

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

En appui du datacamp



Plan

Généralités

Généralités

Pallier le sur-apprentissage

Pallier le
sur-apprentissage

Evaluer la performance des modèles

Evaluer la
performance des
modèles

Plan

Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Généralités

Présentation

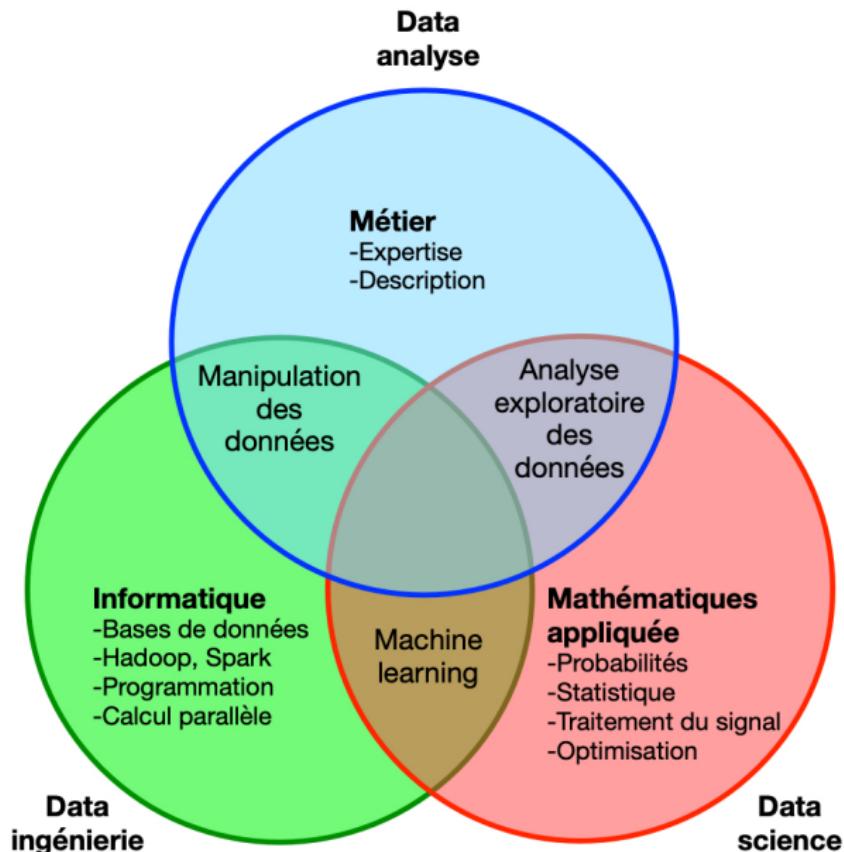
Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

- ▶ **Objectif** : traiter un cas d'étude réel, de grande dimension, dans sa complétude, de l'analyse exploratoire à la modélisation/prévision.
- ▶ **Jeu de données** : proposé par Engie dans le cadre d'un challenge sous l'égide du Collège de France en 2017 (chaire Sciences des données, Stéphane Mallat).
- ▶ **Enjeu métier** : prévoir la production d'énergie éolienne à partir des données opérationnelles des éoliennes afin de détecter les écarts anormaux entre la production prévue et la production réelle.

Data - science, analyse et ingénierie

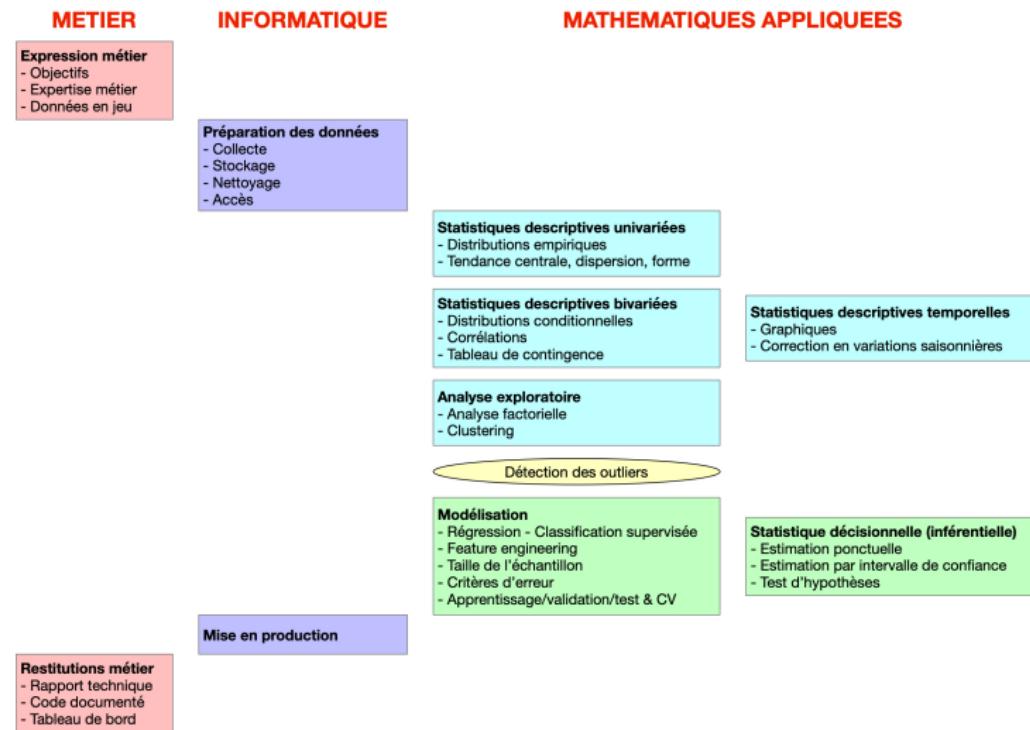


Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Un parcours « data » classique



