IFT3395 Devoir2 Théorique

Byungsuk Min (20234231)
Shayan Nicolas Hollet (20146766)

October 2024

1 Question 1

$$
\begin{aligned}
E\left[(h_p(x') - y)^2\right] &= E\left[(h_p(x') - f(x') - \epsilon)^2\right] \\
&= E\left[(h_p(x') - f(x'))^2\right] - 2E\left[(h_p(x') - f(x'))\epsilon\right] + E[\epsilon^2] \\
&= E\left[(h_p(x') - f(x'))^2\right] - 2E\left[(h_p(x') - f(x'))\right] E[\epsilon] + E[\epsilon^2]
\end{aligned}
$$

car E[x + y] = E[x] + E[y] et E[ax] = a E[x].

$$
\begin{aligned}
&= E\left[(h_D(x') - f(x'))^2\right] - 2E\left[(h_D(x') - f(x'))\right] E[\epsilon] + \sigma^2 \\
&\quad \text{car } \epsilon \sim \mathcal{N}(0, \sigma^2). \\
&= E\left[(h_D(x') - f(x'))^2\right] + \sigma^2 \\
&\quad \text{car } h_D(x') \text{ dépend uniquement de } D \text{ mais pas de } \epsilon, \\
&\quad f(x') \text{ est une distribution inconnue et indépendante de } \epsilon, \\
&\quad E[\epsilon] = 0 \text{ où } \epsilon \sim \mathcal{N}(0, \sigma^2).
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[(h_D(x') - E[h_D(x')] + E[h_D(x')] - f(x'))^2\right] + \sigma^2 \\
&\quad \text{car } E[(h_D(x'))] - E[h_D(x')] = 0 \\
&= E\left[(h_D(x') - E[h_D(x')])^2 + 2\left(h_D(x') - E[h_D(x')]\right)\left(E[h_D(x')] - f(x')\right)\right. \\
&\quad \left. + (E[h_D(x')] - f(x'))^2\right] + \sigma^2 \\
&= E\left[(h_D(x') - E[h_D(x')])^2\right] + 2E\left[(h_D(x') - E[h_D(x')])\left(E[h_D(x')] - f(x')\right)\right] \\
&\quad + E\left[(E[h_D(x')] - f(x'))^2\right] + \sigma^2.
\end{aligned}
$$

Pour $2E\left[(h_D(x') - E[h_D(x')])\left(E[h_D(x')] - f(x')\right)\right]$,
$E[h_D(x')] - f(x') = \text{constant}$, car $E[h_D(x')]$: moyenne de $h_D(x')$ sur D,
$f(x')$: valeur de la fonction pour x'.

$$
\begin{aligned}
&= 2\left(E[h_D(x')] - f(x')\right) E\left[h_D(x') - E[h_D(x')]\right] \\
&= 2\left(E[h_D(x')] - f(x')\right)\left(E[h_D(x')] - E[h_D(x')]\right) \\
&\quad \text{car } E[x + y] = E[x] + E[y], E[E[x]] = E[x], \\
&= 0 \quad \text{car } E[h_D(x')] - E[h_0(x')] = 0.
\end{aligned}
$$

$$\Rightarrow E\left[(h_D(x') - y)^2\right] = E\left[\left(h_D(x') - E[h_D(x')]\right)^2\right] + \left(E[h_D(x')] - f(x')\right)^2 + \sigma^2$$
$$= \text{variance} + (\text{biais})^2 + \text{bruit}.$$

2 Question 2

a) i)

