IFT 6390A Fundamentals of Machine Learning
Guillaume Rabusseau

Devoir 2 - Partie Théorique

Clément DETRY
Hamed Nazim MAMACHE

- Ce devoir doit être fait et envoyé sur Gradescope par équipes d'au plus 2 étudiant.e.s. Vous pouvez discuter avec d'autres étudiants mais les réponses que vous soumettez doivent être les vôtres. A noter que nous utiliserons l'outil de détection de plagiat de Gradescope. Tous les cas suspectés de plagiat seront enregistrés et transmis à l'Université pour vérification.
- Vous devez soumettre vos solutions au format pdf sur Gradescope en utilisant le devoir intitulé `Devoir 2 - Partie Théorique`.

1. **Décomposition biais/variance** [5 points]

   Considérons les données générées de la manière suivante: une donnée x est échantillonnée à partir d'une distribution inconnue, et nous observons la mesure correspondante y générée d'après la formule

   $$y = f(x) + \epsilon,$$

   où f est une fonction déterministe inconnue et $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Ceci définit une distribution sur les données x et mesures y, nous notons cette distribution p.

   Étant donné un ensemble d'entraînement $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ échantillonné i.i.d. à partir de p, on définit l'hypothèse h_D qui minimise le risque empirique donné par la fonction de coût erreur quadratique. Plus précisément,

   $$h_D = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} (y_i - h(x_i))^2$$

   où $\mathcal{H}$ est l'ensemble d'hypothèses (ou classe de fonction) dans lequel nous cherchons la meilleure fonction/hypothèse.

L'erreur espérée[1] de h_D sur un point donné (x', y') est notée $\mathbb{E}[(h_D(x') - y')^2]$. Deux termes importants qui peuvent être définis sont:

- Le biais, qui est la différence entre l'espérance de la valeur donnée par notre hypothèse en un point x' et la vraie valeur donnée par $f(x')$. Plus précisément,

$$biais = \mathbb{E}[h_D(x')] - f(x')$$

- La variance, est une mesure de la dispersion des hypothèse apprises sur des ensemble de données différents, autour de la moyenne $\mathbb{E}[h_D(x')]$. Plus précisément,

$$variance = \mathbb{E}[(h_D(x') - \mathbb{E}[h_D(x')])^2]$$

Montrez que l'erreur espérée pour un point donné (x', y') peut être décomposée en une somme de 3 termes: $(biais)^2$, *variance*, et un terme de *bruit* qui implique ϵ. Vous devez justifier toutes les étapes de dérivation.

Réponse :
On cherche à calculer l'erreur espérée pour un point donné (x', y'), soit $\mathbb{E}[(h_D(x') - y')^2]$:

$$\mathbb{E}[(h_D(x') - y')^2] = \mathbb{E}[h_D(x')^2 - 2h_D(x')y' + y'^2]$$

Or $y' = f(x') + \epsilon$. D'où :

$$\mathbb{E}[(h_D(x') - y')^2] = \mathbb{E}[h_D(x')^2 - 2h_D(x')f(x') - 2h_D(x')\epsilon + f(x')^2 + 2f(x')\epsilon + \epsilon^2]$$

Sachant que f est une fonction déterministe,
on a alors : $\mathbb{E}[f(x')] = f(x')$.
D'où :

$$\mathbb{E}[(h_D(x') - y')^2] = \mathbb{E}[h_D(x')^2] - 2f(x')\,\mathbb{E}[h_D(x')] - 2\,\mathbb{E}[h_D(x')\epsilon] + f(x')^2 + 2f(x')\,\mathbb{E}[\epsilon] + \mathbb{E}[\epsilon^2]$$

Or $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Donc : $\mathbb{E}[\epsilon] = 0$ et $\mathbb{E}[\epsilon^2] = \sigma^2$.
En effet :

$$\begin{aligned}\mathbb{E}[\epsilon^2] &= \mathbb{E}[\epsilon]^2 + \mathrm{Var}(\epsilon) \\ &= 0^2 + \sigma^2 = \sigma^2\end{aligned}$$

[1] Ici l'espérance porte sur le choix aléatoire d'un ensemble d'entraînement D de n points tirés à partir de la distribution inconnue p. Par exemple (et plus formellement) : $\mathbb{E}[(h_D(x')] = \mathbb{E}_{(x_1,y_1)\sim p} \cdots \mathbb{E}_{(x_n,y_n)\sim p}\, \mathbb{E}[(h_{\{(x_1,y_1),\ldots,(x_n,y_n)\}}(x')]$.

