

Cours de bootstrap et ré-échantillonnage

April 22, 2020

Plan

Introduction

Bootstrap

Intervalles de confiance par bootstrap

Application à la régression

Sub-sampling

Erreur de généralisation, cross-validation et bootstrap

Introduction

But de l'inférence statistique

On a

- ▶ $\mathcal{X}_n = (X_1, \dots, X_n)$ un échantillon i.i.d. de fonction de répartition F
- ▶ $\theta(F)$ une quantité d'intérêt, qui dépend de F
- ▶ $T(\mathcal{X}_n)$ une statistique, estimateur de $\theta(F)$,

on voudrait connaître

- ▶ le biais : $\mathbb{E}_F(T(\mathcal{X}_n)) - \theta(F)$
- ▶ la variance : $\mathbb{E}_F(T^2(\mathcal{X}_n)) - \mathbb{E}_F^2(T(\mathcal{X}_n))$
- ▶ le MSE (EQM) : $\mathbb{E}_F((T(\mathcal{X}_n) - \theta(F))^2)$
- ▶ la loi de $T(\mathcal{X}_n)$: $G^n(x) = \mathbb{P}_F(T(\mathcal{X}_n) \leq x)$, $\forall x$.
- ▶ etc

pour comparer des estimateurs, connaître leurs qualités, construire des intervalles de confiance...

Problème : toutes ses quantités dépendent de la loi F inconnue !

Ce que l'on sait

On a à disposition la fonction de répartition empirique des X_i .

Fonction de répartition empirique

Pour $\mathcal{X}_n = (X_1, \dots, X_n)$, la fonction de répartition empirique est définie par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq x}, \quad \forall x.$$

On va considérer les estimateurs par plug-in.

Principe de plug-in

Pour tout paramètre $\theta(F)$ et tout échantillon $\mathcal{X}_n = (X_1, \dots, X_n)$, on considère l'estimateur par plug-in

$$T(\mathcal{X}_n) = \theta(F_n) = \hat{\theta}$$

→ exemples : espérance, variance, médiane

Bootstrap

Bootstrap d'Efron 1982; Efron 1992

Conditionnellement à $\mathcal{X}_n = (X_1, \dots, X_n)$, on construit des échantillons

$$\mathcal{X}_1^* = (X_{1,1}^* = X_{m_1}, \dots, X_{1,n}^* = X_{m_n})$$

...

$$\mathcal{X}_b^* = (X_{b,1}^* = X_{m_{(b-1)n+1}}, \dots, X_{b,n}^* = X_{m_{bn}})$$

...

où les m_k ont été tirés aléatoirement et avec remise dans $\{1, \dots, n\}$.

Loi des $X_{b,j}^*$ conditionnelle à \mathcal{X}_n

Conditionnellement \mathcal{X}_n , $X_{b,j}^*$ est une v.a. de fonction de répartition F_n , fonction de répartition empirique des X_1, \dots, X_n .

Estimateurs du bootstrap classique

Soit un paramètre inconnu $\theta(F)$

▶ Monde réel

- ▶ avec l'échantillon initial \mathcal{X}_n , on définit l'estimateur $\hat{\theta} = \theta(F_n) = T(\mathcal{X}_n)$
- ▶ on note G^n la f.d.r. inconnue de $\hat{\theta}$, qui dépend de F (et de n !), inconnue

▶ Monde bootstrap

- ▶ pour chaque échantillon bootstrapé \mathcal{X}_b^* , on définit l'estimateur $\hat{\theta}_b^* = T(\mathcal{X}_b^*)$
- ▶ conditionnellement à F_n , de loi $G^{n,*}$ qui dépend de F_n
- ▶ on estime $G^{n,*}$ par

$$\widehat{G}_{n,B}^*(t) = \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{\hat{\theta}_b^* \leq t}$$

Exemples d'estimateurs bootstrap (1)

Estimation de la loi de $\hat{\theta}$

La f.d.r. G^n de $\hat{\theta} = T(\mathcal{X}_n)$ est définie pour $t \in \mathbb{R}$ par

$$G^n(t) = \int \mathbb{1}_{x \leq t} dG^n(x)$$

elle est estimée par (1ere approximation du bootstrap)

$$G^{n,*}(t) = \int \mathbb{1}_{x \leq t} dG^{n,*}(x) = \mathbb{P}_{F_n}(\hat{\theta}_b^* \leq t)$$

puis (2ieme approximation du bootstrap)

$$\hat{G}_{n,B}^*(t) = \int \mathbb{1}_{x \leq t} d\hat{G}_{n,B}^*(x).$$

Exemples d'estimateurs bootstrap (2)

