Fouille de données

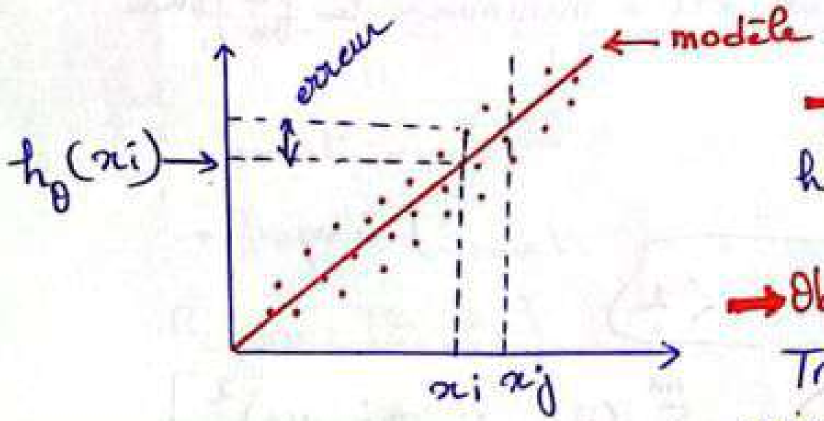
10/09 2023 ①

1 - Méthodes de regression

1.1 Regression linéaire Univariée (avec une seule variable)

- data $\theta = \{(x_1, y_1), \dots, (x_n, y_n)\}$

Exemple de l'agence immobilière on a $\begin{cases} x = \text{surface} \\ y = \text{prix du bien} \end{cases}$



→ hypothèse :

$h_\theta(x) = \theta_0 + \theta_1$ avec $\theta = \{\theta_0, \theta_1\}$ paramètres $\in \mathbb{R}^2$

→ Objectif

Trouver les valeurs θ_0 et θ_1 qui permettent de définir le modèle (droite) qui s'appelle le plus des données "= modélisé"

$$\text{Erreur} = h_\theta(x_i) - y$$

→ Minimiser l'erreur moyenne sur tous les pts de θ.

$$J_\theta(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=0}^{m} \left| (h_\theta(x_i) - y_i) \right|$$

$$\min_{\theta_0 \theta_1} J(\theta_i, \theta_1) = \frac{1}{m} \sum^{m} \dots\dots\dots$$

La fct absolue n'est pas dérivable!
→ c'est pour ça on a ajouté le carré dans la formule de Σ.

$\min f(x) = ?$ $f'(x) = 0$?

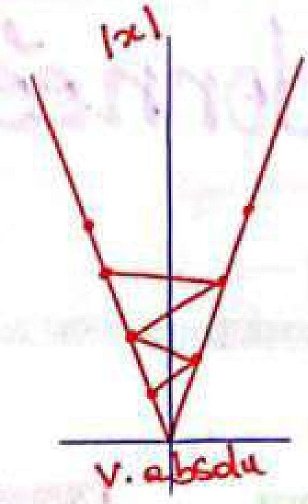
$\min f(x) = x^2$, $f'(x) = 2x = 0$ ssi $\boxed{x = 0}$

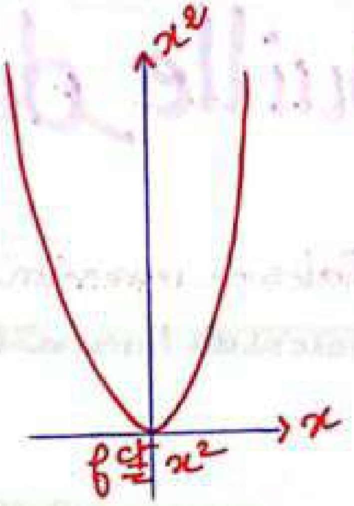
on pose

pt	1	2	3	4	5
erreur	2	-2	3	1	-3

$\Rightarrow \sum_{\text{erreur}} = 1.$

$|h_\theta(x_i) - y_i| = 11.$





L'algorithme du Gradient = algo qui sert à minimiser la f° coût $J(\theta_0, \theta_1)$

REGLE DE MISE À JOUR:

