

# Lesson 5 : (Non-regularized) Regression

Motivation : regressor

Features

Solution

to train the model

#	Living area [ $m^2$ ] $X_1$	# Room [-] $X_2$	Y Price [1000 Euros] $Y_1$
1	45	3	200
2	60	2	300
⋮	⋮	⋮	⋮
70			
71			
⋮			
100	90	4	1000

Training  
to find the optimal values for the model parameters.  
(Validation)  
For simplicity

Test  
to test the optimality of the model with test data (which were not used for training)

$N \triangleq$  total # of features

$J \triangleq$  Dimension of output variable

$I \triangleq$  total # of training examples

$X_{\text{training}}$

$Y_{\text{training}}$

$X_{\text{test}}$

$Y_{\text{test}}$

$$X_{\text{training}} = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \vdots \\ (x^{(i)})^T \\ \vdots \\ (x^{(I)})^T \end{bmatrix}$$

transposed

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_N^{(i)} \end{bmatrix}$$

Ex.: 2

$$(x^{(i)})^T = [x_1^{(i)} \ x_2^{(i)} \ \dots \ x_N^{(i)}]$$

Ex:

$$X_{\text{training}} = \begin{bmatrix} 1 & 45 & 3 \\ 2 & 60 & 2 \\ \vdots & \vdots & \vdots \\ I = 70 & x & x \end{bmatrix} \quad (x^{(2)})^T$$

$I = 70$

$J = 1$

$$Y_{\text{training}} = \begin{bmatrix} 1 & 200 \\ 2 & 300 \\ \vdots & \vdots \\ I = 70 & x \end{bmatrix}$$

$I = 70$

Idem for  $X_{\text{test}}$

$$Y_{\text{training}} = \begin{bmatrix} (y^{(1)})^T \\ (y^{(2)})^T \\ \vdots \\ (y^{(i)})^T \\ \vdots \\ (y^{(I)})^T \end{bmatrix}$$

$$y^{(i)} = \begin{bmatrix} y_1^{(i)} \\ y_2^{(i)} \\ \vdots \\ y_J^{(i)} \end{bmatrix}$$

Ex.: 1

$$(y^{(i)})^T = [y_1^{(i)} \ y_2^{(i)} \ \dots \ y_J^{(i)}]$$

Idem for  $Y_{\text{test}}$

$$X_{transposed} = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \vdots \\ (x^{(n)})^T \end{bmatrix} \quad \text{transposed} \quad x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_N^{(i)} \end{bmatrix} \quad (x^{(i)})^T = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & \dots & x_N^{(i)} \end{bmatrix}$$

Regression pb = problem

Find the hyperplane that optimally represents the relation between some given data and their solutions.

similar  $\swarrow$  mathematically

Hypothesis f or = function.

an example of features

$$Y \sim \hat{y}^{(i)} = h_{\theta}(x^{(i)})$$

Actual output variable

Estimated output variable

model parameters

$$x^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \\ \vdots \\ x_N^{(i)} \end{bmatrix} \in \mathbb{R}^{N+1}$$

$$y^{(i)} = \begin{bmatrix} y_1^{(i)} \\ \vdots \\ y_N^{(i)} \end{bmatrix}$$

variable  $\neq$  parameters

model parameters  $\theta$   
hyper parameters

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_N \end{bmatrix}$$

Intercept or Bias

For linear regression

$$\hat{y}^{(i)} = h_{\theta}(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \dots + \theta_N x_N^{(i)}$$

(Intercept term)

Aside comment :

Linear hypothesis  $\hat{y}^{(i)} = h_{\theta}(x^{(i)}) = \theta^T x^{(i)}$

(logistic regression) sigmoid hypothesis  $\hat{y}^{(i)} = h_{\theta}(x^{(i)}) = \frac{1}{1 + e^{-\theta^T x^{(i)}}}$

Regression (pb) (mathematically defined) :