Estimation du biais de $\hat{\theta}$

Le biais de $\hat{\theta}$ est défini par

$$\mathbb{E}_F(T(\mathcal{X}_n)) - \theta(F) = \int x dG^n(x) - \theta(F)$$

est estimé (1ere approximation) par

$$\int x dG^{n,*}(x) - \theta(F_n)$$

puis (2ieme approximation) par

$$\int x d\hat{G}_{n,B}^*(x) - \theta(F_n) = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^* - \theta(F_n).$$

Exemples d'estimateurs bootstrap (3)

Estimation de la variance de $\hat{\theta}$

Le biais de $\hat{\theta}$ est défini par

$$\mathbb{E}_F((T(\mathcal{X}_n) - \mathbb{E}_F(T(\mathcal{X}_n)))^2) = \int (x - \int x dG^n)^2 dG^n(x)$$

est estimé (1ere approximation) par

$$\int (x - \int x dG^{n,*})^2 dG^{n,*}(x)$$

puis (2ieme approximation) par

$$\int (x - \int x d\hat{G}_{n,B}^*)^2 d\hat{G}_{n,B}^*(x) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^* - \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^*)^2.$$

etc

Éléments de validation asymptotique du bootstrap (1)

Le bootstrap fait deux approximations

$$G^n \xrightarrow{(1)} G^{n,*} \xrightarrow{(2)} \widehat{G}_{n,B}^*$$

Pour contrôler la deuxième approximation, on utilise une borne de Dvoretzky-Kiefer-Wolfowitz.

Borne DKW, DKW (1956) - Massart (1990)

Si Z_1, \dots, Z_N est un échantillon i.i.d. de f.d.r. H et H_N est la f.d.r. empirique associée alors

$$\mathbb{P}\left(\sqrt{N} \sup_{x \in \mathbb{R}} |H_N(x) - H(x)| > \epsilon\right) \leq 2 \exp(-2\epsilon^2).$$

Application pour le choix de B :

Si on veut que $\sup_{x \in \mathbb{R}} |\widehat{G}_{n,B}^*(x) - G^{n,*}(x)| \leq 0.02$ avec une probabilité plus grande que 0.05, comment choisir B ?

Éléments de validation asymptotique du bootstrap (2)

La première approximation est contrôlée par les développements d'Edgeworth (voir Hall 2013). Si $\hat{\theta}$ est asymptotiquement normal :

$$S = \sqrt{n} \frac{\hat{\theta} - \theta}{\sigma(F)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

avec quelques conditions supplémentaires, on peut montrer que G^n admet un développement d'Edgeworth

$$\mathbb{P}(S \leq x) = G^n(\theta + \sigma x / \sqrt{n}) = \Phi(x) + \frac{1}{n^{1/2}} p(x) \phi(x) + \mathcal{O}\left(\frac{1}{n}\right).$$

Dans le monde bootstrap, on peut montrer que si

$$S^* = \sqrt{n} \frac{\hat{\theta}^* - \hat{\theta}}{\sigma(F)}$$

on a

$$\mathbb{P}_{F_n}(S^* \leq x) = G^{n,*}(\theta + \sigma x / \sqrt{n}) = \Phi(x) + \frac{1}{n^{1/2}} \hat{p}(x) \phi(x) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Éléments de validation asymptotique du bootstrap (3)

Le point clé est que $p(x) - \hat{p}(x) = \mathcal{O}_{\mathbb{P}}(\frac{1}{n^{1/2}})$. Un simple calcul montre alors :

- ▶ Approximation gaussienne

$$\mathbb{P}(S \leq x) - \Phi(x) = \frac{1}{n^{1/2}} p(x) \phi(x) + \mathcal{O}(\frac{1}{n}) = \mathcal{O}(\frac{1}{n^{1/2}})$$

- ▶ Approximation bootstrap

$$\mathbb{P}(S \leq x) - \mathbb{P}_{F_n}(S^* \leq x) = \mathcal{O}_{\mathbb{P}}(\frac{1}{n}).$$

Exemples

Ca marche

- ▶ pour la moyenne quand

$$\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

- ▶ pour la médiane quand

$$\sqrt{n}(F_n^-(1/2) - F^-(1/2)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{4f^2(F^-(1/2))}\right)$$

Ca ne marche pas pour les extrêmes

- ▶ par exemple X_1, \dots, X_n i.i.d. $\mathcal{U}(\theta, \theta + 1)$ alors $X_{(1)} \xrightarrow{\mathbb{P}} \theta$ et

$$n(X_{(1)} - \theta) \xrightarrow{\mathcal{L}} \mathcal{E}(1).$$

Intervalles de confiance par bootstrap

Intervalle de confiance du bootstrap basique

On définit les statistiques d'ordre des estimateurs bootstrap

$$\hat{\theta}_{(1)}^* \leq \hat{\theta}_{(2)}^* \leq \dots \leq \hat{\theta}_{(n_b)}^*$$

IC du bootstrap basique

$$\widehat{IC}_{basic}^*(1 - \alpha) = \left[2\hat{\theta} - \hat{\theta}_{(\lceil B(1-\alpha/2) \rceil)}^*, 2\hat{\theta} - \hat{\theta}_{(\lceil B\alpha/2 \rceil)}^* \right]$$

Détails (1)

Détails (2)

Intervalle de confiance du percentile

S'il existe une fonction h monotone telle que la loi de $h(T)$ est symétrique autour de $h(\theta)$.