$\theta_i = \theta_i - \alpha \frac{\partial J}{\partial \theta_i}$ avec $\boxed{0 < i < 1}$

on a
$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0} = \frac{\partial}{\partial \theta_0}\left[\frac{1}{2m}\sum_{i=1}^{m}(\theta_0 + \theta_1 x_i - y_i)^2\right]$$
$$= \frac{1}{2m}\sum_{i=1}^{m}\frac{\partial}{\partial \theta_0}(\theta_0 + \boxed{\theta_1 x_i} - y_i)^{\boxed{2}}$$
$$= \frac{1}{\cancel{2}m}\sum_{i=1}^{m}\cancel{\times 2}(\theta_0 + \theta_1 x_i - y_i)\underline{\times 1}$$

$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} = \frac{1}{2m}\sum_{i=1}^{m}\frac{\partial}{\partial \theta_1}(\theta_0 + \theta_1 x_i - y_i)^2$$
$$= \frac{1}{\cancel{2}m}\sum_{i=1}^{m}\cancel{2}(\theta_0 + \theta_1 x_i - y_i)\underline{x_i}$$
$$= \frac{1}{m}\sum_{i=1}^{m}(\theta_0 + \theta_1 x_i - y_i)\underline{x_i}$$

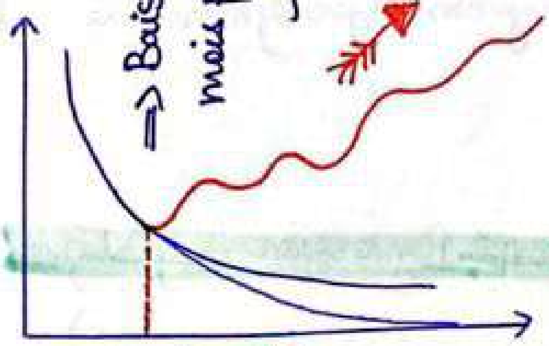
Algorithme de Gradient :

```
init $\theta_0^{old}$  $\theta_1^{old}$ (aléatoire)
$J_{old} \leftarrow J(\theta_0^{old}, \theta_1^{old})$
Répéter
    $\theta_0 \leftarrow \theta_0^{old} - \alpha \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0}$
    $\theta_1 \leftarrow \theta_1^{old} - \alpha \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1}$
    $J \leftarrow J(\theta_0, \theta_1)$ ; $\theta_0^{old} \leftarrow \theta_0$ ; $\theta_1^{old} \leftarrow \theta_1$
    $J_{avant} \leftarrow J_{old}$ ; $J_{old} \leftarrow J$.
Jusqu'à $(J_{avant} - J) \leq \Sigma$
```

hyperparamètre

paramètre

Retrouver (θ_0, θ_1)



⇒ Baisser α, mais pas trop, sinon l'algo de Gradient devient long

on s'arrête là !!

La machine Apprend ?

$\alpha \in \mathbb{R}$: le taux d'Apprentissage

1.2 : Regression lineaire multivariée

Donnéés

$$a_1 \begin{pmatrix} x_1 & x_2 & x_n & y \\ x_{11} & x_{12} & x_{1n} & y_1 \\ & & & \\ \dots & & & y_2 \end{pmatrix} \Big\} \boxed{i=2}$$

$$\theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2 + \theta_3 x_3^2 + \dots + \theta_n x_n^n$$

Hypothèse

$$h_0(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

$$= \theta_0 + \sum_{i+1}^{n} \theta_i x_i \quad \underline{\text{dont}}\ \theta = \{\theta_0, \theta_1, \theta_2, \dots, \theta_n\} \in \mathbb{R}^{n+1}$$

Erreur $= h_\theta(x_i) - y_j$

coût $= J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$ est le $i^{ème}$ élement de l'instance i

