Apprentissage Statistique en Grande Dimension

F. Panloup
LAREMA-Université d'Angers

—

INTRODUCTION

—

Plan du cours

- Apprentissage Statistique : Fondements
- Arbres/Agrégation de modèles/Bootstrap/Boosting
- L'effet de la dimension
- Méthodes pénalisées (LASSO/Ridge. . . et extensions)
- Support Vector Machine/Méthodes à noyau
- Réseaux de Neurones et Deep Learning
- Modèles graphiques

Des références

- Trevor Hastie, Robert Tibshirani, Gareth James, Daniela Witten. Introduction to Statistical Learning. `http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf`
- 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/`.
- Statistical Learning with Sparsity: the Lasso and Generalizations. `https://web.stanford.edu/~hastie/StatLearnSparsity/`.
- Christophe Giraud. Introduction to High-Dimensional Statistics. Chapman & Hall.
- Sylvain Arlot. Fondamentaux de l'apprentissage statistique. Disponible sur Moodle.
- Quelques exemplaires disponibles à la BU.

Introduction à l'apprentissage statistique

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,...

Généralités

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,...

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 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 : $Y = f(X) + \varepsilon$ où f est une fonction appartenant à une classe de fonctions fixée au départ.

Supervisé

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.

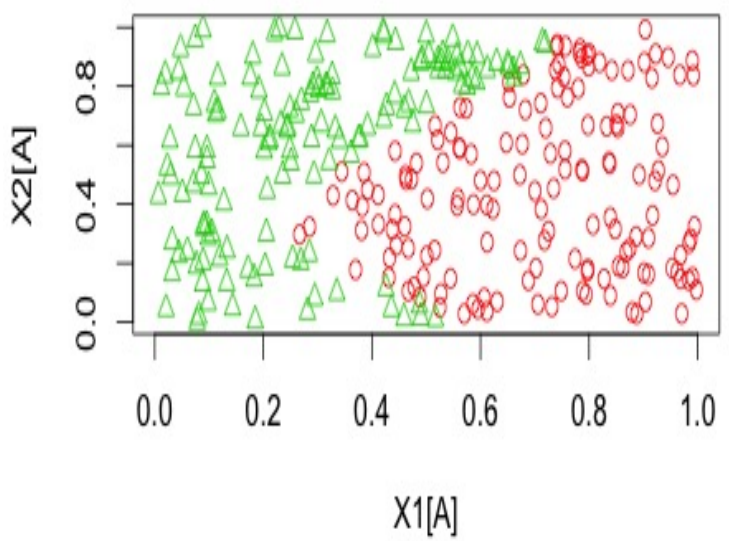


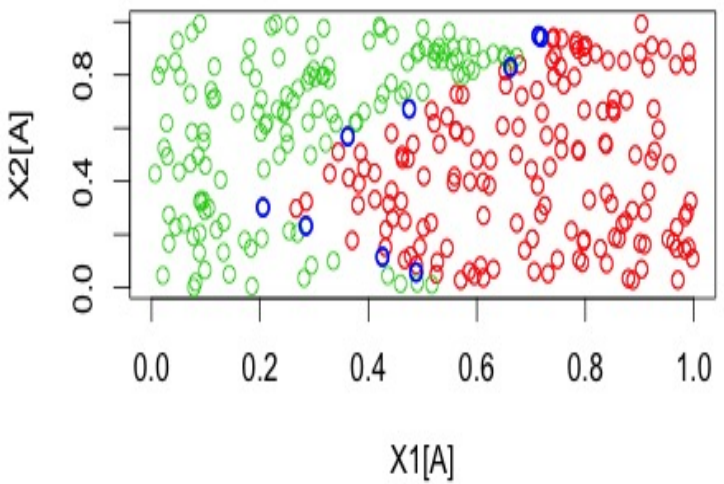
Figure: Deux groupes de couleur différente

