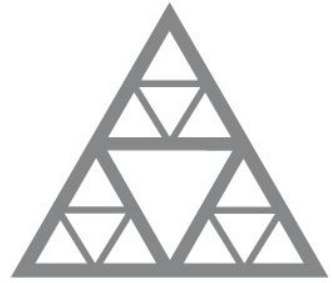
Pratique de l'apprentissage statistique

6. Bagging

V. Lefieux



École des Ponts
ParisTech

Plan

Plan

Données considérées

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

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

  On note :

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

- On considère dans la suite que :
  - $X \in \mathbb{R}^p$.
    Toutes les covariables sont considérés quantitatives.
    Mais il est également possible de considérer des covariables qualitatives.
  - $Y \in \{-1, 1\}$ dans le cas de la classification supervisée.
    On se place dans le cadre d'une classification supervisé binaire
    Mais il est également possible de considérer des classifications supervisées avec K classes.
  - $Y \in \mathbb{R}$ dans le cas de la régression.

Agrégation de modèles

Le principe est le suivant :

1. Estimer plusieurs modèles :
   - Cas de la classification supervisée : $(\widehat{g}_b)_{b\in 1,\dots,B}$.
   - Cas de la régression : $(\widehat{m}_b)_{b\in 1,\dots,B}$.

2. Agréger ces modèles :
   - Cas de la classification supervisée binaire (à valeurs dans $\{-1,1\}$) :

$$\widehat{g} = \text{signe}\left(\sum_{b=1}^{B} \alpha_b \, \widehat{g}_b\right) .$$

   - Cas de la régression :

$$\widehat{m} = \sum_{b=1}^{B} \alpha_b \, \widehat{m}_b .$$

Différentes méthodes d'agrégation

- Bagging
  - Du bagging « de base » : (Breiman, 1996).
    Ajuster des modèles (LASSO, arbres. . .) sur des échantillons bootstrappés, et les agréger.
  - Aux random forests : (Breiman, 2001).
    Ajuster des arbres décorrélés sur des échantillons bootstrappés, et les agréger.
- Boosting : (Freund et Schapire, 1997).
  Ajuster de petits arbres sur un échantillon repondéré de manière récursive, et les agréger.

Deux cas de figure

- Bagging : pour des modèles à forte variance et faible biais.
- Boosting : pour des modèles à faible variance et fort biais.

Un point commun

Ces méthodes sont réputées pour leur efficacité numérique.

Plan

Origines du bootstrap

- A l'origine du mot, The surprising adventures of Baron Munchausen de Rudolph Erich Raspe :
  « *The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.* »

- En France, Cyrano de Bergerac d'Edmond Rostand (pour atteindre la lune) :
  « *Enfin, me plaçant sur un plateau de fer,*
  Prendre un morceau d'aimant et le lancer en l'air !
  Ça, c'est un bon moyen : le fer se précipite,
  Aussitôt que l'aimant s'envole, à sa poursuite ;
  On relance l'aimant bien vite, et cadédis !
  On peut monter ainsi indéfiniment. »

- Une amélioration de l'idée du jacknife : (Efron, 1979).

Objectifs et principe du bootstrap

- Objectif : approcher par simulation la distribution d'un estimateur statistique de loi inconnue.
- Tandis que la validation croisée considère des ensembles de données indépendantes, le bootstrap fonctionne sur des échantillons de la même taille que l'original, en rééchantillonnant les observations via un tirage aléatoire avec remise.
- Dans un échantillon bootstrappé, certaines observations peuvent apparaître plusieurs fois, et d'autres ne pas apparaître du tout.
- Si n est grand, la loi de l'échantillon bootstrappé est proche de la loi de l'échantillon d'origine.

Plan

Principe

- Bagging : **B**ootstrap **agg**regat**ing**.
- Agréger des modèles estimés sur des échantillons bootstrappés.