Généralités

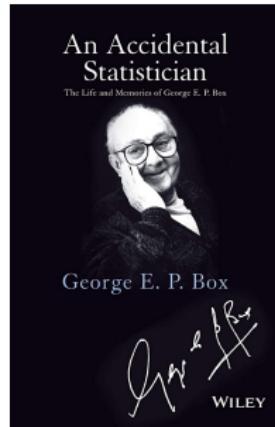
Pallier le sur-apprentissage

Evaluer la performance des modèles

En guise de préambule

All models are wrong,
but some are useful

George E. P. Box



Less is more
Mies van der Rohe



Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Plan

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Pallier le sur-apprentissage

Plusieurs stratégies

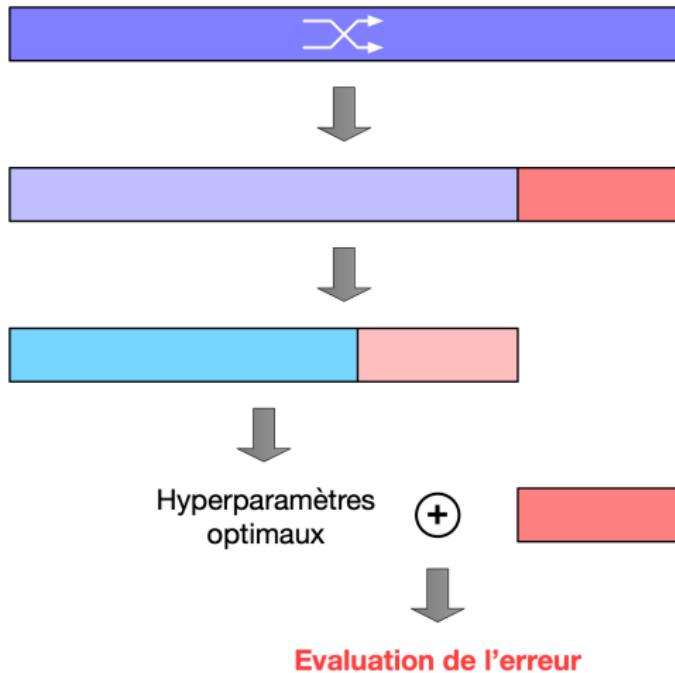
- ▶ Utilisation de **critères d'information** tels que l'AIC et le BIC (ou le C_p de Mallows) pour les **modèles linéaires** (régression linéaire et régression logistique).
- ▶ Méthodes de rééchantillonage :
 - ▶ Les procédures sont :
 - ▶ **Apprentissage / Validation / Test** : dans le cas où on dispose de **gros échantillons**.
 - ▶ **Validation croisée** : dans le cas où on dispose d'**échantillons de taille moyenne**.
 - ▶ **Validation croisée itérée** : dans le cas où on dispose de **petits échantillons**.
 - ▶ L'échantillon de **validation** permet de calibrer les **hyperparamètres optimaux**.
 - ▶ L'échantillon **test** permet d'évaluer l'**erreur de prévision** à partir des données disponibles.

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Apprentissage/validation/test : erreur