Pour $j \in \{1, \dots, d\}$:

$$\frac{\partial L(w_i(x_i, y_i))}{\partial w_j} = \frac{\partial L(w_i(x_i, y_i))}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_j} \quad \text{par la règle de la chaîne.}$$

$$\begin{aligned}\text{Pour } \frac{\partial L(W; (x_i, y_i))}{\partial z_i} &= \frac{\partial}{\partial z_i}\left(-y_i \log(\sigma(z_i))\right) + \frac{\partial}{\partial z_i}\left(-(1-y_i)\log(1-\sigma(z_i))\right) \\ &= \underbrace{\frac{\partial}{\partial z_i}\left(-y_i \log(\sigma(z_i))\right)}_{A} + \underbrace{\frac{\partial}{\partial z_i}\left(-(1-y_i)\log(1-\sigma(z_i))\right)}_{B}.\end{aligned}$$

$$\begin{aligned}A : \frac{d}{dz_i}\left(-y_i \log(\sigma(z_i))\right) &= -y_i \cdot \frac{d\log(\sigma(z_i))}{d\sigma(z_i)} \cdot \frac{d\sigma(z_i)}{dz_i} \\ &= -y_i \cdot \frac{1}{\sigma(z_i)} \cdot \sigma(z_i)(1-\sigma(z_i)) \\ &= -y_i\left(1-\sigma(z_i)\right)\end{aligned}$$

$$\begin{aligned}B : \frac{d}{dz_i}\left(-(1-y_i)\log(1-\sigma(z_i))\right) \\ &= -(1-y_i) \cdot \frac{d\log(1-\sigma(z_i))}{d(1-\sigma(z_i))} \cdot \frac{d(1-\sigma(z_i))}{dz_i} \\ &= -(1-y_i) \cdot \frac{-1}{1-\sigma(z_i)} \cdot \sigma(z_i)(1-\sigma(z_i)) \\ &= (1-y_i)\sigma(z_i)\end{aligned}$$

$$\begin{aligned}\Rightarrow \frac{\partial L(w_i; (x_i, y_i))}{\partial z_i} &= \left(-y_i\left(1-\sigma(z_i)\right) + (1-y_i)\sigma(z_i)\right) \\ &= \left(-y_i + y_i\sigma(z_i) + \sigma(z_i) - y_i\sigma(z_i)\right) \\ &= \sigma(z_i) - y_i\end{aligned}$$

$$\text{Pour } \frac{\partial z_i}{\partial w_j} = \sum_{j=1}^{d} \frac{\partial}{\partial w_j}(w_j x_{ij}) = x_{ij}$$

$$\begin{aligned}\Rightarrow \frac{\partial}{\partial w_j} L(w; (x_i, y_i)) &= (\sigma(z_i) - y_i)x_{ij} \quad \text{pour } j \text{ quelconque} \\ \Rightarrow \nabla_w L(w; (x_i, y_i)) &= (\sigma(z_i) - y_i)x_i \quad \text{pour tous les } w_j.\end{aligned}$$

(ii)

$$\begin{aligned}
\frac{d\mathcal{J}(w)}{dw} &= \frac{d}{dw}\left(\frac{1}{n}\sum_{i=1}^{n} L(w;(x_i,y_i))\right) + \frac{d}{dw}\left(\frac{\lambda}{2}\|w\|^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{dw}\left(L(w;(x_i,y_i))\right) + \frac{\lambda}{2}\sum_{j=1}^{d}\frac{d}{dw}(w_j^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}\nabla_w L(w;(x_i,y_i)) + \frac{\lambda}{2}\sum_{j=1}^{d} 2w_j \\
&= \frac{1}{n}\sum_{i=1}^{n}(\sigma(z_i)-y_i)x_i + \lambda\sum_{j=1}^{d} w_j \quad \text{par la question (i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\sigma(z_i)-y_i)x_i + \lambda w \\
\Rightarrow \nabla_w\mathcal{J}(w) &= \frac{1}{n}\sum_{i=1}^{n}(\sigma(z_i)-y_i)x_i + \lambda w.
\end{aligned}$$

b)

La règle de mise à jour est

$$\text{w} \leftarrow w - \eta\nabla_w\mathcal{J}(w)$$

En utilisant les résultats obtenus dans la question précédente, on a :

$$\text{w} \leftarrow w - \eta\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\sigma(z_i)-y_i)x_i + \lambda w\right)$$

$$\text{ou} \quad \text{w} \leftarrow w - \eta\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\sigma(z_i)-y_i)x_i\right) + \eta\lambda w$$

c) Le paramètre λ dans la régularisation L_2 contrôle la taille des poids et joue un rôle clé dans la prévention du surapprentissage.

