IFT 6390, Homework 2

Adel Nabli, Myriam Laiymani

05 Octobre 2018

1 Régression linéaire et non linéaire régularisée

1.1 Régression linéaire

1. On a $f(x) = \omega^T x + b$, avec $x \in \mathbb{R}^d$ et $f(x) \in \mathbb{R}$. On observe $D_n = \{(x^{(1)}, y^{(1)})...(x^{(n)}, y^{(n)})\}$, ainsi les paramètres de la régression sont ω **et** b, ce qui donne:

$$\theta = \{\omega, b\} \text{ avec } \omega \in \mathbb{R}^d \text{ et } b \in \mathbb{R}$$

2. Par définition du risque empirique et de part la fonction de perte choisie, on a:

$$\hat{R} = \sum_{i=1}^n L((x_i, t_i), f) = \sum_{i=1}^n (\omega^T x_i + b - t_i)^2$$

3. Pour faciliter les notations et calculs, on va se ramener au cas linéaire (et non affine). On va ainsi noter θ le vecteur de $\mathbb{R}^{d+1}$ résultant de la concaténation de b et ω ($\theta = \begin{pmatrix} b \\ \omega \end{pmatrix}$). De même, nous allons dorénavant travailler avec les $x'_i = \begin{pmatrix} 1 \\ x_i \end{pmatrix}$, $x'_i \in \mathbb{R}^{d+1}$ de sorte que $\forall x \in \mathbb{R}^d,\ f(x) = \theta^T x'$.

   Nous voulons trouver θ^* le paramètre minimisant le risque empirique. Ainsi, notre problème d'optimisation est le suivant:

$$\theta^* = \underset{\theta}{\text{argmin}} \sum_{i=1}^n L((x_i, t_i) = \underset{\theta}{\text{argmin}} \sum_{i=1}^n (\theta^T x'_i - t_i)^2$$

4. On a:

$$\hat{R} = \sum_{i=1}^n (\theta^T x'_i - t_i)^2 = \sum_{i=1}^n (\theta_0 x'_{i0} + ... + \theta_d x'_{id} - t_i)^2$$

   Ainsi, on calcule que:

$$\forall j \in [\![0, d]\!], \frac{\partial \hat{R}}{\partial \theta_j} = 2 \sum_{i=1}^n x'_{ij} (\theta^T x'_i - t_i)$$

   Or, on sait que:

$$\nabla_\theta(\hat{R}) = \begin{pmatrix} \frac{\partial \hat{R}}{\partial \theta_0} \\ \vdots \\ \frac{\partial \hat{R}}{\partial \theta_d} \end{pmatrix}$$

En introduisant la matrice $X \in \mathbb{R}^{n\times(d+1)}$ tel que $X = \begin{pmatrix} 1 & x_{11} & ... & x_{1d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & ... & x_{nd} \end{pmatrix}$, et le vecteur $t = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}$, on peut ainsi écrire:

$$\nabla_\theta(\hat{R}) = 2X^T \underbrace{(X\theta - t)}_{=(f(X)-T)}$$

5. On peut ainsi voir le gradient du risque empirique comme un vecteur dont chaque dimension contiendrait la somme des erreurs du modèle sur l'ensemble d'entraînement pondérée par les valeurs des "x" d'entraînements sur cette dimension.

1.2 Régression linéaire régularisée ("ridge regression")

1. On a :

$$\tilde{R} = \sum_{i=1}^{n}(\omega^T x_i + b - t_i)^2 + \lambda \sum_{j=1}^{d} \omega_j^2$$

Ainsi,

$$\forall j \in [\![0, d]\!], \frac{\partial \tilde{R}}{\partial \theta_j} = 2\sum_{i=1}^{n} x'_{ij}(\theta^T x'_i - t_i) + 2\lambda\theta_j \mathbf{1}_{j>0}$$

Ce qui donne:

$$\nabla_\theta(\tilde{R}) = 2X^T(X\theta - t) + 2\lambda \begin{pmatrix} 0 \\ \omega \end{pmatrix}$$

La différence est donc, par rapport au gradient du risque non régularisé, l'ajout d'un terme proportionel au vecteur ω au gradient du risque.

