EXERCICES DU COURS FONDAMENTAUX DE L'APPRENTISSAGE (M2 IASD)

YANN CHEVALEYRE

1. Notions de risque

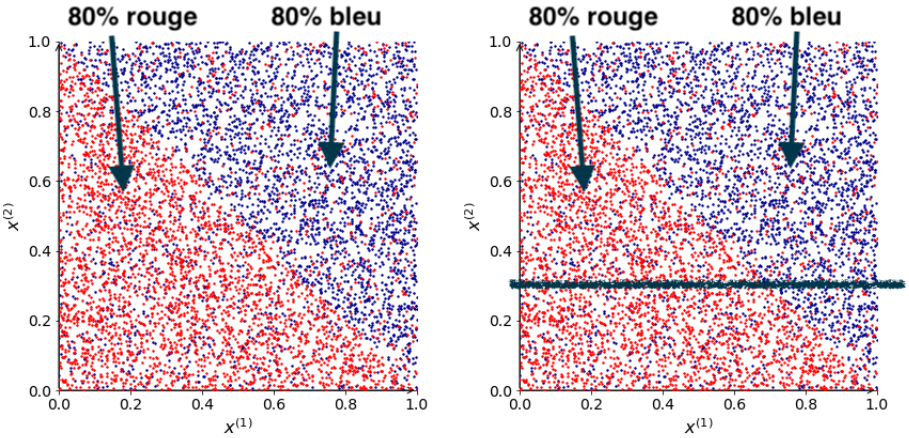


Figure 1.1

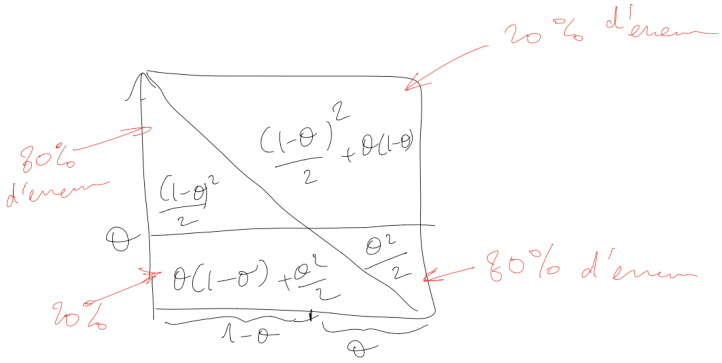
Exercise 1. (Risque de Bayes) On se place dans le cadre de l'apprentissage en ligne supervisé bi-classe. Soit $\mathcal{X} = [0,1]^2$. La distribution des données, est représentée sur la figure 1.1 de gauche (la couleur bleue représente la classe 1, et le rouge représente -1). Cette distribution est décrite ainsi: $P\left(Y = 1 \mid x^{(1)} + x^{(2)} \geq 1\right) = 0.8$, $P\left(Y = 1 \mid x^{(1)} + x^{(2)} < 1\right) = 0.2$, $P(Y = 1) = \frac{1}{2}$ et $P(X)$ est uniforme sur $\mathcal{X}$.

- On s'intéresse d'abord à la perte 0/1
  - Donnez le classifieur optimal de Bayes et calculez le risque de Bayes (détaillez vos calculs).
  - Calculez le risque quand $N \to \infty$ du 1-plus proche voisin.
  - On s'intéresse à la classe $\mathcal{H} = \left\{ h_\theta(x) = \begin{cases} 1 & \text{si } x^{(2)} \geq \theta \\ -1 & \text{sinon} \end{cases} : \theta \in [0,1] \right\}$. Sur la figure 1.1 de droite, on a représenté $h_{0.3}$. Calculez le risque de h_θ pour θ quelconque.