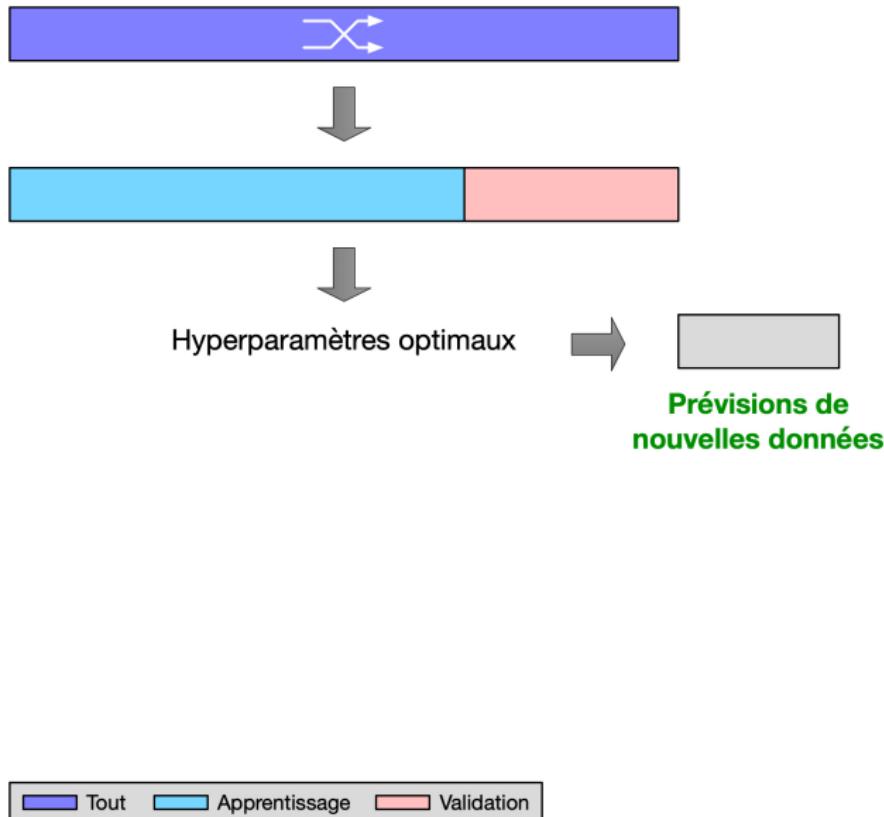


Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Apprentissage/validation/test : prévision



Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Validation croisée I

- Diviser aléatoirement les données en K blocs (égaux ou équivalents).

Le bloc k contient n_k observations : $n_k = \frac{n}{K}$ si n est un multiple de K .

- Pour $k \in \{1, \dots, K\}$:
 - Retirer le bloc k de la base d'apprentissage.
 - Estimer la fonction de prévision sur la base d'apprentissage.
 - Calculer un critère d'erreur de prévision sur le bloc k : CV_k (ex : MSE pour la régression).
- Calculer le critère de validation croisée :

$$\text{CV} = \sum_{k=1}^K \frac{n_k}{n} \text{CV}_k .$$

Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Validation croisée II

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

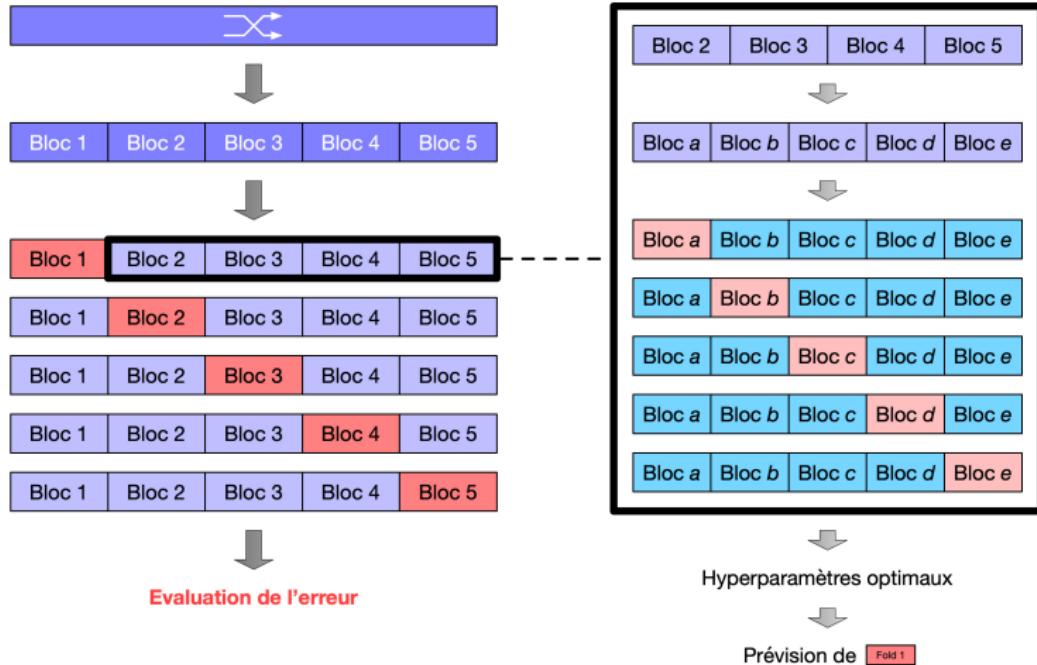
- ▶ Usuellement $K = 5$ ou $K = 10$.
- ▶ Lorsque $K = n$, on parle d'estimateur « **leave one out** » (LOO)