Ex: 
$$z = \begin{matrix} \text{Index} & \text{value} \\ \begin{matrix} 0 \\ 1 \\ 2 \end{matrix} & \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \end{matrix}$$

$\min z = 1$   
 $\text{argmin } z = 2$

How?

$$\left\{ \begin{array}{l} \min_{\theta} E(x, y, \theta) \Rightarrow E_{\min} \\ \theta^* = \underset{\theta}{\text{argmin}} E(x, y, \theta) \end{array} \right.$$

Def: Error  $\hat{y}^{(i)} = h_{\theta}(x^{(i)})$  :  $E \triangleq$  Sum of squared Errors (SSE)  $= \frac{1}{2} \sum_{i=1}^I (\hat{y}^{(i)} - y^{(i)})^2 = \frac{1}{2} \sum_{i=1}^I (h_{\theta}(x^{(i)}) - y^{(i)})^2$

$\uparrow$   
in this course

if  $I=1$   $\rightarrow$   $\#$   $\text{MSE: Mean Square Error}$

Approaches to solve the regression pb. ( $\min_{\theta} E(X, y, \theta)$ )

$$\theta^{(itera)} = \begin{bmatrix} \theta_0^{(itera)} \\ \theta_n^{(itera)} \\ \theta_N^{(itera)} \end{bmatrix}$$

$$\nabla_{\theta} E = \begin{bmatrix} \frac{\partial E}{\partial \theta_0} \\ \frac{\partial E}{\partial \theta_n} \\ \frac{\partial E}{\partial \theta_N} \end{bmatrix}$$

- 1) Batch Gradient Descent (BGD)  
2) Stochastic Gradient Descent (SGD)  
3) closed-form solution (CFS).

1) Batch Gradient Descent (BGD)

"pack"  
"group"

"partial  
derivatives"

Pseudo-code (vectorial version)

Initialize  $\theta$  randomly  $\rightarrow \theta^{(0)}$   
For  $itera = 1, 2, \dots, itera_{max}$ :

$$\theta^{(itera)} \leftarrow \theta^{(itera-1)} - \alpha \nabla_{\theta} E$$

learning rate

iteration  
epoch  
episode

Gradient of  $E$   
w.r.t.  $\theta$

with small values  
(0.001 or 0.0001)

index for intercept

Pseudo-code (scalar version)

Initialize  $\theta$  randomly  $\rightarrow \theta^{(0)}$

For  $itera = 1, 2, \dots, itera_{max}$

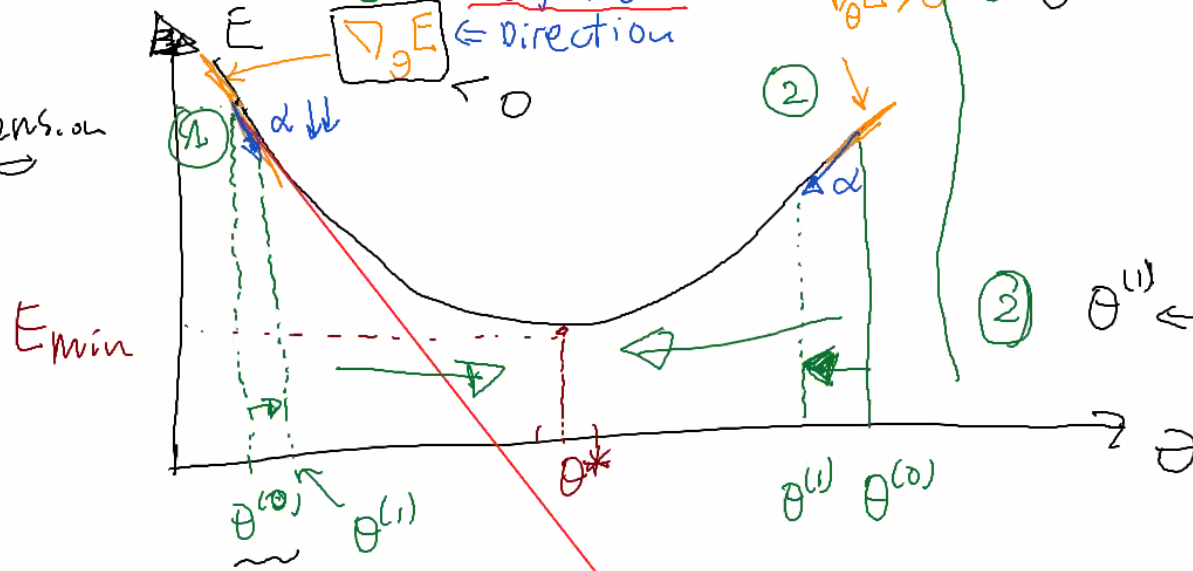
For  $n = 0, 1, \dots, N$ :