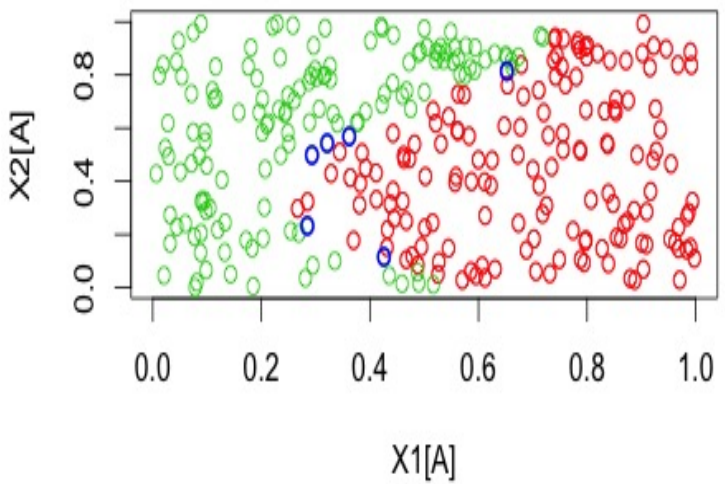
On reviendra sur cet exemple de base dans la suite.

Supervisé

Un exemple 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 :
  - 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". Les méthodes de **K-means** ou de **clustering hiérarchique** en sont des exemples classiques.
  - ceux qui permettent d'approximer/de générer des lois 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. . .

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" : c'est un grand débat ! Faut-il choisir la médiane, faut-il tirer la variable au hasard, faut-il "l'apprendre" ?
  - 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*.
  - 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, ce transparent est de type "cuisine". Il n'y a pas de vraies garanties théoriques ...

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).

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 : $Y = f(X) + \varepsilon$ où f vit dans un espace de fonctions de dimension infinie (alors que la régression linéaire par exemple est clairement paramétrique).

Classification supervisée - Règle et qualité de prévision

Soit $(\mathbf{X}_1, Y_1), \ldots (\mathbf{X}_n, Y_n)$, un *échantillon d'apprentissage* issu d'une loi conjointe $\mathbb{P}$ sur $\mathcal{X} \times \mathcal{Y}$.

Definition

- Une règle de prévision/régression/décision/discrimination est une fonction $f : \mathcal{X} \to \mathcal{Y}$ qui à $\mathbf{x}$ associe la sortie $f(\mathbf{x})$.
- 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 : $\ell(y, y') = |y - y'|^2$ (qui donnera le risque quadratique utilisée en régression) et en classification : $\mathbf{1}_{y \neq y'}$.

Risque

Definition

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.

Espérance conditionnelle : rappels rapides

Loi/espérance conditionnelle : notions importantes en apprentissage.

Définition (et proposition)

Soit X et Y deux variables aléatoires. Alors, si Y est intégrable, il existe une fonction ϕ (mesurable) tel que $U = \phi(X)$ satisfait

$$\mathbb{E}[Yh(X)] = \mathbb{E}[Uh(X)] \quad \text{pour toute fonction } h \text{ mesurable bornée.} \qquad (1)$$

Cette variable U est unique au sens suivant : s'il existe $\tilde{U} = \tilde{\phi}(X)$ tel que (1) est vraie alors $U = \tilde{U}$ presque sûrement. U est alors appelée l'espérance conditionnelle de X sachant S et notée : $U = \mathbb{E}[Y|X]$.

- Si $\mathcal{X}$ discret, X v.a. à valeurs dans $\mathcal{X}$ telle que $\mathbb{P}(X = x) > 0$ pour tout $x \in \mathcal{X}$,

$$\mathbb{E}[Y|X] = \Phi(X) \quad \text{où,}$$

$$\Phi(x) = \mathbb{E}[Y|X = x] = \frac{1}{\mathbb{P}(X = x)} \mathbb{E}[Y 1_{\{X=x\}}].$$

Espérance conditionnelle : rappels rapides

Proposition

1. *En particulier,* $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$ *(prendre* $g = 1$*).*
2. *Si* Y *est* $\sigma(X)$*-measurable,* $\mathbb{E}[Y|X] = Y$*.*
3. *Si* Y *est indépendant de* X*, alors* $\mathbb{E}[Y|X] = \mathbb{E}[Y]$*.*
4. *Si* Y *est dans* L^2*,*

$$\mathbb{E}[Y|X] = \inf_{Z\ \sigma(X)\text{-mesurable tq } \mathbb{E}[|Z|^2] < +\infty} \mathbb{E}[(Y-Z)^2]. \tag{2}$$

C'est une projection orthogonale sur. . .