    Correction

    Soit $\eta(x) = P(Y = 1 \mid X = x)$

* Classifieur optimal de bayes: $h^*(x) = \begin{cases} 1 & si\ x^{(1)} + x^{(2)} \geq 1 \\ -1 & sinon \end{cases}$.
* $R^{0/1}_{bayes} = \mathbb{E}_{XY}\left[\mathbf{1}\left[h^*(X) \neq Y\right]\right] = \mathbb{E}_X \min\left(\eta(X), 1-\eta(X)\right) = \frac{1}{2} * 0.2 + \frac{1}{2} * 0.2 = 0.2$
* Risque du 1-plus proche voisin. Soit (X, Y) un exemple étiqueté sous la diagonale. Dans cette zone, il y a 20% de classe 1 et 80% de classe -1. Donc $P(Y = 1 \mid X) = 20\%$. Lorsque $N \to \infty$, le plus proche voisin de x, qu'on appelle $(\hat{X}, \hat{Y})$ a 20% de chance d'être de classe 1 et 80% de classe -1. Donc $P(\hat{Y} = 1 \mid X) = 20\%$. Donc la probabilité de commettre une erreur de classification est $P(Y \neq \hat{Y} \mid X) = P(Y = 1 \mid X)P(\hat{Y} = -1 \mid X) + P(Y = -1 \mid X)P(\hat{Y} = 1 \mid X) = 0.2 * 0.8 + 0.8 * 0.2 = 0.32$. Si X est au dessus de la diagonale, on obtient la même probabilité d'erreur. Donc, pour le 1-plus proche voisin, le risque tend vers 0.32 quand $N \to \infty$.
* Risque de h_θ. Le dessin ci-dessous nous montre l'aire des différentes parties de la figure en fonction de θ. La probabilité qu'un exemple X tiré depuis la distribution des données tombe dans une de ces zones est égal à son aire. Le risque de h_θ est donc égal au risque dans chacun de ces zones multiplié par son aire. Donc $R^{0/1}(h_\theta) = \frac{\theta^2}{2} \times 0.8 + \left(\theta(1-\theta) + \frac{\theta^2}{2}\right) \times 0.2 + \left(\frac{(1-\theta)^2}{2} + \theta(1-\theta)\right) \times 0.2 + \frac{(1-\theta)^2}{2} \times 0.8$.



- Supposons qu'on s'intéresse maintenant à la perte "mean square error" $\ell^{MSE}(\hat{y}, y) = (\hat{y} - y)^2$
  - Ecrivez la formule de ce risque sur cette distribution
  - Donnez la formule du risque de Bayes avec ℓ^{MSE} dans le cas général
  - Donnez le classifieur optimal de Bayes (en considérant les classifieurs à valeur dans $\mathbb{R}$), et le risque de Bayes associé

  Correction

$$
\begin{aligned}
R^{MSE}(h) &= \mathbb{E}_{XY}\left[(h(X)-Y)^2\right] \\
&= \mathbb{E}_X\left[0.8\times(h(X)-1)^2 + 0.2\times(h(X)+1)^2 \mid X^{(1)}+X^{(2)}\geq 1\right] \\
&+ \mathbb{E}_X\left[0.2\times(h(X)-1)^2 + 0.8\times(h(X)+1)^2 \mid X^{(1)}+X^{(2)}< 1\right]
\end{aligned}
$$

* Le classifieur de Bayes est $h^{*MSE}(x) = \arg\min_{\hat{y}} \mathbb{E}_Y\left[(Y-\hat{y})^2 \mid X=x\right] = \mathbb{E}\left[Y \mid X=x\right] = 2\eta(X)-1$
* Dans le cas général le risque de Bayes est:

$$
\begin{aligned}
R^{MSE}_{bayes} &= \mathbb{E}_{XY}\left[(h^{*MSE}(X)-Y)^2\right] \\
&= \mathbb{E}_X\mathbb{E}_Y\left[(Y-\mathbb{E}\left[Y \mid X\right])^2\right] \\
&= \mathbb{E}_X Var\left(Y \mid X\right) \\
&= 4\mathbb{E}_X\eta(X)\left(1-\eta(X)\right)
\end{aligned}
$$

En effet, $Y\in\{-1,1\}$, et $Var(Y \mid X) = 4Var\left(\frac{Y+1}{2}\right) = 4Var\left(\frac{Y}{2}\right)$. Ici, $\frac{Y}{2}$ est une var aléatoire à valeur dans $\{0,1\}$ de paramètre $\eta(X)$ (donc une Bernouilli), dont la variance est $\eta(X)(1-\eta(X))$. D'où $Var(Y \mid X) = 4\eta(X)(1-\eta(X))$.