IC du percentile

$$\widehat{IC}_{perc}^*(1 - \alpha) = \left[\hat{\theta}_{(\lceil B\alpha/2 \rceil)}^*, \hat{\theta}_{(\lceil B(1-\alpha/2) \rceil)}^* \right]$$

Détails (1)

Détails (2)

Intervalle de confiance du t-bostrap

Si on connaît un estimateur $\hat{\sigma} = \sigma(F_n) = \sigma(\mathcal{X})$ de la variance asymptotique $\sigma^2(F)$ de $T = \hat{\theta}$, on considère la statistique studentisée

$$S = \sqrt{n} \frac{\hat{\theta} - \theta}{\sigma(F_n)}$$

et ses versions bootstrapées

$$S_b^* = \sqrt{n} \frac{\hat{\theta}_b^* - \hat{\theta}}{\sigma(\mathcal{X}_b^*)}.$$

IC du t-bostrap

$$\hat{IC}_{tboot}^*(1 - \alpha) = \left[\hat{\theta} - \frac{\sigma(F_n)}{\sqrt{n}} S_{(\lceil B(1-\alpha/2) \rceil)}^*, \hat{\theta} - \frac{\sigma(F_n)}{\sqrt{n}} S_{(\lceil B(\alpha/2) \rceil)}^* \right]$$

Détails (1)

Détails (2)

Test via l'intervalle de confiance du t-bootstrap

On considère le problème de test de $\mathcal{H}_0 : \theta = \theta_0$ v.s. $\mathcal{H}_1 : \theta \neq \theta_0$. On peut faire ce test par bootstrap en comparant la statistique de test

$$\bar{S} = \left| \sqrt{n} \frac{\hat{\theta} - \theta_0}{\sigma(F_n)} \right|$$

aux statistiques bootstrapées

$$\bar{S}_b^* = \left| \sqrt{n} \frac{\hat{\theta}_b^* - \hat{\theta}}{\sigma(\mathcal{X}_b^*)} \right|.$$

On définit alors la p-value bootstrapée

$$\hat{p}_B = \frac{\#\{b : \bar{S}_b^* > \bar{S}\} + 1}{B + 1}$$

Si l'estimateur de la variance n'est pas disponible, on peut utiliser la statistique $|\hat{\theta} - \theta_0|$ et sa version bootstrapée $|\hat{\theta}_b^* - \hat{\theta}|$.

Application à la régression

Introduction

Problème de la régression

On considère l'échantillon i.i.d. $\mathcal{S} = \left((Y_1, X_1), \dots, (Y_n, X_n) \right)$ avec

- ▶ $Y_i(\Omega) \subset \mathbb{R}$
- ▶ $X_i(\Omega) \subset \mathbb{R}^p$.

On veut estimer $\mathbb{E}(Y_i|X_i) = g(X_i)$. On se donne un estimateur $\hat{g} = \hat{g}_{\mathcal{S}}$

Remarque : la fonction g est entièrement déterminée par la loi conditionnelle de Y_i sachant X_i et donc par la loi de (Y_i, X_i) .

Exemples

Régression linéaire

$$Y_i = X_i\beta + \epsilon_i$$

donc $g(x) = x\beta$ et si $\epsilon \sim \mathcal{L}_\epsilon(0, \sigma^2)$ alors $Y_i|X_i \sim \mathcal{L}_\epsilon(X_i\beta, \sigma^2)$. On se donne $\hat{g}(x) = x\hat{\beta}$ avec $\hat{\beta}$ estimateur des moindres carrés de β .

Régression logistique

$Y_i(\Omega) = \{0, 1\}$ et

$$\mathbb{E}(Y_i|X_i) = \mathbb{P}(Y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)} = \pi(X_i\beta)$$

alors $Y_i|X_i \sim \mathcal{L}(\pi(X_i\beta), \pi(X_i\beta)(1 - \pi(X_i\beta)))$. On se donne $\hat{g}(x) = \pi(x\hat{\beta})$ avec $\hat{\beta}$ estimateur au maximum de vraisemblance de β .