$$\theta_n^{(itera)} \leftarrow \theta_n^{(itera-1)} - \alpha \frac{\partial E}{\partial \theta_n^{(itera-1)}}$$

$$\theta^{(itera)} \leftarrow \theta^{(itera-1)} - \alpha \nabla_{\theta} E$$

Example:  
 $\theta$ : 1 dimension

$\nabla_{\theta} E$   $\leftarrow$  direction  
 $\alpha$   $\leftarrow$  magnitude



$$\textcircled{1} \theta^{(1)} \leftarrow \theta^{(0)} - \alpha \nabla_{\theta} E$$

$\nabla_{\theta} E > 0$

$$\textcircled{2} \theta^{(1)} \leftarrow \theta^{(0)} - \alpha \nabla_{\theta} E$$

$\nabla_{\theta} E < 0$

it can cause  
 $\Rightarrow$  oscillations or instability

BC Intelligence artificielle (OA)  
 vous pouvez remettre le tableau  
 d'avant 2sec svp

Répondre



$$\boxed{\frac{\partial E}{\partial \theta_n}} = \frac{\partial}{\partial \theta_n} \left[ \frac{1}{2} \sum_{i=1}^I (\hat{y}^{(i)} - y^{(i)})^2 \right] = \frac{\partial}{\partial \theta_n} \left[ \frac{1}{2} \sum_{i=1}^I (\underbrace{h_{\theta}(x^{(i)})}_{\theta^T \bar{x}^{(i)}} - y^{(i)})^2 \right]$$

$$\boxed{\text{if } j=1}$$

$$= \frac{\partial}{\partial \theta_n} \left[ \frac{1}{2} \sum_{i=1}^I (\theta^T \bar{x}^{(i)} - y^{(i)})^2 \right]$$

In general,

$$E = \frac{1}{2} \sum_{i=1}^I \|\hat{y}^{(i)} - y^{(i)}\|_2^2$$

$$= \frac{1}{2} \sum_{i=1}^I \sum_{j=1}^J (\hat{y}_j^{(i)} - y_j^{(i)})^2$$

$$= \frac{1}{2} \sum_{i=1}^I \frac{\partial}{\partial \theta_n} \left[ (\theta^T \bar{x}^{(i)} - y^{(i)})^2 \right] \frac{\partial (\theta^T \bar{x}^{(i)})}{\partial \theta_n}$$

$$= \sum_{i=1}^I (h_{\theta}(x^{(i)}) - y^{(i)}) x_n^{(i)}$$

Alternative:  $\frac{\partial}{\partial \theta_n} (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \dots + \theta_n x_n^{(i)} + \dots + \theta_N x_N^{(i)})$

Annotations:  $\downarrow 0$ ,  $\downarrow 0$ ,  $\downarrow x_n^{(i)}$ ,  $\downarrow 0$

- Pseudo-code (scalar version) :

$$\theta_n^{(iter+1)} \leftarrow \theta_n^{(iter-1)} - \alpha \left( \sum_{i=1}^I (h_{\theta}(x^{(i)}) - y^{(i)}) x_n^{(i)} \right)$$

