INTRODUCTION AU MACHINE LEARNING

RÉGRESSION LINÉAIRE

Théo Lopès-Quintas

BPCE Payment Services,
Université Paris Dauphine

2023

Régression linéaire

Problème

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

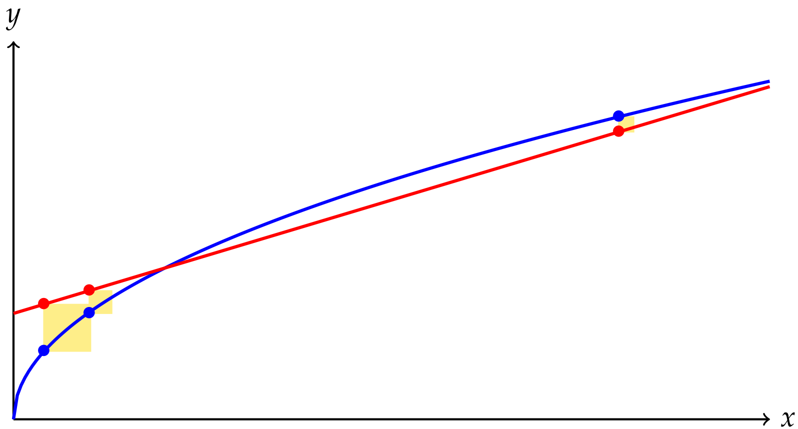

Figure – Visualisation de la MSE entre la Régression linéaire et vraie courbe

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

On peut réécrire notre problème (1) comme :

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

Exercice 1 (Régression linéaire avec une seule information)

On suppose que l'on dispose d'un dataset $\mathcal{D} = \left\{(x^{(i)}, y_i) \mid \forall i \leqslant n : x^{(i)} \in \mathbb{R}, y_i \in \mathbb{R}\right\}$. *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 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$. *On rappelle avec cette convention que pour* $u, v \in \mathbb{R}^n$ *:*

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

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

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

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

$$\iff$$

$$Y = X\theta + \varepsilon$$, avec ε un vecteur de bruit.

On peut réécrire notre problème (1) comme :

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

Proposition 1

Si la matrice X est de rang plein, alors :

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

Mesurer la performance d'une régression

Erreur quadratique moyenne

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

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

$$\mathrm{MSE}(y, \hat{f}(x)) = \left(\mathrm{Bias}\left[\hat{f}(x)\right]\right)^2 + \mathbb{V}\left[\hat{f}(x)\right] + \underbrace{\sigma^2}_{\textbf{Erreur incompressible}} \tag{2}$$

Mesurer la performance d'une régression
RMSE

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

Exercice 2 (Ordre de grandeur)

Montrer que :

$$\mathrm{RMSE}(y, \overline{y}) = \overline{y^2} - \overline{y}^2$$

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

$$R^2 = 1 - \frac{\displaystyle\sum_{i=1}^{n} (y_i - \overbrace{\hat{y}_i}^{\textbf{Prédiction du modèle}})^2}{\displaystyle\sum_{i=1}^{n} (y_i - \underbrace{\overline{y}}_{\textbf{Moyenne de la cible}})^2}$$

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

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

$$\theta^*_{\text{Ridge}} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \|Y - X\theta\|^2 + \lambda \|\theta\|^2 \tag{Ridge}$$

$$\theta^*_{\text{Ridge}} = \left({}^t XX + n\lambda \mathbb{I}_d\right)^{-1} {}^t XY$$

Régressions pénalisées
LASSO

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

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

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.