Rééchantillonnage des cas (“cases sampling”)

Cases sampling

Puisque les (Y_i, X_i) sont supposés i.i.d.,

1. on échantillonne aléatoirement avec remise dans l'échantillon $\mathcal{S} = \left((Y_1, X_1), \dots, (Y_n, X_n) \right)$ pour obtenir

$$\mathcal{S}_1^* = \left((Y_{1,1}^*, X_{1,1}^*), \dots, (Y_{1,n}^*, X_{1,n}^*) \right)$$

...

$$\mathcal{S}_B^* = \left((Y_{B,1}^*, X_{B,1}^*), \dots, (Y_{n,B}^*, X_{n,B}^*) \right)$$

2. on calcule sur chaque échantillon bootstrappé $\hat{g}_{\mathcal{S}_b^*}$.

Rééchantillonnage des erreurs (“errors sampling”) dans le modèle linéaire

Dans le modèle de régression linéaire, on définit les résidus par

$$e_i = Y_i - X_i \hat{\beta} \quad \forall i = 1, \dots, n.$$

On montre que

$$\mathbb{E}(e_i) = 0 \text{ et } \mathbb{V}(e_i) = (1 - H_{ii})\sigma^2$$

avec $H = X(X^T X)^{-1} X^T$. On définit alors les résidus studentisés

$$e_i^S = \frac{e_i}{\sqrt{1 - H_{ii} \hat{\sigma}_{-i}^2}}.$$

Les e_i^S sont censés être proches en loi des e_i/σ^2 . Si, par exemple, les e_i sont gaussiens, on peut montrer

$$e_i^* \sim \mathcal{T}(n - p - 1).$$

Errors sampling

1. On calcule les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ à partir de \mathcal{S} , puis les résidus studentisés e_1^S, \dots, e_n^S .
2. on échantillonne aléatoirement avec remise dans l'échantillon (e_1^S, \dots, e_n^S) pour obtenir

$$(e_{1,1}^{S,*}, \dots, e_{1,n}^{S,*}) \dots (e_{B,1}^{S,*}, \dots, e_{n,B}^{S,*})$$

3. on reconstruit pour chaque b et chaque i

$$Y_{b,i}^* = X_i \hat{\beta} + \hat{\sigma} e_{b,i}^{S,*}$$

4. on calcule dans chaque échantillon bootstrapé des estimateurs de β et σ^2 .

Rééchantillonnage des erreurs (“errors sampling”) dans le modèle logistique

Dans le modèle logistique, il n'y a pas d'erreur ϵ . On définit cependant

- ▶ les résidus de Pearson

$$e_i^P = \frac{Y_i - \hat{\pi}(X^i)}{\sqrt{\hat{\pi}(X^i)(1 - \hat{\pi}(X^i))}}.$$

- ▶ les résidus de déviance

$$\begin{aligned} e_i^D &= \sqrt{-2 \log(\hat{\pi}(X^i))} \text{ si } Y_i = 1 \\ &= -\sqrt{-2 \log(1 - \hat{\pi}(X^i))} \text{ si } Y_i = 0. \end{aligned}$$

Ils jouent le même rôle que les résidus studentisés e_i^S du modèle linéaire. En particulier, on peut écrire “à la louche”

$$Y_i = \pi(X_i\beta) + Y_i - \pi(X_i\beta) = \pi(X_i\beta) + “\epsilon_i'' \quad \forall i = 1, \dots, n$$

avec ϵ_i à valeurs sur $\{-\pi(X_i\beta), 1 - \pi(X_i\beta)\}$, qui vérifient

$$\mathbb{E}(\epsilon_i) = 0 \text{ et } \mathbb{V}(\epsilon_i) = \pi(X_i\beta)(1 - \pi(X_i\beta)).$$

Errors sampling

1. On calcule les estimateurs $\hat{\beta}$ à partir de \mathcal{S} , puis les résidus de Pearson (ou de déviance) e_1^P, \dots, e_n^P .
2. on échantillonne aléatoirement avec remise dans l'échantillon (e_1^P, \dots, e_n^P) pour obtenir

$$(e_{1,1}^{P,*}, \dots, e_{1,n}^{P,*}) \dots (e_{B,1}^{P,*}, \dots, e_{n,B}^{P,*})$$

3. on calcule pour chaque b et chaque i

$$\zeta_{b,i} = \pi(X_i \hat{\beta}) + \sqrt{\pi(X_i \hat{\beta})(1 - \pi(X_i \hat{\beta}))} e_{b,i}^{P,*}.$$

4. si $\zeta_{b,i} < 1/2$, on fixe $Y_{b,i}^* = 0$ et si $\zeta_{b,i} \geq 1/2$, on fixe $Y_{b,i}^* = 1$
5. on calcule dans chaque échantillon bootstrapés des estimateurs de β .

