Validation croisée (*cross-validation*)

Marie-Anne.Poursat@math.u-psud.fr

Master 1 Mathématiques Appliquées
Université Paris-Saclay

4 avril 2019

Évaluation de la qualité de prédiction

L'objectif est de construire à partir des données

$$\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}, \quad \mathbf{X}_i = (X_{i1}, \ldots, X_{ip}),$$

un prédicteur $\widehat{f}(.) = \widehat{f}(.; \mathcal{D})$ (fonction de régression ou de classification).

1. Avant d'utiliser les prévisions que $\widehat{f}(.)$ fournit, il faut évaluer son risque.
2. Bien souvent on dispose de plusieurs règles d'apprentissage (différents choix d'hyperparamètres, différentes méthodes), comment choisir la meilleure méthode et ses paramètres ?

Évaluation de la qualité de prédiction

Risque en régression (*Mean Squared Error*) :

$$MSE = \mathrm{E}\left[(Y - \widehat{f}(X))^2\right]$$

ou en Classification (*Probabilité de mauvais classement*) :

$$\mathrm{P}\left(Y \neq \widehat{f}(X)\right)$$

Comment l'estimer ?

Approche naïve : par l'erreur d'apprentissage.

$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{f}(\mathbf{X}_i)\right)^2$ (RSS) ou $\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{Y_i \neq \widehat{f}(\mathbf{X}_i)\right\}$ (taux de mal classés)

$\hookrightarrow$ sur-apprentissage

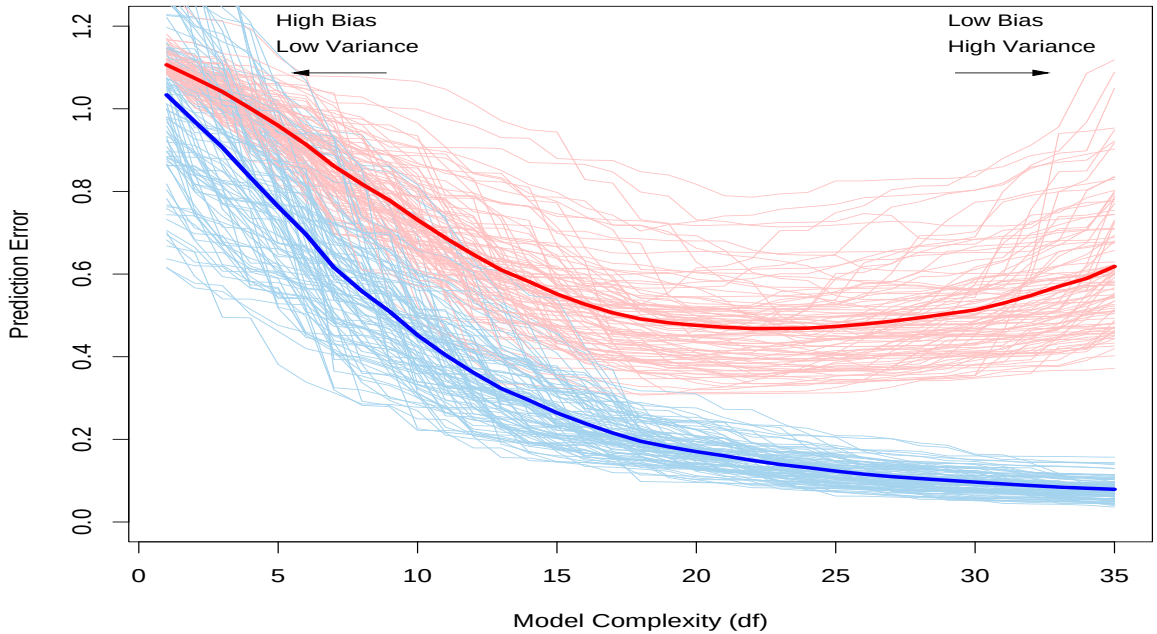


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$*.*

Évaluation de la qualité de prédiction

Plus un modèle est complexe (plus de paramètres) et flexible (s'adapte mieux aux données), plus son erreur d'apprentissage est faible.

MAIS ce modèle pourra s'avérer mauvais dans un but de prédiction ou de généralisation (s'adapter à de nouvelles données).

$\hookrightarrow$ Minimiser l'erreur d'apprentissage conduit au sur-apprentissage.

Comment obtenir un bon estimateur de l'erreur de prédiction ?

$\hookrightarrow$ si l'on dispose de beaucoup de données, on partage le jeu de données en un ensemble d'apprentissage et un ensemble test :

- le modèle est ajusté sur les données d'apprentissage : calcul de $\widehat{f}$.
- le modèle est évalué sur les données test : calcul de l'erreur test

Approche apprentissage/test

$n = n_{app} + n_{test}$, $\{(x_i, y_i), i = 1, \ldots, n_{app}\} \hookrightarrow \widehat{f}$

$$\text{Erreur test} = \frac{1}{n_{test}} \sum_{t=1}^{n_{test}} \left(y_t - \widehat{f}(x_t) \right)^2$$

Avantages :

- simple,
- facile et rapide à implémenter

Inconvénients :

- l'erreur test peut être variable, dépendre du découpage des données
- l'erreur test peut sur-estimer l'erreur de prédiction en n'utilisant qu'une partie des données pour estimer f.