Validation croisée ($K = 5$) : erreur

Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

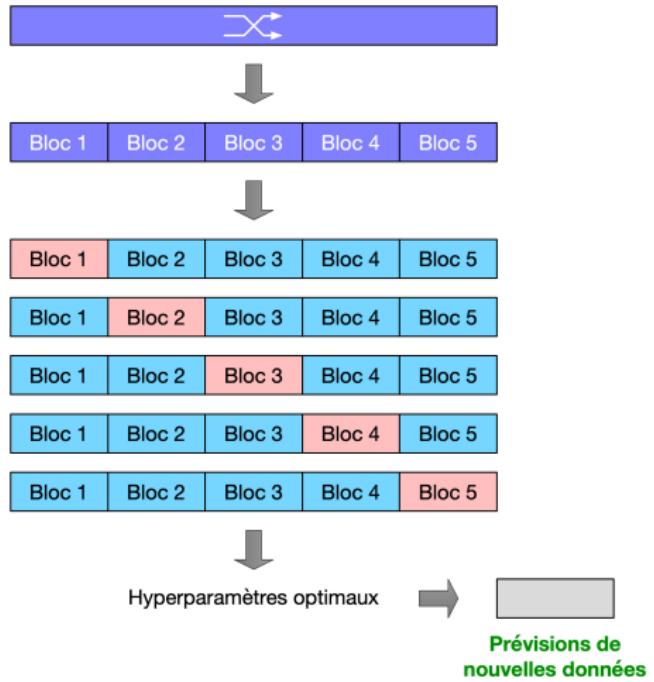


Validation croisée ($K = 5$) : prévision

Généralités

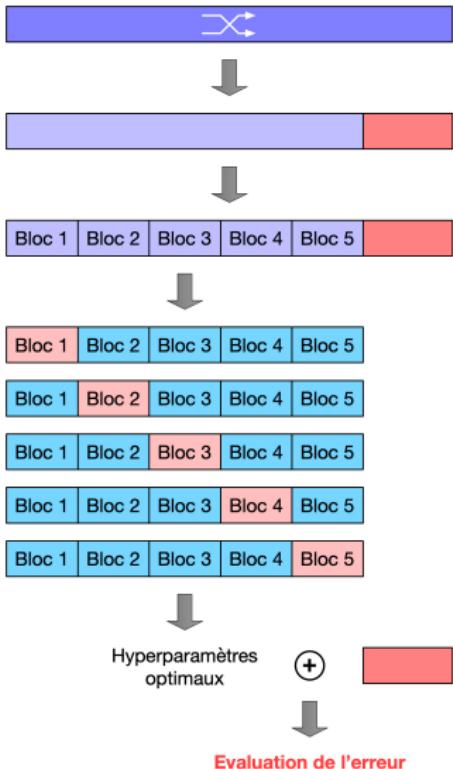
Pallier le sur-apprentissage

Evaluer la performance des modèles



Tout Apprentissage Validation

Validation croisée ($K = 5$) partielle : erreur

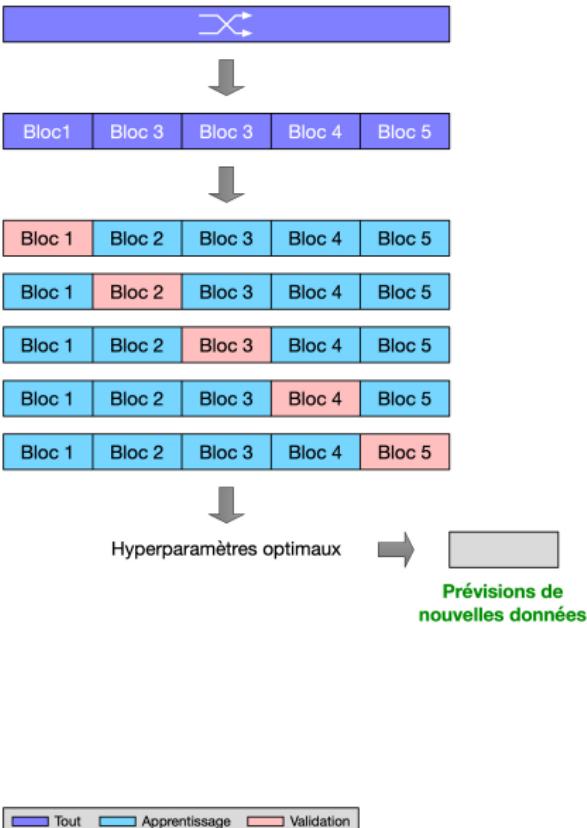


Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Validation croisée ($K = 5$) partielle : prévision