NB : cet algorithme s'étend simplement à tous les modèles linéaires généralisés.

Application aux tests par bootstrap: errors sampling

Grâce à l'algorithme "errors sampling", on peut construire des tests par bootstrap. On note, pour chaque i , $X_i = (X_i^a, X_i^b)$ avec $X_i^a \in \mathbb{R}^{p^a}$ et $X_i^b \in \mathbb{R}^{p^b}$ de sorte que

$$X_i \beta = X_i^a \gamma + X_i^b \delta.$$

Supposons qu'on veut tester $H_0 : \delta = \vec{0}$ v.s. \bar{H}_0 . Les statistiques de tests à considérer sont

- ▶ la statistique de Fisher dans le modèle de régression linéaire

$$F = \frac{(n-p)(\|Y - X^a \hat{\gamma}\|^2 - \|Y - X \hat{\beta}\|^2)}{\|Y - X \hat{\beta}\|^2}$$

- ▶ la statistique du rapport de vraisemblance dans le modèle logistique

$$\Lambda = -2 [\log(\mathcal{V}(\hat{\gamma})) - \log(\mathcal{V}(\hat{\beta}))].$$

On a besoin pour garantir un niveau α au test (ou pour calculer une p-value) de connaître la loi de F et Λ sous H_0 .

Bootstrap sous H_0

1. On calcule l'estimateur $\hat{\gamma}$ à partir de \mathcal{S} dans le modèle sous H_0 , puis les résidus studentisés (ou de Pearson ou de déviance) $e_1^{S,a}, \dots, e_n^{S,a}$.
2. on échantillonne aléatoirement avec remise dans l'échantillon $(e_1^{S,a}, \dots, e_n^{S,a})$ pour obtenir

$$(e_{1,1}^{S,a,*}, \dots, e_{1,n}^{S,a,*}) \dots (e_{B,1}^{S,a,*}, \dots, e_{n,B}^{S,a,*})$$

3. on calcule pour chaque b et chaque i

$$Y_{b,i}^* = ???$$

4. ???

Bootstrap sous H_0

1. On calcule les estimateurs $\hat{\gamma}$ et $\hat{\sigma}^{a,2}$ à partir de \mathcal{S} dans le modèle sous H_0 , puis les résidus studentisés (ou de Pearson ou de déviance) $e_1^{S,a}, \dots, e_n^{S,a}$.
2. on échantillonne aléatoirement avec remise dans l'échantillon $(e_1^{S,a}, \dots, e_n^{S,a})$ pour obtenir

$$(e_{1,1}^{S,a,*}, \dots, e_{1,n}^{S,a,*}) \dots (e_{B,1}^{S,a,*}, \dots, e_{n,B}^{S,a,*})$$

3. on calcule pour chaque b et chaque i

$$Y_{b,i}^* = X_i^a \hat{\gamma} + \hat{\sigma}^{a,2} e_{b,i}^{S,a,*}$$

4. on calcule les valeurs bootstrapées $F_1^{*,H_0}, \dots, F_B^{*,H_0}$

Exercice: prédiction dans le modèle linéaire

- ▶ On dispose d'un échantillon $\mathcal{S} = \left((Y_1, X_1), \dots, (Y_n, X_n) \right)$ (échantillon d'apprentissage)
- ▶ On dispose des variables explicatives X_+ pour un nouvel individu. On note Y_+ la valeur non-observée de sa réponse.
- ▶ On note l'erreur de prédiction

$$\text{Ep}(+) = \hat{Y}_+ - Y_+ = X_+ \hat{\beta} - Y_+ = X_+ \hat{\beta} - (X_+ \beta + \epsilon_+)$$

1. En notant G la f.d.r. (inconnue) de $\text{Ep}(+)$, donner un IP exact pour Y_+ .
2. Proposer des versions bootstrapées de $\text{Ep}(+)$
3. Donner un IP par bootstrap basique pour Y_+ .

Détails

Sub-sampling

Conditionnellement à $\mathcal{X} = (X_1, \dots, X_n)$, on construit des échantillons

$$\mathcal{X}_{S_1} = (X_{S_1(1)}, \dots, X_{S_1(n_b)})$$

...

$$\mathcal{X}_{S_k} = (X_{S_k(1)}, \dots, X_{S_k(n_b)})$$

...

où les S_k ont été tirés aléatoirement uniformément sur

$\{S \subset \{1, \dots, n\}, |S| = n_b\}$. Il y a donc $\binom{n}{n_b}$ sous-échantillons possibles.