Batch!  $\leftarrow$

Mini Batch

Stochastic choice

## 2) Stochastic Gradient Descent (SGD)

pseudo-code (scalar version)

Initialize  $\theta$  randomly  $\rightarrow \theta^{(0)}$

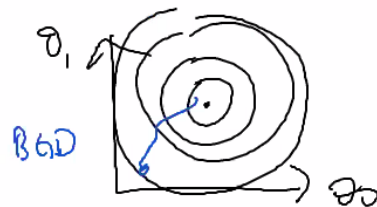
For  $itera = 1, 2, \dots, Itera\_max$ :

For  $n = 0, 1, \dots, N$ :

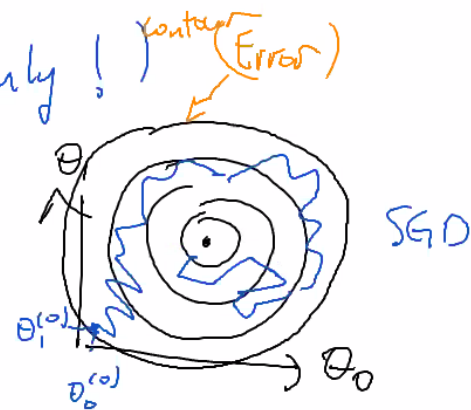
$$i \sim U[1, Itera\_max]$$

$$\theta_n^{(itera)} \leftarrow \theta_n^{(itera-1)} - \alpha (h_{\theta}(x^{(i)}) - y^{(i)}) x_n^{(i)}$$

Note: Precision  $\downarrow$ , Efficiency  $\uparrow$



choose randomly following the uniform distribution between 1 and  $Itera\_max$  (integer values only!)





### 3) Closed-form solution (CFS)

Purely mathematic

Note: Precision  $\uparrow$ , Efficiency  $\uparrow$  (if N  $\downarrow$ )

$$\theta^* = (X^T X)^{-1} X^T y$$

$$X = \overline{X}_{\text{training}} =$$

$$y = y_{\text{training}} =$$

$$X = \begin{bmatrix} 1 & 0 & 1 & \dots & N \\ (\vec{x}^{(1)})^T \\ \vdots \\ (\vec{x}^{(I)})^T \end{bmatrix} = \begin{bmatrix} 1 & (x^{(1)})^T \\ \vdots & \vdots \\ 1 & (x^{(I)})^T \end{bmatrix}$$

$l = x_0^{(1)}$

$l = x_0^{(I)}$

$$y = \begin{bmatrix} 1 & \dots & I \\ (y^{(1)})^T \\ \vdots \\ (y^{(I)})^T \end{bmatrix}$$

P

De Paul Coanet à tout le monde

on peut remettre 1 sec la page précédente svp ?