1. **Cas de λ élevé :**
   - La régularisation est forte, ce qui réduit la taille des poids. Cela permet de limiter la complexité du modèle et de prévenir le surapprentissage.
   - Il introduit un biais plus fort car les poids sont contraints, mais cela réduit la variance du modèle, le rendant moins sensible aux petites fluctuations des données.
2. **Cas de λ faible ou nul :**
   - Les poids peuvent croître sans pénalité, ce qui peut conduire à un modèle plus complexe, mais avec un risque de surapprentissage.
   - Il réduit le biais, mais peut augmenter la variance, car le modèle devient plus flexible et peut s'adapter aux bruits des données.

3 Question 3

a)

On sait que :

$$\ell(f(x), y) = \mathbf{1}_{[f(x)\neq y]} = \begin{cases} 1 & \text{si } f(x) \neq y \\ 0 & \text{sinon} \end{cases}$$

Et l'espérance d'une fonction indicatrice est définie par :

$$E\left[\mathbf{1}_{[A]}\right] = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$$

Dans notre cas :

$$E_{(x,y)\sim\rho}\left[\ell(f(x), y)\right] = 1 \cdot P_{(x,y)\sim\rho}(f(x) \neq y) + 0 \cdot P_{(x,y)\sim\rho}(f(x) = y)$$

$$= P_{(x,y)\sim\rho}(f(x) \neq y)$$

b)

L'événement $g(x) = y$ signifie une bonne prédiction.
On obtient une bonne prédiction lorsque :

$$\begin{aligned} Y = 1 \text{ et } g(x) = 1 \quad &\text{avec} \quad P(Y = 1 \mid x) = \eta(x) \\ Y = 0 \text{ et } g(x) = 0 \quad &\text{avec} \quad P(Y = 0 \mid x) = 1 - \eta(x). \end{aligned}$$

Alors, on peut écrire pour l'événement $g(x) \neq y$ (qui est l'erreur) comme suit :

$$P(g(x) \neq y) = 1 - P(g(x) = y)$$

$$= 1 - \left[\mathbf{1}_{\{g(x)=1\}} P(Y = 1 \mid x) + \mathbf{1}_{\{g(x)=0\}} P(Y = 0 \mid x)\right]$$

$$= 1 - \left[\mathbf{1}_{\{g(x)=1\}} \eta(x) + \mathbf{1}_{\{g(x)=0\}} (1 - \eta(x))\right].$$

c)

Pour l'événement $g(x) \neq Y$:

- Pour $g(x) = 1$ et $Y = 0$, $g(x) \neq Y$,

$$P(g(x) \neq Y \mid x) = P(Y = 0 \mid x) = 1 - \eta(x)$$

- Pour $g(x) = 0$ et $Y = 1$, $g(x) \neq Y$,

$$P(g(x) \neq Y \mid x) = P(Y = 1 \mid x) = \eta(x)$$

D'où,

$$P(g(x) \neq Y \mid x) = \begin{cases} 1 - \eta(x) & \text{si } g(x) = 1 \\ \eta(x) & \text{si } g(x) = 0 \end{cases}$$

Ce qui peut être exprimé en utilisant des indicateurs :

$$P(g(x) \neq Y \mid x) = 1_{\{g(x)=1\}}(1 - \eta(x)) + 1_{\{g(x)=0\}}\eta(x)$$

Pour l'événement $f^*(x) \neq Y$:

