Pratique de l'apprentissage statistique

1. Apprentissage statistique

V. Lefieux

École des Ponts
ParisTech

Plan

Plan

Apprentissages supervisé et non-supervisé

- Apprentissage supervisé :
  Inférer (prédire) une fonction ou une relation à partir de données d'apprentissage labellisées (ex : classification supervisée, régression).

- Apprentissage non-supervisé :
  Trouver une « structure » dans des données non-labellisées (ex : clustering).

  Même s'il est plus « subjectif » que l'apprentissage supervisé, il peut être utile comme étape de pré-traitement pour l'apprentissage supervisé.

Quelques méthodes d'apprentissage supervisé

- Statistique « classique » : régression linéaire, régression paramétrique non-linéaire, régression logistique, méthodes de régularisation (Ridge, Lasso, Lars), PLS.
- Méthodes bayésiennes.
- Méthodes de moyennage local : plus proches voisins, noyau de lissage, CART.
- Méthodes à bases de splines : régression spline, GAM.
- Méthodes à directions révélatrices : SIM, SIR, etc.
- Agrégation : bagging, boosting.
- Méthodes à noyau : SVM.
- Réseaux de neurones.

Exemples d'apprentissage supervisé

- Régression :
  - Pollution.
  - Vente de produits.
  - Prix de marché.
- Classification supervisée :
  - Médecine.
  - Credit scoring.
  - Reconnaissance de texte.
  - Reconnaissance d'images.

Modélisation et/ou prévision

- On peut distinguer modélisation et prévision, par exemple compression d'image vs reconnaissance d'images.
- Un modèle s'appuie sur la régularité des phénomènes sous-jacents.
- La prévision consiste à généraliser un modèle.

Un exemple

- Source : http://yann.lecun.com/exdb/mnist/.
- Modéliser :

- Prévoir :

Quelques enjeux en prévision

- Compromis entre la qualité de la prévision et l'interprétabilité (notion de « boîte noire »).
- Privilégier des modèles parcimonieux (« sparse ») qui éviteront le sur-apprentissage : *less is more*.

Plan

Introduction

Formalisation du problème

Pertes et risques

Biais et variance

Validation croisée

Données

- On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1,\dots,n\}}$$

  où $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$.

- On note :

$$d_n = (x_i, y_i)_{i \in \{1,\dots,n\}} \ .$$

Objectif

On se placera dans le cadre de la prévision : on souhaite prévoir y pour une nouvelle valeur x.

Covariables

On considèrera très souvent dans la suite que :

$$X \in \mathbb{R}^p \ .$$

Par défaut les covariables seront considérées comme quantitatives mais on indiquera régulièrement comme traiter les variables qualitatives.

Régression et classification supervisée

- Régression : la variable Y est quantitative.
  Dans la suite on considèrera que $Y \in \mathbb{R}$.
  Mais il est possible de considérer plus généralement $Y \in \mathbb{R}^d$.
- Classification supervisée : la variable Y est qualitative.
  Dans la suite on considèrera que $Y \in \{-1, 1\}$.
  Par défaut la classification supervisée sera considérée binaire mais on indiquera régulièrement comme traiter plus de 2 modalités.

Prévision

- On suppose que (x_i, y_i) est la réalisation d'une v.a.r (X_i, Y_i) de loi de probabilité inconnue $P_{X,Y}$ (modèle statistique non-paramétrique).
- La fonction de prévision de Y est une fonction $f : \mathcal{X} \rightarrow \mathcal{Y}$.
- On suppose que $f \in \mathcal{F}$.
- Dans la suite, de manière plus spécifique que f, on désignera la fonction de lien par :
  - Cas de la classification supervisée : g .
  - Cas de la régression : m .
- On cherche à estimer f par $\widehat{f}$.

Plan

Qualité d'un prédicteur

- La qualité d'un prédicteur $\widehat{f}$ est évaluée par le risque R (ou encore erreur de généralisation) qui :
  - permet de sélectionner un modèle,
  - fournit un indice de la confiance qu'on peut avoir en une prévision.

- Le risque est définie à partir d'une fonction de coût (ou encore fonction de perte).

