Apprentissage supervisé $\{x^{(i)} \in \mathbb{R}^D, t^{(i)}\}_{i=1}^{N}$

→ Objectif : apprendre un modèle h_β

→ 1ère solution $\underset{OLS}{\ell(\beta)} = \frac{1}{N} \sum_{i=1}^{N} (t^{(i)} - h_\beta(x^{(i)}))^2$

$h_\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_D x_D$

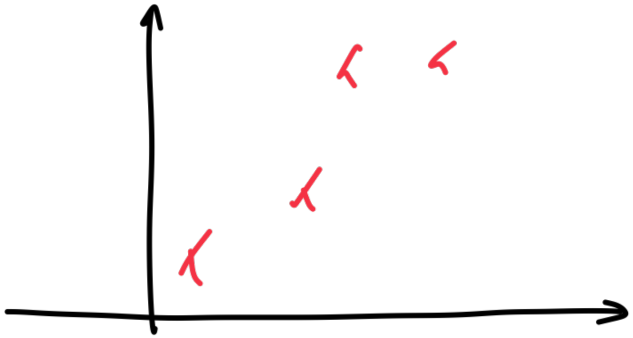
↳ 2 approches : descente de gradient

→ Equations normales : $\beta = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T t$

→ Dans le cas de caractéristiques corrélées / matrice $\tilde{X}$ singulière → on peut toujours coder un processus d'orthogonalisation

$$\beta_D = \frac{\langle t_\varepsilon , z_D \rangle}{\langle z_D , z_D \rangle}$$