2. Algorithme de *batch gradient descent* pour la régression linéaire régularisée:

ALGORITHME :

- Demande de la valeur d'un ϵ et l'hyperparamètre η à l'utilisateur
- Initialisation aléatoire du vecteur $\theta^{(0)} = \begin{pmatrix} b^{(0)} \\ \omega^{(0)} \end{pmatrix}$
- Calcul de $\nabla_\theta^{(0)} = 2X^T(X\theta^{(0)} - t) + 2\lambda \begin{pmatrix} 0 \\ \omega^{(0)} \end{pmatrix}$
- **Tant que** $||\nabla_\theta|| > \epsilon$ **faire:**
  - Calcul de $\theta^{(k+1)} = \theta^{(k)} - \eta\nabla_\theta^{(k)}$
  - Calcul de $\nabla_\theta^{(k+1)} = 2X^T(X\theta^{(k+1)} - t) + 2\lambda \begin{pmatrix} 0 \\ \omega^{(k+1)} \end{pmatrix}$
- Retourne le dernier θ calculé

3. On trouve, en reprenant les expressions déjà écrites dans les questions précédentes:

$$\tilde{R} = (X\theta - t)^T(X\theta - t) \text{ et } \nabla_\theta(\tilde{R}) = 2X^T(X\theta - t) + 2\lambda\theta$$

4. $\nabla_\theta(\tilde{R}) = 0 \iff (X^TX + \lambda\mathbf{I_d})\theta = X^Tt$ et si $(X^TX + \lambda\mathbf{I_d})$ est inversible (on pourra toujours trouver un λ pour que cela soit le cas), cela donne:

$$\theta = (X^TX + \lambda\mathbf{I_d})^{-1}X^Tt$$

Si $N < d$ et $\lambda = 0$, X^TX ne peut être inversible et il y une infinité de solutions θ à l'équation $X^TX\theta = X^Tt$ (l'ensemble des solutions à cette équation est un espace affine de direction $Ker(X^TX)$ qui est de dimension non nulle).

1.3 Régression avec un pré-traitement non-linéaire fixe

1. En posant $\omega \in \mathbb{R}^k$ et $b \in \mathbb{R}$, on a:

$$\forall x \in \mathbb{R}, \tilde{f}(x) = f(\phi(x)) = \omega^T\phi(x) + b = \sum_{i=1}^k w_i x^i + b$$

2. Les paramètres sont donc ω (vecteur de dimension k) et b (scalaire).

3. $\phi_{poly^1}(x) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\phi_{poly^2}(x) = \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1x_2 \end{pmatrix}$, $\phi_{poly^3}(x) = \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1x_2 \\ x_1^3 \\ x_2^3 \\ x_1^2x_2 \\ x_1x_2^2 \end{pmatrix}$

4. Soit $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$. Pour une puissance p donnée, $p \in [\![1, k]\!]$, le nombre de termes d'ordre p qu'il est possible de construire avec les composantes de x correspond au *nombre de possibilité qu'il y a de choisir p éléments dans un ensemble de d éléments* ***sans tenir compte de l'ordre*** (la multiplication étant commutative, $x_1^2x_2 = x_1x_2x_1 = x_2x_1^2$).

Ainsi, cela correspond au nombre de suites d'entiers $(p_i)_{i\in[\![1,d]\!]}$ différentes tel que $\sum_{i=1}^d p_i = p$ qu'il est possible de construire:

- On choisit p_1 fois le termes x_1

  $\vdots$
- On choisit p_d fois le termes x_d

On peut encoder cela sous la forme suivante:

$$\underbrace{+...+}_{p_1\ fois} \mid \underbrace{+...+}_{p_2\ fois} \mid ... \mid \underbrace{+...+}_{p_d\ fois}$$

Sur l'encodage ci-dessus, nous avons p "+" et $d-1$ "|". Ainsi, le *nombre de possibilité qu'il y a de choisir p éléments dans un ensemble de d éléments sans tenir compte de l'ordre* est exactement égal au *nombre de possibilité qu'il y a de placer $d-1$ "|" dans une liste de $p+d-1$ éléments.*