Fonctions de perte

- On appelle fonction de perte une fonction $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ telle que :
  - $\ell(y, y) = 0$,
  - $\forall y \neq y' : \ell(y, y') > 0$.
- Exemples de fonctions de perte :
  - Cas de la classification supervisée binaire :

$$\ell(y, y') = \mathbb{1}_{y \neq y'} = \frac{|y - y'|}{2} = \frac{(y - y')^2}{4} \ .$$

  - Cas de la régression :

$$\ell(y, y') = |y - y'|^q$$

avec $q \in \mathbb{R}^+$.

Risque (erreur de généralisation)

Le risque (ou erreur de généralisation) d'un predicteur $\widehat{f}$ est défini par :

$$R\left(\widehat{f}\right) = \mathbb{E}\left[\ell\left(\widehat{f}(X), Y\right)\right] .$$

Oracle

Si on connaissait $P_{X,Y}$, on pourrait déterminer le prédicteur optimal, appelé oracle :

$$f^\star = \arg\min_{f \in \mathcal{F}} R(f) \; .$$

Exemples d'oracles

- Cas de la classification supervisée binaire :
  Si $\ell(y, y') = \mathbb{1}_{\{y \neq y'\}}$ alors :

$$g^\star(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 / X = x) \geq \mathbb{P}(Y = -1 / X = x) \\ -1 & \text{sinon} \end{cases} .$$

- Cas de la régression :
  - Si $\ell(y, y') = |y - y'|$ alors :

$$m^\star(x) = \text{Med}(Y / X = x) .$$

  - Si $\ell(y, y') = (y - y')^2$ alors :

$$m^\star(x) = \mathbb{E}(Y / X = x) .$$

Enjeu

- L'objectif du data scientist est de déterminer une estimation $\widehat{f}$ de f, à partir de l'échantillon, telle que :

$$R\left(\widehat{f}\right) \approx R\left(f^{\star}\right) .$$

- En pratique, pour estimer $f \in \mathcal{F}$:
  1. On restreint $\mathcal{F}$ à $\mathcal{S}$.
  2. On considère le risque empirique R_n (et non le risque).

D'où :

$$\widehat{f} = \arg\min_{f \in \mathcal{S}} R_n\left(f\right) .$$

Risque empirique

Le risque empirique est défini par :

$$R_n\left(\widehat{f}\right) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{f}(X_i), Y_i\right) .$$

C'est un estimateur de $R\left(\widehat{f}\right)$.

Plan

Biais et variance d'un estimateur I

- Dans le cas général ($\mathcal{F}$), on cherche :

$$f^\star = \arg\min_{f \in \mathcal{F}} R(f) \ .$$

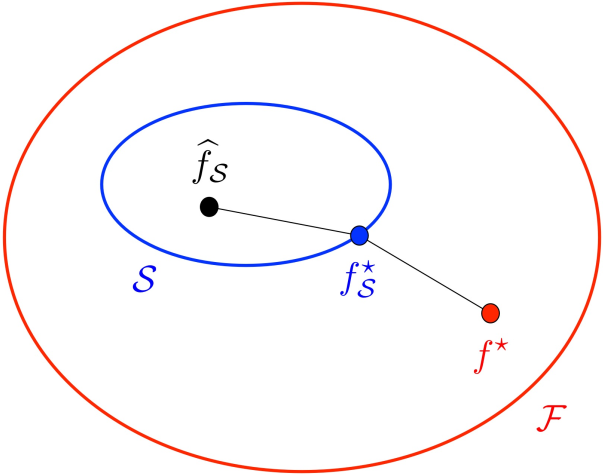
- Dans le cas restreint ($\mathcal{S} \subset \mathcal{F}$), on cherche :

