Méthodes de ré-échantillonage bootstrap

Master 1 Mathématiques Appliquées
Université Paris-Saclay

28 mars 2019

Qu'est-ce que le bootstrap ?

"*To pull oneself up by one's own bootstraps*"

Le bootstrap est une méthode statistique pour évaluer la précision d'une méthode : estimer une variance, une erreur de prédiction, calculer un intervalle de confiance,...

Pour estimer la variabilité d'un estimateur (sa loi), il faut observer plusieurs réalisations.

- ↪ une seule expérience en général
- ↪ Idée : générer des échantillons qui ressemblent à l'échantillon de départ !

Principe

$T(Y_1, \ldots, Y_n)$ une statistique basée sur n observations i.i.d.
Quelle est sa loi ?

$\hookrightarrow$ calcul (asymptotique), souvent difficile

$\hookrightarrow$ estimation par simulation

La loi des Y_i n'est pas connue, il faut l'estimer : $\widehat{P} = P_{\widehat{\theta}(Y)}$.
$\hookrightarrow$ hyp. *robustesse* et *régularité* du modèle.

Contexte non-paramétrique :

$$P_n = \frac{1}{n} \sum_i \delta_{Y_i}$$

P_n est la mesure empirique (la loi uniforme sur l'échantillon), c'est un bon estimateur de P (th. de Glivenko-Cantelli).

Principe

Méthode empirique : soit $\theta = T(P)$ un paramètre ; on estime $T(P)$ par $T(P_n)$.

Exemples : $\mu(P) = \int x dP(x)$ ou $\sigma^2(P) = \int x^2 dP(x) - (\mu(P))^2$.

La donnée de l'échantillon $(Y_1, \ldots, Y_n)$ est équivalente à celle de P_n. On interprète la statistique comme une fonction de P_n :

$$\widehat{\theta} = T(P_n),$$

est un estimateur de $T(P)$.

Exemple : $\sigma^2(P) = E_P[(Y - \mu(P))^2]$ a pour estimateur

$$\sigma^2(P_n) = E_{P_n}[(Y - \mu(P_n))^2] = \frac{1}{n}\sum_i (Y_i - \bar{Y})^2$$

Echantillon bootstrap

C'est un échantillon (en général de taille n) tiré selon P_n parmi $Y_1, \ldots, Y_n$. On le note $Y^* = (Y_1^*, \ldots, Y_n^*)$.

- Dans le monde réel, on estime $\theta = T(P)$ par $\widehat{\theta} = T(P_n) = T(Y_1, \ldots, Y_n)$;
- Dans le monde bootstrap, on estime θ par $T(Y_1^*, \ldots, Y_n^*)$.

Le bootstrap est valide si on peut montrer que la loi de $T(Y^*) = \widehat{\theta}^*$ est une bonne approximation de la loi de $T(Y) = \hat{\theta}$, c'est à dire si $\mathcal{L}_P(\hat{\theta})$ et $\mathcal{L}_{P_n}(\hat{\theta}^* | Y_1, \ldots, Y_n)$ ont même loi limite.

Un mot de théorie

Conditions générales de validité :

- la loi limite de $T(P_n)$ dépend "continûment" de P
- la convergence est uniforme en P dans un voisinage de la vraie loi P_0
- la loi utilisée pour générer l'échantillon bootstrap estime bien P_0

Théorème (Bickel et Freedman, Annals of Statistics 1981) :
si $\theta = \int Y dP$,

$$\lim_{n\to\infty} \sqrt{n} \sup_x |P(T_n \leq x) - P(T_n^* \leq x | Y_1, \ldots, Y_n)| = 0$$

p.s. conditionnellement à $(Y_1, \ldots, Y_n)$, avec $T_n = \sqrt{n}(\hat{\theta} - \theta)$ et $T_n^* = \sqrt{n}(\hat{\theta}^* - \hat{\theta})$.

Ce résultat a été étendu aux fonctions de la moyenne, modèles de régression, séries temporelles...

Bootstrap dans un cas non régulier

$Y_1, \ldots, Y_n$ de loi uniforme $(0, \theta)$.

$$\widehat{\theta} = \max(Y_1, \ldots, Y_n)$$

$\hookrightarrow$ $n(\widehat{\theta} - \theta)/\theta$ de loi exponentielle $\mathcal{E}(1)$
$\hookrightarrow$ $\Pr(\widehat{\theta}^* = \widehat{\theta}) \simeq 0.632$. *Pourquoi ?*

En pratique : une méthode bootstrap doit toujours être validée par un théorème, ou par une étude de simulations.

Ré-échantillonnage bootstrap

$Y = (Y_1, \ldots, Y_n)$ échantillon initial

Échantillon bootstrap : n observations tirées uniformément avec replacement parmi les observations de départ

$\hookrightarrow \quad Y^* = Y_1^*, Y_2^*, \ldots, Y_n^*$

Problème : en général, on ne connait pas mieux la loi de $T(Y^*)$ que celle de $T(Y)$!

Avantage : on peut générer autant d'échantillons Y^* que l'on veut !

$\hookrightarrow \quad Y^{*1} = (Y_1^{*1}, \ldots, Y_n^{*1})$

$\vdots$

$\hookrightarrow \quad Y^{*B} = (Y_1^{*B}, \ldots, Y_n^{*B})$

En théorie, $B = n^n$; en pratique $B = 200$ à 1000.