Loi conditionnelle

Définition

Supposons que $\mathbb{P}(X = x) \neq 0$. *Dans ce cas, la loi de* Y *sachant* $X = x$ *est la mesure de probabilité* $\mu_{Y|X=x}$ *définie par: pour tout* f *borélienne bornée :*

$$\int f d\mu_{Y|X=x} = \mathbb{E}[f(Y)|X = x].$$

Plus généralement, la loi de Y *sachant* X *est une variable aléatoire à valeurs dans l'espace des probabilités définie par : pour tout* f *borélienne bornée, p.s.,*

$$\int f d\mu_{Y|X} = \mathbb{E}[f(Y)|X].$$

Exercice

- Déterminez la loi de Y lorsque $\mathcal{L}(Y|X) = \mathcal{B}(e^{-X})$ et que X suit la loi uniforme sur $[0, 1]$.
- Supposons que (X, Y) a une densité $f_{(X,Y)}$ et que f_X est strictement positive. Quelle forme prend la densité de la loi conditionnelle $\mathcal{L}(Y|X)$?

Risque

Definition

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,

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

Une telle règle est appelée prédicteur/règle de Bayes. Pour une règle de prévision f, la quantité $R_f - R_{f^\star}$ est appelée l'excès de risque tandis que le risque minimal est appelé risque de Bayes.

Remarque. En pratique, on cherche souvent la règle de décision au sein d'une sous-classe $\mathcal{S}$ de l'ensemble des fonctions. Par exemple, lorsque l'on fait de la régression linéaire, on cherche à trouver la meilleure fonction f parmi les fonctions de la forme $f(x) = \langle x, \theta \rangle$, $\theta \in \mathbb{R}^p$. Cela génère le *biais* ou l'*erreur d'approximation*

$$\inf_{f \in \mathcal{S}} R_f - R_{f^\star}.$$

Risque

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

Théorème

(i) En régression avec $\mathcal{Y} = \mathbb{R}$ *et* $\ell(y, y') = |y - y'|^2$, $f^\star(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}]$ *est la règle de Bayes.*
(ii) En régression avec $\mathcal{Y} = \mathbb{R}$ *et* $\ell(y, y') = |y - y'|$, *la fonction* $f^\star(\mathbf{x}) = \text{médiane}(\mathcal{L}(Y|X = \mathbf{x}))$ *est la règle de Bayes.*
(iii) En classification (avec $\mathcal{Y}$ *de cardinal fini) et* $\ell(y, y') = 1_{y \neq y'}$, *la fonction* $f^\star$ *définie pour tout* $\mathbf{x} \in \mathcal{X}$ *par*

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

Notons

$$\eta(x, y) = \mathbb{P}(Y = y|\mathbf{X} = x).$$

A x *fixé, la probabilité de se tromper avec la règle* $f^\star$ *est égale à* $1 - \max_{y \in \mathcal{Y}} \eta(x, y)$. *On a donc*

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

Risque (suite)

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 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.

Remark

Dans certains algorithmes récents (XGBOOST, Réseaux de neurones), approche utilisée en classification moins basée sur l'erreur dite de classification, accuracy en anglais (voir plus loin softmax/entropie croisée). Par ailleurs, dans les approches classiques, comme cela a été signalé précédemment, l'erreur de classification est généralement complétée par d'autres mesures telles que le F_1-score afin de mieux mesurer la prédiction associée à chaque classe (et d'éviter par exemple des désagréments liés au déséquilibre des classes...).

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,

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

- En régression linéaire standard $\mathbf{y} = \langle \mathbf{x}, \theta \rangle + \varepsilon$, $\hat{f}_n(\mathbf{x}) = \mathbf{x}^T \hat{\theta}_n$ où

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

- 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.

Definition

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 $\ell(y, y') = (y - y')^2$:
$$\mathbb{E}_{(\mathbf{X},Y)}[R_{\hat{f}_n}] = \mathbb{E}[(Y - \mathbf{X}^T \hat{\theta}_n)^2].$$

Consistance et algorithme par moyennisation locale

Definition

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 :

Definition

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.