Algorithme : classification supervisée (binaire)

1. Pour $b \in \{1, \ldots, B\}$:
   1.1 Tirer un échantillon bootstrappé à partir de l'échantillon $\mathcal{D}_n$: $\mathcal{D}_n^{\star b}$.
   1.2 Estimer un modèle sur l'échantillon bootstrappé $\mathcal{D}_n^{\star b}$: $\widehat{g}_b$.
2. Agréger les modèles :

$$\widehat{g} = \text{signe}\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{g}_b\right) .$$

Illustration I

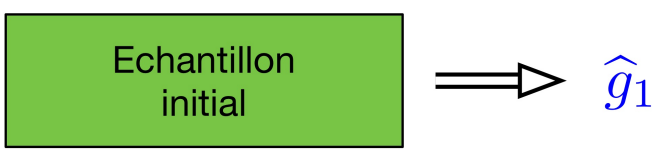


Illustration II

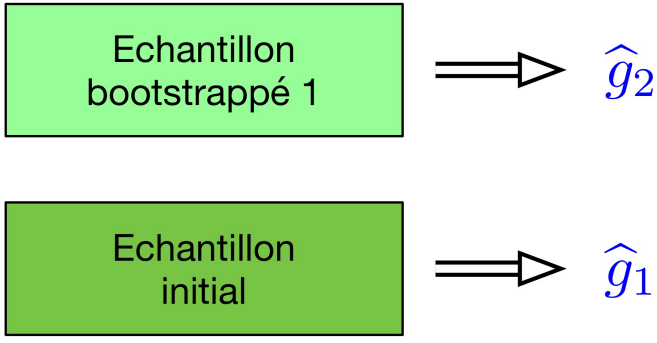


Illustration III

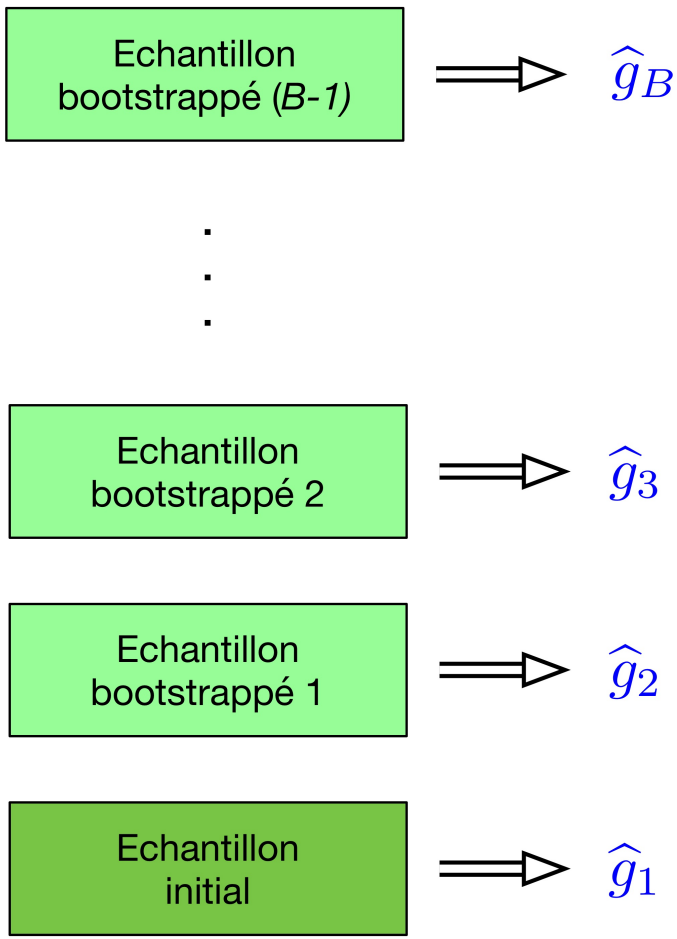
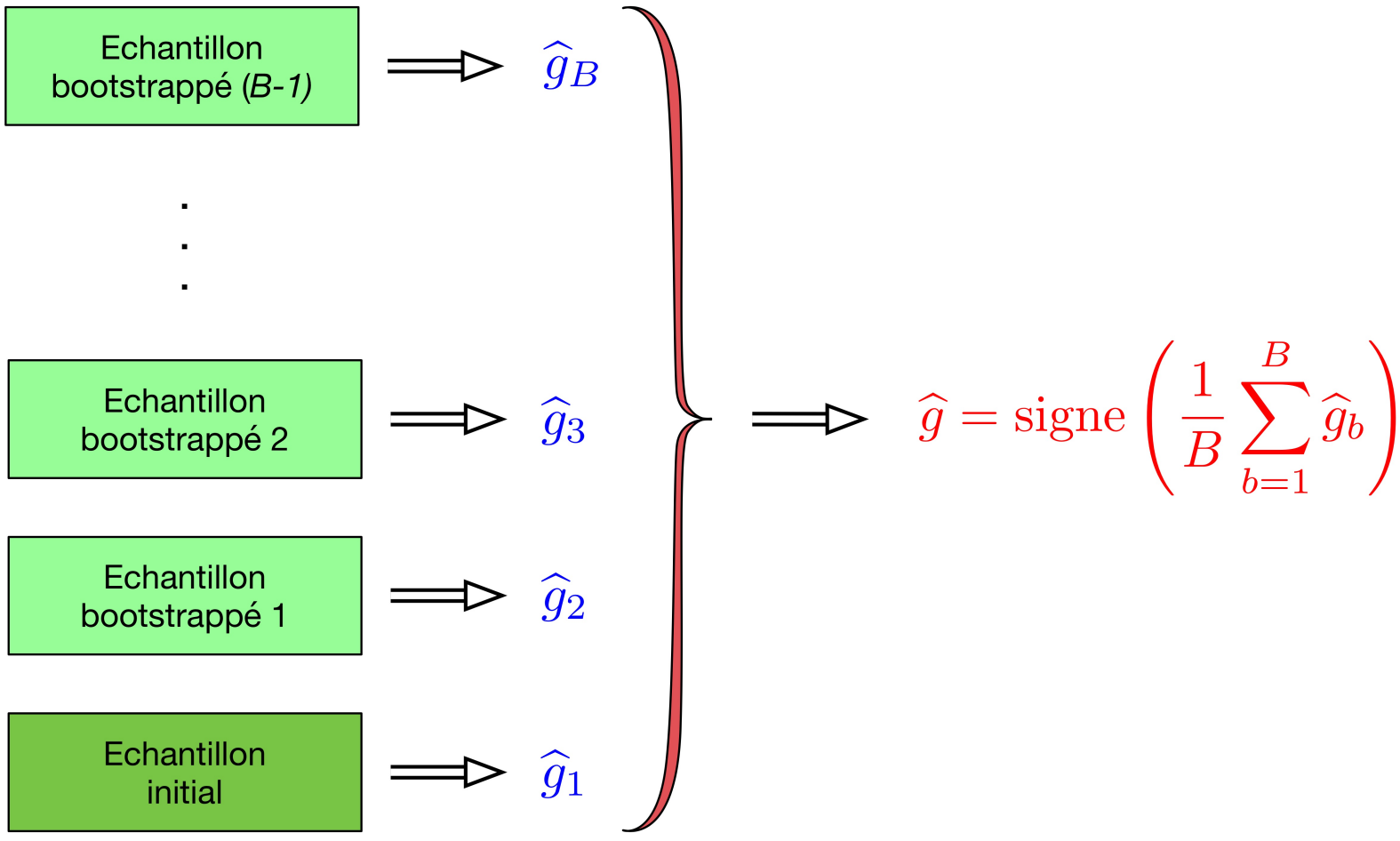


Illustration IV



Algorithme : régression

1. Pour $b \in \{1, \ldots, B\}$:
   - 1.1 Tirer un échantillon bootstrappé à partir de l'échantillon $\mathcal{D}_n$: $\mathcal{D}_n^{\star b}$.
   - 1.2 Estimer un modèle sur l'échantillon bootstrappé $\mathcal{D}_n^{\star b}$: $\widehat{m}_b$.
2. Agréger les modèles :

$$\widehat{m} = \frac{1}{B} \sum_{b=1}^{B} \widehat{m}_b \ .$$

Plan

Cas de la régression

On se place ici dans le cadre de la régression :

$$Y = m(X) + \varepsilon \ .$$

Considérons $\widehat{m}$ l'estimateur de m obtenu en agrégeant B estimateurs $\widehat{m}_1, \ldots, \widehat{m}_B$:

$$\widehat{m}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{m}_b(x) \ .$$

Biais et variance : sous hypothèse i.i.d

- Hypothèse : $\widehat{m}_1, \ldots, \widehat{m}_B$ sont i.i.d.
- Biais :

$$\mathbb{E}\left[\widehat{m}(x)\right] = \mathbb{E}\left[\widehat{m}_b(x)\right] .$$

  Agréger ne modifie pas le biais.