De plus :

$$\begin{aligned} -2\,\mathbb{E}[h_D(x')\epsilon] &= -2\,\mathbb{E}[h_D(x')]\,\mathbb{E}[\epsilon] \\ &= -2\,\mathbb{E}[h_D(x')] \times 0 \qquad\qquad = 0 \end{aligned}$$

Ainsi, on a à ce stade:

$$\mathbb{E}[(h_D(x') - y')^2] = \mathbb{E}[h_D(x')^2] - 2f(x')\,\mathbb{E}[h_D(x')] + f(x')^2 + \sigma^2 \quad (1)$$

De plus, on sait que :

$$\begin{aligned} \mathit{variance} &= \mathbb{E}[(h_D(x') - \mathbb{E}[h_D(x')])^2] \\ &= \mathbb{E}[h_D(x')^2 - 2h_D(x')\,\mathbb{E}[h_D(x')] + \mathbb{E}[h_D(x')]^2] \\ &= \mathbb{E}[h_D(x')^2] - 2\,\mathbb{E}[h_D(x')\,\mathbb{E}[h_D(x')]] + \mathbb{E}[\mathbb{E}[h_D(x')]^2] \\ &= \mathbb{E}[h_D(x')^2] - 2\,\mathbb{E}[h_D(x')]\,\mathbb{E}[\mathbb{E}[h_D(x')]] + \mathbb{E}[\mathbb{E}[h_D(x')]^2] \\ &= \mathbb{E}[h_D(x')^2] - 2\,\mathbb{E}[h_D(x')]^2 + \mathbb{E}[h_D(x')]^2 \end{aligned}$$

Puisque : $\mathbb{E}[\mathbb{E}[h_D(x')]] = \mathbb{E}[h_D(x')]$ et $\mathbb{E}[\mathbb{E}[h_D(x')]^2] = \mathbb{E}[h_D(x')]^2$.
On a alors :

$$\begin{aligned} \mathit{variance} &= \mathbb{E}[h_D(x')^2] - 2\,\mathbb{E}[h_D(x')]^2 + \mathbb{E}[h_D(x')]^2 \\ &= \mathbb{E}[h_D(x')^2] - \mathbb{E}[h_D(x')]^2 \end{aligned}$$

D'où :

$$\mathbb{E}[h_D(x')^2] = \mathit{variance} + \mathbb{E}[h_D(x')]^2$$

.
Ainsi, en remplaçant dans (1) :

$$\begin{aligned} \mathbb{E}[(h_D(x') - y')^2] &= \mathbb{E}[h_D(x')^2] - 2f(x')\,\mathbb{E}[h_D(x')] + f(x')^2 + \sigma^2 \\ &= \mathit{variance} + \mathbb{E}[h_D(x')]^2 - 2f(x')\,\mathbb{E}[h_D(x')] + f(x')^2 + \sigma^2 \end{aligned}$$

En remarquant que :

$$\begin{aligned} \mathbb{E}[h_D(x')]^2 - 2f(x')\,\mathbb{E}[h_D(x')] &= \mathbb{E}[h_D(x')](\mathbb{E}[h_D(x')] - 2f(x')) \\ &= \mathbb{E}[h_D(x')](\mathit{biais} - f(x')) \end{aligned}$$

On a au final :

$$\begin{aligned} \mathbb{E}[(h_D(x') - y')^2] &= \mathit{variance} + \mathbb{E}[h_D(x')](\mathit{biais} - f(x')) + f(x')^2 + \sigma^2 \\ &= \mathit{variance} + \mathbb{E}[h_D(x')]\mathit{biais} - \mathbb{E}[h_D(x')]f(x') + f(x')^2 + \sigma^2 \\ &= \mathit{variance} + \mathbb{E}[h_D(x')]\mathit{biais} + f(x')(f(x') - \mathbb{E}[h_D(x')]) + \sigma^2 \\ &= \mathit{variance} + \mathbb{E}[h_D(x')]\mathit{biais} - f(x')\mathit{biais} + \sigma^2 \\ &= \mathit{variance} + \mathit{biais}(\mathbb{E}[h_D(x')] - f(x')) + \sigma^2 \\ &= \mathit{variance} + \mathit{biais}^2 + \sigma^2 \end{aligned}$$