Consistance faible et algorithme par moyennisation locale (suite)

Si l'on revient maintenant à la définition des règles optimales, on peut montrer sous des conditions assez générales (cf TD) que

Théorème

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).

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.

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).

Overfitting

Ce phénomène est appelé "surapprentissage" (overfitting). Ci-dessous, deux situations simples pour bien comprendre (où le risque est nul à droite !!).

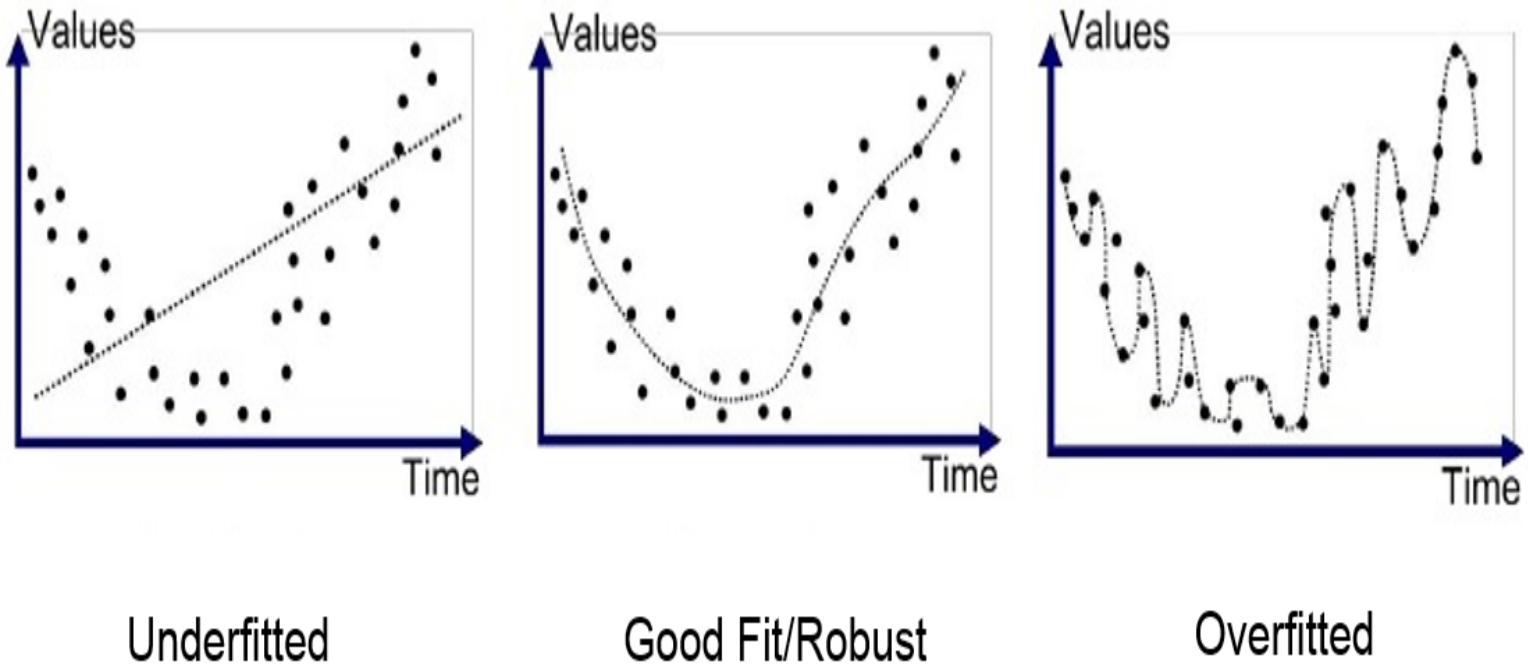


Figure: (taken from `https://medium.com/greyatom`)

Overfitting (2)

Ce phénomène est appelé "surapprentissage" (overfitting). Ci-dessous, deux situations simples pour bien comprendre (où le risque est nul à droite !!).