$$f^\star_{\mathcal{S}} = \arg\min_{f \in \mathcal{S}} R(f) \ .$$

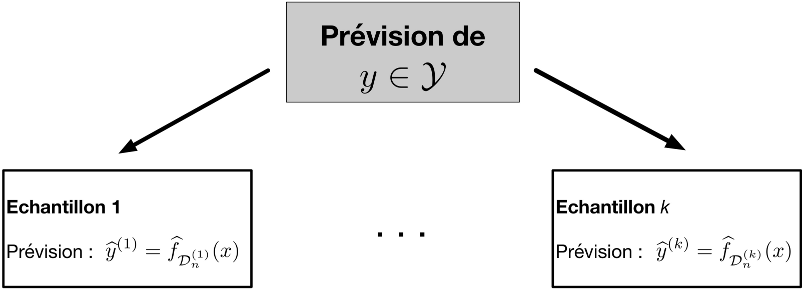
La décomposition biais (erreur d'approximation)-variance (erreur d'estimation) s'écrit :

$$R\left(\widehat{f_{\mathcal{S}}}\right) - R(f^\star) = \underbrace{R(f^\star_{\mathcal{S}}) - R(f^\star)}_{\text{erreur d'approximation}} + \underbrace{R\left(\widehat{f_{\mathcal{S}}}\right) - R(f^\star_{\mathcal{S}})}_{\text{erreur d'estimation}} \ .$$