Conclusion :
L'erreur espérée pour un point donné (x', y') est égale à :

$$\boxed{\mathbb{E}[(h_D(x') - y')^2] = variance + biais^2 + \sigma^2}$$

2. **Fonctions de transformation des données** [6 points]

Dans cet exercice, vous allez concevoir des fonctions de transformation depuis l'espace de features original vers un espace où les données sont linéairement séparables. Pour les questions suivantes, si vous répondez '*oui*', écrivez l'expression de la transformation correspondante; et si votre réponse est '*non*', ajoutez une courte justification de votre réponse. Vous devez donner les formules explicites des transformations, et ces formules doivent utiliser uniquement des opérations mathématiques simples.

(a) Soient les données 1-D suivantes (Figure 1). Pouvez-vous proposer une transformation 1-D (i.e. vers un espace de dimension 1) qui rend les points linéairement séparables?

Figure 1: Jeu de données 1D. Les points entre $2k$ et $2k+1$ sont étiquetés par X. Les points entre $2k+1$ et $2k+2$ sont étiquetés par O.

Réponse :
Une transformation 1-D qui rend les points linéairement séparables est la suivante : $\phi(x) = |x|\ sin(\pi x)$.

(b) Soient les données 2-D suivantes (Figure 2). Pouvez-vous proposer une transformation 1-D qui rend les points linéairement séparables?

Réponse :
Une transformation 1-D qui rend les points linéairement séparables est la suivante : $\phi(x) = x_1 x_2$.

(c) En utilisant les idées que vous avez utilisées pour les deux questions précédentes, pouvez-vous proposer une transformation des données suivantes (Figure 3) qui les rendent linéairement séparables? Si votre réponse est '*oui*', donnez l'expression du noyau qui correspond à la transformation proposée. Souvenez-vous que

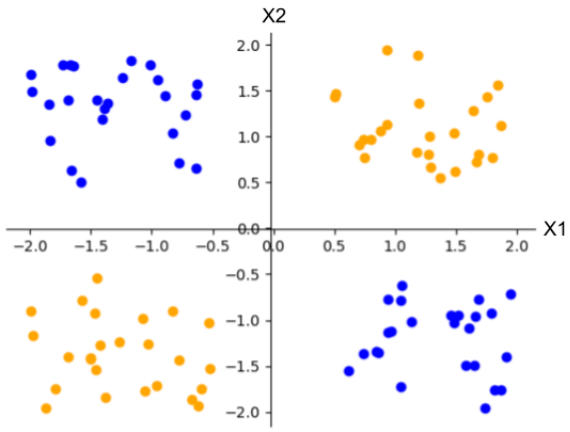

Figure 2: Jeu de données 2D.

$K(x, y) = \langle \phi(x), \phi(y) \rangle$, donc trouvez ϕ et faites le produit scalaire pour obtenir le noyau.

Réponse :

Une transformation qui rend les points linéairement séparables est la suivante : $\phi(x) = \sqrt{x_1^2 + x_2^2}\ sin(\pi\sqrt{x_1^2 + x_2^2})$.
L'expression du noyau qui correspond à la transformation proposée est la suivante :

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$
$$= (\sqrt{x_1^2 + x_2^2})(\sqrt{y_1^2 + y_2^2})\ sin(\pi\sqrt{x_1^2 + x_2^2})\ \ sin(\pi\sqrt{y_1^2 + y_2^2})$$

3. **Validation croisée "k-fold"** [10 points]

Soit $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un ensemble de données échantillonné i.i.d. à partir d'une distribution inconnue p. Pour estimer le risque (erreur de test) d'un algorithme d'apprentissage en utilisant D, la validation croisée "k-fold" utilise la *i-ème* portion des données $D_i = \{(x_j, y_j) \mid j \in \text{ind}[i]\}$ (où ind$[i]$ sont les indices des points de données dans la *i-ème* portion) pour estimer le risque de l'hypothèse retournée par un algorithme d'apprentissage entraîné sur toutes les données sauf la *i-ème* portion : $D_{\setminus i} = \{(x_j, y_j) \mid j \notin \text{ind}[i]\}$.