- Pour $\eta(x) = P(Y = 1 \mid x) \geq \frac{1}{2}$, alors $f^*(x) = 0 \neq Y$,

$$P(f^*(x) \neq Y \mid x) = P(Y = 1 \mid x) = \eta(x)$$

- Pour $\eta(x) = P(Y = 0 \mid x) < \frac{1}{2}$, alors $f^*(x) = 1 \neq Y$,

$$P(f^*(x) \neq Y \mid x) = P(Y = 0 \mid x) = 1 - \eta(x)$$

D'où,

$$P(f^*(x) \neq Y \mid x) = \begin{cases} 1 - \eta(x) & \text{si } f^*(x) = 1 \\ \eta(x) & \text{si } f^*(x) = 0 \end{cases}$$

Ce qui peut aussi être écrit en utilisant des indicateurs :

$$P(f^*(x) \neq Y \mid x) = 1_{\{f^*(x)=1\}}(1 - \eta(x)) + 1_{\{f^*(x)=0\}}\eta(x)$$

Alors on a :

$$P(g(x) \neq Y \mid x) - P(f^*(x) \neq Y \mid x)$$

$$= 1_{\{g(x)=1\}}(1-\eta(x)) + 1_{\{g(x)=0\}}\eta(x) - 1_{\{f^*(x)=1\}}(1-\eta(x)) - 1_{\{f^*(x)=0\}}\eta(x)$$

$$= 1_{\{g(x)=1\}}(1-\eta(x))+(1-1_{\{g(x)=1\}})\eta(x)-1_{\{f^*(x)=1\}}(1-\eta(x))-(1-1_{\{f^*(x)=1\}})\eta(x)$$

par l'énoncé

$$= 1_{\{g(x)=1\}}(1-\eta(x))-1_{\{f^*(x)=1\}}(1-\eta(x))+\eta(x)-1_{\{g(x)=1\}}\eta(x)-\eta(x)+1_{\{f^*(x)=1\}}\eta(x)$$

$$= (1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}})\eta(x) + (1_{\{g(x)=1\}} - 1_{\{f^*(x)=1\}})(1-\eta(x))$$

$$= (1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}})\eta(x) - (1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}})(1-\eta(x))$$

$$= (1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}})\left(\eta(x) - (1-\eta(x))\right)$$

$$= (2\eta(x)-1)\left(1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}}\right)$$

(d)

Si $\eta(x) \geq \frac{1}{2}$, alors $(2\eta(x)-1) \geq 0$, et $f^*(x) = 1$:

$$\Rightarrow 1_{\{f^*(x)=1\}} = 1$$

$$\Rightarrow 1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}} \geq 0$$

$$\Rightarrow (2\eta(x)-1)(1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}}) \geq 0$$

Si $\eta(x) < \frac{1}{2}$, alors $(2\eta(x)-1) < 0$, et $f^*(x) = 0$:

$$\Rightarrow 1_{\{f^*(x)=1\}} = 0$$

$$\Rightarrow 1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}} \leq 0$$

$$\Rightarrow (2\eta(x)-1)(1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}}) \geq 0$$

Donc, pour tout $g : X \to Y$:

$$(2\eta(x)-1)(1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}}) \geq 0$$

(e)

D'après la question précédente :

$$\forall g : X \to Y, \quad (2\eta(x) - 1)\left(1_{\{f^*(x)=1\}} - 1_{\{g(x)=1\}}\right) \geq 0$$

Cela implique :

$P(g(x) \neq Y \mid X = x) - P(f^*(x) \neq Y \mid X = x) \geq 0$ par la question 3.c.

Ce qui donne :

$$P(g(x) \neq Y \mid X = x) \geq P(f^*(x) \neq Y \mid X = x)$$

En intégrant sur toutes les valeurs de x (loi des probabilités totales) :

$$\int P(g(x) \neq Y \mid X = x)p(x)dx \geq \int P(f^*(x) \neq Y \mid X = x)p(x)dx$$

D'où :

$$P(g(x) \neq Y) \geq P(f^*(x) \neq Y)$$

On en conclut que :

$$\forall g : X \to Y, \quad R(g) \geq R(f^*) \quad \text{par la question 3.a.}$$

4 Question 4

a)
Le risque d'une hypothèse h pour un problème de régression avec la fonction

de coût carrée quadratique est défini par :

$$R(h) = E_{(x,y)\sim p}[(y - h(x))^2]$$

où y est la variable sortie et $h(x)$ la prédiction de y pour une entrée x.

