DEVLOG RÉSEAU DU DEVELOPPEMENT LOGICIEL Recherche et Enseignement Supérieur

cnrs dépasser les frontières www.cnrs.fr

INSTITUT de MATHEMATIQUES de TOULOUSE

ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

cimi Centre International de Mathématiques et d'Informatique TOULOUSE

Introduction au Deep Learning avec PyTorch

Partie 1 : Introduction

Laurent Risser

Ingénieur de Recherche à l'Institut de Mathématiques de Toulouse et au 3IA ANITI

lrisser@math.univ-toulouse.fr

Partie 1 : Apprentissage supervisé et optimisation ($\approx$ *30 minutes*)

Partie 2 : Architecture et apprentissage des réseaux de neurones ($\approx$ *40 minutes*)

Partie 3 : Apprentissage avec PyTorch ($\approx$ *30 minutes*)

Partie 4 : TPs sous PyTorch ($\approx$ *60 minutes*)

On connais les notes de n élèves au cours de l'année scolaire 2019-2020 ainsi que leur notes à un concours de fin d'année.

L'année suivante, **on aimerait prédire** les notes au concours de la promotion 2020-2021 en fonction de leurs notes au cours de l'année.

	Maths	Info	Français	Concours
Elève 1	12	15	09	14
Elève *2*	05	09	12	07
…	…	…	…	…
Elève *i*	x_i^1	x_i^2	x_i^3	y_i
…	…	…	…	…
Elève *n*	10	12	15	11

	Maths	Info	Français	Concours
Nouvel élève	13	14	11	?

On connais les notes de n élèves au cours de l'année scolaire 2019-2020 ainsi que leur notes à un concours de fin d'année.

L'année suivante, **on aimerait prédire** les notes au concours de la promotion 2020-2021 en fonction de leurs notes au cours de l'année.

	Maths	Info	Français	Concours
Elève 1	12	15	09	14
Elève *2*	05	09	12	07
…	…	…	…	…
Elève *i*	x_i^1	x_i^2	x_i^3	y_i
…	…	…	…	…
Elève *n*	10	12	15	11

	Maths	Info	Français	Concours
Nouvel élève	13	14	11	?

Posons les notations :

- Notes de l'élève $i \in \{1,\dots,n\}$ durant l'année 2019-2020 : $x_i = \left(x_i^1, x_i^2, \dots, x_i^p\right) \in \mathbb{R}^p$
- Notes de l'élève i au concours 2019-2020 : $y_i \in \mathbb{R}$
- Notes de l'élève new durant l'année 2020-2021 : $x_{new} = \left(x_{new}^1, x_{new}^2, \dots, x_{new}^p\right) \in \mathbb{R}^p$
- Fonction pour prédire y_{new} en fonction des x_{new} : $\widehat{y_{new}} = h_\Theta(x_{new})$

Prédiction de y_{new}

Fonction $\mathbb{R}^p \to \mathbb{R}$ avec les paramètres Θ que l'on va apprendre avec les $(x_i, y_i)_{i=1,\dots,n}$

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

1ere question clé : Quel modèle choisir pour h_Θ ?

- Nous allons commencer un modèle très simple : La **régression linéaire** !

$$\widehat{y_{new}} = h_\Theta(x_{new}) = w_0 + \sum_{j=1}^{p} w_j x_{new}^j \quad \text{dont les paramètres sont} \quad \Theta = \{w_0, w_1, \ldots, w_p\}$$

- Prédiction pour le nouvel élève ?

	Maths	Info	Français	Concours
Nouvel élève	13	14	11	?

Prenons $w_0 = 0.$, $w_1 = 0.33$, $w_2 = 0.33$ et $w_3 = 0.33$ → Alors $\widehat{y_{new}} = 12.54$

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

2eme question clé : Critère pour apprendre les paramètres Θ ?