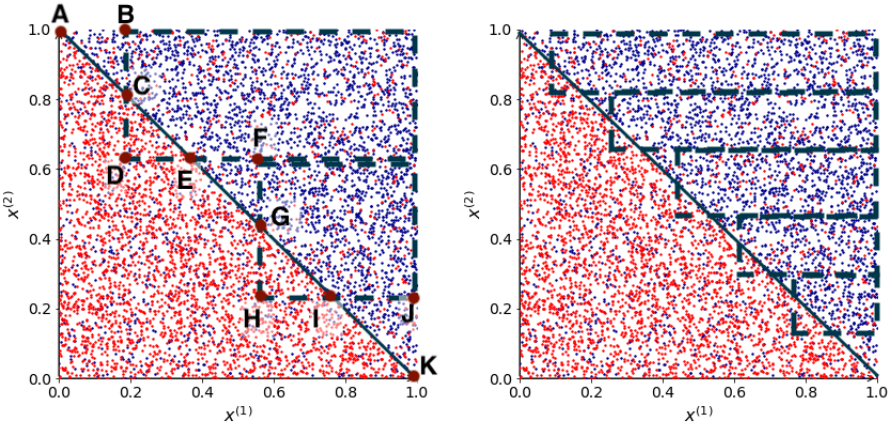


FIGURE 1.2

Exercise 2. (Erreur d'approximation). On suppose qu'on a la même distribution d'exemples que dans l'exercice 1.1. Cette fois, on s'intéresse à un autre type de classifieurs: les unions de rectangles disjoints. Soit $\mathcal{H}_k$, les classifieurs composées d'union de k rectangles. Dans la figure 1.2, on a représenté le classifieur optimal pour $\mathcal{H}_2$ à gauche et le classifieur optimal $\mathcal{H}_5$ à droite. On admettra ce fait: pour n'importe quelle valeur de k, le classifieur optimal possède toujours une frontière de décision en escalier autour de la diagonale, de telle sorte que les triangles rectangles formés par la diagonale et la frontière en escalier soient tous identiques (sur

la figure 1.2 à gauche, les triangles ABC, DCE, EFG, GHI, IJK sont identiques, et les segments AC, CE, EG, GI, IK sont de même longueur). Calculez le risque de Bayes (avec perte 0/1) du classifieur optimal de $\mathcal{H}_k$. Que se passe-t-il quand $k \to \infty$?

Correction

Le risque de Bayes est de 20% comme dans l'exercice précédent.

Prenons la partie gauche de la figure 1.2.

Si h est le classifieur en escalier et f^* le classifieur de Bayes, On voit que

$$\begin{aligned} P(h \neq f^\star) &= Aire(IJK) + Aire(GHI) + Aire(EFG) + Aire(CDE) + Aire(ABC) \\ &= (2k+1).Aire(ABC) = (2k+1).\frac{1}{2(2k+1)^2} = \frac{1}{2(2k+1)} \end{aligned}$$

On a donc

$$\begin{aligned} P(h(X) \neq Y) =& P\left(h(X) \neq Y \mid h(X) = f^\star(X)\right).P(h(X) = f^\star(X)) \\ &+ P\left(h(X) \neq Y \mid h(X) \neq f^\star(X)\right).P(h(X) \neq f^\star(X)) \\ =& 0.2 \times P(h(X) = f^\star(X)) \\ &+ 0.8 \times P(h(X) \neq f^\star(X)) \\ =& 0.2 \times \left(1 - \frac{1}{2(2k+1)}\right) + 0.8 \times \frac{1}{2(2k+1)} \\ =& 0.2 + \frac{0.6}{2(2k+1)} \end{aligned}$$

Ce risque tend vers zéro quand $k \to \infty$.

