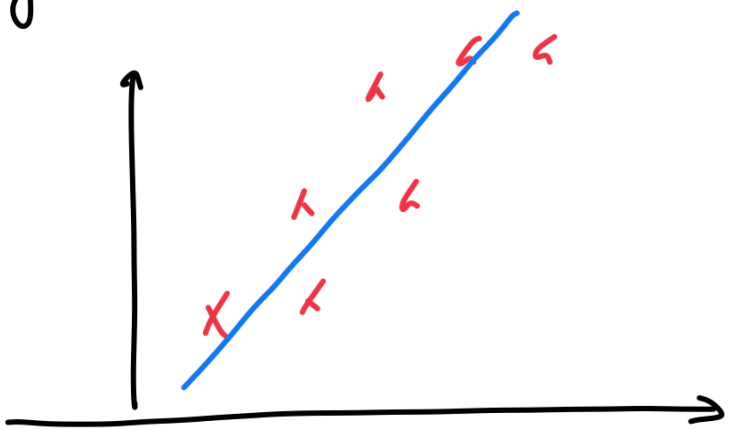
Introduction à l'apprentissage

→ Régression linéaire



$\{x^{(i)}, t^{(i)}\}$ $x^{(i)} \in \mathbb{R}^D$

$t^{(i)} \in \mathbb{R}$

→ Modèle linéaire

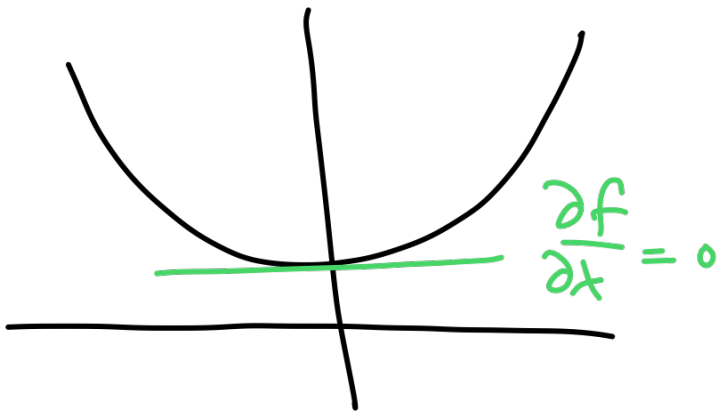
$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_D x_D$$

$$= \beta^T \tilde{x} = [\beta_0 \dots \beta_D] \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix} \leftarrow \tilde{x}$$

→ Fonction de coût (loss)

→ Mesure de l'erreur (Sum of Squares)

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



⟶ Fonction de coût Convexe

↳ 2 approches : → descente de gradient

→ Résolution des équations normales

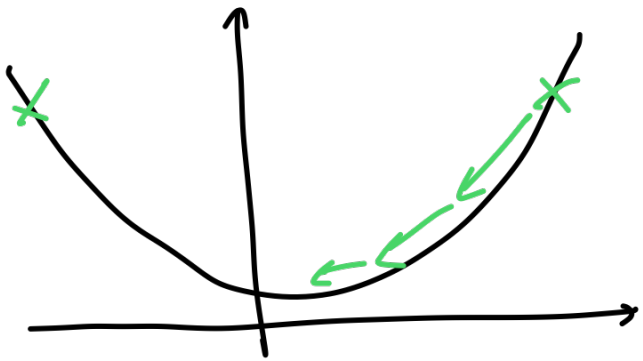
$$\ell(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2 \longleftarrow$$

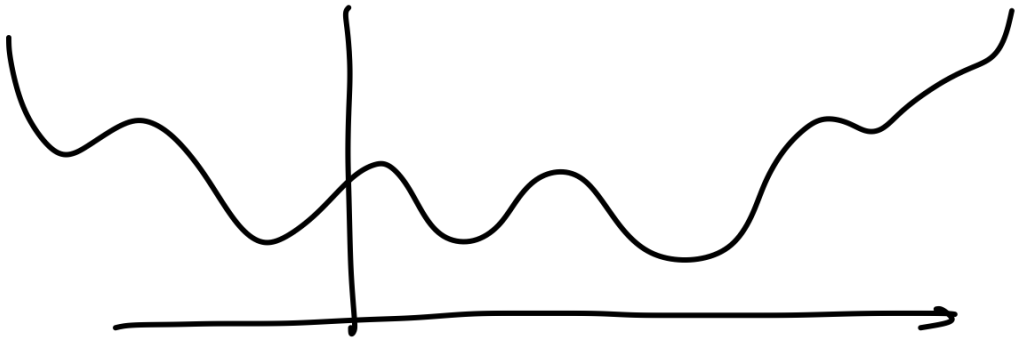
$$\text{gradient} \rightarrow \frac{\partial \ell}{\partial \beta_j} = \frac{2}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\underline{\beta_0} + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right) \tilde{x}_j^{(i)}$$