Calcul des derivés

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h\theta(x_i)-y_i)\times 1$$

$$\frac{\partial J(\theta)}{\partial \theta_i} = \frac{1}{m}.$$

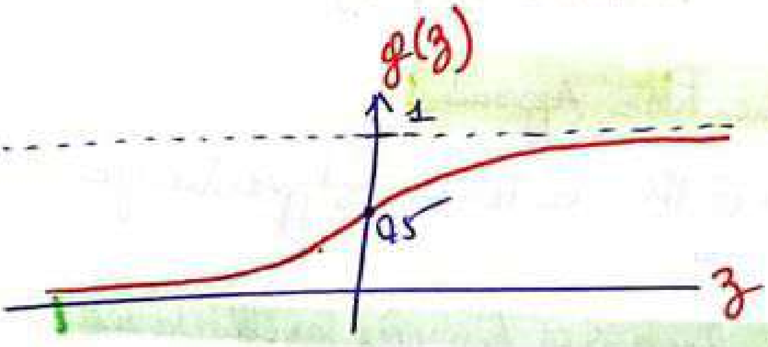
$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$

→ Choisir l'hyperparamètres (dont dépend la méthode)

1.3. Sur/Sous Apprentissage

Regression polynomiale $h\theta(x) = \theta_0 + \sum_{i=1}^{p}\theta_i x^p$

1.4. Regression Logistique



pb : spam (mails)

$$\text{sigmoid}(z) = \frac{1}{1-e^{-z}}$$

F^{ct} coût

$$h\theta(x) = \underbrace{g(\underbrace{\theta_0 + \sum_{i=1}^{m}\theta_i x_i}_{\in \mathbb{R}})}_{\in]0,1[}$$

$$\cos(h_\theta(x_i), y_i) = \frac{1}{2}(h_\theta(x_i) - y_i)^2$$

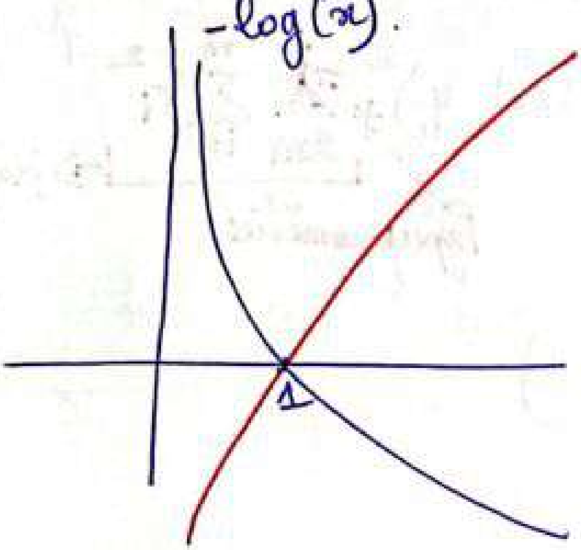
$$\cos(h\theta(x_i, y_i)) = \frac{1}{2}(h\theta(x_i) - y_i)^2 = \cos(h\theta(x_i, y_i))$$

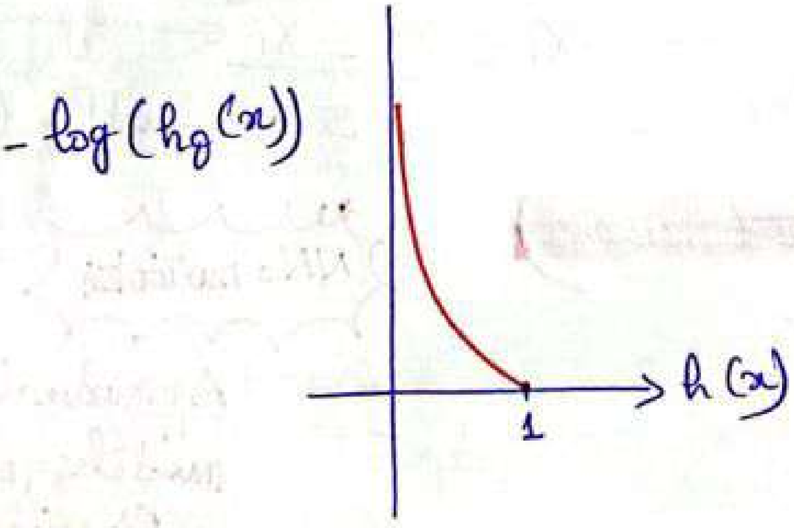
$$= -\log(h_\theta(x)) \quad \boxed{\text{si } y_i = 1}$$

