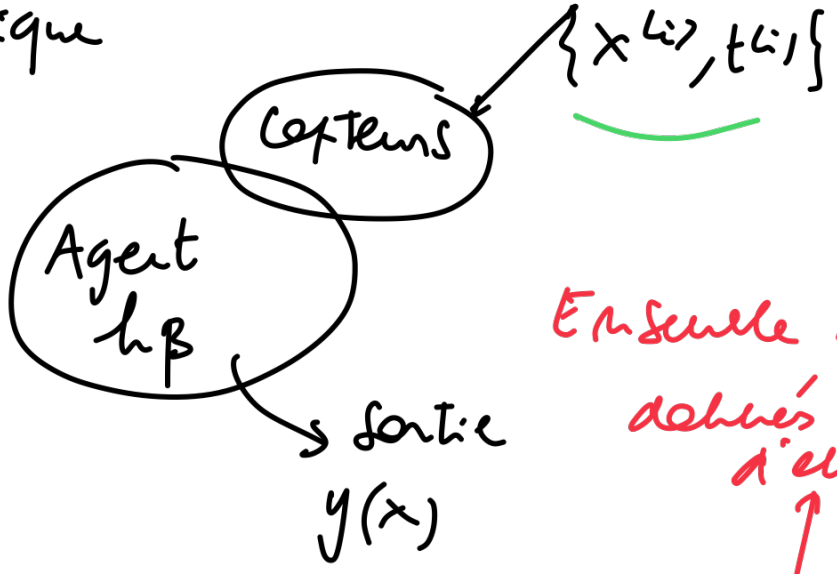
Intelligence Artificielle

- Apprentissage
  - Supervisé
  - non-supervisé
- Raisonnement logique
- Renforcement



Apprentissage Supervisé

Ensemble des données d'entraînement

agent a accès à un ensemble de données $\left\{x^{(i)}, t^{(i)}\right\}_{i=1}^{N}$

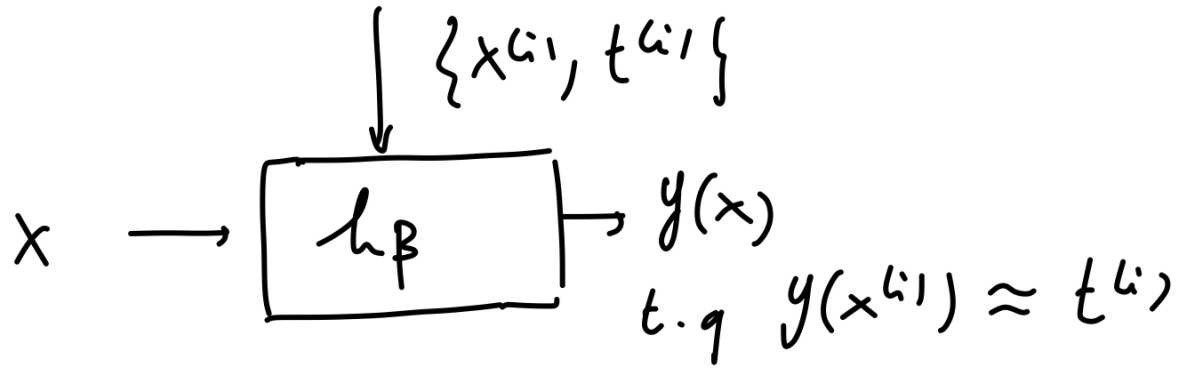
observables $\vec{x}^{(i)} \in \mathbb{R}^D$ $\quad x^{(i)} = (x_1, x_2, x_3, \cdots, x_D)$

prédiction à produire à produire à partir de observations

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

$\vec{x}^{(i)}$ = vecteur caractéristique $x_j^{(i)}$ → caractéristiques

Objectif : apprendre un modèle : h_β

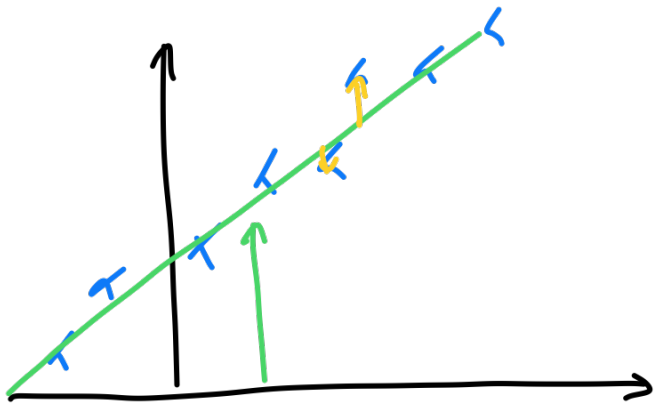


X : espace des données d'entrée

Y : espace de sortie

Une première approche : choisir $h_\beta(x)$ comme une combinaison linéaire

$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_D x_D \qquad \vec{\beta} \in \mathbb{R}^{D+1}$$



