Sélection et évaluation de modèle - Partie 1

Véronique Tremblay

Objectifs

- Comprendre les difficultés liées à la mesure de l'erreur d'un modèle
- Connaître et utiliser une méthode d'évaluation d'un modèle (la validation croisée)

Rappel

Modèle prédictif

$$Y = f(x) + \epsilon$$

En général, on trouve $\hat{f}$ en minimisant l'espérance d'une certaine fonction de perte

$$L(Y, \hat{f}(x))$$

Décomposition de l'EQM

$$E[(Y_0 - \hat{f}(x_0))^2] = [Biais(\hat{f}(x_0))]^2 + Var(\hat{f}(x_0)) + \sigma_\epsilon^2$$

Mesurer l'erreur d'un modèle

Pourquoi?

1. Choisir le meilleur modèle

Pourquoi?

1. Choisir le meilleur modèle

2. Avoir une idée de la confiance qu'on peut accorder à notre modèle

1. Choisir le meilleur modèle

2. Avoir une idée de la confiance qu'on peut accorder à notre modèle

- L'erreur de généralisation

$$Err_\tau = \mathbb{E}_{X^0,Y^0}[L(Y^0, \hat{f}(X^0))|\tau]$$

- L'erreur de généralisation

$$Err_\tau = \mathbb{E}_{X^0,Y^0}[L(Y^0, \hat{f}(X^0))|\tau]$$

- L'erreur sur l'échantillon d'entraînement

$$e\bar{r}r = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

- L'erreur de généralisation

$$Err_{\tau} = \mathbb{E}_{X^0,Y^0}[L(Y^0, \hat{f}(X^0))|\tau]$$

- L'erreur sur l'échantillon d'entraînement

$$e\bar{r}r = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

- L'erreur *in sample*

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{Y^0}(L(Y^0, \hat{f}(x_i)))$$

Optimisme de la mesure d'erreur sur l'échantillon d'entraînement[1]

$$\mathbb{E}_Y(Err_{in}) - \mathbb{E}_Y(e\bar{r}r) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{C}\mathsf{ov}(y_i, \hat{y})$$

$$\mathbb{E}_Y(Err_{in}) - \mathbb{E}_Y(e\bar{r}r) = \frac{2d\sigma_\epsilon^2}{N}$$

[1] Lire la section 7.4 de ESL

Estimer Err_{in} [2]

- AIC
- BIC
- C_p

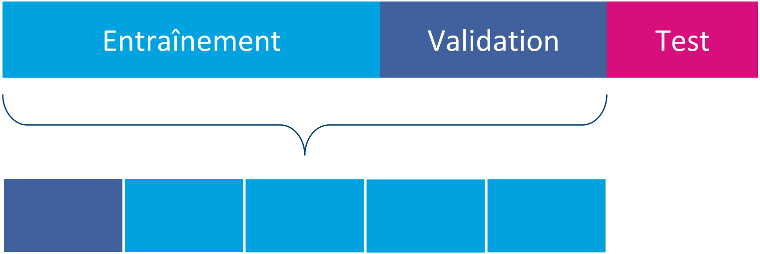
[2] Les sections 7.5, 7.6 et 7.7 sont intéressantes mais ne font pas partie de la matière.

Estimer l'erreur de généralisation (Err)

- Bootstrap
- Validation-croisée

Validation croisée [3]

[3] Lisez la section 7.10 de ESL au complet, particulièrement 7.10.2 et 7.10.3.