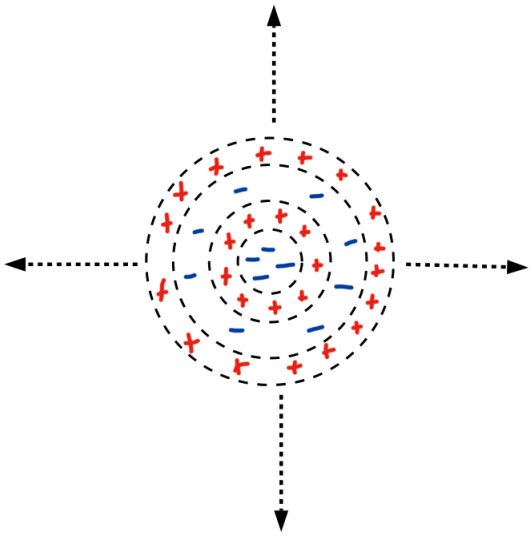

Figure 3: Un autre jeu de données 2D. Les points entre les aires de rayon $2k$ et $2k+1$ sont étiquetés par -. Les points entre les aires de rayon $2k+1$ et $2k+2$ sont étiquetés par +.

Plus précisément, si on note $h_{D\setminus i}$ l'hypothèse obtenue par l'algorithme d'apprentissage entraîné sur les données $D_{\setminus i}$, l'erreur de validation croisée k-fold est donnée par:

$$\text{error}_{k-fold} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\sum_{j\in \text{ind}[i]} l(h_{D\setminus i}(x_j), y_j)$$

où l est la fonction de perte.

Dans cet exercice, nous nous intéressons à certaines des propriétés de cet estimateur

k-fold est non biaisé

(a) Rappelez la définition du risque d'une hypothèse h pour un problème de régression avec la fonction de coût erreur quadratique

(b) En utilisant D' pour dénoter un ensemble de données de taille

$n - \frac{n}{k}$, montrez que

$$\mathop{\mathbb{E}}_{D\sim p}[\text{error}_{k-fold}] = \mathop{\mathbb{E}}_{\substack{D'\sim p,\\(x,y)\sim p}}[(y - h_{D'}(x))^2]$$

où la notation $D \sim p$ signifie que D est échantillonné i.i.d. à partir de la distribution p et où h_D est l'hypothèse obtenue par l'algorithme d'apprentissage sur les données D. Expliquez en quoi cela montre que erreur_{k-fold} est un estimateur (presque) non-biaisé du risque de h_D.

Complexité de k-fold Nous étudions maintenant la validation croisée k-fold pour la régression linéaire où les données d'entrées $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sont des vecteurs à d dimensions. Nous utilisons $\mathbf{X} \in \mathbb{R}^{n\times d}$ et $\mathbf{y} \in \mathbb{R}^n$ pour représenter la matrice des données d'entrée et le vecteur des sorties correspondantes.

(c) En considérant que la complexité en temps pour inverser une matrice de taille $m \times m$ est en $\mathcal{O}(m^3)$, quelle sera la complexité du calcul de la solution de la régression linéaire sur l'ensemble de données D?

(d) Soient $\mathbf{X}_{-i} \in \mathbb{R}^{(n-\frac{n}{k})\times d}$ and $\mathbf{y}_{-i} \in \mathbb{R}^{(n-\frac{n}{k})}$ la matrice des données d'entrées et le vecteurs des sorties obtenus en supprimant les lignes de la *i-ème* portion de $\mathbf{X}$. En utilisant la formule pour $error_{k-fold}$ mentionnée précédemment, écrivez l'expression de l'erreur de validation croisée "k-fold" pour la régression linéaire. Quelle est la complexité algorithmique du calcul de cette formule?

(e) Dans le cas particulier de la régression linéaire, l'erreur k-fold peut être calculée de manière plus efficace. Montrez que dans le cas de la régression linéaire, on a:

$$\text{error}_{k-fold} = \frac{1}{k}\sum_{i=1}^{k}\left\|\frac{\mathbf{y}_i - \mathbf{X}_i\mathbf{w}^*}{\mathbf{I} - \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top}\right\|^2$$

où $\mathbf{w}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ est la solution de la régression linéaire calculée sur tout l'ensemble de données D, la notation $\frac{\mathbf{A}}{\mathbf{B}}$ veut dire $\mathbf{A}\mathbf{B}^{-1}$ et $\mathbf{X}_i \in \mathbb{R}^{(\frac{n}{k})\times d}$ et $\mathbf{y}_i \in \mathbb{R}^{(\frac{n}{k})}$ sont la matrice des données d'entrées et le vecteur des sorties obtenus en ne gardant que les lignes de la *i-ème* portion de $\mathbf{X}$. Quelle est la complexité du calcul de cette expression?