→ Itérations de (descente de) gradient

$\beta^{(0)}$

$\beta^{(k+1)} \leftarrow \beta^{(k)} - \eta \cdot \underset{\beta}{\text{grad}}\, \ell(\beta)$





→ Résolution de equations

$$t^{(\cdot)} = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} \in \mathbb{R}^N \qquad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_D \end{bmatrix} \in \mathbb{R}^{D+1}$$

$$\underline{\underline{X}} = \begin{bmatrix} 1 & \text{—} & \vec{x}^{(1)} & \text{—} \\ \vdots & & & \\ 1 & \text{—} & \vec{x}^{(N)} & \text{—} \end{bmatrix} \in \mathbb{R}^{N \times (D+1)}$$

$$\vec{\varepsilon} = \left(\vec{t} - \tilde{\underline{X}} \vec{\beta} \right) \qquad \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} \approx \begin{bmatrix} 1 & - \vec{x}^{(1)} - \\ \vdots & \vdots \\ 1 & - \vec{x}^{(N)} - \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_1 \end{bmatrix}$$

$$\rightarrow \mathcal{L}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i^2 = \| \vec{\varepsilon} \|^2 = \langle \vec{\varepsilon}, \vec{\varepsilon} \rangle = \vec{\varepsilon}^T \vec{\varepsilon}$$

Dérivée de $\mathcal{L}(\beta)$ par rapport à $\vec{\beta}$

$$\mathcal{L}(\beta) = \frac{1}{N} \quad (\vec{t} - \tilde{\underline{X}} \vec{\beta})^T (\vec{t} - \tilde{\underline{X}} \beta)$$

$$\mathcal{L}(\beta) = \frac{1}{N}\,\vec{\varepsilon}^{\,T}\vec{\varepsilon} + \vec{\beta}^{\,T}\underline{\underline{\tilde{X}}}^{T}\underline{\underline{\tilde{X}}}\,\vec{\beta} - \vec{\varepsilon}^{\,T}\underline{\underline{\tilde{X}}}\,\vec{\beta} - \vec{\beta}^{\,T}\underline{\underline{\tilde{X}}}^{T}\vec{\varepsilon}$$

transposé du produit $(A\,B)^T = B^T A^T$

$-2\,\vec{\varepsilon}^{\,T}\underline{\underline{\tilde{X}}}$

dérivé = 0

$2\,\underline{\underline{\tilde{X}}}^{T}\underline{\underline{\tilde{X}}}\,\vec{\beta}$

$$\vec{\varepsilon}^{\,T}\underline{\underline{\tilde{X}}} = \underbrace{\left[\varepsilon^{(1)} \dots \varepsilon^{(N)}\right]}_{\in \mathbb{R}^{1\times N}} \underbrace{\begin{bmatrix} 1 & -\vec{x}^{(1)}- \\ \vdots & \vdots \\ 1 & -\vec{x}^{(N)}- \end{bmatrix}}_{\mathbb{R}^{N\times(D+1)}}$$

$$= \vec{v}^{\,T}\vec{\beta} \quad \text{ou} \quad v = \vec{\varepsilon}^{\,T}\underline{\underline{\tilde{X}}}$$

$$\frac{\partial(v^T\beta)}{\partial\beta} = \frac{\partial}{\partial\beta}\left(v_0\beta_0 + v_1\beta_1 + \dots + v_D\beta_D\right) = \left[v_0\ v_1\ \dots\ v_D\right]$$

$$\frac{\partial(v^T \vec{\beta})}{\partial \beta} = \vec{v}^T$$

$$\Rightarrow \frac{\partial}{\partial \beta}\left(\vec{E}^T \underline{\underline{X}}\, \vec{\beta}\right) = 2\vec{E}^T \underline{\underline{X}}$$

$$\left[\beta_1 A_{11} + \beta_2 A_{12} \quad \beta_1 A_{12} + \beta_2 A_{22}\right]$$

pour le terme d'ordre 2

$$\frac{\partial}{\partial \beta}\left(\beta^T A \beta\right) \rightarrow \begin{bmatrix}\beta_1 & \beta_2\end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22}\end{bmatrix} \begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix}$$

$$= A_{11}\beta_1^2 + A_{22}\beta_2^2 + 2A_{12}\beta_1\beta_2 \Big)$$

$$\frac{\partial}{\partial \beta_1} = 2 A_{11} \beta_1 + 2 A_{12} \beta_2$$

$$\frac{\partial}{\partial \beta_2} = 2 A_{22} \beta_2 + 2 A_{12} \beta_1$$

$$= \frac{\partial (\beta^T A \beta)}{\partial \beta} = 2 \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$= 2 A \beta$$