Méthodes pour débiaser l'erreur d'apprentissage

- Utiliser un jeu de données test de grande taille, rarement disponible.
- Corriger mathématiquement l'erreur d'apprentissage par pénalisation : AIC, BIC, Cp de Mallows,...
- Technique de simulation très utilisée : validation croisée (*cross-validation*)

Critères pénalisés

$Y = f(x) + \varepsilon$, $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2$.

$$\mathrm{E}\left[(Y - \widehat{f}(x))^2\right] = \sigma^2 + \mathrm{Biais}^2(\widehat{f}(x), f(x)) + \mathrm{Var}(\widehat{f}(x)).$$

$\longrightarrow$ compromis biais-variance qui minimise l'erreur de prédiction ?
(*cost-complexity*)

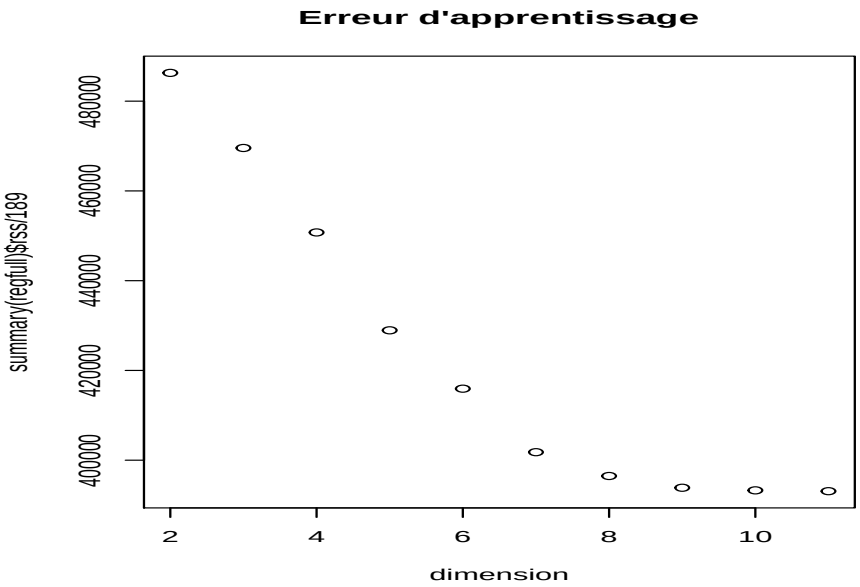
$$\mathrm{E}\left[(Y - \widehat{f}(x))^2\right] \simeq \mathrm{E}\left[RSS/n\right] + \frac{2}{n}\sum \mathrm{Cov}(y_i, \widehat{y}_i)$$

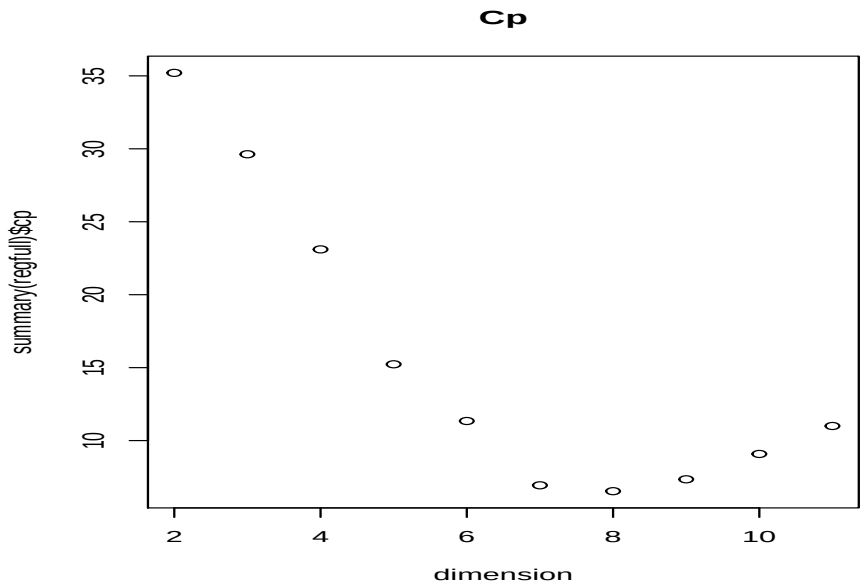
et $\sum \mathrm{Cov}(y_i, \widehat{y}_i) = (p+1)\sigma^2$ pour une régression linéaire gaussienne à p covariables.
D'où

