①

Leçon n°2 : Régression Linéaire, Régression Ridge

Contexte : $(x_1, y_1), \dots, (x_n, y_n) \in (\mathcal{X} \times \mathcal{Y})^n$

Minimisation du risque $R(\theta) = \mathbb{E}_{(x,y)}\left[(y - f_\theta(x))^2\right]$, $\theta \in \Theta$

Estimateur de minimisation du risque empirique

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2 = \hat{R}(\theta)$$

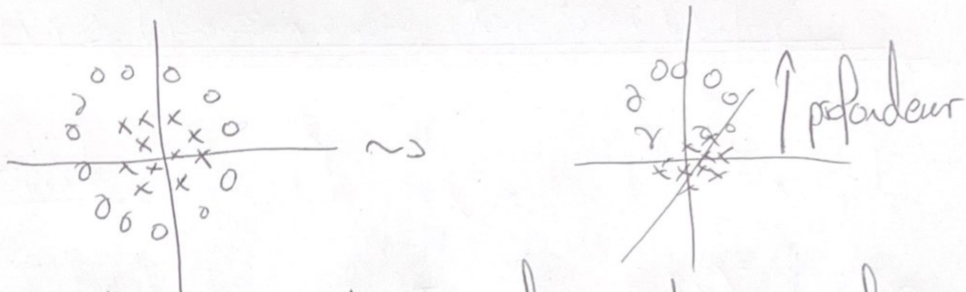
Modèles linéaires $f_\theta(x) = \underbrace{\varphi(x)}_{\text{feature map}}^T \theta$; $\theta \in \mathbb{R}^d$

Exemples :

- $\varphi(x) = x$: Régression linéaire
- $\varphi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}$: Régression affine (Souvent simplement appelé Régression linéaire)
- $\varphi(x) = \begin{pmatrix} \text{monomes en } x \text{ jusqu'au degré } q \end{pmatrix}$:

  Régression polynômiale de degré q

- Avec de bonnes features, il est possible de rendre séparable linéairement des formes complexes (pour plus tard en classification)



- Avec des noyaux : peut se généraliser en dimension infinie

<u>Notation Matricielle :</u> En notant $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ et $\Phi = \begin{pmatrix} \varphi(x_1) \\ \vdots \\ \varphi(x_n) \end{pmatrix}$,

$$\boxed{\hat{\mathcal{R}}(\theta) = \frac{1}{n} \| y - \Phi \theta \|_2^2}$$

I. Moindres Carrés :

<u>Question</u> Comment trouver $\hat{\theta} \in \underset{\theta}{\operatorname{argmin}}\, \hat{\mathcal{R}}(\theta)$?

<u>Hypothèse 1 :</u> Φ a ses colones indépendantes

i.e. Il n'y a pas de dépendance linéaire entre les features du jeu de données.

En pratique, c'est le presque sûrement le cas (sous certaines hypothèses sur φ et sur la distribution) dès lors que n est assez grand.

Définition - Proposition :

Sous l'hypothèse 1, ~~le problème~~ $\hat{R}(\cdot)$ admet un unique minimum sur $\mathbb{R}^d$ appelé estimateur des moindres carrés ordinaire.

Cet estimateur satisfait l'équation normale :

$$\Phi^T \Phi \hat{\theta} = \Phi^T y$$

Il a donc pour expression

$$\boxed{\hat{\theta} = \left(\Phi^T \Phi\right)^{-1} \Phi^T y}$$

preuve : Sous H1, $\hat{R}(\cdot)$ est coercive, et elles est continue et même C^1.

$\hat{R}(\cdot)$ admet donc un minimiseur global qui est solution de l'équation $\nabla \hat{R}(\hat{\theta}) = 0$.

i.e. $\frac{2}{n} \Phi^T \left(\Phi \theta - y\right) = 0$

ou encore $\Phi^T \Phi \hat{\theta} = \Phi^T y$.