Les meilleurs paramètres $\hat{\Theta}$ minimisent un **risque empirique** sur les $(x_i, y_i)_{i=1,\ldots,n}$, par exemple :

$$\hat{\Theta} = \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} loss\left(h_\Theta(x_i), y_i\right)$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(h_\Theta(x_i) - y_i\right)^2$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{p} w_j x_i^j - y_i\right)^2}_{\text{Risque empirique } R_\Theta((x_i, y_i)_{i=1,\ldots,n})}$$

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

2eme question clé : Critère pour apprendre les paramètres Θ ?

Les meilleurs paramètres $\hat{\Theta}$ minimisent un **risque empirique** sur les $(x_i, y_i)_{i=1,\ldots,n}$, par exemple :

$$\hat{\Theta} = \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} loss\left(h_\Theta(x_i), y_i\right)$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(h_\Theta(x_i) - y_i\right)^2$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{p} w_j x_i^j - y_i\right)^2$$

	Maths	Info	Français	Concours
Eleve 1	12	15	09	14
Eleve 2	05	09	12	07
Eleve 3	13	12	13	12
…	…	…	…	…
Eleve *n*	10	12	15	11

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

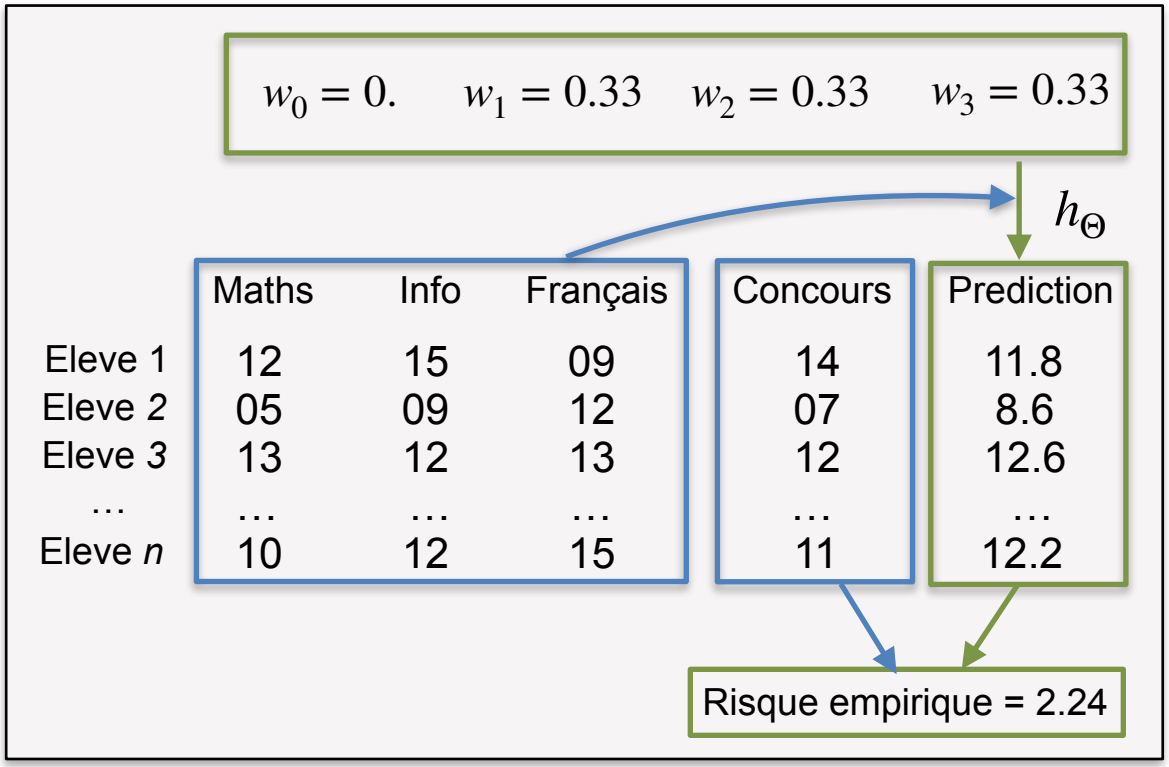
2eme question clé : Critère pour apprendre les paramètres Θ ?