$\cos(h_\theta(x_i, y_i)) = -\log(1 - h_\theta(x))$

$\boxed{\text{si } y_i = 0}$

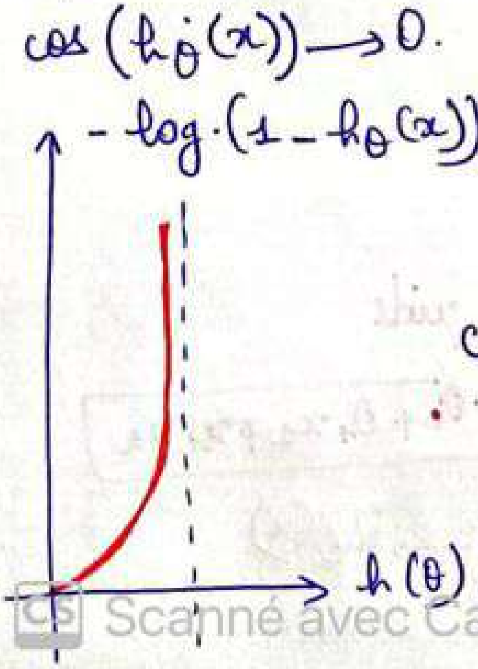
$f(x) = -\log(1 - x) \in x\ ]0,1[$





$h_\theta(x) \to 1$

$\cos(h_\theta(x)) \to 0.$



$$\cos(h_\theta(x_i), y_i) = \begin{cases} -\log(h_\theta(x) \text{ si } y_i = 1 \\ -\log(1 - h_\theta(x)) \text{ si } y_i = 0 \end{cases}$$

$\text{cost}(h_\theta(x_i), y_i)$

$$= -y_i \underbrace{\log(h_\theta(x_i))}_{y_i = 1} - (1 - y_i) \underbrace{\log(1 - h_\theta(x))}_{y_i = 0}$$

$$\mathcal{F}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \cos\left(h_\theta(x_i), y_i\right)$$

7. Régularisation

Situation : $J(\theta) \sim 0$, mais le modèle ne généralise pas bien (mauvaises performances)

$$\Rightarrow \text{coût} = J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x_i) - y_i\right)^2 + \underbrace{\frac{\lambda}{2m} \sum_{i=1}^{m} \theta_i^2}_{\Rightarrow \text{Régularisation}}$$

Hyperparamètres

8. Feature Scaling (Normalisation)

Data $\{x_1, x_2, \dots, x_n\}$

$$X_i = \frac{X_i - \min(X_i)}{\max(X) - \min(X_i)} \quad \text{ou} \quad X_i = \frac{X_i - \overline{X_i}}{\sigma}$$

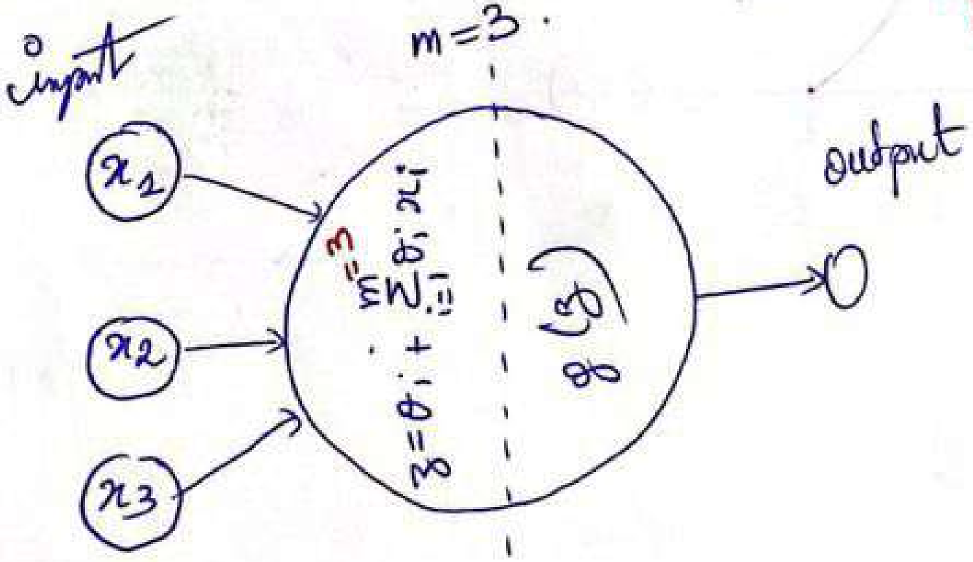
$\overline{X_i} \leftarrow$ moyenne(X_i)