Sous H1, cette équation admet une seule solution

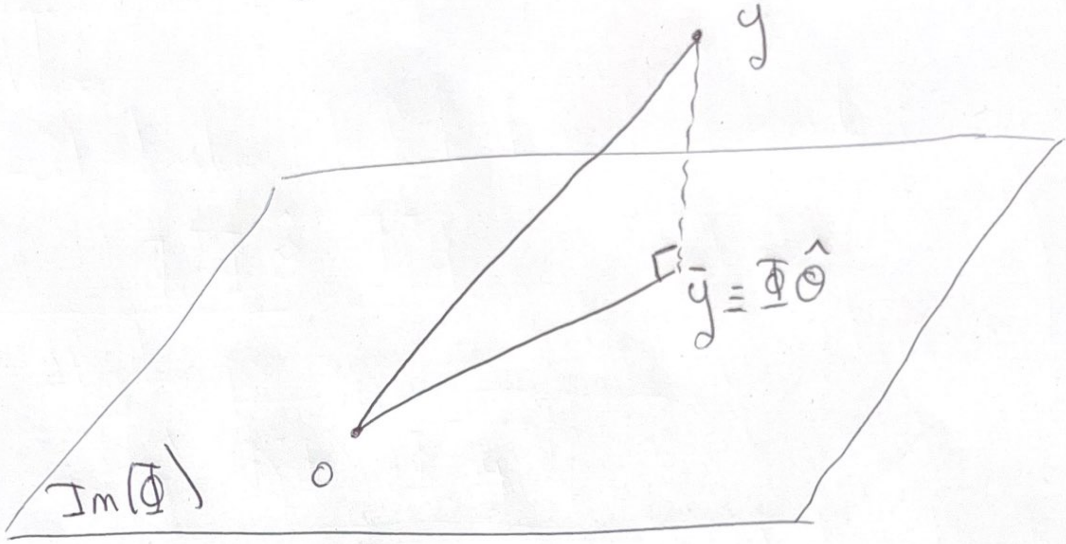
$$\hat{\theta} = \left(\Phi^T \Phi\right)^{-1} \Phi^T y$$

qui est à la fois un minimum local et global.

$\square$

Interprétation géométrique :

- Nous avons l'orthogonalité suivante :



preuve : $\hat{\theta} \in \underset{\theta}{\arg\min}\ \frac{1}{n}\|y - \Phi\theta\|_2^2$

iff $\bar{y} \in \underset{z \in Im(\Phi)}{\arg\min}\ \frac{1}{n}\|y - z\|_2^2$

iff $\bar{y} \in \underset{z \in Im(\Phi)}{\arg\min}\ \frac{1}{n}\|\overbrace{(y - p(y))}^{\in Im(\Phi)^\perp} + \overbrace{(p(y) - z)}^{\in Im(\Phi)}\|_2^2$

où $p(y)$ est la projection orthogonale de y sur $Im(\Phi)$.

iff $\bar{y} \in \underset{z \in Im(\Phi)}{\arg\min}\ \frac{1}{n}\left\{\|y - p(y)\|_2^2 + \|p(y) - z\|_2^2\right\}$

d'après Pythagore.

iff $\bar{y} = p(y)$. $\square$

⑤

II. Analyse statistique - design fixe

Simplification : Dans les couples (x, y), seul le y est aléatoire. Le x peut être vu comme déterministe.

Modèle :

- $\forall i, \quad y_i = \varphi(x_i)^T \theta_* + \varepsilon_i$
- $(\varepsilon_i)_i$ sont indépendantes
- $\forall i, \ \mathbb{E}(\varepsilon_i) = 0, \ \mathbb{E}(\varepsilon_i^2) = \sigma^2$

Risque : $\mathcal{R}(\theta) = \frac{1}{n} \mathbb{E}\left(\| y - \Phi \theta \|_2^2 \right)$

Proposition : $\mathcal{R}^* = \sigma^2$. De plus, pour tout $\hat{\theta}$,

$$\mathcal{R}(\hat{\theta}) - \mathcal{R}^* = \| \hat{\theta} - \theta_* \|_{\hat{\Sigma}}^2 \quad \text{où} \quad \hat{\Sigma} = \frac{1}{n} \Phi^T \Phi .$$

Alors, $$\mathbb{E}(\mathcal{R}(\hat{\theta})) - \mathcal{R}^* = \underbrace{\| \mathbb{E}(\hat{\theta}) - \theta_* \|_{\hat{\Sigma}}^2}_{\text{Biais}^2} + \underbrace{\mathbb{E}\left(\| \hat{\theta} - \mathbb{E}[\hat{\theta}] \|_{\hat{\Sigma}}^2 \right)}_{\text{Variance}}.$$

Preuve :

$$\mathcal{R}(\hat{\theta}) = \frac{1}{n} \mathbb{E}_y \left(\| y - \Phi \hat{\theta} \|_2^2 \right)$$

$$= \frac{1}{n} \mathbb{E}_\varepsilon \left(\frac{1}{n} \| \Phi \theta_* + \varepsilon - \Phi \hat{\theta} \|_2^2 \right)$$

$$= \frac{1}{n} \mathbb{E}_\varepsilon \left(\| \Phi (\theta_* - \hat{\theta}) \|_2^2 + \| \varepsilon \|_2^2 + 2 \left(\Phi (\theta_* - \hat{\theta}) \right)^T \varepsilon \right)$$

$$= \frac{1}{n} \| \hat{\theta} - \theta_* \|_{\hat{\Sigma}}^2 + \underbrace{\sigma^2}_{= \mathcal{R}(\theta_*) = \mathcal{R}_*} + 0$$

