Introduction à l'apprentissage

$$\mathcal{D} = \{(x^{(i)}, t^{(i)})\}_{i=1}^{N}$$

→ Regression linéaire

$\vec{x} \in \mathbb{R}^D \quad \vec{\beta} \in \mathbb{R}^{D+1}$

$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_D x_D$$

$\tilde{x} = [1, \vec{x}]$

$$\rightarrow J(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2$$

→ 2 approches → Descente de gradient

→ Résolution des équations normales

$$\rightarrow \vec{\beta} = (\underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}})^{-1} \underline{\underline{\tilde{X}}}^T \vec{t}$$

$\underline{\underline{X}} \in \mathbb{R}^{N \times D+1}$

→ Cas des caractéristiques dépendantes

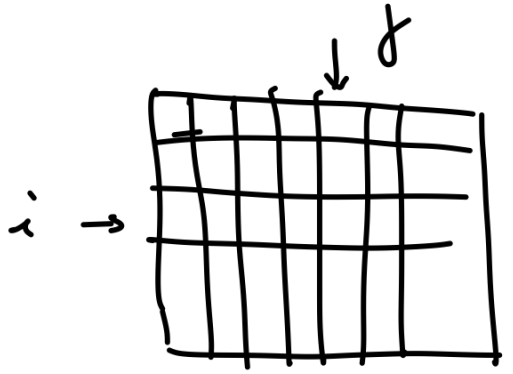
→ $y(x) = \tilde{\underline{\underline{X}}}\beta = \tilde{\underline{\underline{X}}}(\tilde{\underline{\underline{X}}}^T\tilde{\underline{\underline{X}}})^{-1}\tilde{\underline{\underline{X}}}^T t$

Si caractéristiques linéairement dépendantes → $\exists \vec{v} \quad \tilde{\underline{\underline{X}}}^T\tilde{\underline{\underline{X}}} v = 0$

→ $\exists \lambda_i \approx 0$

pour rappel $\tilde{\underline{\underline{X}}}^T\tilde{\underline{\underline{X}}} = U \Lambda U^T$ $\quad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \ddots \\ \vdots & & \lambda_{D+1} \\ 0 & & \end{bmatrix}$

$(\tilde{\underline{\underline{X}}}^T\tilde{\underline{\underline{X}}})^{-2} = U \Lambda^{-2} U^T$ $\quad \Lambda^{-2} = \begin{bmatrix} \lambda_1^{-2} & & \\ & \ddots & \\ & & \lambda_{D+1}^{-2} \end{bmatrix}$



Si caractéristiques sont dépendantes $\Rightarrow \exists\ \lambda_i \ll$

$$\Rightarrow (\tilde{\underline{\underline{X}}}^T \tilde{\underline{\underline{X}}})^{-1} \gg \quad \vec{\beta}$$

$$\Rightarrow y(x) = \vec{X}_{kl} (\tilde{\underline{\underline{X}}}^T \tilde{\underline{\underline{X}}})^{-1} \tilde{\underline{\underline{X}}}^T \vec{t}$$

$$\vec{X}_{kl} (\tilde{\underline{\underline{X}}}^T \underline{\underline{X}})^{-1} \tilde{\underline{\underline{X}}}^T t + \delta$$

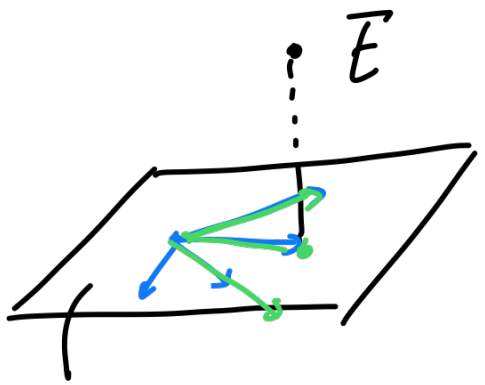
→ On peut montrer que l'entré (i,j) de la matrice inverse $(\tilde{\underline{\underline{X}}}^T \tilde{\underline{\underline{X}}})^{-1}$ est donnée par

$$\frac{(\varepsilon_i)^T (\varepsilon_j)}{\|\varepsilon_i\|^2 \|\varepsilon_j\|^2}$$

où $\varepsilon_i = X_i - \hat{X}_i$ et $\hat{X}_i$ est la projection de X_i (i^ème col. de $\tilde{\underline{\underline{X}}}$) sur l'espace engendré par toutes les colonnes restantes.