Estimateurs par sub-sampling

Soit un paramètre inconnu $\theta(F)$

- ▶ Monde réel

- ▶ avec l'échantillon initial \mathcal{X}_n , on définit l'estimateur $\hat{\theta} = \theta(F_n) = T(\mathcal{X}_n)$ que l'on suppose symétrique en X_1, \dots, X_n (invariant par permutation)
- ▶ on note G^n la f.d.r. inconnue de $\hat{\theta}$, qui dépend de F , inconnue

- ▶ Sous-échantillonnage

- ▶ pour chaque sous-échantillon \mathcal{X}_{S_k} , on définit l'estimateur $\hat{\theta}_k^S = T(\mathcal{X}_{S_k})$
- ▶ conditionnellement à F_n , de loi G_n^S uniforme sur

$$\{\hat{\theta}^S, S \subset \{1, \dots, n\}, |S| = n_b\}$$

Estimateur re-normalisé

En pratique (intervalles de confiance, tests), on s'intéresse plutôt à la variable

$$r_n(T(\mathcal{X}_n) - \theta(F))$$

avec $r_n \rightarrow \infty$ dont les équivalents sous-échantillonnés sont

$$r_b(T(\mathcal{X}_k^S) - T(\mathcal{X}_n)).$$

Conditions

Si $r_n(T(\mathcal{X}_n) - \theta(F)) \xrightarrow{\mathcal{L}^*} \mathcal{L}^*$ alors $r_b(T(\mathcal{X}_k^S) - \theta(F)) \xrightarrow{\mathcal{L}^*} \mathcal{L}^*$ quand $r_b \rightarrow \infty$. Il faut donc s'assurer que

$$r_b(T(\mathcal{X}_n) - \theta(F)) \xrightarrow{\mathbb{P}} 0$$

soit $r_b/r_n \rightarrow 0$.

Exemple de la moyenne

On suppose X_1, \dots, X_n i.i.d. de f.d.r F inconnue, on choisit $\theta(F) = \mathbb{E}(X_i)$ et on suppose $\mathbb{V}(X_i) = \sigma^2$.

On pose $T(\mathcal{X}_n) = \bar{X}_n$, dans ce cas $\sqrt{n}(\bar{X}_n - \theta(F)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$. Les versions sous-échantillonnées sont alors

$$\sqrt{n_b}(T(\mathcal{X}_k^S) - \bar{X}_n).$$

Les conditions sont vérifiées si $n_b \rightarrow \infty$ et $n_b/n \rightarrow 0$ puisqu'on a alors bien

$$\sqrt{n_b}(T(\mathcal{X}_n) - \theta(F)) \xrightarrow{\mathbb{P}} 0.$$

On peut montrer que, conditionnellement à \mathcal{X}_n

$$\begin{aligned}\mathbb{E}(\sqrt{n_b}(T(\mathcal{X}_k^S) - \bar{X}_n)) &= 0 \\ \mathbb{E}(n_b(T(\mathcal{X}_k^S) - \bar{X}_n)^2) &= \frac{n_b}{n} \hat{\sigma}^2.\end{aligned}$$

Détails

Jackknife et leave-one-out

Historiquement cette méthode a été introduite par Quenouille et Tuckey Quenouille 1956; Tukey 1958 qui proposaient de travailler avec des échantillons de taille $n - 1$.

Estimateur du jackknife de la variance de $\hat{\theta}$

On définit

$$\widehat{\mathbb{V}}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^n \left(T(\mathcal{X}_k^S) - \frac{1}{n} \sum_k T(\mathcal{X}_k^S) \right)^2$$

Exercice

1. A quelle renormalisation de la statistique correspond ce choix ?
2. Dans le cas de la moyenne, quelle est la loi limite de cette statistique ?

Application au minimum

Le bootstrap ne marche pas pour les extrêmes, le sub-sampling si. Par exemple X_1, \dots, X_n i.i.d. $\mathcal{U}(\theta, \theta + 1)$ alors $X_{(1)} \xrightarrow{\mathbb{P}} \theta$ et

$$n(X_{(1)} - \theta) \xrightarrow{\mathcal{L}} \mathcal{E}(1).$$

Erreur de généralisation, cross-validation et bootstrap

Vrai modèle, sélection de modèle

Modèles, vrai modèle

On se donne une famille de modèles \mathcal{M} , par exemple $\mathcal{M} = \mathcal{P}\{1, \dots, p\}$. On suppose qu'il existe un vrai modèle $m^* \in \mathcal{M}$ tel que :

$$Y = X^{(m^*)} \beta^{(m^*)} + \epsilon^*$$

avec ϵ_i i.i.d., $\mathbb{E}(\epsilon_i) = 0$ et $\mathbb{V}(\epsilon_i) = \sigma^2$.

sélection de modèle : on veut retrouver m^* .