Généralités

Pallier le sur-apprentissage

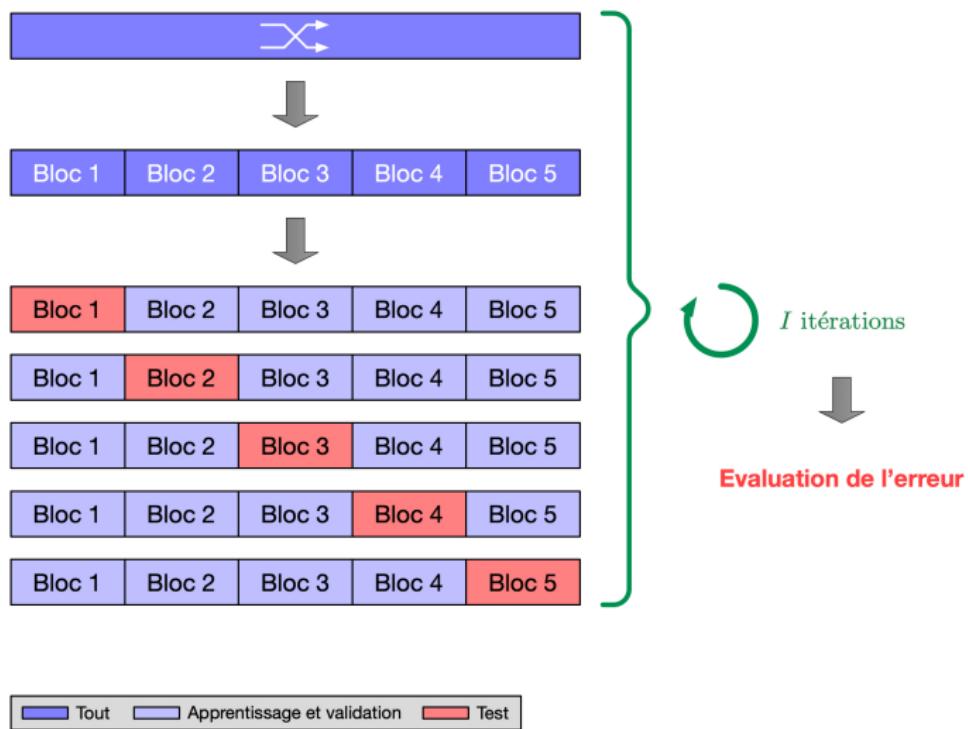
Evaluer la performance des modèles

Validation croisée ($K = 5$) itérée : erreur

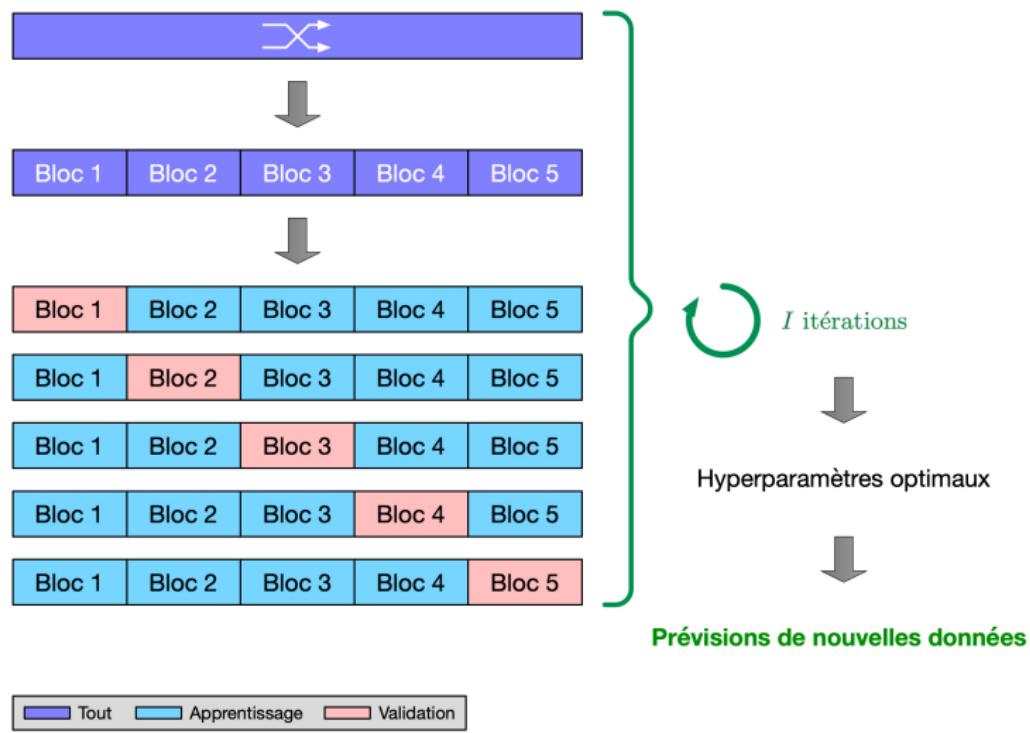
Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles



Validation croisée ($K = 5$) itérée : prévision



Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Plan

Evaluer la performance des modèles

Classification supervisée

Régression

Généralités

Pallier le
sur-apprentissage

**Evaluer la
performance des
modèles**

Classification
supervisée

Régression

Plan

Evaluer la performance des modèles

Classification supervisée

Régression

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Classification
supervisée

Régression

Prévision (cas de la classification supervisée binaire)

Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Classification supervisée

Régression

- ▶ On procède de la manière suivante :
 - ▶ On **estime** un modèle sur l'**échantillon d'apprentissage**.
 - ▶ On utilise ce modèle estimé pour calculer les **probabilités prévues** $\widehat{\mathbb{P}}(Y = 1 / X = x_i)$ sur l'**échantillon test**.
 - ▶ On en déduit les **prévisions** $\widehat{y}_i^{(p)}$ sur l'**échantillon test** :

$$\widehat{y}_i^{(p)} = \begin{cases} 1 & \text{si } \widehat{\mathbb{P}}(Y = 1 / X = x_i) \geq s \\ 0 \text{ (ou -1)} & \text{sinon} \end{cases}$$

où s est un seuil de décision fixé par l'utilisateur (usuellement $s = \frac{1}{2}$).

- ▶ S'il existe de nombreux critères de qualité possibles, il faut choisir celui(ceux) qui répond(ent) au mieux à la problématique métier.