(b)

On a :

$$E_D[\text{erreur}_{\text{LOO}}] = E_D\left[\frac{1}{n}\sum_{i=1}^{n} L(h_{D\setminus\{i\}}(x_i), y_i)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E_D\left[(y_i - h_{D\setminus\{i\}}(x_i))^2\right]$$

Pour chaque terme $E_D\left[(y_i - h_{D\setminus\{i\}}(x_i))^2\right]$, chaque point (x_i, y_i) est échantillonné indépendamment et que $h_{D\setminus\{i\}}$ est un modèle entraîné sur les $n-1$ autres points, cette espérance est équivalente à l'erreur moyenne d'un modèle entraîné sur un ensemble D' de $n-1$ points, évalué sur un point indépendant (x, y). Ainsi :

$$E_D\left[(y_i - h_{D\setminus\{i\}}(x_i))^2\right] = E_{D',(x,y)\sim p}\left[(y - h_{D'}(x))^2\right]$$

Puisque cette espérance est identique pour chaque i, on a :

$$E_D[\text{erreur}_{\text{LOO}}] = \frac{1}{n}\sum_{i=1}^{n} E_{D',(x,y)\sim p}\left[(y - h_{D'}(x))^2\right]$$

Comme cette quantité est identique pour tous les i, on peut simplifier comme suit :

$$E_D[\text{erreur}_{\text{LOO}}] = E_{D',(x,y)\sim p}\left[(y - h_{D'}(x))^2\right]$$

L'erreur *leave-one-out* (VCLOO) imite la performance du modèle sur des données non vues en excluant successivement chaque point du jeu d'entraînement. L'égalité démontrée montre que VCLOO estime en moyenne le même type d'erreur que le risque réel $R(h_D)$, c'est-à-dire l'erreur sur des données futures.

Elle est dite presque non-biaisée car l'exclusion d'un point modifie légèrement l'ensemble d'entraînement, mais cet effet diminue lorsque n devient grand.

L'espérance de l'erreur *leave-one-out* (VCLOO) est un estimateur presque non-biaisé du risque réel $R(h_D)$, car elle simule l'évaluation du modèle sur des

données non vues. À mesure que la taille des données augmente, l'erreur *leave-one-out* devient de plus en plus proche du risque réel du modèle.

c)
La solution optimale de la régression linéaire est donnée par :

$$w^* = (X^T X)^{-1} X^T y \quad \text{où} \quad w \in R^d, \quad X \in R^{n \times d}, \quad y \in R^n$$

- Pour $X^T X$ avec $X^T \in R^{d \times n}$ et $X \in R^{n \times d}$, on obtient une matrice de taille $d \times d$ et sa complexité de calcul est $O(nd^2)$.
- Pour $X^T y$ avec $y \in R^n$, on obtient un vecteur de taille d et sa complexité de calcul est $O(nd)$.
- Pour $(X^T X)^{-1}$ avec $(X^T X) \in R^{d \times d}$, on obtient une matrice de taille $d \times d$ et sa complexité de calcul est $O(d^3)$ par l'énoncé.
- Pour $(X^T X)^{-1} X^T y$ avec $(X^T X)^{-1} \in R^{d \times d}$ et $X^T y \in R^d$, on obtient un vecteur de taille d et sa complexité est $O(d^2)$.