PC

Intelligence artificielle (OA)  
on peut remettre la page précédente 1 sec svp

Répondre



Activer



Vidéo



Participants



Chat

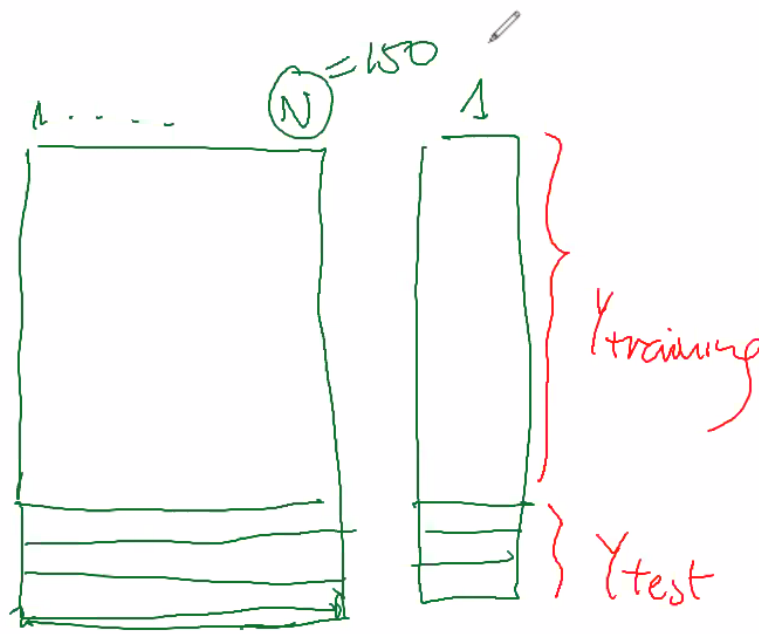
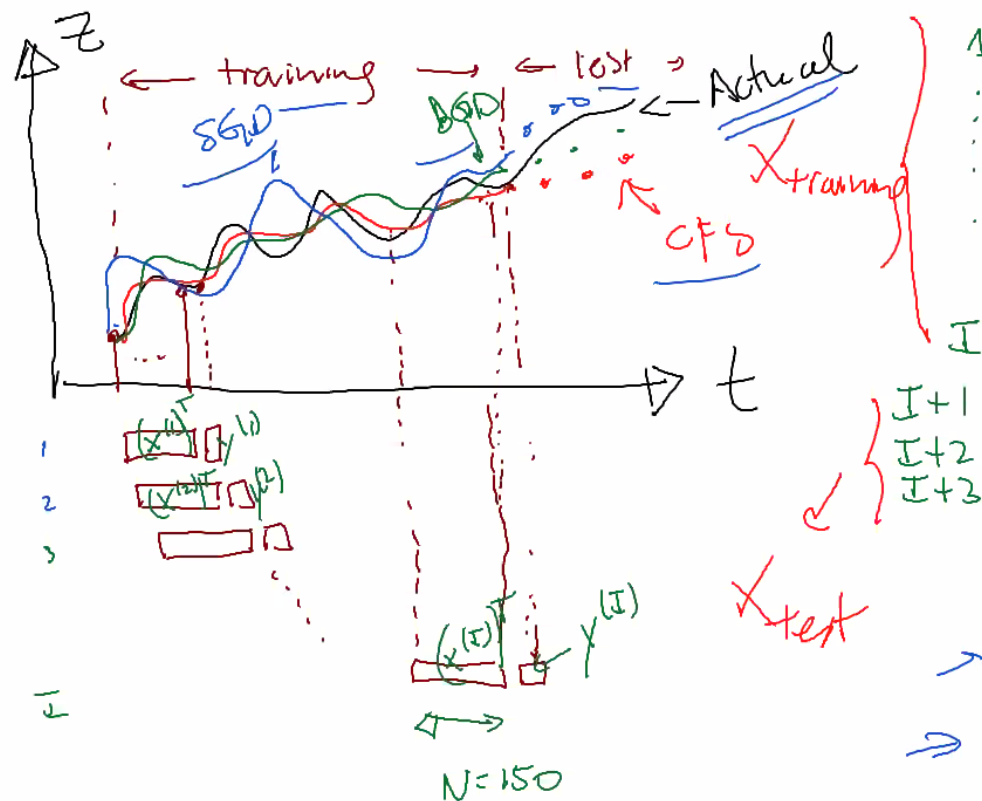


Écran partagé



Enregistrer

### Example (Homework)



1) BGD  $\Rightarrow \hat{\theta}_{BGD}^* \Rightarrow \hat{y}_{BGD} \begin{cases} \nearrow \text{training} \\ \searrow \text{test} \end{cases}$

2) SGD  $\Rightarrow \hat{\theta}_{SGD}^* \Rightarrow \hat{y}_{SGD} \begin{cases} \nearrow \text{training} \\ \searrow \text{test} \end{cases}$

3) CFS  $\Rightarrow \hat{\theta}_{CFS}^* \Rightarrow \hat{y}_{CFS} \begin{cases} \nearrow \text{training} \\ \searrow \text{test} \end{cases}$