⑥ De plus,

$$\mathbb{E}_{\hat\theta}(M(\hat\theta)) - M^* = \mathbb{E}_{\hat\theta}\left(\|\hat\theta - \mathbb{E}(\hat\theta) + \mathbb{E}(\hat\theta) - \theta_*\|^2_{\hat\Sigma}\right)$$

$$= \mathbb{E}_{\hat\theta}\left(\|\hat\theta - \mathbb{E}(\hat\theta)\|^2_{\hat\Sigma}\right) + 2\underbrace{\mathbb{E}_{\hat\theta}\left((\hat\theta - \mathbb{E}(\hat\theta))^T \hat\Sigma (\mathbb{E}(\hat\theta) - \theta_*)\right)}_{=0}$$

$$+ \mathbb{E}_{\hat\theta}\left(\|\mathbb{E}(\hat\theta) - \theta_*\|^2_{\hat\Sigma}\right)$$

$\square$

Avec l'estimateur des moindres carrés ordinaire

Ici: $\hat\theta = (\Phi^T\Phi)^{-1}\Phi^T y = \hat\Sigma^{-1}\left(\frac{1}{n}\Phi^T y\right)$

Analyse du biais :

$$\mathbb{E}(\hat\theta) = \mathbb{E}\left(\hat\Sigma^{-1}\left(\frac{1}{n}\Phi^T y\right)\right) = \hat\Sigma^{-1}\left(\frac{1}{n}\Phi^T \mathbb{E}(y)\right)$$

or : $y = \Phi^T\theta_* + \varepsilon$, donc $\mathbb{E}(y) = \Phi^T\theta_*$ et

$$\mathbb{E}(\hat\theta) = \underbrace{\hat\Sigma^{-1}\hat\Sigma}_{I}\,\theta_* = \theta_*$$

$\hat\theta$ est non biaisé

⑦ Analyse de la variance

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\left((\hat{\theta}-\theta_*)(\hat{\theta}-\theta_*)^T\right)$$

$$= \mathbb{E}\left((\Phi^T\Phi)^{-1}\Phi^T \varepsilon\varepsilon^T \Phi (\Phi^T\Phi)^{-1}\right)$$

$$= (\Phi^T\Phi)^{-1}\Phi^T \underbrace{\mathbb{E}(\varepsilon\varepsilon^T)}_{=\sigma^2 I} \Phi(\Phi^T\Phi)^{-1}$$

$$= \sigma^2 (\Phi^T\Phi)^{-1}$$

$$= \frac{\sigma^2}{n}\hat{\Sigma}^{-1}$$

Risque de l'estimateur des moindres carrés :

Nous pouvons donc appliquer la proposition précédente et obtenir :

$$\mathbb{E}(R(\hat{\theta})) - R^* = \mathbb{E}\left(\|\hat{\theta}-\theta_*\|_{\Sigma}^2\right)$$

$$= \mathbb{E}\left((\hat{\theta}-\theta_*)^T\hat{\Sigma}(\hat{\theta}-\theta_*)\right)$$

$$= \mathbb{E}\left(\mathrm{tr}\left(\quad " \quad \right)\right)$$

$$= \mathbb{E}\left(\mathrm{tr}\left(\hat{\Sigma}(\hat{\theta}-\theta_*)(\hat{\theta}-\theta_*)^T\right)\right)$$

$$= \mathrm{tr}\left(\hat{\Sigma}\,\mathbb{E}\left((\hat{\theta}-\theta_*)(\hat{\theta}-\theta_*)^T\right)\right)$$

$$= \frac{\sigma^2}{n}\mathrm{tr}(I_d) = \frac{\sigma^2 d}{n}$$

$$\boxed{\mathbb{E}(R(\hat{\theta})) - R^* = \frac{\sigma^2 d}{n}}$$

III Régression de Ridge - Design fixe

Problèmes de la régression linéaire en grande dimension ($d \geq n$)

- Overfitting : l'estimateur apprend le jeu d'apprentissage par coeur
- Solutions complexes car l'équation normale n'admet plus une unique solution.

Solutions :

- Réduction de dimension
- Régularisation (Régression Ridge, Régression Lasso).

Régression Ridge : Pour $\lambda > 0$,

$$\hat{\theta}_\lambda \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_2^2$$

1) Expression de la solution :

Comme pour l'estimateur des moindres carrés ordinaires, il est facile de trouver une condition nécessaire sur $\hat{\theta}_\lambda$ avec de l'analyse de base. Comme $\theta \mapsto \underbrace{\frac{1}{n}\|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_2^2}_{= \hat{R}_\lambda(\theta)}$ est coercive et C^1,

$\hat{\theta}_\lambda$ existe et satisfait $\nabla \hat{R}_\lambda(\hat{\theta}_\lambda) = 0$,

i.e. $\frac{2}{n}(\Phi^T \Phi \hat{\theta}_\lambda - \Phi^T y) + 2\lambda \hat{\theta}_\lambda = 0$.