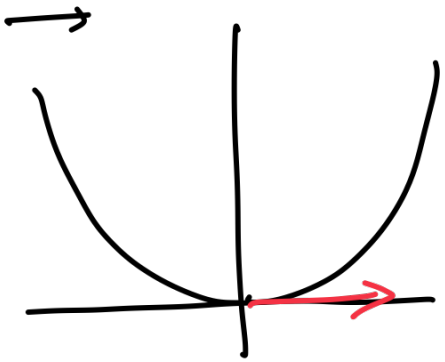
Question : comment déterminer les coefficients β_j de façon à ce que

$$h_\beta(\vec{x}^{(i)}) \cong t^{(i)}$$

$$\rightarrow J(\vec{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_\beta(x^{(i)}) \right)^2 \quad \text{(fonction de coût / loss)}$$

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

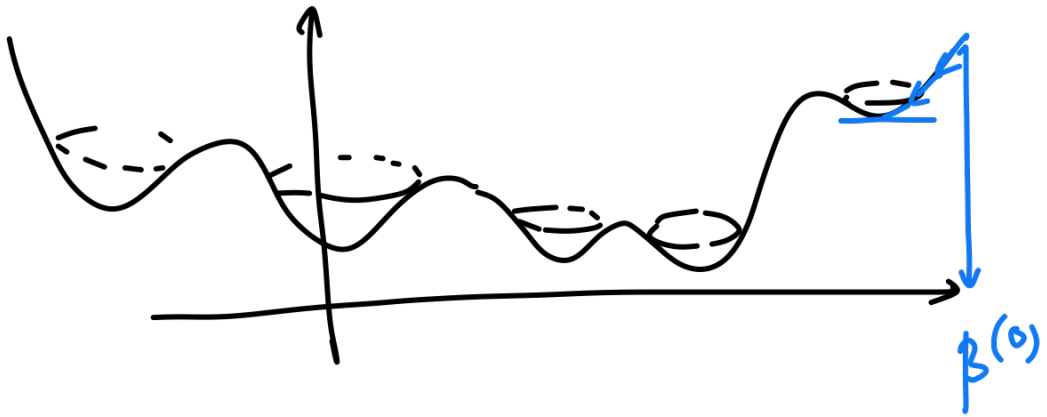
→ Pour trouver les β_j on minimise la fonction de coût $l(\beta)$ par rapport à $\vec{\beta}$

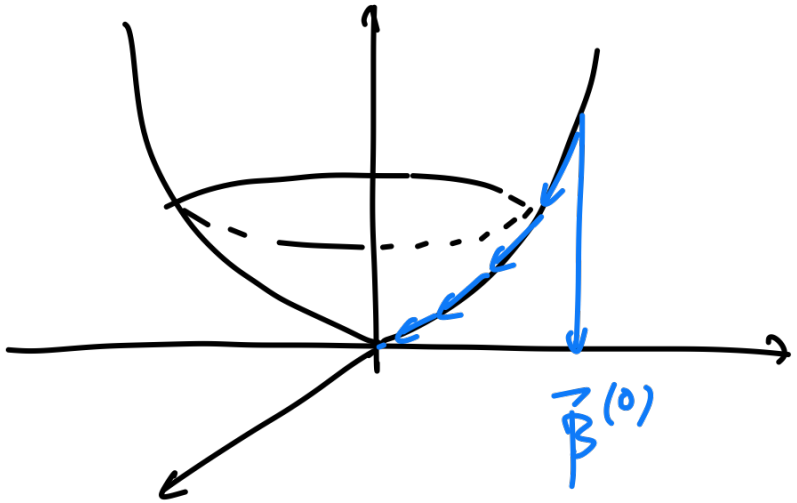


$f(x) = x^2$

$f'(x) = 2x$

direction de plus forte décroissance de la fonction





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

$$\operatorname*{grad}_{\beta} \mathcal{L}(\beta) = \left[\frac{\partial \mathcal{L}}{\partial \beta_0}, \dots, \frac{\partial \mathcal{L}}{\partial \beta_D} \right]$$

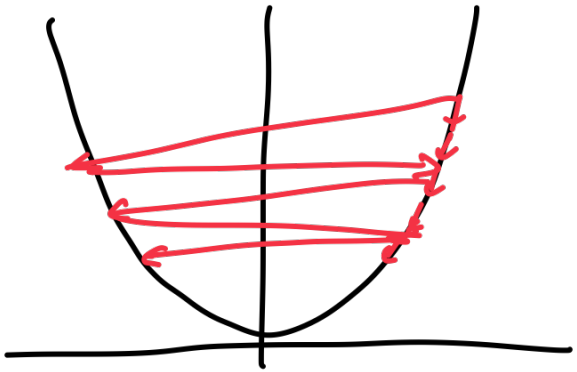
$$\frac{d}{dx} f(x)^2 = 2 f(x) \cdot f'(x)$$

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

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \frac{1}{N} \sum_{i=1}^{N} 2 \left(\underbrace{t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})}_{\varepsilon^{(i)}} \right) \cdot (-x_j^{(i)})$$