→ Dans le cas de corrélation → amplitude élevée des β_j



→ Pour déterminer la complexité optimale du modèle

↳ On peut étudier l'erreur quadratique moyenne

$$MSE(x) = \text{bias}^2 + \text{Variance}$$

Sélection automatique de modèles / caractéristiques

→ Best subset selection

→ Regression Ridge

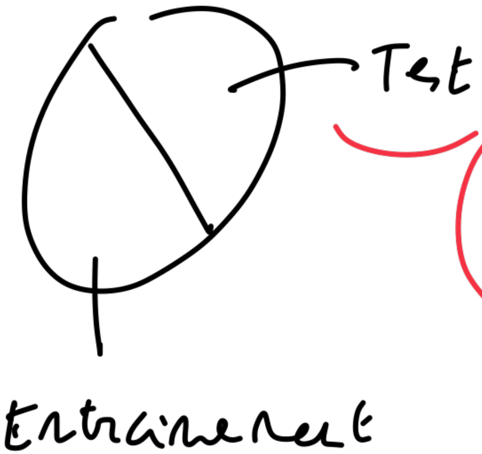
→ Regression LASSO

→ Classification → 2 classes via moindres carrés

→ >2 classes

→ regression logistique

Question : Comment sélectionner les caractéristiques les plus représentatives



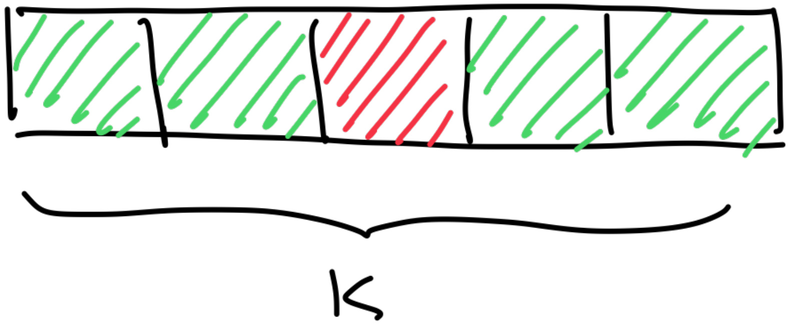
$x^{(i)} \in \mathbb{R}^D$

$$\sum_{k=1}^{D} \binom{D}{k}$$

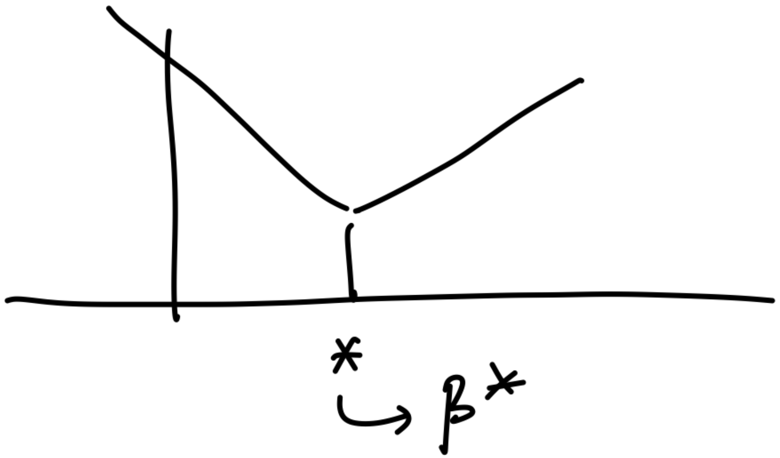
$$(x_1, x_2, x_3, \dots, x_D)$$

→ Si peu de données → validation croisée (cross validation)

K-fold cross validation



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

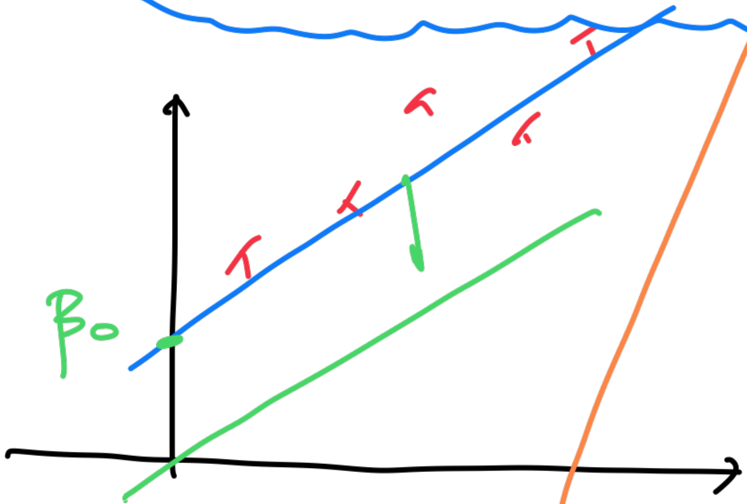


$$\ell(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - h_\beta(x^{(i)})\right)^2 + \lambda \sum_{j=1}^{D} |\beta_j|^2$$

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

RIDGE

pénalité complexité



Fidélité aux données

2 possibilités

→ descente du gradient

→ Résolution des équations normales

Equations normales

Étape 1 centrer les vecteurs $x^{(i)} \in \mathbb{R}^D$

$$\boxed{x^{(i)} \leftarrow x^{(i)} - \frac{1}{N}\sum_{i=1}^{N} x^{(i)}}$$

$$\frac{1}{N}\sum x_j^{(i)} = 0$$

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N} \vec{x}^{(i)} = 0}$$

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

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

$$= -\underbrace{\frac{2}{N}\sum_{i=1}^{N} t^{(i)}} + \underbrace{\frac{2}{N}\sum_{i=1}^{N}} \beta_0 + \underbrace{\frac{2}{N}\sum_{i=1}^{N}\cancel{(\beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})}}_{0}$$

$$2\beta_0 = \frac{2}{N}\sum t^{(i)} \rightarrow \beta_0 = \frac{1}{N}\sum_{i=1}^{N} t^{(i)}$$

Étape 2 : centrer le $t^{(i)}$ $\qquad t_c^{(i)} = t^{(i)} - \frac{1}{N}\sum_{i=1}^{N} t^{(i)}$