On peut aussi voir la minimisation de la somme des carrés des résidus comme le calcul d'une projection orthogonale sur l'espace engendré par les colonnes de la matrice $\underline{\underline{\tilde{X}}}$

$$\min_{\beta} \| \vec{t} - \underline{\underline{\tilde{X}}} \beta \|^2$$

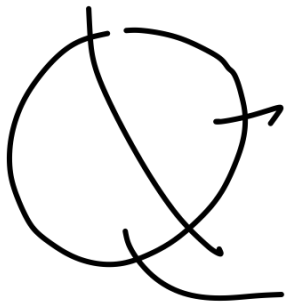


span engendré par les colonnes de $\underline{\underline{\tilde{X}}}$

Solution #1 → Gram Schmidt

#2 → Selection du meilleur sous-ensemble

$\begin{pmatrix} D \\ k' \end{pmatrix}$ → Option a) → $\mathcal{D}$



→ test $\mathcal{D}_{test}$

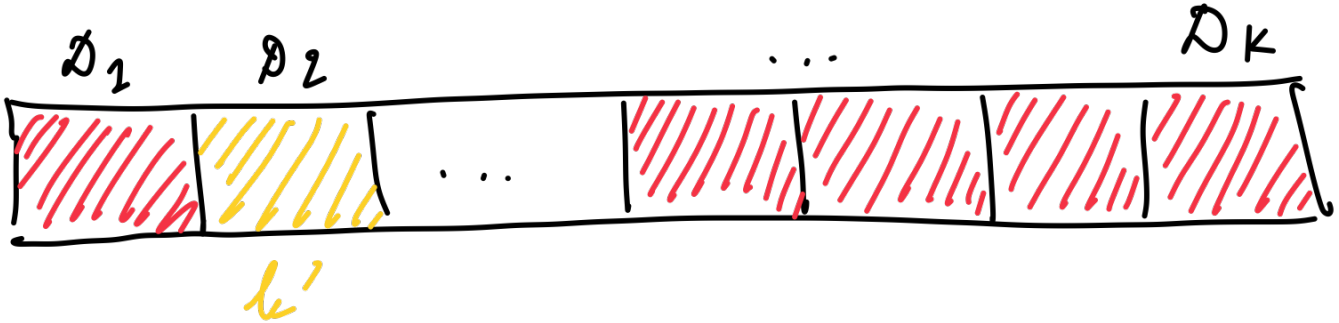
entrainement $\mathcal{D}_{train}$

→ pour chaque sous ensemble parmi les $\binom{D}{K'}$ entraîner le modèle sur $\mathcal{D}_{train}$ et l'évaluer sur $\mathcal{D}_{test}$.

Option b) Validation croisé à K compartiments

pour chaque sous ensemble de caractéristique $\binom{D}{K'}$

on subdivise l'ensemble $\mathcal{D}$ en K compartiments



$$\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$$

pour chaque compartiment k', on entraîne le modèle sur les K-1 compartiments restant

Lorsque les modèles associés à chaque compartiment ont été entraînés, calculer l'erreur pour le modèle défini sur le sous ensemble de caractéristiques

$$Err_{CV} = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_{\beta}^{-k'(i)}(x^{(i)}) \right)^2$$

où $h_{\beta}^{-k'(i)}(x)$ est le modèle entraîné sur tous les compartiments sauf celui contenant la donnée i

→ Caractéristiques dépendantes → $h_\beta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ avec $x_2 = x_1$

β_1 $\quad -\beta_1 + \Delta$

→ Alternative #3 Pénalité Ridge / Régression de type Ridge

$$\min_\beta \underbrace{\ell(\beta)}_{\text{RIDGE}} = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2 + \lambda \sum_{j=1}^{D} \beta_j^2$$

contrôle l'équilibre entre fidélité aux données et complexité du modèle

→ Pour résoudre les équations normales, on commence par centrer les données

$$t^{(i)} \leftarrow t^{(i)} - \frac{1}{N} \sum_{i=1}^{N} t^{(i)} \qquad \forall i$$

$$x_j^{(i)} \leftarrow x_j^{(i)} - \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} \qquad \forall i, j$$

Si on calcule la dérivée par rapport à β_0 on trouve

$$\frac{\partial \ell}{\partial \beta_0} = \frac{2}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)(-1)$$

Si on isole β_0, on trouve

Si données centrées $\beta_0 = 0$

$$2\beta_0 = \frac{2}{N} \sum_{i=1}^{N} t^{(i)} - \frac{2}{N} \sum_{i=1}^{N} (\beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})$$

(0 sous $\sum t^{(i)}$ et 0 sous $\sum \beta_1 x_1^{(i)}$)

On peut donc réécrire $\ell_{RIDGE}(\beta)$ en fonction des $\beta_1, \dots, \beta_D$ uniquement