$j \neq 0$

#1 Descente de gradient (BATCH)



Soit $\beta^{(0)}$ un itéré initial

for k in np.arange(maxIter):

$$\begin{cases} \beta_0 \leftarrow \beta_0 - \eta \dfrac{\partial \ell}{\partial \beta_0} \\ \beta_j \leftarrow \beta_j - \eta \dfrac{\partial \ell}{\partial \beta_j} \end{cases}$$

taux d'apprentissage ("learning rate")

Descente de gradient stochastique (SGD)

→ for iter in maxIter

→ shuffle $\{x^{(i)}, t^{(i)}\}$

→ for i in $\{x^{(i)}, t^{(i)}\}$

$$\beta \leftarrow \beta - \eta \underset{\beta}{\text{grad}} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2$$

Descente de type MINI BATCH

(à chaque itération on échantillonne un sous ensemble de données de taille $m < N$)

#2 Annulation de la dérivée (Résolution des equations normales)

$$\rightarrow \vec{t} = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} \qquad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_D \end{bmatrix} \rightarrow$$

Paramètres du modèle / coefficients de regression

$$\tilde{x} = [1, x_1, x_2, \dots, x_D]$$
$$= [1, \vec{x}]$$

$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_D x_D = [\beta_0 \cdots \beta_D] \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix} \Big\} D$$
$$= \vec{\beta}^T \tilde{x}$$

$$\underline{\underline{\tilde{X}}} = \begin{bmatrix} 1 & \text{---} & \vec{x}^{(1)} & \text{---} \\ \vdots & & \vdots & \\ 1 & \text{---} & \vec{x}^{(N)} & \text{---} \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_D^{(1)} \\ \vdots & & & & \\ 1 & x_1^{(N)} & x_2^{(N)} & \cdots & x_D^{(N)} \end{bmatrix}$$

vecteur d'erreur $\vec{\varepsilon} = \vec{t} - \underline{\underline{\tilde{X}}}\,\vec{\beta}$

$$\begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(N)} \end{bmatrix} = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} - \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_D^{(1)} \\ \vdots & \vdots & & \\ 1 & x_1^{(N)} & \cdots & x_D^{(N)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}$$

$$\begin{bmatrix} \varepsilon^{(1)} \\ \vdots \\ \varepsilon^{(N)} \end{bmatrix} = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} - \begin{bmatrix} \beta_0 + \beta_1 x_1^{(1)} + \beta_2 x_2^{(1)} + \dots + \beta_D x_D^{(1)} \\ \vdots \\ \beta_0 + \beta_1 x_1^{(N)} + \beta_2 x_2^{(N)} + \dots + \beta_D x_D^{(N)} \end{bmatrix}$$

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

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

$$\rightarrow ? \underset{\beta}{\text{gradient}}\ \ell(\beta) = 0$$

$$l(\beta) = \frac{1}{N}\left(\vec{t}^{\,T}\vec{t} - \vec{\beta}^{\,T}\tilde{\tilde{X}}^{T}\vec{t} - \vec{t}^{\,T}\tilde{\tilde{X}}\beta + \vec{\beta}^{\,T}\tilde{\tilde{X}}^{T}\tilde{\tilde{X}}\vec{\beta}\right)$$

$$\frac{\partial l}{\partial \beta}$$

0

$\beta^T v \qquad -2\,\tilde{\tilde{X}}^T\vec{t}$

$\frac{\partial}{\partial \beta}\left(2\,\tilde{\tilde{X}}^T\tilde{\tilde{X}}\beta\right)$

$$\frac{\partial}{\partial \beta}\left(\beta_0 v_0 + \beta_1 v_1 + \cdots \beta_\Delta v_\Delta\right) \rightarrow [v_0, v_1, \ldots, v_\Delta]$$

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

$$\beta^T \tilde{\tilde{X}}^T\tilde{\tilde{X}}\,\vec{\beta} \Rightarrow \frac{2}{\partial \beta}(\beta^T A\,\beta)$$

$$\begin{bmatrix}\beta_1 & \beta_2\end{bmatrix}\begin{bmatrix}A_{11} & A_{12}\\ A_{12} & A_{22}\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} \rightarrow \begin{bmatrix}\beta_1 A_{11} + \beta_2 A_{12} & \beta_1 A_{12} + \beta_2 A_{22}\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$$

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

$$+ \beta_1\beta_2 A_{12} + \beta_2^2 A_{22}$$

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

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

$$= 2\begin{bmatrix}A_{11} & A_{12}\\ A_{12} & A_{22}\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$$

En regroupant chacun des termes on obtient

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

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

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

Equations normales

Vecteur des coefficients du modèle

$$A = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_{D+1} \end{bmatrix} \rightarrow A^{-1} = \begin{bmatrix} \lambda_1^{-2} & & \\ & \ddots & \\ & & \lambda_{D+1}^{-2} \end{bmatrix}$$