(9). Comme $\lambda > 0$, cette équation admet une unique solution

$$\boxed{\hat{\theta}_\lambda = \frac{1}{n}\left(\hat{\Sigma} + \lambda I\right)^{-1} \Phi^T y}$$

2) Analyse statistique

Analyse du biais

$$\mathbb{E}(\hat{\theta}_\lambda) = \frac{1}{n}\left(\hat{\Sigma} + \lambda I\right)^{-1} \Phi^T \mathbb{E}(y)$$

$$= \frac{1}{n}\left(\hat{\Sigma} + \lambda I\right)^{-1} \Phi^T \Phi \theta_*$$

$$= \left(\hat{\Sigma} + \lambda I\right)^{-1} \hat{\Sigma} \theta_*$$

$$= \left(\hat{\Sigma} + \lambda I\right)^{-1} \left[\left(\hat{\Sigma} + \lambda I\right) - \lambda I\right] \theta_*$$

$$= \theta_* - \lambda \left(\hat{\Sigma} + \lambda I\right)^{-1} \theta_*$$

De plus, remarquons que $\hat{\Sigma}$ et $\left(\hat{\Sigma} + \lambda I\right)^{-1}$ commutent

($\hat{\Sigma}$ est PSD)

~~Alors~~ Biais$^2 = \|\mathbb{E}(\hat{\theta}_\lambda) - \theta_*\|_{\hat{\Sigma}}^2 = \lambda^2 \theta_*^T \left(\Sigma + \lambda I\right)^{-2} \hat{\Sigma} \theta_*$

Analyse de la variance:

$$\text{Variance} = \mathbb{E}\left(\|\hat{\theta}_\lambda - \mathbb{E}(\hat{\theta}_\lambda)\|^2_{\hat{\Sigma}}\right)$$

$$= \mathbb{E}\left(\|\frac{1}{n}(\hat{\Sigma}+\lambda I)^{-1}\Phi^T\varepsilon\|^2_{\Sigma}\right)$$

$$= \mathbb{E}\left(\frac{1}{n^2}\operatorname{tr}\left(\varepsilon^T\Phi(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\Phi^T\varepsilon\right)\right)$$

$$= \mathbb{E}\left(\frac{1}{n^2}\operatorname{tr}\left(\varepsilon\varepsilon^T\Phi(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\right)\right)$$

$$= \frac{\sigma^2}{n}\operatorname{tr}\left(\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\right)$$

$$= \frac{\sigma^2}{n}\operatorname{tr}\left(\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\right)$$

Analyse du risque

Par la décomposition Biais-Variance, nous avons

$$\mathbb{E}(R(\hat{\theta}_\lambda)) - R^* = \lambda^2\theta_*^T(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\theta_*$$
$$+ \frac{\sigma^2}{n}\underbrace{\operatorname{tr}\left(\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\right)}_{\text{degrees of freedom}}$$

Simplification de la borne supérieure

Remarquons que μ est une valeur propre de $\hat{\Sigma}$ ssi

$\frac{\lambda\mu}{(\mu+\lambda)^2}$ est une valeur propre de $(\hat{\Sigma}+\lambda I)^{-2}\lambda\hat{\Sigma}$.

Comme pour tous λ, μ nous avons $2\lambda\mu \leq (\lambda+\mu)^2$, toutes les valeurs propres de $(\hat{\Sigma}+\lambda I)^{-2}\lambda\hat{\Sigma}$ sont inférieures à $\frac{1}{2}$.

Alors $\text{Biais}^2 \leq \frac{\lambda}{2}\|\theta_*\|^2$

et $\text{Variance} \leq \frac{\sigma^2 \operatorname{tr}(\hat{\Sigma})}{2\lambda n}$

Donc

$$\mathbb{E}\left(R(\hat{\theta}_\lambda)\right) - R^* \leq \frac{\lambda}{2}\|\theta_*\|_2^2 + \frac{\sigma^2 \operatorname{tr}(\hat{\Sigma})}{2\lambda n}.$$

Minimiser cette expression en λ donne

$$\lambda_* = \frac{\sigma \operatorname{tr}(\hat{\Sigma})^{1/2}}{\|\theta_*\|_2\sqrt{n}}$$ et

$$\mathbb{E}\left(R(\hat{\theta}_{\lambda^*})\right) - R^* \leq \frac{\sigma \operatorname{tr}(\hat{\Sigma})^{1/2}\|\theta_*\|_2}{\sqrt{n}}$$