Matrice de confusion

- ▶ Dans le cas de la classification supervisée binaire, la **matrice de confusion** vaut (avec les notations TP et TN pour True Positive et True negative) :

		Prévision	
		1	0 (ou -1)
Vérité	1	Vrai Positif (TP)	Faux Négatif (FN)
	0 (ou -1)	Faux Positif (FP)	Vrai Négatif (TN)

- ▶ Dans le cas de la classification supervisée non-binaire, on peut établir la matrice de confusion, avec autant de lignes et de colonnes que de classes, et en déduire les nombres TP, FP, TN et FN.

Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Classification supervisée

Régression

Critères de qualité

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Classification
supervisée

Régression

- ▶ On considère usuellement :
 - ▶ L'exactitude.
 - ▶ La spécificité.
 - ▶ La précision.
 - ▶ La sensibilité
 - ▶ Le F_1 .
 - ▶ L'AUC.
- ▶ Ces indicateurs prennent leurs valeurs sur $[0, 1]$: plus ils sont proches de 1, meilleur est le modèle.

Exactitude (et erreur de classification) et spécificité

- L'**exactitude** (*accuracy*) vaut :

$$\text{exactitude} = \frac{TP + TN}{TP + FP + TN + FN}.$$

Notons que l'erreur de classification (*classification error*) vaut :

$$\text{erreur} = \frac{FP + FN}{TP + FP + TN + FN} = 1 - \text{exactitude}.$$

- La **spécificité** (*specificity*), le taux de négatifs classés négatifs (« taux de vrais négatifs » noté **TNR** : True Negative Rate), vaut :

$$\text{spécificité} = \frac{TN}{FP + TN}.$$

L'**anti-spécificité** est le taux de négatifs classés positifs (« taux de faux positifs » noté **FPR** : False Positive Rate).

Précision, sensibilité et F_1

- ▶ La **précision** (*precision*), ou valeur prédictive positive, vaut :

$$\text{précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

- ▶ La **sensibilité** (*sensitivity*), ou rappel (*recall*), est le taux de positifs classés positifs (« taux vrais positifs », noté **TPR** : *True Positive Rate*) :

$$\text{sensibilité} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

- ▶ Le score **F_1** est la moyenne harmonique de la précision et de la sensibilité :

$$\begin{aligned}\mathbf{F}_1 &= \frac{2}{\frac{1}{\text{précision}} + \frac{1}{\text{sensibilité}}} \\ &= 2 \frac{\text{précision} \cdot \text{sensibilité}}{\text{précision} + \text{sensibilité}}.\end{aligned}$$

Courbe ROC I

- ▶ La courbe **ROC** (Receiver Operating Characteristic) représente la sensibilité (taux de vrais positifs : TPR) en fonction de l'anti-spécificité (taux de faux positifs : FPR) pour différents seuils de décision s :

$$\hat{y}_i^{(p)} = \begin{cases} 1 & \text{si } \hat{\mathbb{P}}(Y = 1 / X = x_i) \geq s \\ 0 \text{ (ou -1)} & \text{sinon} \end{cases}.$$

- ▶ Plus le seuil s est important :
 - ▶ plus le taux de vrais positifs est important,
 - ▶ moins le taux de faux positifs est important.
- ▶ La courbe ROC est croissante et au-dessus de la première bissectrice (correspondant à une prédiction de type « tirage au sort »).
- ▶ La prédiction « optimale » fournirait une courbe ROC égale à 0 pour $s = 0$ et égale à 1 pour $s \in]0, 1]$.

Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Classification supervisée

Régression

Courbe ROC II

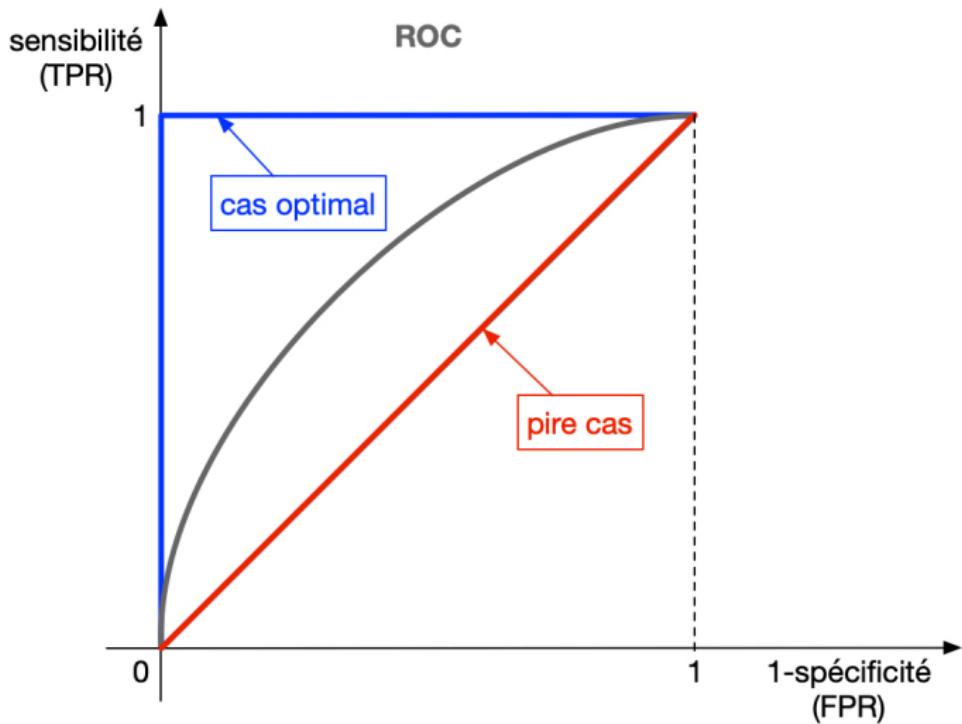
Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Classification supervisée