Les meilleurs paramètres $\hat{\Theta}$ minimisent un **risque empirique** sur les $(x_i, y_i)_{i=1,\ldots,n}$, par exemple :

$$\hat{\Theta} = \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} loss\left(h_\Theta(x_i), y_i\right)$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(h_\Theta(x_i) - y_i\right)^2$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{p} w_j x_i^j - y_i\right)^2$$



Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

2eme question clé : Critère pour apprendre les paramètres Θ ?

Les meilleurs paramètres $\hat{\Theta}$ minimisent un **risque empirique** sur les $(x_i, y_i)_{i=1,\ldots,n}$, par exemple :

$$\hat{\Theta} = \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} loss\left(h_\Theta(x_i), y_i\right)$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(h_\Theta(x_i) - y_i\right)^2$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{p} w_j x_i^j - y_i\right)^2$$

$w_0 = 0.$ $w_1 = 0.40$ $w_2 = 0.40$ $w_3 = 0.20$

h_Θ

	Maths	Info	Français	Concours	Prediction
Eleve 1	12	15	09	14	12.6
Eleve *2*	05	09	12	07	8.0
Eleve *3*	13	12	13	12	12.6
…	…	…	…	…	…
Eleve *n*	10	12	15	11	11.8

Risque empirique = 0.99

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

2eme question clé : Critère pour apprendre les paramètres Θ ?

Les meilleurs paramètres $\hat{\Theta}$ minimisent un **risque empirique** sur les $(x_i, y_i)_{i=1,\ldots,n}$, par exemple :

$$\hat{\Theta} = \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} loss\left(h_\Theta(x_i), y_i\right)$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(h_\Theta(x_i) - y_i\right)^2$$

$$= \arg\, min_{\Theta=\{w_0,\ldots,w_p\}} \frac{1}{n} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{p} w_j x_i^j - y_i\right)^2$$

$w_0 = 0.$ $w_1 = 0.50$ $w_2 = 0.50$ $w_3 = 0.00$

h_Θ

	Maths	Info	Français	Concours	Prediction
Eleve 1	12	15	09	14	13.5
Eleve *2*	05	09	12	07	7.0
Eleve *3*	13	12	13	12	12.5
…	…	…	…	…	…
Eleve *n*	10	12	15	11	11.0

Risque empirique = 0.13

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

3eme question clé : Minimisation automatique du risque en fonction de Θ

$$R_\Theta((x_i, y_i)_{i=1,\ldots,n}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_\Theta(x_i) - y_i\right)^2$$

$$\nabla R_\Theta(\ldots) = \left(\frac{\partial R_\Theta(\ldots)}{\partial w_0}, \frac{\partial R_\Theta(\ldots)}{\partial w_1}, \ldots, \frac{\partial R_\Theta(\ldots)}{\partial w_p}\right)^\intercal = \frac{1}{n}\sum_{i=1}^{n}\left(2\left(h_\Theta(x_i) - y_i\right), 2x_i^1\left(h_\Theta(x_i) - y_i\right), \ldots, 2x_i^p\left(h_\Theta(x_i) - y_i\right)\right)^\intercal$$

Gradient de R_Θ

	Maths	Info	Français	Concours
Eleve 1	12	15	09	14
Eleve 2	05	09	12	07
Eleve 3	13	12	13	12
…	…	…	…	…
Eleve n	10	12	15	11

Paramètres Θ

$w_0 = 0.$

$w_1 = 0.33$

$w_2 = 0.33$

$w_3 = 0.33$

h_Θ

Predictions

11.8

8.6

12.6

…

12.2

$R_\Theta(\ldots) = 2.24$

$\nabla R_\Theta(\ldots) = (0.81, 2.93, 7.23, 12.03)^\intercal$

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

3eme question clé : Minimisation automatique du risque en fonction de Θ

$$R_\Theta((x_i, y_i)_{i=1,\ldots,n}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_\Theta(x_i) - y_i\right)^2 \qquad \nabla R_\Theta(\ldots) = \frac{2}{n}\sum_{i=1}^{n}\left(\left(h_\Theta(x_i) - y_i\right), \ldots, x_i^p\left(h_\Theta(x_i) - y_i\right)\right)^\intercal$$