Estimation dans le modèle m

Dans le modèle m , on note $|m|$ le nombre de covariables qu'il contient et

$$\hat{\beta}^{(m)} = ((X^{(m)})^\top X^{(m)})^{-1} (X^{(m)})^\top Y$$

$$\hat{Y}^{(m)} = X^{(m)} \hat{\beta}^{(m)}$$

$$\widehat{(\sigma^m)^2} = \frac{\|Y - \hat{Y}^{(m)}\|_2^2}{n - |m|}$$

Moindres carré, risque quadratique et validation interne

Le risque quadratique de $\hat{Y}^{(m)}$ pour l'estimation de $X^* \beta^*$ est donné par

$$\begin{aligned}\mathbb{E}(\|\hat{Y}^{(m)} - X^* \beta^*\|^2) &= \underbrace{\mathbb{E}(\|\hat{Y}^{(m)} - X^{(m)} \beta^{(m)}\|^2)}_{\text{variance}} + \underbrace{\|X^{(m)} \beta^{(m)} - X^* \beta^*\|^2}_{\text{biais}^2} \\ &= \sigma^2 |m| + \|X^{(m)} \beta^{(m)} - X^* \beta^*\|^2\end{aligned}$$

où $X^{(m)} \beta^{(m)}$ est la projection de $X^* \beta^*$ sur $\text{vect}(X^{(m)})$. Pour l'espérance de l'erreur de prédiction (erreur apparente), on montre que

$$\mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2) = (n - |m|)\sigma^2 + \|X^{(m)} \beta^{(m)} - X^* \beta^*\|^2.$$

Finalement

$$\begin{aligned}\mathbb{E}(\|\hat{Y}^{(m)} - X^* \beta^*\|^2) &= \sigma^2 |m| - (n - |m|)\sigma^2 + \mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2) \\ &= 2\sigma^2 |m| + \mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2) - n\sigma^2.\end{aligned}$$

Cp de Mallows

On choisit $\hat{m}_{Cp} \in \mathcal{M}$ tel que :

$$\hat{m}_{Cp} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} Cp(m),$$

avec

$$Cp(m) = \frac{\widehat{(\sigma^m)^2}}{\widehat{(\sigma^{m_{tot}})^2}} + 2 \frac{|m|}{n}$$

Validation externe

Si on avait à disposition d'autres données, on aurait

- ▶ des données d'apprentissage (training, learning set)

$$\mathcal{S}_L = \{(Y_1, X_1), \dots, (Y_n, X_n)\}$$

- ▶ des données de validation, de test (testing, validation set)

$$\mathcal{S}_T = \{(Y_{+,1}, X_{+,1}), \dots, (Y_{+,n'}, X_{+,n'})\} \text{ avec } Y_+ = X_+ \beta^* + \epsilon_+ \text{ ET } \epsilon \text{ et } \epsilon_+ \text{ indépendants.}$$

On choisirait alors le modèle qui minimise l'erreur de généralisation.

Generalization error, erreur de généralisation

$$\mathbb{E}(\|Y_+ - \hat{Y}_+^{(m)}\|^2) = \mathbb{E}(\|Y_+ - X_+ \hat{\beta}^{(m)}\|^2) = n' \sigma^2 + |m| \sigma^2 + \|X^{(m)} \beta^{(m)} - X^* \beta^*\|^2.$$

Théoriquement, on choisit le même modèle qu'avec le risque quadratique :

$$\mathbb{E}(\|\hat{Y}^{(m)} - X^* \beta^*\|^2) = \sigma^2 |m| + \|X^{(m)} \beta^{(m)} - X^* \beta^*\|^2$$

Excès d'erreur

Si on minimisait

$$\mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2) = (n - |m|)\sigma^2 + \|X^{(m)}\beta^{(m)} - X^*\beta^*\|^2.$$

on choisirait toujours le plus grand modèle.

Excess error, excès d'erreur

On définit l'excès d'erreur comme

$$\mathbb{E}(\|Y_+ - \hat{Y}^{(m)}\|^2) - \mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2)$$

Estimation de l'erreur de généralisation et de l'excès d'erreur

On estime l'erreur de généralisation $\mathbb{E}(\|Y_+ - \hat{Y}_+^{(m)}\|^2) = \mathbb{E}(\|Y_+ - X_+ \hat{\beta}^{(m)}\|^2)$ à partir de l'échantillon $\mathcal{S}_T = \{(Y_{+,1}, X_{+,1}), \dots, (Y_{+,n'}, X_{+,n'})\}$ par

$$\frac{1}{n'} \sum_{i=1}^{n'} (Y_{+,i} - X_{+,i} \hat{\beta}^{(m)})^2,$$

où $\hat{\beta}^{(m)}$ a été calculé sur l'échantillon d'apprentissage \mathcal{S}_L et dans le modèle m .