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

$$\frac{\partial l}{\partial \beta_j} = \frac{2}{N}\sum_{i=1}^{N}\left(\underbrace{t_c^{(i)} - (\beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})}_{e_i}\right)(-x_j^{(i)}) + 2\lambda\beta_j$$

$$= -\frac{2}{N}\sum_{i=1}^{N} e_i \cdot x_j^{(i)} + 2\lambda\beta_j \leftarrow$$

$$X = \begin{bmatrix} — & x^{(1)} & — \\ — & x^{(2)} & — \\ & \vdots & \\ — & x^{(N)} & — \end{bmatrix}$$

$$\operatorname{grad}_\beta = \left[\frac{\partial l}{\partial \beta_1}, \frac{\partial l}{\partial \beta_2}, \dots, \frac{\partial l}{\partial \beta_D}\right] =$$

$$= -\frac{2}{N}\sum_{i=1}^{N} e_i\, \vec{x}^{(i)} + 2\lambda\vec{\beta}$$

$$e_i = \vec{t}_i - (\underline{\underline{X}}\vec{\beta})_i$$

$$= \left(-\frac{2}{N}(t - \underline{\underline{X}}\beta)^T X\right)^T + 2\lambda\beta \leftarrow$$

$$\vec{t} = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix}$$

$$= -\frac{2}{N} X^T(t - X\beta) + 2\lambda\beta \leftarrow$$

$$= -\frac{2}{N} X^T t + \frac{2}{N} X^T X\beta + 2\lambda\beta$$

$$\Rightarrow \left(\frac{2}{N} X^T X + 2\lambda I\right)\beta = \frac{2}{N} X^T t$$

$$\sum e_i \vec{x}^{(i)} + 2\lambda\beta$$

$$\sum_i e_i \left[\overrightarrow{x^{(i)}} \right]$$

$$\begin{bmatrix} — x^{(1)} — \\ \vdots \\ — x^{(N)} — \end{bmatrix}$$

$$\beta_{RIDGE} = \left(\frac{2}{N} X^T X + 2\lambda I\right)^{-1} \frac{2}{N} X^T t$$

$$X^T X v = \alpha v$$

$$(\underbrace{X^T X + \lambda}_{} I) v = \alpha' v$$

$$X^T X v + \lambda v = \alpha v + \lambda v$$

$$= (\alpha + \lambda) v$$

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

formulations contraintes

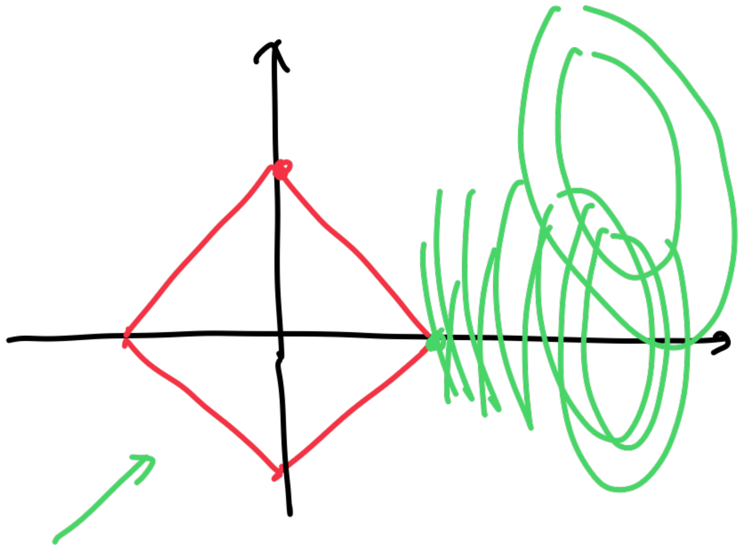
$$\min_{\beta} \ \ell(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\right)\right)^2$$