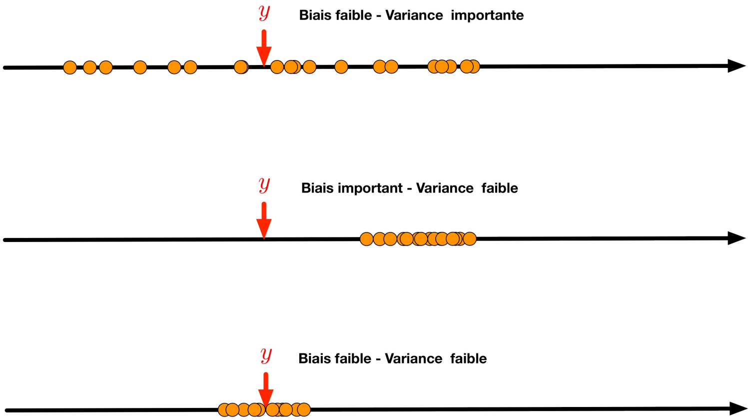
Biais et variance d'un estimateur II



Biais et variance d'un estimateur III

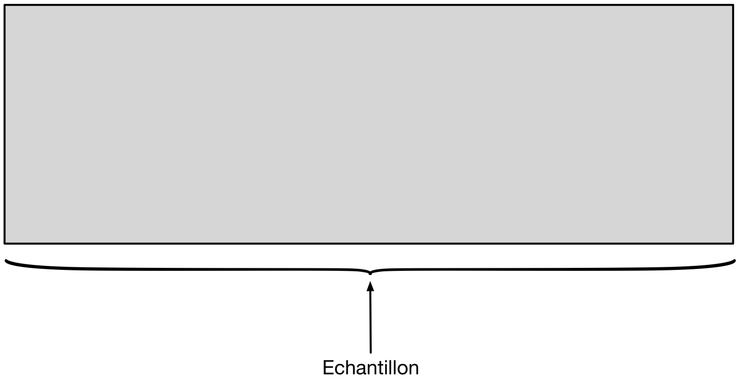


Biais et variance d'un estimateur IV

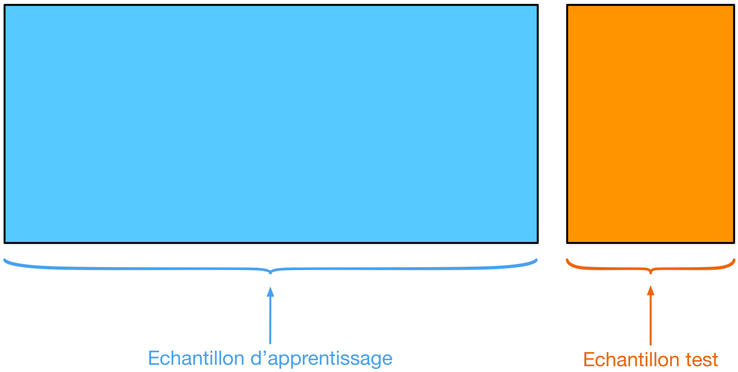


En orange : $\widehat{y}^{(1)}, \ldots, \widehat{y}^{(k)}$

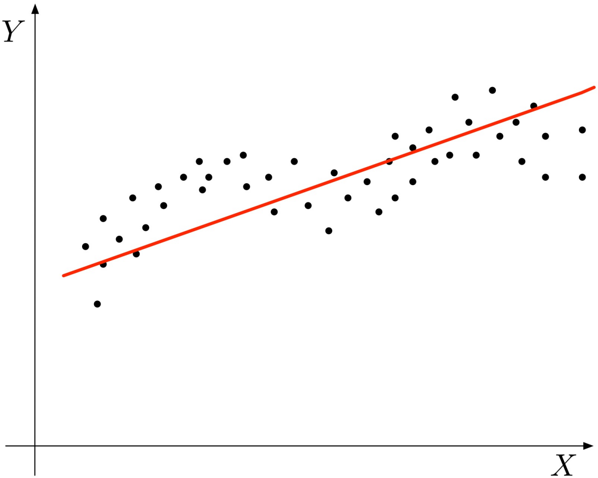
Echantillons d'apprentissage et de test I



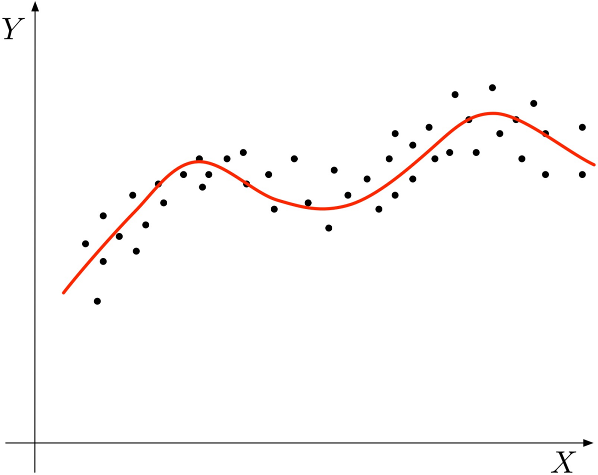
Echantillons d'apprentissage et de test II