Distribution d'échantillonage bootstrap

C'est l'étape de Monte-Carlo :

$$\begin{array}{lll} (Y_1^{*(1)}, \ldots, Y_n^{*(1)}) & \to & \hat{\theta}_n^{*(1)} \\ \ldots & \ldots & \ldots \\ (Y_1^{*(b)}, \ldots, Y_n^{*(b)}) & \to & \hat{\theta}_n^{*(b)} \\ \ldots & \ldots & \ldots \\ (Y_1^{*(B)}, \ldots, Y_n^{*(B)}) & \to & \hat{\theta}_n^{*(B)} \end{array}$$

Il y a donc deux étapes d'approximation dans le bootstrap :

1. l'approximation statistique (asymptotique) ($n \to \infty$)
2. l'approximation numérique (Monte-Carlo) (B grand).

$$\hat{\theta}_n^{*(1)}, \hat{\theta}_n^{*(2)}, \ldots, \hat{\theta}_n^{*(B)}$$

est appelée loi d'échantillonage bootstrap de $T(Y) = \hat{\theta}$.

Validité du bootstrap

Dans un cadre paramétrique, si $\widehat{\theta} = T(\mathbf{Y})$ estime θ :

$$\begin{aligned} \text{Loi } (\widehat{\theta} - \theta) &\approx \text{Loi bootstrap}(\widehat{\theta}^* - \widehat{\theta}) \\ &\approx \text{Loi d'échantillonage } (\widehat{\theta}^{*1} - \widehat{\theta}, \widehat{\theta}^{*2} - \widehat{\theta}, \ldots, \widehat{\theta}^{*B} - \widehat{\theta}) \end{aligned}$$

- Dans certains cas, l'approximation bootstrap est meilleure que l'approximation asymptotique gaussienne de la loi de l'estimateur du Maximum de Vraisemblance

Avantages du bootstrap

- Permet d'apprécier par simulation la variabilité donc la précision d'un estimateur
- Simplicité du principe et de la mise en oeuvre : pas de formules mathématiques compliquées
- Méthode très générale dans un cadre paramétrique ou non-paramétrique (sous réserve de validité)
- le bootstrap "de base" est valide typiquement dans des situations où l'approximation gaussienne est valide.
  Contre exemples : valeurs extrêmes (max, quantiles), U-statistiques. Il existe alors des solutions par sous-échantillonnage.
- le bootstrap présenté est le bootstrap *non-paramétrique*. Il existe des bootstraps paramétriques.

Estimation du biais

Le biais de $\hat{\theta} = \theta(P_n)$ est :

$$B_P(\hat{\theta}, \theta) = E_P(\hat{\theta}) - \theta.$$

La contrepartie bootstrap du biais est

$$B_{P_n}(\hat{\theta}^*, \hat{\theta}) = E_{P_n}(\hat{\theta}^*) - \hat{\theta}.$$

Le biais est estimé par bootstrap par

$$\widehat{\text{Biais}}^* = \frac{1}{B} \sum_{b=1}^{b=B} \hat{\theta}^{*(b)} - \hat{\theta}.$$

Estimation de la variance

1. Génerer $Y_1^*, \ldots, Y_n^* \sim P_n$
2. Calculer $T^* = T(Y_1^*, \ldots, Y_n^*)$
3. Répéter les étapes 1 et 2 B fois pour obtenir $T^{*(1)}, T^{*(2)}, \ldots, T^{*(B)}$
4. Estimateur de la variance de la loi de T_n :

$$V_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{b=B} \left(T^{*(b)} - \frac{1}{B} \sum_{b=1}^{b=B} T^{*(b)} \right)^2$$

Intervalles de confiance bootstrap

Principe : on approche la loi de $\widehat{\theta} - \theta$ par la loi de $\widehat{\theta}^* - \widehat{\theta}$.

$\hookrightarrow$ on calcule les quantiles bootstrap $b^*_{\alpha/2}$ et $b^*_{1-\alpha/2}$ de la loi d'échantillonage bootstrap

$$(\widehat{\theta}^{*1} - \widehat{\theta}, \widehat{\theta}^{*2} - \widehat{\theta}, \ldots, \widehat{\theta}^{*B} - \widehat{\theta})$$

L'intervalle bootstrap de niveau (asymptotique) $1 - \alpha$ est

$$IC(\theta) = \left[\widehat{\theta} - b^*_{1-\alpha/2};\ \widehat{\theta} - b^*_{\alpha/2}\right]$$

Pourquoi ?

Bootstrap et régression

$\{(x_i, y_i),\ i = 1, \ldots, n\}$ données d'apprentissage

$$y_i = g(x_i) + \varepsilon_i$$

- $\hookrightarrow$ bootstrap des résidus $\widehat{\varepsilon}_i = Y_i - \widehat{g}(x_i)$
- $\hookrightarrow$ bootstrap des couples
- $\hookrightarrow$ bootstrap paramétrique des résidus

Références

- *An Introduction to Statistical Learning, with Applications in R*, Gareth, Witten, Hastie and Tibshirani
- *Bootstrap Methods and their Application*, Davison and Hinkley