2. PAC Learning

Exercise 3. Soit $\mathcal{H}$ une classe finie de fonctions avec $m = |\mathcal{H}|$ et $conv(\mathcal{H}) = \left\{x \mapsto \alpha_1 h_1(x) + \ldots \alpha_m h_m(x) : \alpha_i \geq 0, \sum_{i=1}^M \alpha_i = 1, h_i \subseteq \mathcal{H}\right\}$, la fermeture convexe de $\mathcal{H}$. Montrer que $\mathcal{R}ad\left(conv\left(\mathcal{H}\right)\right) \leq \mathcal{R}ad\left(\mathcal{H}\right)$

Exercise 4. (Rademacher pour les classes linéaires avec L_1 bornée). This is Massart's finite lemma:

Finite Class Lemma (Massart). *Let $\mathcal{A}$ be some finite subset of $\mathbb{R}^m$ and $\epsilon_1, \ldots, \epsilon_m$ be independent Rademacher random variables. Let $r = \sup_{a \in \mathcal{A}} \|a\|$. Then, we have,*

$$\mathbb{E}\left[\sup_{a \in \mathcal{A}} \frac{1}{m} \sum_{i=1}^m \epsilon_i a_i\right] \leq \frac{r\sqrt{2 \ln |\mathcal{A}|}}{m} .$$

The goal of this exercise is to proove a bound on the empirical rademacher complexity for linear functions with bounded ℓ_1 norm: $\mathcal{R}ad(\mathcal{F}) \leq X_\infty W_1 \sqrt{\frac{2 \log d}{m}}$ where $\|x_i\|_\infty \leq X_\infty$ for all i and $\mathcal{F} = \left\{x \mapsto x^T\theta : \theta \in \mathbb{R}^d, \|\theta\|_1 \leq W_1\right\}$.

(1) Write the formula of empirical rademacher complexity
(2) Use the fact[1] that $\|x\|_\infty = \sup_{\|y\|_1 \leq 1} \langle y, x \rangle$. The "sup" term should disappear here

[1] All dual norms share this property

(3) Express the $\|\cdot\|_\infty$ as a supremum over coordinates.
(4) Use Massart's finite lemma
(5) Conclude

Correction:

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}) &= \frac{1}{m}\mathbb{E}\left[\sup_{w:||w||_1\leq W_1}\sum_{i=1}^m \epsilon_i w\cdot x_i\right]\\
&= \frac{1}{m}\mathbb{E}\left[\sup_{w:||w||_1\leq W_1} w\cdot\sum_{i=1}^m \epsilon_i x_i\right]\\
&= \frac{W_1}{m}\mathbb{E}\left[||\sum_{i=1}^m \epsilon_i x_i||_\infty\right]\\
&= \frac{W_1}{m}\mathbb{E}\left[\sup_j \sum_{i=1}^m \epsilon_i [x_i]_j\right]\\
&\leq \frac{W_1\sqrt{2\log d}}{m}\sup_j\sqrt{\sum_{i=1}^m [x_i]_j^2}\\
&\leq X_\infty W_1\sqrt{\frac{2\log d}{m}}
\end{aligned}$$

3

(see for example `https://ttic.uchicago.edu/~tewari/lectures/lecture17.pdf`)

3. Apprentissage en ligne

Exercise 5. (Halving et VC-dim). Assume we have a dataset $(x_1, y_1)\dots(x_N, y_N)$ where $x_i \in \mathcal{X} \subset \mathbb{R}^2$ and $y_i \in \{0, 1\}$. Assume this dataset is perfectly linearly separable.

(1) What is the VC-dim of linear classifiers here ? What do Sauer's lemma tell us ?
(2) Can we apply Halving's algorithm here (without any modification on the algorithm) ? How many classifiers do we need to consider ? How many mistakes would we make ?