Régression



AUC I

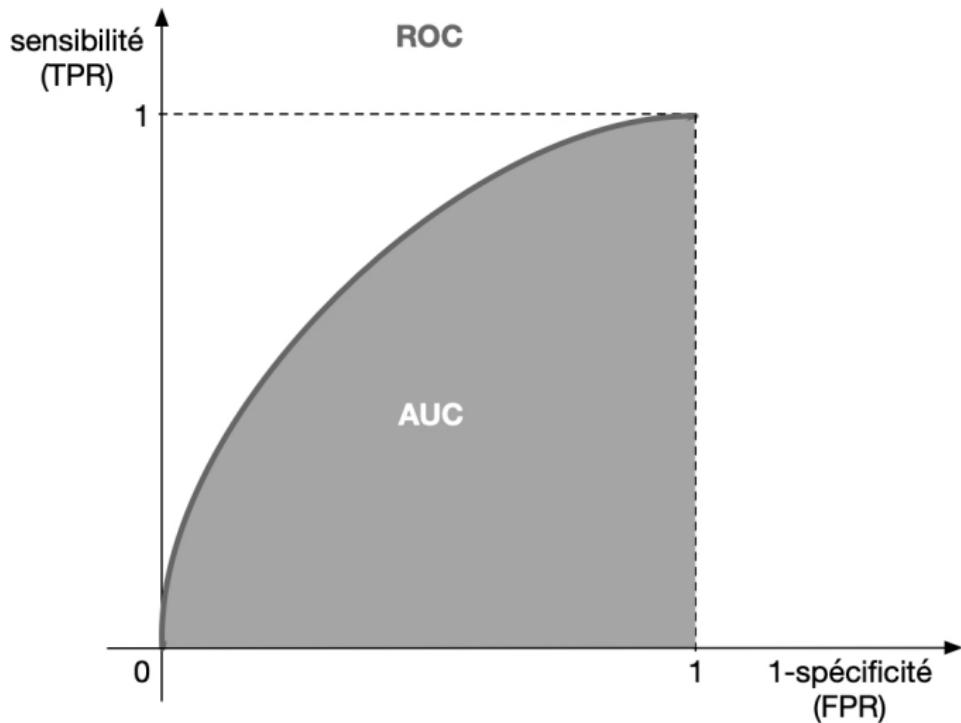
Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Classification
supervisée

Régression



Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Classification
supervisée

Régression

- ▶ L'aire sous la courbe ROC, l'**AUC** (Area Under the ROC), est une mesure de la qualité de la classification et varie entre :
 - ▶ $\text{AUC} = \frac{1}{2}$: le pire des cas (prédiction de type « tirage au sort »),
 - ▶ $\text{AUC} = 1$: le meilleur des cas (prédiction « optimale »).

Exemple ROC/AUC I

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Classification
supervisée

Régression

Individus	Label	Probabilité d'être positif
1	0	0.10
2	0	0.15
3	1	0.20
4	0	0.30
5	0	0.35
6	1	0.35
7	0	0.40
8	1	0.55
9	1	0.60
10	1	0.75
11	1	0.75
12	0	0.80
13	1	0.90
14	1	0.95

Exemple ROC/AUC II

Individus	Label	Probabilité d'être positif	Label prévu (avec $s = 0.25$)
1	0	0.10	0
2	0	0.15	0
3	1	0.20	0
4	0	0.30	1
5	0	0.35	1
6	1	0.35	1
7	0	0.40	1
8	1	0.55	1
9	1	0.60	1
10	1	0.75	1
11	1	0.75	1
12	0	0.80	1
13	1	0.90	1
14	1	0.95	1