Principe de la descente de gradient :

$$\Theta_{it1} = (0.0, 0.33, 0.33, 0.33)^\intercal \rightarrow \begin{array}{l} R_{\Theta_{it1}}(\ldots) = 2.24 \\ \nabla R_{\Theta_{it1}}(\ldots) = (0.81, 2.93, 7.23, 12.03)^\intercal \end{array}$$

$$\Theta_{it2} = \Theta_{it1} - \lambda \nabla\Theta_{it1}(\ldots) = (-0.00, 0.33, 0.32, 0.32)^\intercal \rightarrow \begin{array}{l} R_{\Theta_{it2}}(\ldots) = 2.09 \\ \nabla R_{\Theta_{it2}}(\ldots) = \left(0.28, -2.47, 1.23, 5.10\right)^\intercal \end{array}$$

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

3eme question clé : Minimisation automatique du risque en fonction de Θ

$$R_\Theta((x_i, y_i)_{i=1,\ldots,n}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_\Theta(x_i) - y_i\right)^2 \qquad \nabla R_\Theta(\ldots) = \frac{2}{n}\sum_{i=1}^{n}\left(\left(h_\Theta(x_i) - y_i\right), \ldots, x_i^p\left(h_\Theta(x_i) - y_i\right)\right)^\intercal$$

Principe de la descente de gradient :

$$\Theta_{it1} = (0.0, 0.33, 0.33, 0.33)^\intercal \rightarrow \begin{array}{l} R_{\Theta_{it1}}(\ldots) = 2.24 \\ \nabla R_{\Theta_{it1}}(\ldots) = (0.81, 2.93, 7.23, 12.03)^\intercal \end{array}$$

$$\Theta_{it2} = \Theta_{it1} - \lambda \nabla\Theta_{it1}(\ldots) = (-0.00, 0.33, 0.32, 0.32)^\intercal \rightarrow \begin{array}{l} R_{\Theta_{it2}}(\ldots) = 2.09 \\ \nabla R_{\Theta_{it2}}(\ldots) = \left(0.28, -2.47, 1.23, 5.10\right)^\intercal \end{array}$$

$$\Theta_{it3} = \Theta_{it2} - \lambda \nabla\Theta_{it2}(\ldots) = (-0.00, 0.33, 0.32, 0.31)^\intercal \rightarrow \begin{array}{l} R_{\Theta_{it3}}(\ldots) = 2.04 \\ \nabla R_{\Theta_{it3}}(\ldots) = \left(0.17, -3.48, 0.05, 3.71\right)^\intercal \end{array}$$

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

3eme question clé : Minimisation automatique du risque en fonction de Θ

$$R_\Theta((x_i, y_i)_{i=1,\ldots,n}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_\Theta(x_i) - y_i\right)^2 \qquad \nabla R_\Theta(\ldots) = \frac{2}{n}\sum_{i=1}^{n}\left(\left(h_\Theta(x_i) - y_i\right), \ldots, x_i^p\left(h_\Theta(x_i) - y_i\right)\right)^\intercal$$

Principe de la descente de gradient :

$$\Theta_{it1} = (0.0, 0.33, 0.33, 0.33)^\intercal \rightarrow \begin{array}{l} R_{\Theta_{it1}}(\ldots) = 2.24 \\ \nabla R_{\Theta_{it1}}(\ldots) = (0.81, 2.93, 7.23, 12.03)^\intercal \end{array}$$

$$\Theta_{it2} = \Theta_{it1} - \lambda \nabla \Theta_{it1}(\ldots) = (-0.00, 0.33, 0.32, 0.32)^\intercal \rightarrow \begin{array}{l} R_{\Theta_{it2}}(\ldots) = 2.09 \\ \nabla R_{\Theta_{it2}}(\ldots) = \left(0.28, -2.47, 1.23, 5.10\right)^\intercal \end{array}$$