→ Regroupant les dérivés on obtient

$$\frac{\partial l}{\partial \beta} = 2 \underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}} \vec{\beta} - 2 \underline{\underline{\tilde{X}}}^T \vec{t} \overset{\frown}{=} 0$$

pour trouver le minimum

$$Ax - b = 0$$

→ $\tilde{\underline{X}}^T \tilde{\underline{X}}\, \vec{\beta} = \tilde{\underline{X}}^T \vec{E}$ — Equations Normales

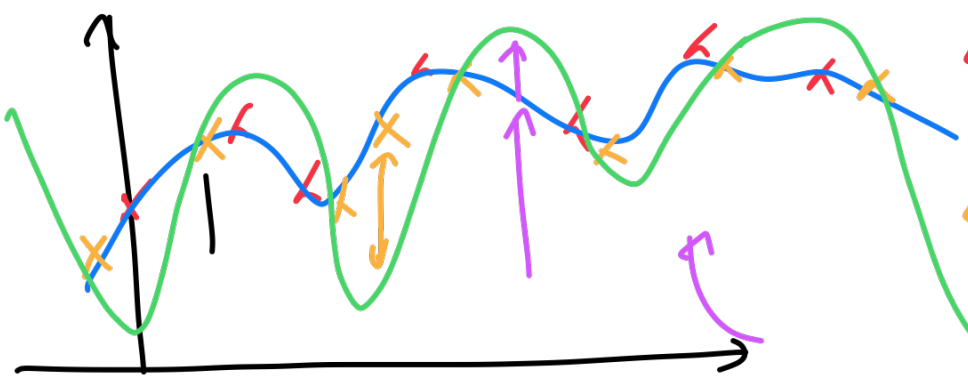
$\vec{\beta} = (\tilde{\underline{X}}^T \tilde{\underline{X}})^{-1} \tilde{\underline{X}}^T \vec{E}$ → solution du problème de minimisation de la somme des carrés

$\vec{E} + \delta$

→ OK si la matrice est inversible.

→ ? si la matrice est mal conditionnée

$$\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{b+1} \end{bmatrix}^{-1} \rightarrow \begin{bmatrix} \lambda_1^{-2} & & \\ & \ddots & \\ & & \lambda_{b+1}^{-2} \end{bmatrix}$$



$$\underline{\underline{\tilde{X}}} = \begin{bmatrix} 1 & — & \tilde{x}^{(1)} & — \\ \vdots & & \vdots & \\ 1 & — & x^{(N)} & — \end{bmatrix}$$

→ Si colonnes sont linéairement dépendantes ?

→ x_j sont corrélés

→ Première approche : extraire un ensemble de colonnes orthogonales

Soit $\underline{z_0} = c_0$ la première colonne de $\tilde{\underline{X}} = \begin{bmatrix} | & | & & | \\ c_0 & c_1 & \cdots & c_{D+1} \\ | & | & & | \end{bmatrix}$

for $j = 1, \ldots, D+1$

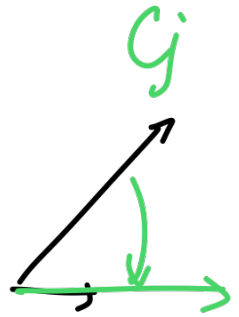
[Calculer les projections $\hat{\gamma}_{kj} = \frac{\langle z_k, c_j \rangle}{\langle z_k, z_k \rangle}$

pour $k = 0, \ldots, j-1$

Définir la nouvelle colonne z_j via

$$\vec{z}_j = \vec{c}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \vec{z}_k$$

$z=0$



Minimiser la fonction de coût $l(\beta) = \frac{1}{N} (\vec{t} - \tilde{\underline{X}}\beta)^T (\quad)$

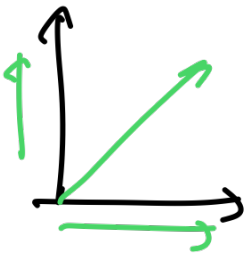
$\beta_0 \vec{c}_0 + \beta_1 c_1 + \cdots$

$\vec{\beta}$ → meilleure combinaison linéaire de colonnes de $\tilde{\tilde{X}}$ au sens des moindres carrés telle que $\vec{t} \simeq \tilde{\tilde{X}}\vec{\beta}$

$\beta = (\tilde{\tilde{X}}^T \tilde{\tilde{X}})^{-1} \tilde{\tilde{X}}^T \vec{t}$ → projection orthogonale sur l'espace des colonnes de $\tilde{\tilde{X}}$

projection sur les z → On peut la calculer via

$$\alpha_l = \frac{\langle \vec{t}, z_l \rangle}{\langle z_l, z_l \rangle}$$



Plusieurs solutions pour améliorer le conditionnement

Best subset selection

→ Nbre de sous ensembles de taille k d'un ensemble de caractéristiques de taille m : $\binom{D+1}{m}$