$$\|v\|_1 = \sum_{i=1}^{N} |v_i|$$

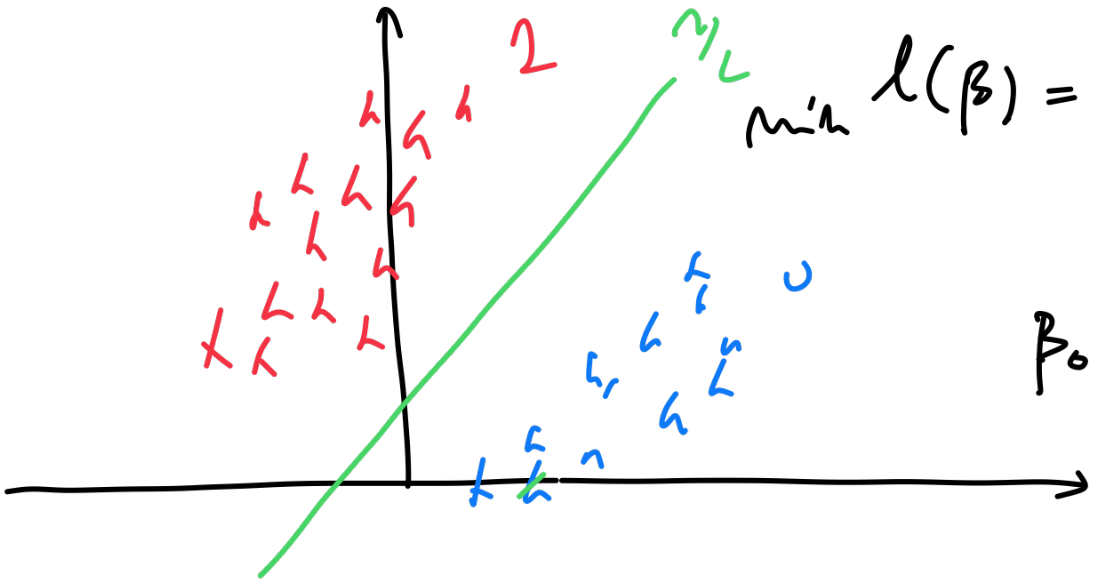
s.t

$$\sum_{j=1}^{D} \beta_j^2 \leq t \qquad\qquad \sum_{j=1}^{D} |\beta_j| \leq t$$

(Ridge) (LASSO)



<u>Classification</u> $\{x^{(i)}, t^{(i)}\}$ $t^{(i)} \in \{1, \dots, K\}$



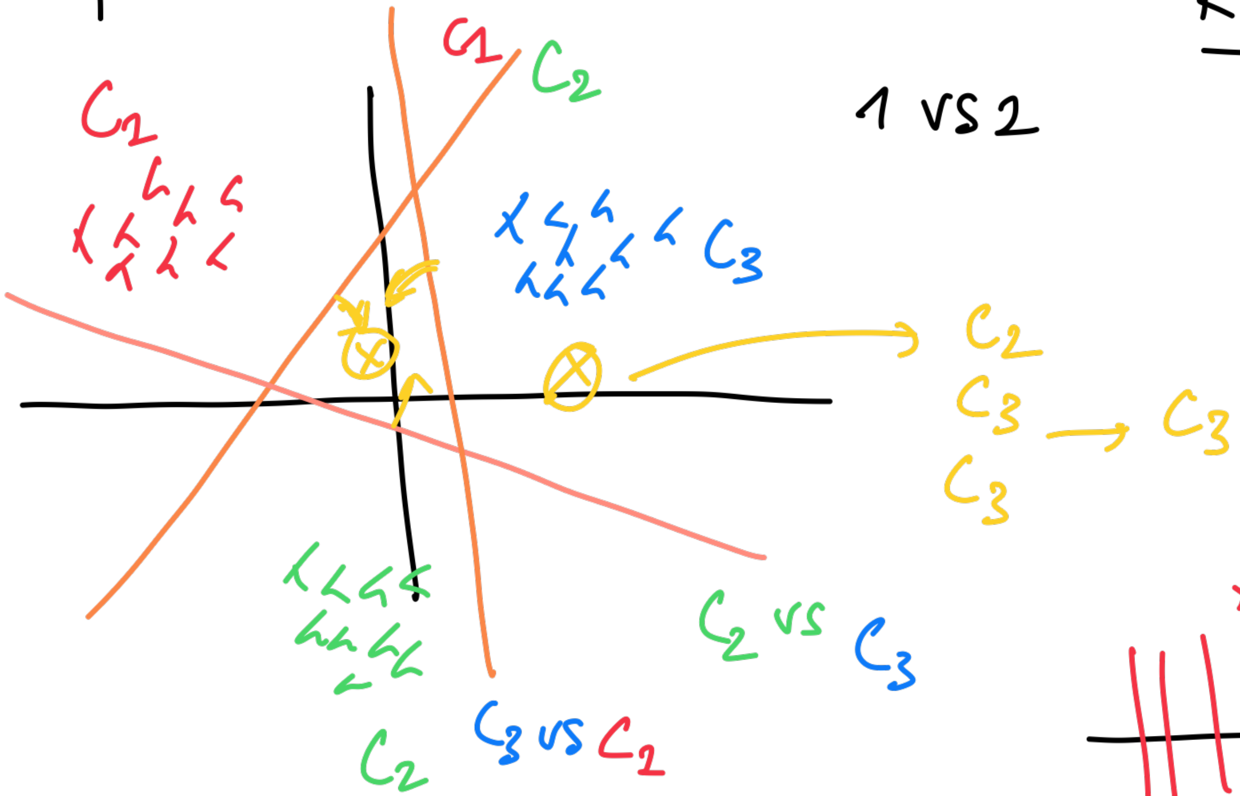
$$\min \ell(\beta) = \frac{1}{N} \sum_{i=1}^{N} \Big(\overset{0}{\underset{1}{t^{(i)}}} - \big(\beta_0 + \beta_1 x_1^{(i)} + \dots \beta_D x_D^{(i)}\big)\Big)^2$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