Given a function class $\mathcal{F}$ of VC-dimension $\text{VC-dim}(\mathcal{F})$, its growth function is bounded for all $n > \text{VC-dim}(\mathcal{F})$ by

$$\Pi_{\mathcal{F}}(n) \leq \sum_{k=0}^{\text{VC-dim}(\mathcal{F})}\binom{n}{k} \leq \left(\frac{en}{\text{VC-dim}(\mathcal{F})}\right)^{\text{VC-dim}(\mathcal{F})}.$$

Correction:

L'algorithme Halving commet $\ln_2 |\mathcal{H}|$ erreurs au pire. Sur N points, on a au plus $\left(\frac{eN}{3}\right)^3$ façon différentes de classifier ces points avec des séparateurs linéaires. Donc, au lieu de considérer tous les classifieurs linéaires, dont la plupart sont équivalents sur ces points, on peut ne considérer que $\left(\frac{eN}{3}\right)^3$ classifieurs (un pour chaque classe

d'équivalence). Avec ce raisonnement, Halving commet $\ln_2 \left(\frac{eN}{3}\right)^3 = O(\ln N)$ erreurs.

Exercise 6. (Infinite Halving). Dans le cadre de l'apprentissage en ligne supervisé bi-classe, supposons que $\mathcal{H}$ soit infinie, et qu'on soit dans le cas réalisable. Normalement, on ne peut pas appliquer l'algorithme Halving, puisqu'il requiert que $\mathcal{H}$ soit fini. Adaptez Halving au cas $|\mathcal{H}|$ infini (en supposant que vous avez des ressources de calcul sans limite) et bornez le nombre d'erreurs du mieux que vous pouvez.

Indice: Halving ne peut pas s'appliquer directement parce que l'étape de prédiction $\hat{y}_t = \arg\max_{k \in \mathcal{Y}} \left|\mathcal{H}_t^k\right|$ revient à comparer la cardinalité d'ensembles infinis. Pour pallier ce problème, *pondérez* les classifieurs de $\mathcal{H}$ de telle sorte que la somme de leurs poids soit finie, et comparez la somme des poids.

Correction. de l'exercice 6. Le principal problème est que je ne peux pas effectuer la prédiction $\hat{y}_t = \arg\max_k |\mathcal{H}_{t,k}|$ puisque ces ensembles sont infinis. Je dois donc *pondérer* les classifieurs de telle sorte que la somme des poids ne soit pas infinie (par exemple $w_i = \frac{6}{\pi^2 i^2}$).

Ordonnons les classifieurs $\mathcal{H} = \{h_1, h_2, \ldots\}$. Construisons $w_1, w_2 \ldots \in \mathbb{R}_+$ de telle sorte que $\sum_{i=1}^{\infty} w_i = 1$ et $w_1 > w_2 > \ldots$. Par exemple, $w_i = \frac{6}{\pi^2 i^2}$.

Au lieu de manipuler les classifieurs $\mathcal{H}$, je vais manipuler leurs indices $\mathcal{I}$.

Algorithme:

(1) $\mathcal{I}_1 = \mathbb{N}^*$
(2) Pour $t = 1 \ldots$
  (a) Je reçois x_t
  (b) Pour $k \in \{-1, 1\}$
    (i) Soit $\mathcal{I}_t^k = \{i \in \mathcal{I}_t : h_i(x) = k\}$
    (ii) Soit $\Omega_t^k = \sum_{\{i \in \mathcal{H}_t^k\}} w_i$
  (c) Je prédis $\hat{y}_t = \arg\max_k \Omega_t^k$
  (d) $\mathcal{I}_{t+1} = \{i \in \mathcal{I}_t : h_i(x_t) = y_t\}$

Quelle est l'erreur commise ? Soit $\Omega_t = \Omega_t^1 + \Omega_t^{-1}$, le poids total des classifieurs de $\mathcal{I}_t$.