$\sigma \leftarrow$ ecartType(X_i)

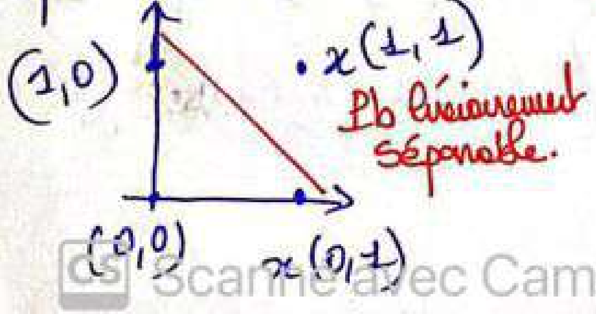
Perception (Approche Réseau de neurones)

NN : motivation

Apprendre des modèles non linéaires



Opérateurs logiques



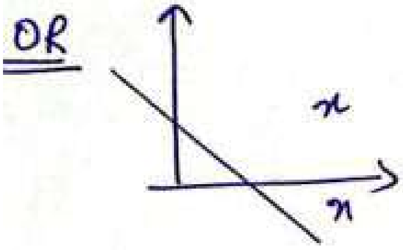
$\{-30, 20, 20\}$: poids

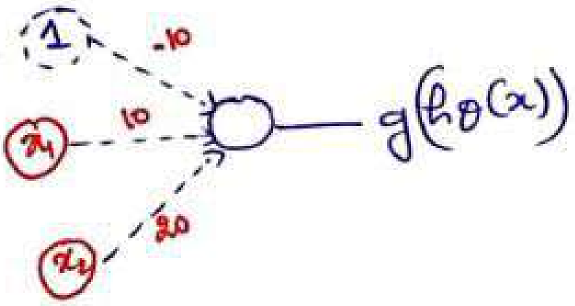
$\boxed{\theta_0 + \theta_1 x_1 + \theta_2 x_2}$

x_1	x_2	$h_\theta(x)$
0	0	$g(-30) \sim 0$
0	1	$g(-10) \sim 0$
1	0	$g(-10) \sim 0$
1	1	$g(10)$

$$\underset{\substack{\uparrow \\ -30}}{\theta_0} + \theta_1 \underset{\substack{\uparrow \\ 20}}{x_1} + \theta_2 \underset{\substack{\uparrow \\ 20}}{x_2}$$

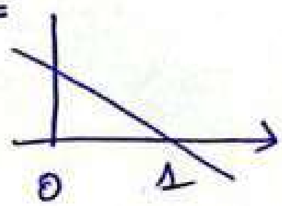
OR

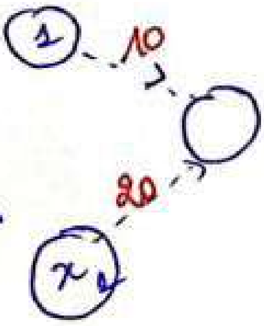




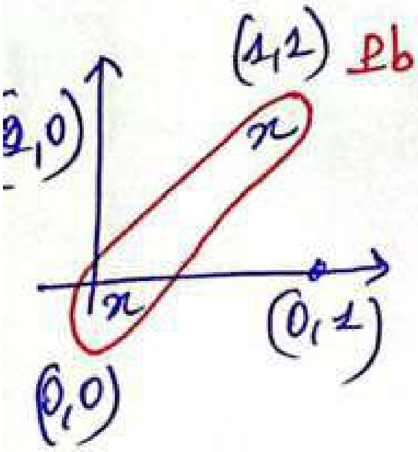
0	0	0
1	0	1
0	1	1
1	1	1

NOT





x_1	$h_\theta(x)$
0	1
1	0



Pb non linéairement séparable.

0	0	1
0	1	0
0	1	0
1	1	1