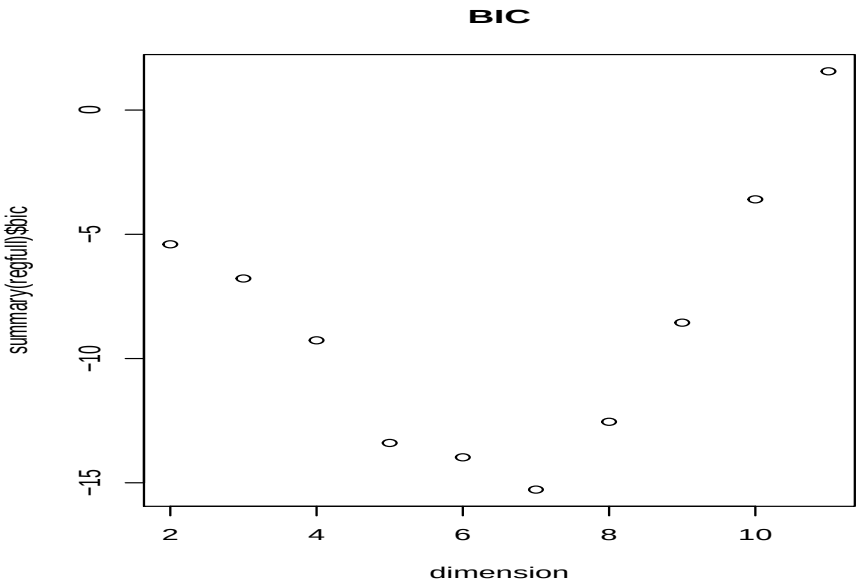
$$C_p = \frac{RSS}{n} + \frac{2}{n}(p+1)\sigma^2$$

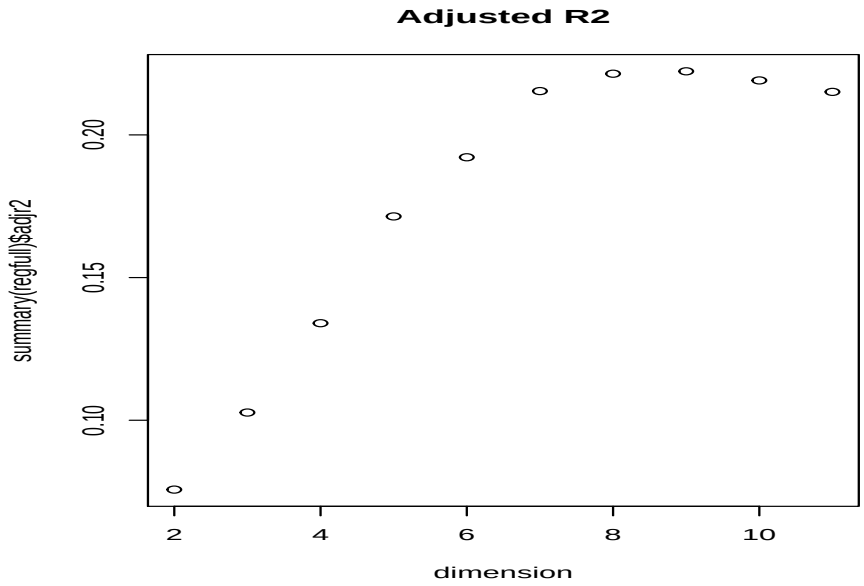
et $AIC = -\frac{2}{n}\log \mathrm{Lik} + \frac{2}{n}(p+1)$.

Exemple Birthwt









Validation croisée

- Méthode universelle : mise en oeuvre dans un cadre statistique général et pour la plupart des procédures d'estimation
- Principe : séparer les données d'*apprentissage* et de *test* ; construire l'estimateur sur l'échantillon d'apprentissage et utiliser l'échantillon test pour calculer un risque de prédiction (*model assessment*).
  Répéter plusieurs fois et moyenner les risques de prédiction obtenus
- Technique très utilisée pour choisir les tuning parameters (*model selection*)
  En particulier : paramètre de complexité d'un arbre CART (cp)

Validation croisée

Schémas de validation croisée

- *K-fold CV* : partition des données en K sous-ensembles. Chaque sous-ensemble sert successivement d'échantillon test, le reste d'échantillon d'apprentissage.
  En pratique : K entre 5 et 10.
- *Leave-one-out* : n-fold CV
- *Leave-q-out* : chaque sous-ensemble de cardinal q est retiré comme échantillon test, le reste servant d'apprentissage

Evaluation/comparaison de prédicteurs

La procédure complète partitionne en 3 les données :

apprentissage + validation (2/3 données) + test (1/3 données)

1. sur les données d'apprentissage : calcul de l'estimateur (algorithme d'apprentissage)
2. sur les données de validation : *model selection*, choix d'un meilleur modèle dans la classe de modèles considérée, par sélection de variables et/ou optimisation des hyper-paramètres de l'algorithme (*tuning parameters*).
3. sur les données test : calcul de l'erreur de prédiction finale

Les 2 dernières étapes sont souvent réalisées par *validation-croisée*.