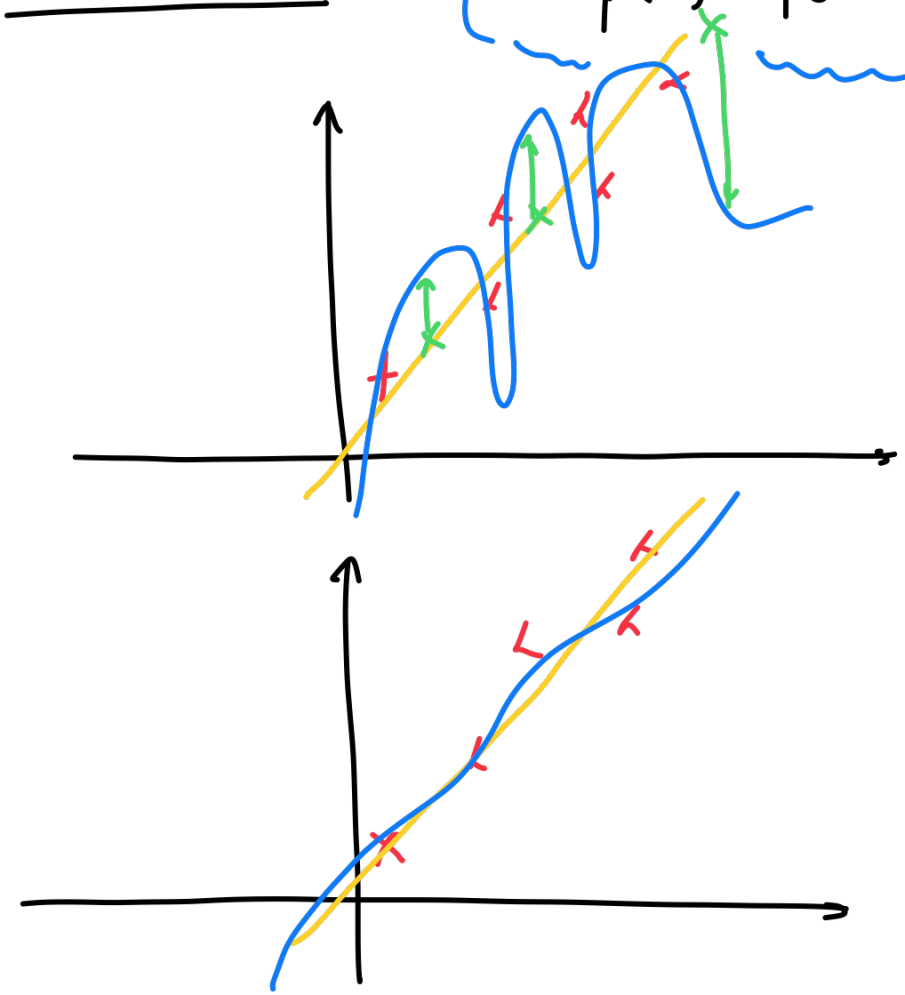
Si on suppose les données centrées

$$\ell_{RIDGE}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2 + \lambda \sum_{j=1}^{D} \beta_j^2$$

Illustration

$$h_\beta(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_D x^D$$

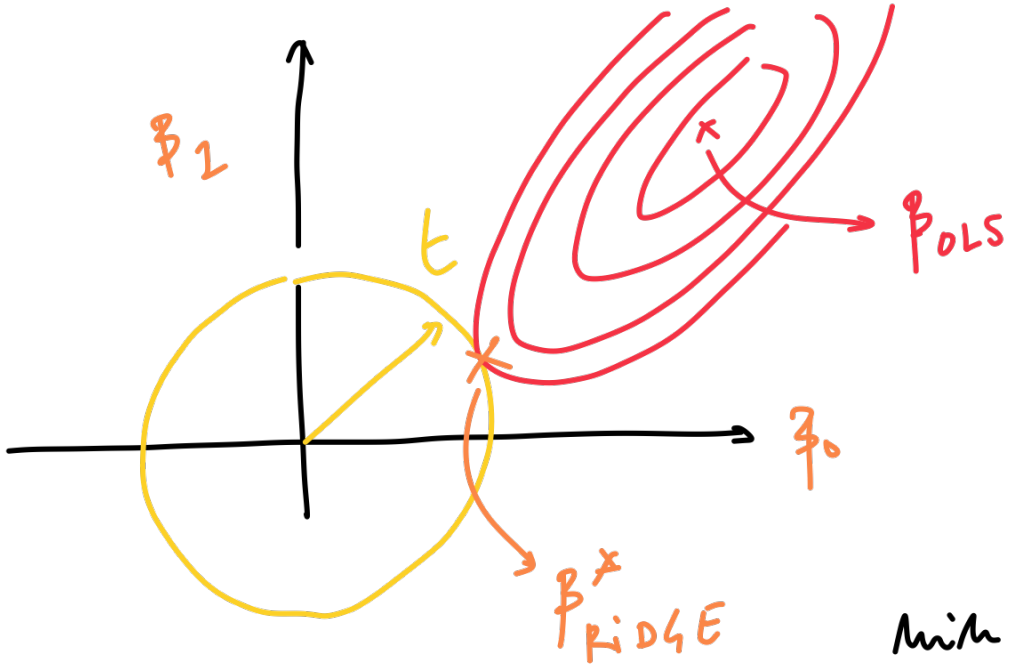
x_1 x_2 x_D



pénalité de type RIDGE

$$h_\beta(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_D x_D^D$$

$\beta_1 \gg \beta_j \quad j \neq 1$



Formulation contrainte

$$\min_{\beta} \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - h_\beta(x^{(i)})\right)^2 + \lambda \sum_{j=1}^{D} \beta_j^2$$

Formulation contrainte

$$\min_{\beta} \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - h_\beta(x^{(i)})\right)^2$$

$$\text{s.t} \quad \sum_{j=1}^{D} \beta_j^2 \leq t$$