- Variance :

$$\mathrm{Var}\left[\widehat{m}(x)\right] = \frac{1}{B}\,\mathrm{Var}\left[\widehat{m}_b(x)\right] .$$

  Agréger réduit la variance.

Biais et variance : sous hypothèse de loi identique

- Hypothèse : $\widehat{m}_1, \ldots, \widehat{m}_B$ sont de même loi mais corrélées, de corrélation $\rho(x)$ en $x \in \mathbb{R}^p$.
- Biais :

$$\mathbb{E}\left[\widehat{m}(x)\right] = \mathbb{E}\left[\widehat{m}_b(x)\right] .$$

  Agréger ne modifie pas le biais.

- Variance :

$$\mathrm{Var}\left(\widehat{m}(x)\right) = \left[\rho\left(x\right)\left(1 - \frac{1}{B}\right) + \frac{1}{B}\right] \mathrm{Var}\left[\widehat{m}_b(x)\right]$$

$$\sim \rho(x)\,\mathrm{Var}\left[\widehat{m}_b(x)\right] \text{ pour } B \text{ grand.}$$

  Agréger réduit la variance, d'autant plus que la corrélation entre les modèles est faible.

Quel type de modèle considérer ?

- Il faut considérer des estimateurs sensibles à de légères perturbations de l'échantillon.
- Il faut considérer des estimateurs avec un biais faible (et donc avec une forte variance).
- Les arbres (CART) ont cette caractéristique.

Choix du nombre d'itérations (d'estimateurs)

- Il n'y a pas de risque de sur-apprentissage avec le nombre d'itérations B.
- La variance tend à se stabiliser avec le nombre d'itérations.
- Il faut choisir un compromis entre le gain sur la prévision et le temps de calcul.

Avantages du bagging

- Simplicité de la mise en oeuvre.
- Qualité de la prévision !

Inconvénients du bagging

- Temps de calcul.
  Mais le calcul est parallélisable.
- Aspect « boîte noire ».
  Mais il est possible d'estimer empiriquement l'importance d'une variable.
- Les différents estimateurs obtenus sur des échantillons bootstrappés, ne peuvent pas être considérés comme indépendants.
  Mais il existe des solutions comme les random forests qui introduisent une nouvelle source d'aléa pour rendre des arbres plus indépendants.

Plan

Erreur Out Of Bag : objectif

L'erreur Out Of Bag (OOB) permet d'estimer l'erreur de prévision (sans procéder par validation croisée) :

- Pour la classification supervisée : $\mathbb{P}\left(g\left(X\right) \neq Y\right)$.
- Pour la régression : $\mathbb{E}\left[\left(Y - m\left(x\right)\right)^2\right]$.

Erreur Out Of Bag : classification supervisée (binaire)

Pour l'observation i :

- Soit $\mathcal{I}_i$ les indices des échantillons bootstrappés $\mathcal{D}_n^{\star b}$ qui ne contiennent pas (X_i, Y_i).
- On agrège les modèles $b \in \mathcal{I}_i$ pour déterminer la prévision de Y_i :

$$\widehat{Y}_i = \text{signe}\left(\frac{1}{\text{Card}\left(\mathcal{I}_i\right)} \sum_{b \in \mathcal{I}_i} \widehat{g}_b\left(X_i\right)\right) .$$

On calcule l'erreur Out Of Bag (EOOB) :

$$\text{EOOB} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\widehat{Y}_i \neq Y_i} .$$

Erreur Out Of Bag : régression

Pour l'observation i :

- Soit $\mathcal{I}_i$ les indices des échantillons bootstrappés $\mathcal{D}_n^{\star b}$ qui ne contiennent pas l'observation (X_i, Y_i).
- On agrège les modèles $b \in \mathcal{I}_i$ pour déterminer la prévision de Y_i :

$$\widehat{Y}_i = \frac{1}{\text{Card}\left(\mathcal{I}_i\right)} \sum_{b \in \mathcal{I}_i} \widehat{m}_b\left(X_i\right) .$$

On calcule l'erreur Out Of Bag (EOOB) :

$$\text{EOOB} = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{Y}_i - Y_i\right)^2 .$$

Importance des variables : cas de la régression I

- Pour $b \in \{1, \ldots, B\}$:

  Soit OOB_b l'ensemble Out Of Bag des observations ne figurant pas dans l'échantillon bootstrappé b.