Entraînement
Validation
Test

1. On sépare l'échantillon en K plis de façon aléatoire.

1. On sépare l'échantillon en K plis de façon aléatoire.
2. Pour k de 1 à K

1. On sépare l'échantillon en K plis de façon aléatoire.
2. Pour k de 1 à K
   i. On estime $\hat{f}^{(-k)}$ en utilisant uniquement les observations qui ne sont pas dans k

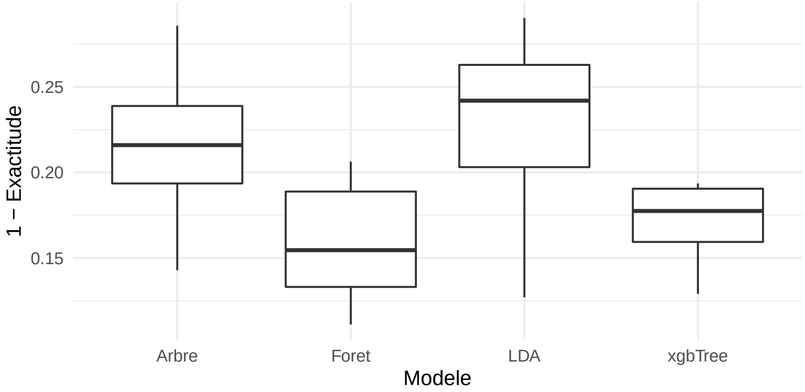
1. On sépare l'échantillon en K plis de façon aléatoire.
2. Pour k de 1 à K
   - i. On estime $\hat{f}^{(-k)}$ en utilisant uniquement les observations qui ne sont pas dans k
   - ii. On prédit $\hat{Y}^{(k)} = \hat{f}^{(-k)}(X^{(k)})$

1. On sépare l'échantillon en K plis de façon aléatoire.
2. Pour k de 1 à K
   i. On estime $\hat{f}^{(-k)}$ en utilisant uniquement les observations qui ne sont pas dans k
   ii. On prédit $\hat{Y}^{(k)} = \hat{f}^{(-k)}(X^{(k)})$
   iii. On calcule $CV^{(k)} = \frac{1}{N_k} \sum_{i=1}^{N_k} L(Y^{(k)}, \hat{Y}^{(k)})$

1. On sépare l'échantillon en K plis de façon aléatoire.
2. Pour k de 1 à K
   i. On estime $\hat{f}^{(-k)}$ en utilisant uniquement les observations qui ne sont pas dans k
   ii. On prédit $\hat{Y}^{(k)} = \hat{f}^{(-k)}(X^{(k)})$
   iii. On calcule $CV^{(k)} = \frac{1}{N_k} \sum_{i=1}^{N_k} L(Y^{(k)}, \hat{Y}^{(k)})$
3. On calcule la moyenne $CV = \frac{1}{K} \sum_{j=1}^{K} CV^{(k)}$

1. On sépare l'échantillon en K plis de façon aléatoire.
2. Pour k de 1 à K
   1. On estime $\hat{f}^{(-k)}$ en utilisant uniquement les observations qui ne sont pas dans k
   2. On prédit $\hat{Y}^{(k)} = \hat{f}^{(-k)}(X^{(k)})$
   3. On calcule $CV^{(k)} = \frac{1}{N_k} \sum_{i=1}^{N_k} L(Y^{(k)}, \hat{Y}^{(k)})$
3. On calcule la moyenne $CV = \frac{1}{K} \sum_{j=1}^{K} CV^{(k)}$
4. On répète 1 à 3 pour tous les modèles et on choisit le modèle dont la valeur de CV est la plus basse.

Distribution de l'erreur mesurée sur chaque pli lors de la validation croisée



Résumé

Vous devriez être en capable de

- Expliquer l'optimisme de l'erreur sur l'échantillon d'entraînement.

Vous devriez être en capable de

- Expliquer l'optimisme de l'erreur sur l'échantillon d'entraînement.
- Coder l'algorithme de validation croisée avec R.

Vous devriez être en capable de

- Expliquer l'optimisme de l'erreur sur l'échantillon d'entraînement.
- Coder l'algorithme de validation croisée avec R.
- Donner des exemples de situation où la validation croisée est plus difficile à mettre en oeuvre.

Vous devriez être en capable de

- Expliquer l'optimisme de l'erreur sur l'échantillon d'entraînement.
- Coder l'algorithme de validation croisée avec R.
- Donner des exemples de situation où la validation croisée est plus difficile à mettre en oeuvre.
- Utiliser la validation croisée pour comparer différents modèles.