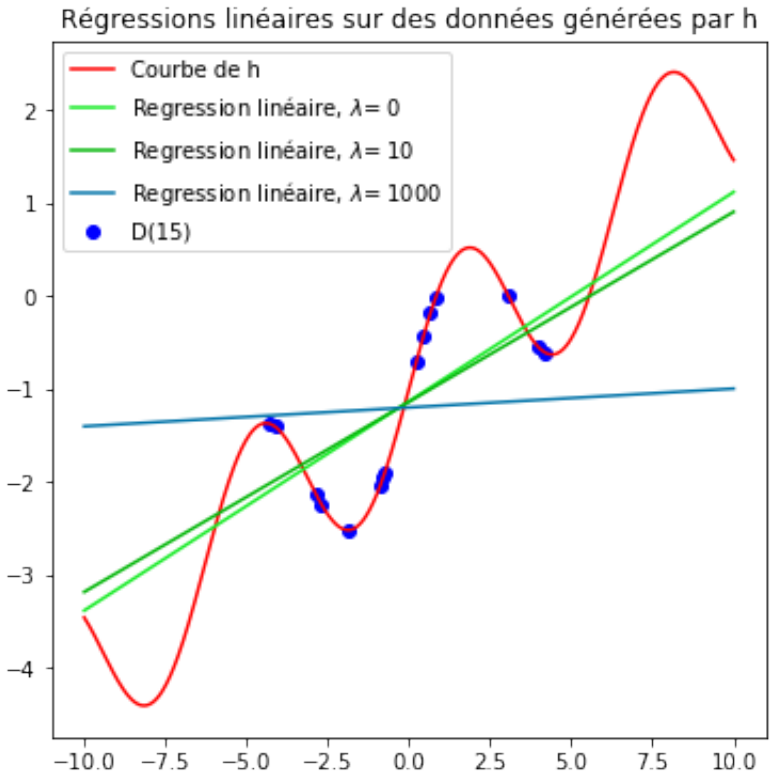
Le nombre cherché est donc $\binom{p+d-1}{d-1}$.

Cela étant pour une puissance p donnée, on en déduit que la dimension de $\phi^k_{poly}(x)$ est:

$$\dim(\phi^k_{poly}(x)) = \sum_{p=1}^{k} \binom{p+d-1}{d-1}$$

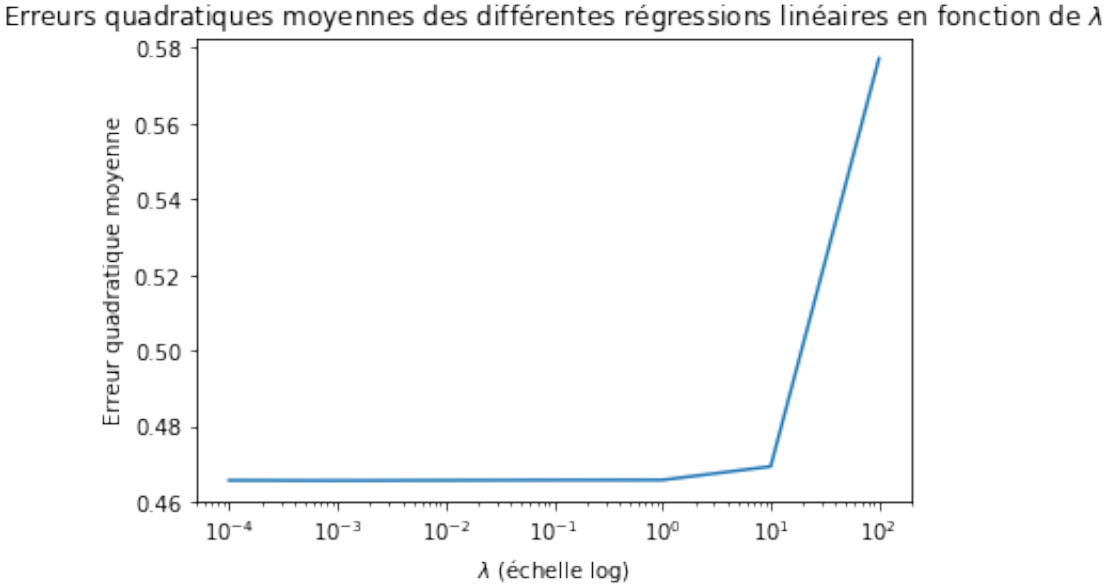
2 Partie pratique

Question 3 et 4:

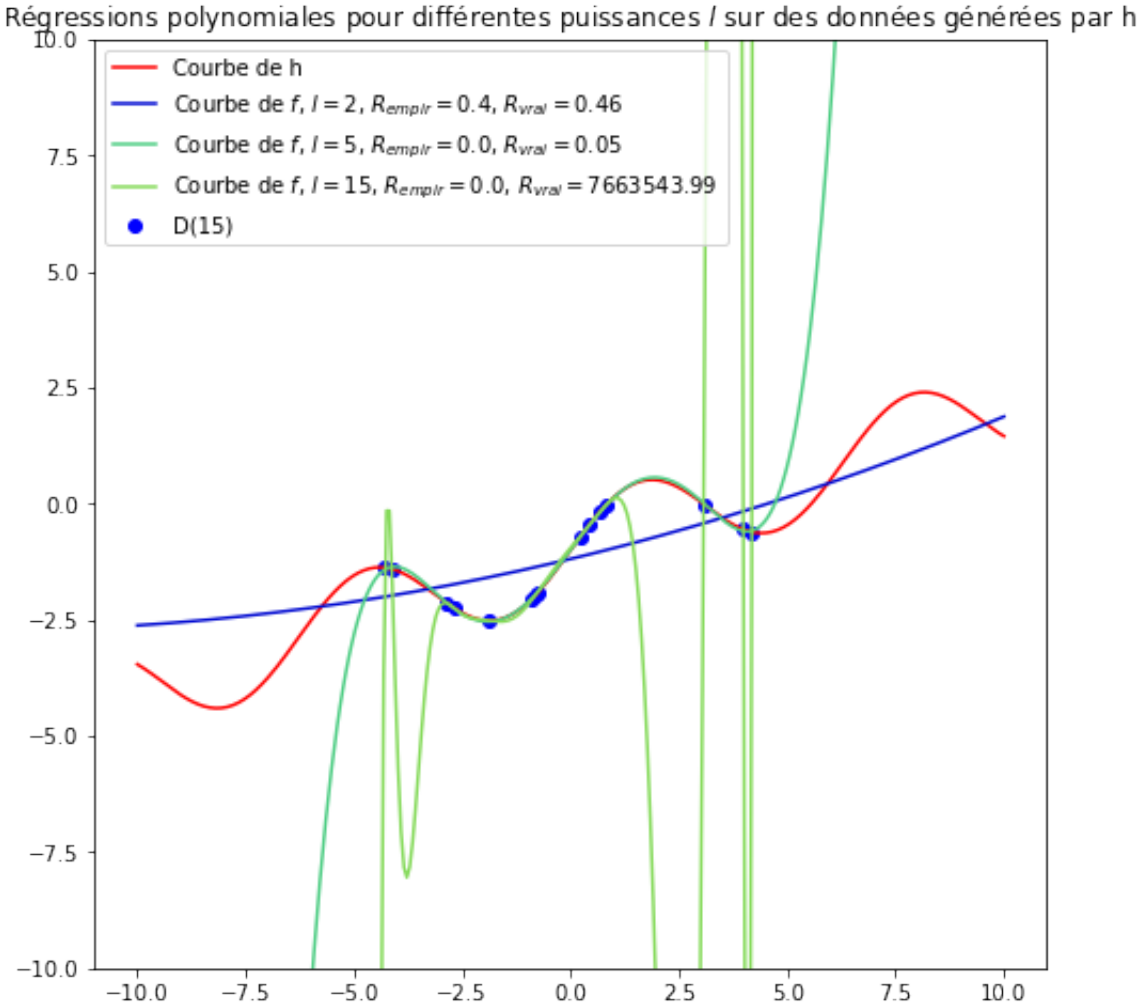


Ainsi, plus λ est grand, plus la "complexité" du modèle est pénalisée et plus on tant vers le modèle "constant".

Question 5:



Question 6:



Question 7:

On observe que plus l augmente, plus le modèle colle aux points de D_{train}.

Quand l est petit ($l = 2$ par exemple), on a une modélisation qui reste trop simpliste et qui n'arrive pas à capturer la complexité du modèle générateur (on a un *risque empirique* et un *vrai risque* relativement élevé).

Quand l n'est pas trop grand ($l = 5$ par exemple), la courbe générée colle presque parfaitement avec la courbe de h, ce qui engendre un *risque empirique* et un *vrai risque* très bas: notre modélisation est suffisamment complexe pour capturer la complexité du modèle générateur, mais ne l'est pas trop car se généralise bien à de nouvelles données.

Cependant, dès que l'on augmente plus en puissance (sur notre graphique, on a pris l'exemple de $l = 15$), les polynomes générés passent certes par tous les points d'apprentissage (le *risque empirique* est proche de 0), mais les polynomes ont des "oscillations" de grande amplitude entre chaque point de D_{train}, ce qui crée un *vrai risque* très élevé. On est clairement dans un cas de sur-apprentissage.