  On calcule EOOB_b l'erreur Out Of Bag des observations de l'ensemble OOB_b :

$$\text{EOOB}_b = \frac{1}{\text{Card}\left(\text{OOB}_b\right)} \sum_{i \in \text{OOB}_b} \left[\widehat{m}_b\left(X_i\right) - Y_i\right]^2 .$$

- Pour $j \in \{1, \ldots, p\}$:

  Soit $\text{OOB}_{b,j}$ l'ensemble des observations obtenu à partir de OOB_b en permutant aléatoirement les valeurs de la variable X_j.

  On calcule $\text{EOOB}_{b,j}$ l'erreur Out Of Bag des observations de l'ensemble $\text{OOB}_{b,j}$:

$$\text{EOOB}_{b,j} = \frac{1}{\text{Card}\left(\text{OOB}_{b,j}\right)} \sum_{i \in \text{OOB}_{b,j}} \left[\widehat{m}_b\left(X_i\right) - Y_i\right]^2 .$$

Importance des variables : cas de la régression II

- On considère comme critère d'importance de la variable X_j la différence moyenne sur tous les échantillons b entre EOOB_b et $\text{EOOB}_{b,j}$.
- Plus la différence est élevée, plus on peut considérer que la variable X_j a eu de l'importance pour l'échantillon b.
- On moyenne ces écarts pour tous les échantillons $b \in \{1, \ldots, B\}$, et on obtient l'importance de la variable X_j :

$$\text{Imp}\,(X_j) = \frac{1}{B} \sum_{b=1}^{B} (\text{EOOB}_{b,j} - \text{EOOB}_b) \ .$$

Plan

Algorithme

Pour $b \in \{1, \ldots, B\}$:

1. Tirer un échantillon bootstrappé à partir de l'échantillon $\mathcal{D}_n$: $\mathcal{D}_n^{\star b}$.
2. Estimer un arbre à partir de d variables tirées au sort parmi les p variables disponibles sur l'échantillon bootstrappé $\mathcal{D}_n^{\star b}$.
3. Agréger les modèles :
   - Pour la classification supervisée : vote majoritaire.
   - Pour la régression : moyenne des prévisions obtenues.

Paramètres par défaut

- Nombre d'observations dans les feuilles :
  - Pour la classification supervisée : 1.
  - Pour la régression : 5.
- Nombre de variables considérées pour chaque arbre :
  - Pour la classification supervisée : $\sqrt{p}$.
  - Pour la régression : $\frac{p}{3}$.

Quelques références internet

- Leo Breiman and Adele Cutler :

  http://www.stat.berkeley.edu/~breiman/RandomForests/

- Andrej Karpathy :

  http://cs.stanford.edu/people/karpathy/svmjs/demo/demoforest.html

Le coin R

On peut utiliser plusieurs packages, parmi lesquels :

- Le package `randomForest` :
  - La fonction de base est `randomForest`.
  - L'option `ntree` permet de choisir le nombre d'arbres.
  - On peut utiliser la fonction `importance` pour calculer l'importance des variables.

- Le package `caret` :
  - La fonction de base est `train`.
  - On utilise l'option correspondant aux random forests : `method="rf"`.
  - On peut utiliser la fonction `varImp` pour calculer l'importance des variables.

Références

Breiman, L. 1996, «Bagging predictors», *Machine Learning*, vol. 24, p. 123–140.

Breiman, L. 2001, «Random forests», *Machine Learning*, vol. 45, p. 5–32.

Efron, B. 1979, «Bootstrap methods : another look at the jackknife», *The Annals of Statistics*, vol. 7, n° 1, p. 1–26.

Efron, B. et R. Tibshirani. 1994, *An introduction to the bootstrap*, Chapman & Hall.

Freund, Y. et R. E. Schapire. 1997, «A decision-theoretic generalization of on-line learning and an application to boosting», *Journal of Computer and System Sciences*, vol. 55, n° 1, p. 119–139.