$$P = 8, \quad N = 6,$$

$$FP = 4, \quad FPR = \frac{FP}{N} = \frac{4}{6},$$

$$TP = 7, \quad TPR = \frac{TP}{P} = \frac{7}{8}.$$

Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Classification supervisée

Régression

Exemple ROC/AUC III

Généralités

Pallier le
sur-apprentissage

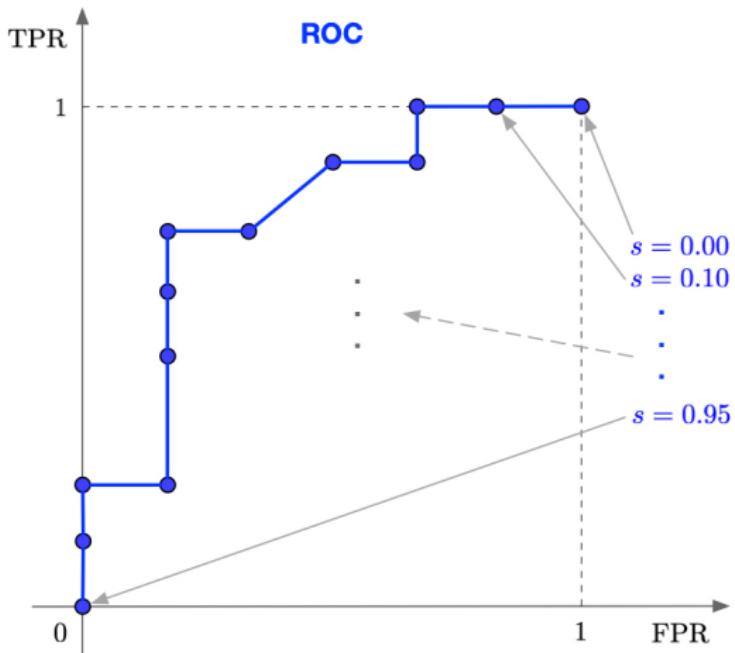
Evaluer la
performance des
modèles

Classification
supervisée

Régression

Seuil	Taux de vrais positifs (TP)	Taux de faux positifs (FP)
0.00	$\frac{8}{8}$	$\frac{6}{6}$
0.10	$\frac{8}{8}$	$\frac{5}{6}$
0.15	$\frac{8}{8}$	$\frac{4}{6}$
0.20	$\frac{7}{8}$	$\frac{4}{6}$
0.30	$\frac{7}{8}$	$\frac{3}{6}$
0.35	$\frac{6}{8}$	$\frac{2}{6}$
0.40	$\frac{6}{8}$	$\frac{1}{6}$
0.55	$\frac{5}{8}$	$\frac{1}{6}$
0.60	$\frac{4}{8}$	$\frac{1}{6}$
0.75	$\frac{2}{8}$	$\frac{1}{6}$
0.80	$\frac{2}{8}$	$\frac{0}{6}$
0.90	$\frac{1}{8}$	$\frac{0}{6}$
0.95	$\frac{0}{8}$	$\frac{0}{6}$

Exemple ROC/AUC IV



$$\text{AUC} = \frac{1}{6} \frac{1}{4} + \frac{1}{6} \frac{3}{4} + \frac{1}{2} \frac{1}{6} \left(\frac{6}{8} + \frac{7}{8} \right) + \frac{1}{6} \frac{7}{8} + \frac{1}{3} = \frac{75}{96} \simeq 0.78.$$

Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Classification supervisée

Régression

Plan

Evaluer la performance des modèles

Classification supervisée

Régression

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Classification
supervisée

Régression

Prévision

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Classification
supervisée

Régression

- ▶ On procède de la manière suivante :
 - ▶ On **estime** un modèle sur l'**échantillon d'apprentissage**.
 - ▶ On utilise ce modèle estimé pour calculer les **prévisions** $\hat{y}_i^{(p)}$ sur l'**échantillon test**.
- ▶ Concernant les critères d'erreur :
 - ▶ S'il existe de nombreux critères possibles, il faut choisir celui(ceux) qui répond(ent) au mieux à la problématique métier.
 - ▶ Les **écart quadratiques** sont classiquement utilisés à l'aide des **RMSE** et **nRMSE** (ils correspondent souvent au critère qui a été optimisé).
 - ▶ On peut utiliser des **écart absolu**s à l'aide du **MAE** ou **absolu**s relatifs à l'aide du **MAPE** (à éviter si Y prend des valeurs proches de 0).

MSE, RMSE et nRMSE

- Le **MSE** (*Mean Squared Error* : erreur quadratique moyenne) vaut :

$$\text{MSE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\hat{y}_i^{(p)} - y_i)^2 .$$

- Le **RMSE** (*Root Mean Squared Error* : racine carrée de l'erreur quadratique moyenne) vaut :

$$\text{RMSE} = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\hat{y}_i^{(p)} - y_i)^2} .$$

- Le **nRMSE** (*normalized RMSE* : RMSE normalisé) vaut :

$$\text{nRMSE} = \frac{\text{RMSE}}{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \hat{y}_i^{(p)}} .$$

Généralités

Pallier le sur-apprentissage

Evaluer la performance des modèles

Classification supervisée

Régression

MAE et MAPE

Généralités

Pallier le
sur-apprentissage

Evaluer la
performance des
modèles

Classification
supervisée

Régression

- Le **MAE** (*Mean Absolute Error* : erreur absolue moyenne) vaut :

$$\text{MAE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left| \hat{y}_i^{(p)} - y_i \right| .$$

- Le **MAPE** (*Mean Absolute Percent Error* : erreur absolue moyenne relative) vaut :

$$\text{MAPE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left| \frac{\hat{y}_i^{(p)} - y_i}{y_i} \right| \times 100 .$$