilon}{n}\|_\infty \geq C\sigma\sqrt{\frac{2\log p + \delta^2}{n}}) \leq 2e^{-\frac{\delta^2}{2}}.$$

Le résultat suit.

3.3.3 Prédiction

Théorème 3.3.3. *Supposons que les ϵ_i sont i.i.d. de loi $\mathcal{N}(0,\sigma^2)$ et que les colonnes de $\mathbf{X}$ ont été standardisés de sorte pour tout $j \in \{1,\dots,p\}$, $n^{-\frac{1}{2}}\|\mathbf{X}_j\| \leq C$. Alors, si $\delta > 0$ et $\lambda = 2C\sigma\sqrt{\frac{2\log p+\delta^2}{n}}$, (i)*

$$\frac{\|X(\hat{\theta}_\lambda - \theta^\star\|_2^2}{n} \lesssim \|\beta^\star\|_1\sigma\sqrt{\frac{s_0\log p}{n}}.$$

(ii) Si la condition RE est satisfaite, alors

$$\frac{\|X(\hat{\theta}_\lambda - \theta^\star\|_2^2}{n} \lesssim \frac{s_0\log p}{n}.$$

3.3.4 Recouvrement du support

Théorème 3.3.4. *Notons $\mathbf{X}_{J_0}$ la sous-matrice obtenue en conservant les colonnes actives de $\theta^\star$. Si $\mathbf{X}_{J_0}^T\mathbf{X}_{J_0}$ est inversible et qu'une condition d'incohérence mutuelle est satisfaite. Alors, si les colonnes sont normalisées comme précédemment le choix $\lambda \approx \sqrt{\frac{\log p}{n}}$ permet d'assurer avec grande probabilité que le support de $\hat{\theta}_\lambda$ est contenu dans celui de $\theta^\star$.*

Bibliographie

[1] Sylvain Arlot. Fondamentaux de l'apprentissage statistique. Disponible sur Moodle.

[2] Christophe Giraud. Introduction to High-Dimensional Statistics. Chapman & Hall.

[3] Trevor Hastie, Robert Tibshirani, Gareth James, Daniela Witten. Introduction to Statistical Learning. http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition. February 2009. https://web.stanford.edu/~hastie/ElemStatLearn/.

[5] Statistical Learning with Sparsity: the Lasso and Generalizations. https://web.stanford.edu/~hastie/StatLearnSparsity/.

[6] Christophe Giraud. Introduction to High-Dimensional Statistics. Chapman & Hall.