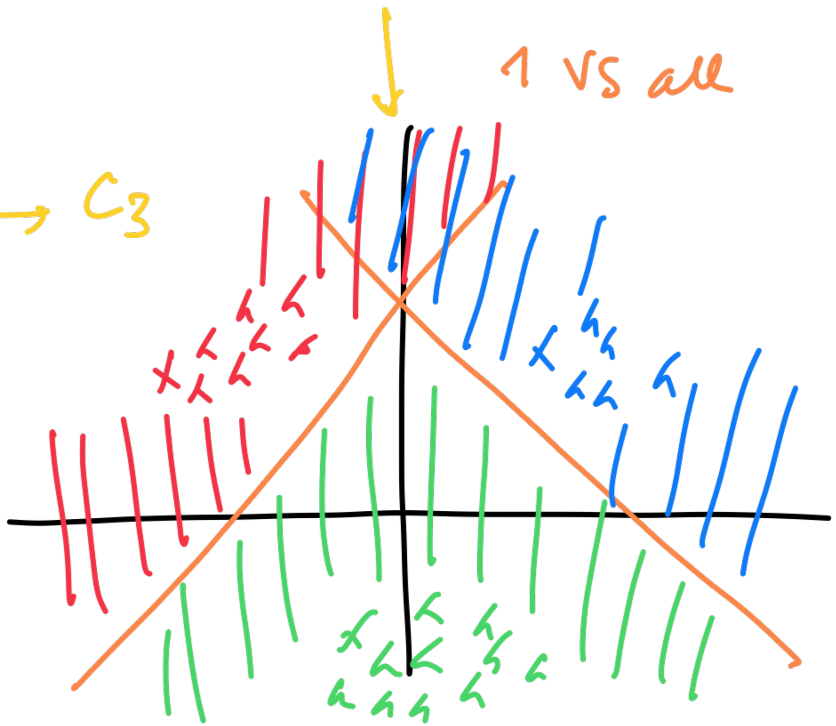
$$x_2 = \frac{-\beta_1 x_1 - \beta_0}{\beta_2}$$

Q1? 2 classes → classe multiples

Q2? points autres



$$\frac{K.K-1}{2}$$



$$T = \begin{bmatrix} — \vec{t}^{(1)} — \\ \vdots \\ — \vec{t}^{(N)} — \end{bmatrix}$$

$$t^{(i)} \in \{1, \dots, k\}$$

$$\vec{t}^{(i)} = [0, 1, 0 \text{ — } 0]$$

$$B = \begin{bmatrix} — \vec{\beta}_1^T — \\ \vdots \\ — \vec{\beta}_k^T — \end{bmatrix} \begin{bmatrix} | & & & \\ x^{(1)} & x^{(2)} & \cdots & x^{(N)} \\ | & & & \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & | \\ \vdots & | \\ 0 & 0 \end{bmatrix}$$

$$T \simeq X B^T$$

$$B X^T$$

$$\min_{B} \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{k} \left(t_k^{(i)} - (\vec{\beta}_k)^T x^{(i)} \right)^2$$

$$B = (X^T X)^{-2} X^T T$$