$$\Theta_{it3} = \Theta_{it2} - \lambda \nabla \Theta_{it2}(\ldots) = (-0.00, 0.33, 0.32, 0.31)^\intercal \rightarrow \begin{array}{l} R_{\Theta_{it3}}(\ldots) = 2.04 \\ \nabla R_{\Theta_{it3}}(\ldots) = \left(0.17, -3.48, 0.05, 3.71\right)^\intercal \end{array}$$

$$\vdots$$

$$\Theta_{it1000} = \Theta_{it999} - \lambda \nabla \Theta_{it999}(\ldots) = \left(-0.00, 0.48, 0.53, -0.01\right)^\intercal \rightarrow \begin{array}{l} R_{\Theta_{it3}}(\ldots) = 0.08 \\ \nabla R_{\Theta_{it3}}(\ldots) = (-0.01, 0.09, -0.12, 0.04)^\intercal \end{array}$$

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1,\ldots,n$ et fonction de prédiction h_Θ de paramètres Θ

3eme question clé : Minimisation automatique du risque en fonction de Θ

$$R_\Theta((x_i, y_i)_{i=1,\ldots,n}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_\Theta(x_i) - y_i\right)^2 \qquad \nabla R_\Theta(\ldots) = \frac{2}{n}\sum_{i=1}^{n}\left(\left(h_\Theta(x_i) - y_i\right), \ldots, x_i^p\left(h_\Theta(x_i) - y_i\right)\right)^\intercal$$

Score à convergence :

$$\Theta_{it1000} = \Theta_{it999} - \lambda \nabla\Theta_{it999}(\ldots) = \left(-0.00, 0.48, 0.53, -0.01\right)^\intercal \rightarrow R_{\Theta_{it3}}(\ldots) = 0.08$$

On a convergé vers une parametrisation qui marche très bien **sur le jeu d'apprentissage** !

Prédiction :

	Maths	Info	Français	Concours
Nouvel élève	13	14	11	$\widehat{y_{new}} = 13.49$

… dans la vrai vie, il faudra aussi valider l'apprentissage avant de l'utiliser sur de nouvelles données

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1,\ldots,n$ et fonction de prédiction h_Θ de paramètres Θ

Pour aller plus loin : Descente de gradient stochastique

$$\nabla R_\Theta(\ldots) = \frac{2}{n}\sum_{i=1}^{n}\left(\left(h_\Theta(x_i) - y_i\right), \ldots, x_i^p\left(h_\Theta(x_i) - y_i\right)\right)^\intercal$$

$$\approx \frac{2}{\#B}\sum_{i\in B}\left(\left(h_\Theta(x_i) - y_i\right), \ldots, x_i^p\left(h_\Theta(x_i) - y_i\right)\right)^\intercal$$

Mini-batch (exemple $B = \{1,2,3\}$)

- Meilleur rapidité !
- Moins de mémoire nécessaire !

Batch-training → descent de gradient avec un mini-batch qui évolue itération après itération

- Meilleur exploration de l'espace des paramètres !

Notations : n observations d'apprentissage : $x_i = \left(x_i^1, x_i^2, \ldots, x_i^p\right)$ et y_i avec $i = 1, \ldots, n$ et fonction de prédiction h_Θ de paramètres Θ

Pour aller plus loin : Descente de gradient stochastique

Initialise $\Theta = (w_0, w_1, \ldots, w_p)$

For $epoch = 1$ to 5

  For $b = 1$ to n with step 10

    $B = \{b, b+1, \ldots, b+9\}$

    Approche $\nabla R_\Theta(\ldots)$ avec $\nabla R_\Theta(\ldots) \approx \frac{2}{\#B} \sum_{i \in B} \left(\left(h_\Theta(x_i) - y_i\right) \ , \ \ldots \ , x_i^p \left(h_\Theta(x_i) - y_i\right) \right)$

    $\Theta = \Theta - \lambda \nabla R_\Theta(\ldots)$

  End For

End For

Descente de gradient avec 5 **epochs** et des **mini-batchs** de taille 10