En pratique

Même en l'absence de données de validation (situation fréquente en pratique), on peut vouloir créer des données qui “ressemblent” à des données de test pour appliquer ce qui précède. Il y a deux grandes méthodes

- ▶ la cross-validation
- ▶ le bootstrap

Leave-one-out (jackknife)

Chaque observation joue à tour de rôle le rôle d'échantillon de validation.

Estimation de l'erreur de généralisation par leave-one-out

$$\mathbb{E}(\|\widehat{Y}_+ - \widehat{Y}_+^{(m)}\|^2)_{loo} = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i \widehat{\beta}_{(-i)}^{(m)})^2,$$

où $\widehat{\beta}_{(-i)}^{(m)}$ a été calculé sur l'échantillon $\mathcal{S}_L \setminus (Y_i, X_i)$ et dans le modèle m .

K-fold cross-validation

On découpe l'échantillon initial en K sous-ensembles pour obtenir la partition $\mathcal{S}_L = \mathcal{S}_{L,1} \cup \dots \cup \mathcal{S}_{L,K}$. Dans le cas, où $n = Kn_K$, on tire aléatoirement et sans remise dans \mathcal{S}_L pour former les $\mathcal{S}_{L,k}$.

Estimation de l'erreur de généralisation par K-fold cross-validation

$$\widehat{eg(m)}_{Kfold-cv} = \mathbb{E}(\|Y_+ - \widehat{Y}_+^{(m)}\|^2)_{Kfold-cv} = \frac{1}{n_K K} \sum_{k=1}^K \sum_{i=1}^{n_K} (Y_{k,i} - X_{k,i} \widehat{\beta}_{(-k)}^{(m)})^2,$$

où $\widehat{\beta}_{(-k)}^{(m)}$ a été calculé sur l'échantillon $\mathcal{S}_L \setminus \mathcal{S}_{L,k}$ et dans le modèle m . On peut préférer l'ajustement

$$\begin{aligned} \widehat{eg(m)}_{A-Kfold-cv} &= \mathbb{E}(\|Y_+ - \widehat{Y}_+^{(m)}\|^2)_{A-Kfold-cv} = \mathbb{E}(\|Y_+ - \widehat{Y}_+^{(m)}\|^2)_{Kfold-cv} \\ &+ \frac{1}{n} \|Y - X\widehat{\beta}\|^2 - \frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n (Y_i - X_i \widehat{\beta}_{(-k)}^{(m)})^2. \end{aligned}$$

Estimation de l'excès d'erreur

Une estimation bootstrap de l'excès d'erreur

$$ee(m) = \mathbb{E}(\|Y_+ - \hat{Y}^{(m)}\|^2 - \|\hat{Y}^{(m)} - Y\|^2)$$

est

$$\widehat{ee(m)}_{boot}^* = \frac{1}{nB} \sum_{b=1}^B \sum_{i=1}^n \left((Y_i - X_i \hat{\beta}_b^*)^2 - (Y_{b,i}^* - X_{b,i}^* \hat{\beta}_b^*)^2 \right).$$

Estimation de l'erreur de généralisation par bootstrap

On obtient alors un estimateur par bootstrap de l'erreur de généralisation

$$\widehat{eg(m)}_{boot} = \widehat{ee(m)}_{boot}^* + \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i^{(m)} - Y_i)^2$$

Introduction

Bootstrap

Définitions et résultats

Exemples d'estimateurs

Éléments de validation asymptotique

Intervalles de confiance par bootstrap

Intervalles de confiance

Test par bootstrap

Application à la régression

Introduction

Rééchantillonnage

Tests par bootstrap

Sub-sampling

Erreur de généralisation, cross-validation et bootstrap

sélection de modèle

Validation interne

Validation externe

Cross-validation

Bootstrap

References I



Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.



Bradley Efron. "Bootstrap methods: another look at the jackknife". In: *Breakthroughs in Statistics*. Springer, 1992, pp. 569–593.



Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.



Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling Springer series in statistics*. 1999.



Maurice H Quenouille. "Notes on bias in estimation". In: *Biometrika* 43.3/4 (1956), pp. 353–360.



John W Tukey. "Bias and confidence in not-quite large samples". In: *Annals of Mathematical Statistics*. Vol. 29. 2. INST MATHEMATICAL STATISTICS IMS BUSINESS OFFICE-SUITE 7, 3401 INVESTMENT BLVD, HAYWARD, CA 94545. 1958, pp. 614–614.