

INTRODUCTION AU MACHINE LEARNING

RÉGRESSION LINÉAIRE

Théo Lopès-Quintas

BPCE Payment Services,
Université Paris Dauphine

2022 - 2024

RÉGRESSION LINÉAIRE

PROBLÈME

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \left(\underbrace{y_i}_{\text{Vraie valeur}} - \underbrace{f_{\theta}(x^{(i)})}_{\text{Valeur prédite}} \right)^2 \quad (1)$$

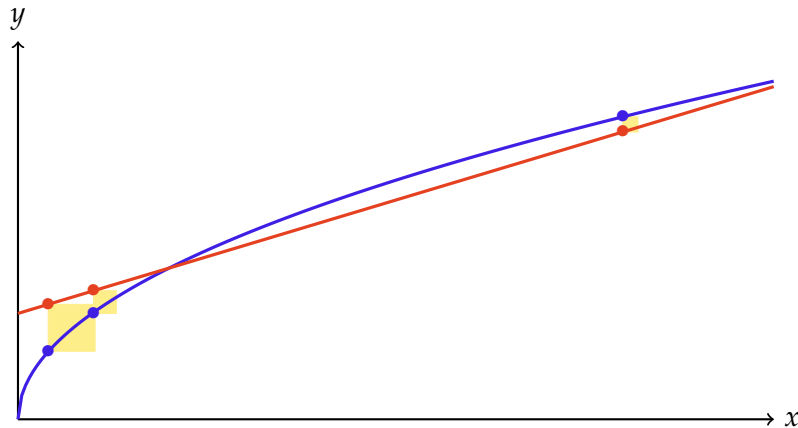


Figure – Visualisation de la MSE entre la Régression linéaire et vraie courbe

RÉGRESSION LINÉAIRE

MODÉLISATION

$$\hat{y} = \theta_0 + \sum_{j=1}^d \theta_j \times x_j$$

On peut réécrire notre problème (1) comme :

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^{d+1}} \sum_{i=1}^n \left[y_i - \left(\theta_0 + \sum_{j=1}^d \theta_j \times x_j^{(i)} \right) \right]^2$$

RÉGRESSION LINÉAIRE

MODÉLISATION

Exercice 1 (Régression linéaire avec une seule information)

On suppose que l'on dispose d'un dataset $\mathcal{D} = \{(x^{(i)}, y_i) \mid \forall i \leq n : x^{(i)} \in \mathbb{R}, y_i \in \mathbb{R}\}$. On a donc une seule information pour prédire la valeur y .

1. Écrire le problème (1) dans le cadre de l'exercice.
2. Donner le meilleur vecteur de paramètre θ .

On note $\bar{x} = \frac{1}{n} \sum_{i=1}^n x^{(i)}$. On rappelle avec cette convention que pour $u, v \in \mathbb{R}^n$:

$$\begin{aligned}\text{Cov}(u, v) &= \overline{uv} - \bar{u} \times \bar{v} \\ \mathbb{V}[u] &= \overline{u^2} - \bar{u}^2\end{aligned}$$

3. Montrer que θ_0^* et θ_1^* les deux paramètres optimaux peuvent s'écrire :

$$\begin{aligned}\theta_0^* &= \bar{y} + \theta_1^* \times \bar{x} \\ \theta_1^* &= \frac{\text{Cov}(x, y)}{\mathbb{V}[x]}\end{aligned}$$

RÉGRESSION LINÉAIRE

RÉSOLUTION

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & x_3^{(n)} & \cdots & x_d^{(n)} \end{pmatrix} \times \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$
$$\Longleftrightarrow$$

$$Y = X\theta + \varepsilon, \text{ avec } \varepsilon \text{ un vecteur de bruit.}$$

On peut réécrire notre problème (1) comme :

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^d} \|Y - X\theta\|^2$$

Proposition 1

Si la matrice X est de rang plein, alors :

$$\theta^* = ({}^tXX)^{-1} {}^tXY$$

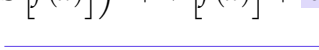
MESURER LA PERFORMANCE D'UNE RÉGRESSION

ERREUR QUADRATIQUE MOYENNE

$$\text{MSE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- $\text{Bias} [\hat{f}(x)] = \mathbb{E} [\hat{f}(x)] - f(x)$: l'écart moyen entre la valeur prédite et la vraie valeur
- $\mathbb{V} [\hat{f}(x)] = \mathbb{E} \left[\left(\mathbb{E} [\hat{f}(x)] - \hat{f}(x) \right)^2 \right]$: la dispersion moyenne des valeurs prédites autour de la moyenne

$$\text{MSE}(y, \hat{f}(x)) = \left(\text{Bias} [\hat{f}(x)] \right)^2 + \mathbb{V} [\hat{f}(x)] + \sigma^2 \quad (2)$$


Erreur incompressible

MESURER LA PERFORMANCE D'UNE RÉGRESSION

RMSE

$$\text{RMSE}(y, \hat{y}) = \sqrt{\sum_{i=1}^n \frac{1}{n} (y_i - \hat{y}_i)^2}$$

Exercice 2 (Ordre de grandeur)

Montrer que :

$$\text{RMSE}(y, \bar{y}) = \sqrt{\overline{y^2} - \bar{y}^2}$$

En déduire une interprétation de la RMSE et un critère de performance d'une régression.

MESURER LA PERFORMANCE D'UNE RÉGRESSION

COEFFICIENT DE DÉTERMINATION R^2

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Prédiction du modèle
↓

↑ Moyenne de la cible

Exercice 3

On suppose que l'on dispose des vecteurs y et \hat{y} .

1. Comment interpréter la valeur 1 pour le R^2 ? Et la valeur 0?
2. Le R^2 peut-il être négatif?

RÉGRESSIONS PÉNALISÉES

RIDGE

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \left(y_i - f_{\theta} \left(x^{(i)} \right) \right)^2 + \mathcal{P}_{\lambda}(\theta) \quad \text{(Pénalisation)}$$

Apprentissage ↓

↑ Pénalisation

$$\theta_{\text{Ridge}}^* = \arg \min_{\theta \in \mathbb{R}^d} \|Y - X\theta\|^2 + \lambda \|\theta\|^2 \quad \text{(Ridge)}$$

$$\theta_{\text{Ridge}}^* = ({}^tXX + n\lambda\mathbb{I}_d)^{-1} {}^tXY$$

RÉGRESSIONS PÉNALISÉES

LASSO

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^d} \underbrace{\sum_{i=1}^n \left(y_i - f_{\theta} \left(x^{(i)} \right) \right)^2}_{\text{Apprentissage}} + \underbrace{\mathcal{P}_{\lambda}(\theta)}_{\text{Pénalisation}} \quad (\text{Pénalisation})$$

$$\theta_{\text{LASSO}}^* = \arg \min_{\theta \in \mathbb{R}^d} \|Y - X\theta\|^2 + \lambda \|\theta\|_1 \quad (\text{LASSO})$$

Exercice 4 (Biais/Variance pour Ridge et LASSO)

Pour la régression Ridge, puis la régression LASSO, comment évolue le biais quand λ augmente ? Même question pour la variance.