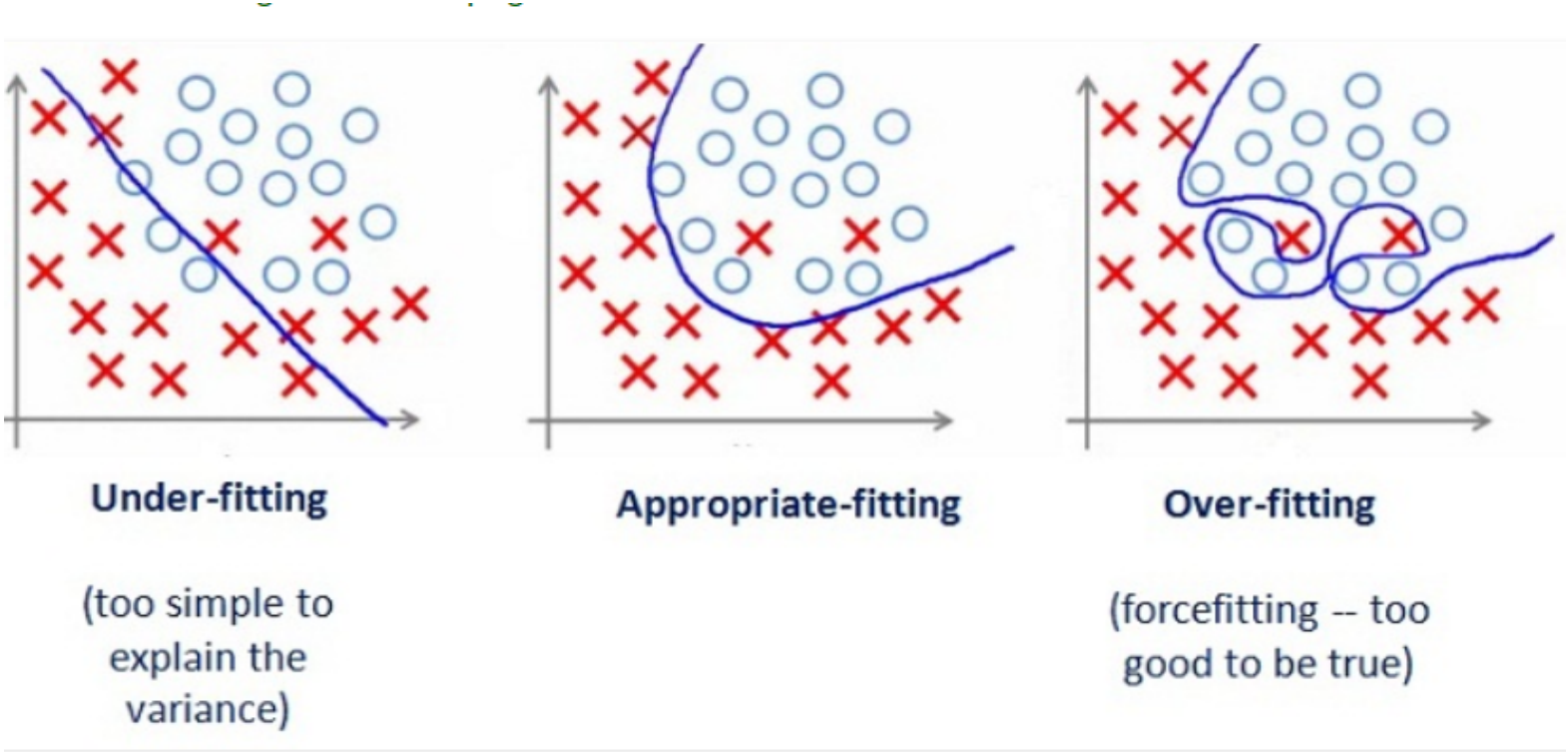


Figure: (taken from `https://medium.com/greyatom`)

Estimation du risque

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

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

où $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ 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...).

Estimation du risque (suite)

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

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

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).

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

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.

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.

Validation croisée

Schématiquement, dans le cas $K = 5$ (K-folds cross-validation),

Validation croisée (suite)

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 (encore) 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. 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.

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, F_1-score, MCC, voir TD).