Lorsqu'il y a une erreur de prédiction commise, alors on supprime de $\mathcal{I}_t$ une partie des classifieurs qui représentent plus de 50% du poids total Ω_t. Donc à chaque erreur, on a $\Omega_{t+1} \leq \frac{1}{2}\Omega_t$. Donc si E est le nombre d'erreur de prédiction au temps T, on a $\Omega_{T+1} \leq 2^{-E}\Omega_1 = 2^{-E}$. Supposons que h_{i^*} soit le classifieur d'index minimum qui donne une erreur nulle sur toute la procédure d'apprentissage. Ce classifieur n'est donc jamais enlevé de $\mathcal{I}_t$. Donc, $w_{i^*} \leq \Omega_{T+1} \leq 2^{-E}$, donc $E \leq \ln_2 \frac{1}{w_{i^*}} \leq 2 \ln_2 i^* + \ln_2 \frac{\pi^2}{6} = O(\ln i^*)$

Exercise 7. Expliquez le lien entre Hedge et Adaboost.

4. Régularisation

Exercise 8. (Régression Logistique). Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$. The simple logistic regression problem can be formulated as $\arg\min_{\theta_0, \theta_1} -\sum_{i=1}^{N} \log p(y_i \mid x_i, \theta)$. Recall that $p(Y = 1 \mid X = x, \theta) = \frac{1}{1+e^{-\theta_0 - \theta_1 x}}$. Assume we have only two examples $(x_1, y_1) = (-1, 0)$ and $(x_2, y_2) = (1, 1)$.

(1) Let $\theta_0 = 0$. What is the optimal value of θ_1 ?
  Correction: c'est l'infini

(2) What is the problem here ? What is the solution ?
Correction: le problème est mal posé, puisqu'on cherche à minimiser une fonction convexe strictement décroissante sur $\mathbb{R}$. Il faut ajouter une régularisation, par exemple $\lambda \|\theta_1\|^2$
(3) Can the same behavior happen in linear regression ? why ?
Correction: ce problème précis n'arrive pas puisque la fonction objectif en régression linéaire n'est pas strictement décroissante sur $\mathbb{R}$. En revanche, le problème peut être sous-déterminé, auquel cas la régularisation va restreindre l'espace des solutions.

5. Arbres de Décision et Ensembles de Classifieurs

Exercise 9. (Arbres de Décision).

(1) Assume we want to build a decision tree on a regression problem, minimizing the absolute deviation loss $l(y, \hat{y}) = |y - \hat{y}|$. Derive the appropriate criterion for decision trees.
Correction
Supposons que k exemples atteigne une feuille f de l'arbre et que ces exemples aient les étiquettes $y_1 \dots y_k \in \mathbb{R}$. La meilleure prédiction que cette feuille peut produire est celle qui minimise la loss cumulée $\arg\min_{\hat{y}} \sum_{i=1}^{k} |y_i - \hat{y}|$, ce qui correspond à la médiane de $y_1 \dots y_k$. Le critère qu'on choisira pour évaluer les feuilles sera donc la perte moyenne subie par cette feuille avec cette prédiction, soit: $\frac{1}{k} \sum_{i=1}^{k} |y_i - mediane\,(y_1 \dots y_k)|$
(2) Apply it on this dataset:

$x^{(1)}$	$x^{(2)}$	y
0	1	10
1	0	15
1	1	0

Exercise 10. (Bagging 1-ppv). Supposons qu'on applique le Bagging au 1-plus-proche-voisin en régression. Quand $N \to \infty$ et que le nombre de sacs du bagging tend aussi vers l'infini, à quoi est ce que ce bagging est équivalent ? Pour vous aider, commencez par calculer la probabilité que le $i^{ième}$ plus proche voisin d'un point x particulier soit dans un sac arbritraire.