→ taille 2 : $\beta_0\beta_1$, $\beta_0\beta_2$, $\beta_0\beta_3$, ⋮

taille 3 : $\beta_0\beta_1\beta_2$, $\beta_0\beta_1\beta_3$, ⋮, $\beta_0\beta_L\beta_4$

→ Nbre total de modèles à tester :

$$\sum_{m=2}^{D+1} \binom{D+1}{m}$$

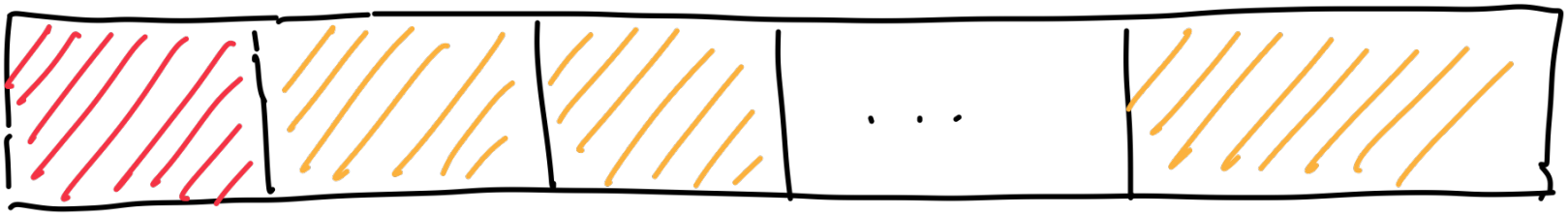
→ Validation croisée

→ ensemble d'entraînement

ensemble de test

→ Le bcp de données → il suffit de séparer l'ensemble complet de données en un ensemble test et un ensemble d'entraînement

→ Si peu de données → K fold cross validation



↳ training

K sous ensemble

test

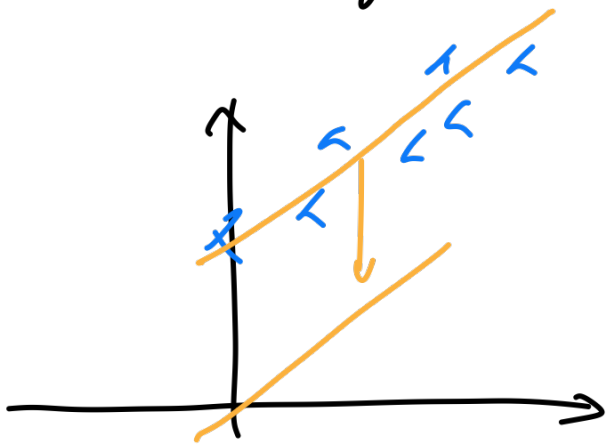
$$CV_k(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_\beta(x^{(i)}) \right)^2$$

n'importe quel sous ensemble des coefficients de regression

donnée du tirage k

↳ Modèle appris sur toutes les données "hors tirage k"

approche #2 → On étend la fonction de coût en ajoutant une pénalité sur les β_j



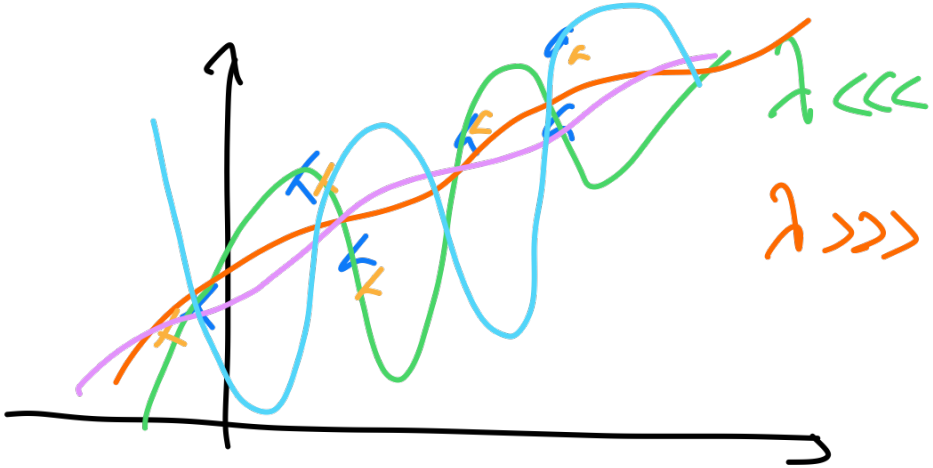
→ Regression Ridge

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

→ Comment calculer β_{RIDGE} ?

fit aux données

Model Complexity


λ <<<
λ >>>