Puisque ces opérations s'effectuent de façon indépendante, la complexité totale du calcul de la solution de la régression linéaire est :

$$O(nd^2 + nd + d^3 + d^2)$$

$$\Rightarrow O(nd^2 + d^3)$$

d)
L'expression de l'erreur VCLOO pour la régression est :

$$\text{erreur}_{\text{VCLOO}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - h_{n-1}(x_i))^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (y_i - w_{D\setminus i}^T x_i)^2$$

Pour chaque point i, on doit recalculer ${w_{D\setminus i}^*}^T = (X_{D\setminus i}^T X_{D\setminus i})^{-1} X_{D\setminus i}^T y_{D\setminus i}$ en réentraînant le modèle sur l'ensemble des points sauf (x_i, y_i).
Pour $w_{D\setminus i}^T x_i$, on a $O((n-1)(d-1)^2 + (d-1)^3) \approx O(nd^2 + d^3)$ par Question 3c)
Pour $w_{D\setminus i}^T x_i$, on a $O(n-1) \approx O(n)$
Pour $y_i - w_{D\setminus i}^T x_i$, on a $O(n-1) \approx O(n)$

$$\Rightarrow \text{Pour chaque } i, \quad O(nd^2 + d^3 + n + n) \approx O(nd^2 + d^3)$$

$$\Rightarrow \text{Pour tout } i, \quad O(n(nd^2 + d^3)) = O(n^2 d^2 + nd^3)$$

$$\Rightarrow \text{Pour chaque } i, \quad O(nd^2 + d^3 + nd + d^2) = O(nd^2 + d^3)$$

$$\Rightarrow \text{Pour tout } i, \quad O(n(nd^2 + d^3)) = O(n^2d^2 + nd^3)$$

e)

En régression linéaire, la prédiction pour chaque point est donnée par :

$$\hat{y} = Xw = Hy \quad \text{où} \quad H = X(X^TX)^{-1}X^T : \text{matrice de projection}$$

$$\hat{y} : \text{vecteur des prédictions}$$

$$y : \text{vecteur des vraies valeurs}$$

La prédiction VCLOO $\hat{y}_i^{(i)}$ est donnée par :

$$\hat{y}_i^{(i)} = \frac{\hat{y}_i - h_i y_i}{1 - h_i} \quad \text{où} \quad h_i : \text{i-ème diagonale de la matrice } H$$

$$\hat{y}_i : \text{prédiction initiale pour } x_i, \text{ calculée avec tous les points}$$

Le résidu VCLOO est la différence entre la vraie valeur y_i et la prédiction VCLOO $\hat{y}_i^{(i)}$ donnée par :

$$\text{résidu}_{\text{VCLOO}} = y_i - \hat{y}_i^{(i)} = y_i - \frac{\hat{y}_i - h_i y_i}{1 - h_i}$$

$$= \frac{y_i - \hat{y}_i}{1 - h_i}$$

L'erreur quadratique VCLOO pour chaque point est donnée par :

$$\text{erreur}_{\text{VCLOO}} = \left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2 \quad \text{pour } i$$

Alors l'erreur totale VCLOO pour tous les points i est donnée par :

$$\text{erreur}_{\text{VCLOO}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2 \quad \forall i$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - w^T x_i}{1 - x_i^T(X^TX)^{-1}x_i}\right)^2$$

Complexité du calcul :

- Pour w^*, sa complexité est $O(nd^2 + d^3)$ par la question 3c).

- Pour $w^T x_i$, avec $w^* \in R^d$ et $x_i \in R^d$, sa complexité est $O(d)$.
- Pour $y_i = w^T x_i$, avec $i \in \{1, ..., n\}$, sa complexité est $O(nd)$.
- Pour $h_i = x_i^T (X^T X)^{-1} x_i$, avec $i \in \{1, ..., n\}$, sa complexité est $O(nd^2)$.
- Pour $1 - h_i$, avec $i \in \{1, ..., n\}$, sa complexité est $O(n)$.
- Pour $\frac{(y_i - w^T x_i)^2}{1-h_i}$, sa complexité est $O(1)$ pour un i
  $\Rightarrow O(n)$ pour tout i.

Donc la complexité totale du calcul est :

$$O(nd^2 + d^3 + d + nd + nd^2 + n + n)$$

$$\Rightarrow O(nd^2 + d^3)$$