Correction. de l'exercice 10. Soit S, un sample de taille N. Soit m échantillons $B_1 \dots B_m$ tiré de S par bootstrap, avec $m \to \infty$ et $N \to \infty$. Soit $x \in \mathcal{X}$, et $(x_{(1)}, y_{(1)}) \dots (x_{(N)}, y_{(N)})$, les points de S triés par ordre croissant de distance à x.

$$
\begin{aligned}
&\text{“}x_{(i)} \text{ est le plus proche voisin de } x \text{ dans } B_1\text{”} \\
\equiv &\text{“}x_{(1)} \notin B_1, \dots, x_{(i-1)} \notin B_1\text{” et “}x_{(i)} \in B_1\text{”} \\
\equiv &\text{“}x_{(1)} \notin B_1, \dots, x_{(i-1)} \notin B_1\text{” et non “}x_{(1)} \notin B_1, \dots x_{(i)} \notin B_1\text{”}
\end{aligned}
$$

Comme le second évennement “$x_{(1)} \notin B_1, \dots x_{(i)} \notin B_1$” est inclus dans le premier “$x_{(1)} \notin B_1, \dots x_{(i-1)} \notin B_1$”, on peut écrire

$$
\begin{aligned}
&P\left(x_{(i)} \text{ est le plus proche voisin de } x \text{ dans } B_1\right) \\
&= P\left(x_{(1)} \notin B_1, \dots x_{(i-1)} \notin B_1\right) - P\left(x_{(1)} \notin B_1, \dots x_{(i)} \notin B_1\right) \\
&= \left(1 - \frac{i-1}{N}\right)^N - \left(1 - \frac{i}{N}\right)^N \\
&\approx e^{-(i-1)} - e^{-i} = (e-1)e^{-i} \approx 1.71e^{-i}
\end{aligned}
$$

(Note: on peut vérifier que cette probabilité somme bien à 1 pour $i = 1 \dots \infty$)

Le nombre de fois que le point $x_{(i)}$ est choisi comme plus proche voisin dans les m sacs est donc $m \times w(i)$ avec $w_{(i)} = 1.71e^{-i}$.

Donc le 1-ppv baggé est équivalent à un plus-proche voisin pondéré $\hat{h}(x) = \sum_{i=1}^N w_{(i)} y_{(i)}$.

Exercise 11. (Gradient Boosting L2). Appliquez le gradient boosting à la perte mean-square-error, en détaillant l'algorithme obtenu.

Correction de l'exercice 11.

Voici l'algorithme générique (avec α=pas d'apprentissage):

(1) Soit $h_0(x) = constant$ le classifieur de base, et soit $f_0 = h_0$
(2) Pour $t = 1 \dots T$ Faire
  (a) Pour tout $i \in 1 \dots N$,
    (i) $\hat{y}_i = f_{t-1}(x_i)$
    (ii) $g_i = \frac{\partial}{\partial \hat{y}_i} \sum_{i=1}^N \ell(\hat{y}_i, y_i)$
  (b) Apprendre le modèle $h = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^N \left(-\alpha g_i - h(x_i)\right)^2$. Cela se fait en lançant une régression sur le dataset $\{(x_i, -\alpha g_i)\}_{i=1}^N$.
  (c) $f_t = f_{t-1} + h$
(3) Retourner f_T

Pour la fonction de perte $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$, on a $\frac{\partial}{\partial \hat{y}_i} \sum_{i=1}^N \ell(\hat{y}_i, y_i) = (\hat{y}_i - y_i)$

Donc, l'algorithme devient:

(1) Soit $h_0(x) = constant$ le classifieur de base, et soit $f_0 = h_0$
(2) Pour $t = 1 \dots T$ Faire
  (a) Apprendre le modèle $h = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^N \left(\alpha\left(y_i - f_{t-1}(x_i)\right) - h(x_i)\right)^2$. Cela se fait en lançant une régression sur le dataset $\{(x_i, \alpha\left(y_i - f_{t-1}(x_i)\right))\}_{i=1}^N$.
  (b) $f_t = f_{t-1} + h$
(3) Retourner f_T