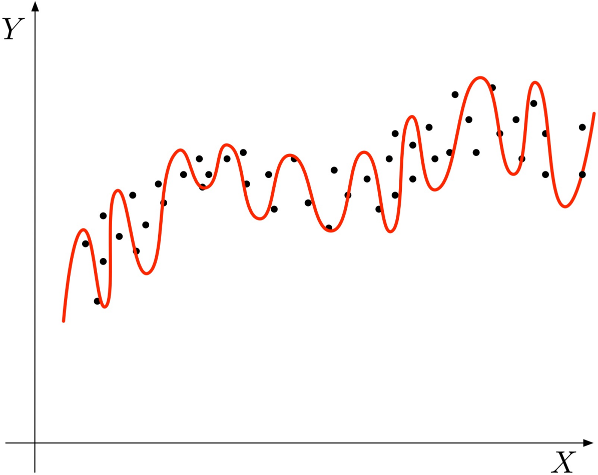
Complexité I



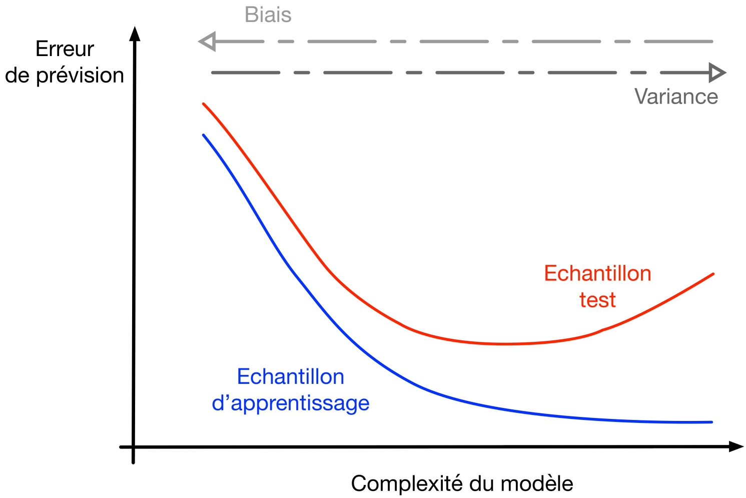
Complexité II



Complexité III



Erreurs de prévision & complexité



Plan

Retour sur le risque empirique

- Le risque empirique sous-estime le risque.
- Cela peut conduire à du sur-apprentissage.
- Il existe plusieurs parades pour obtenir un estimateur non-biaisé du risque :
  - Utilisation de critères tels que l'AIC, le BIC, le C_p de Mallows.
  - Méthodes de rééchantillonage : validation croisée ou bootstrap.

Principe

1. Diviser aléatoirement les données en K blocs (égaux ou équivalents).
   Le bloc k contient n_k observations : $n_k = \frac{n}{K}$ si n est un multiple de K.

2. Pour $k \in \{1, \ldots, K\}$:
   2.1 Retirer le bloc k de la base d'apprentissage.
   2.2 Estimer la fonction de prévision sur la base d'apprentissage.
   2.3 Calculer un critère d'erreur de prévision sur le bloc k : CV_k (ex : MSE pour la régression).

3. Calculer le critère de validation croisée :

$$\mathrm{CV} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{CV}_k \; .$$

Illustration I

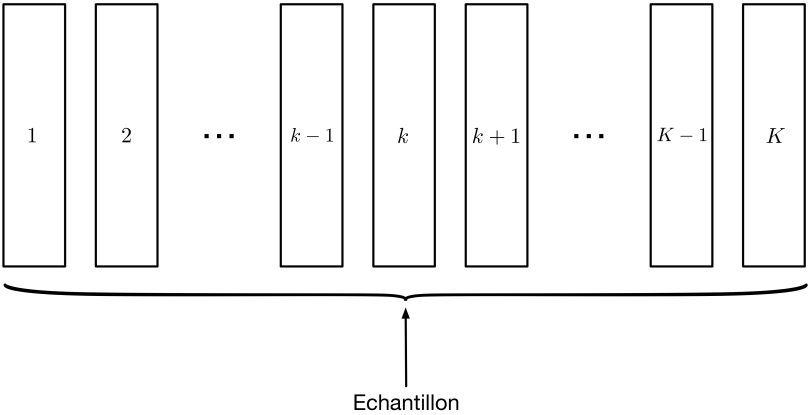
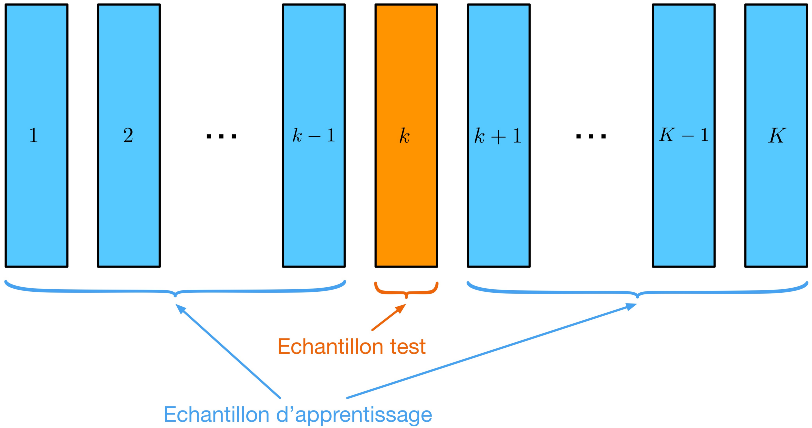


Illustration II



Remarques

- Usuellement : $K = 5$ ou $K = 10$.
- Lorsque $K = n$: on parle d'estimateur « leave one out » (LOO)

Références

Hastie, T., R. Tibshirani et J. H. Friedman. 2009, *The elements of statistical learning. Data Mining, inference, and prediction*, 2e éd., Springer Series in Statistics, Springer.

James, G., D. Witten, T. Hastie et R. Tibshirani. 2015, *An introduction to statistical learning with applications in R*, Springer Texts in Statistics, Springer.