Réponse :

(a) Définition du risque d'une hypothèse h pour un problème de régression avec la fonction de coût erreur quadratique :

$$\text{risque}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2$$

(b) Montrons que : $\mathbb{E}_{D\sim p}[\text{error}_{k-fold}] = \mathbb{E}_{\substack{D'\sim p,\\(x,y)\sim p}}[(y - h_{D'}(x))^2]$.

$$\begin{aligned}\mathop{\mathbb{E}}_{D\sim p}[\text{error}_{k-fold}] &= \mathop{\mathbb{E}}_{D\sim p}[\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\sum_{j\in\text{ind}[i]} l(h_{D\setminus i}(x_j), y_j)] \\ &= \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\sum_{j\in\text{ind}[i]}\mathop{\mathbb{E}}_{D\sim p}[l(h_{D\setminus i}(x_j), y_j)]\end{aligned}$$

Or $\mathbb{E}_{D\sim p}[l(h_{D\setminus i}(x_j), y_j)] = \mathbb{E}_{\substack{D'\sim p,\\(x,y)\sim p}}[(y - h_{D'}(x))^2]$. En effet, au lieu de tirer n échantillons de p et en retenir aléatoirement n/k, k fois, on peut tirer aléatoirment $n - n/k$, k fois, à partir de p. D'où :

$$\begin{aligned}\mathop{\mathbb{E}}_{D\sim p}[\text{error}_{k-fold}] &= \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\sum_{j\in\text{ind}[i]}\mathop{\mathbb{E}}_{\substack{D'\sim p,\\(x,y)\sim p}}[l(h_{D'}(x), y)] \\ &= \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\mathop{\mathbb{E}}_{\substack{D'\sim p,\\(x,y)\sim p}}[l(h_{D'}(x), y)]\sum_{j\in\text{ind}[i]} 1 \\ &= \frac{1}{k}\times k\times\frac{1}{n/k}\times n/k\times\mathop{\mathbb{E}}_{\substack{D'\sim p,\\(x,y)\sim p}}[l(h_{D'}(x), y)] \\ &= \mathop{\mathbb{E}}_{\substack{D'\sim p,\\(x,y)\sim p}}[l(h_{D'}(x), y)]\end{aligned}$$

Donc :

$$\boxed{\mathop{\mathbb{E}}_{D\sim p}[\text{error}_{k-fold}] = \mathop{\mathbb{E}}_{\substack{D'\sim p,\\(x,y)\sim p}}[(y - h_{D'}(x))^2]}$$

(c) La solution d'une régression linéaire s'écrit : $w^* = (X^TX)^{-1}X^Ty$,

si on considère les termes de biais déjà compris dans w^* tel que :

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & ... & x_{1,d-1} & 1 \\ x_{2,1} & x_{2,2} & ... & x_{2,d-1} & 1 \\ x_{3,1} & x_{3,2} & ... & x_{3,d-1} & 1 \\ ... & ... & ... & ... & ... \\ x_{n,1} & x_{n,2} & ... & x_{n,d-1} & 1 \end{bmatrix}$$

et

$$w^* = \begin{bmatrix} w_1 \\ w_2 \\ ... \\ w_{d-1} \\ b \end{bmatrix}$$

Calcul de la complexité :

- Calcul de $X^T X$: Multiplication de 2 matrices de taille $d \times n$ et $n \times d$: $O(nd^2)$.
- Calcul de l'inverse de $X^T X$: inversion d'une matrice de taille $d \times d$: $O(d^3)$.
- Calcul de $X^T y$: Multiplication entre une matrice de taille $d \times n$ et une matrice de taille $n \times 1$: $O(nd)$.
- Calcul de $(X^T X)^{-1}(X^T y)$: multiplication entre une matrice de taille $d \times d$ et une matrice de taille $n \times 1$: $O(d^2)$.

La complexité totale, en temps, est donc de $O(d^3) + O(nd^2)$.

(d) L'expression de l'erreur de validation croisée "k-fold" pour la régression linéaire peut s'écrire :

$$\begin{aligned} \text{error}_{k-fold} &= \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n/k} \left\| y_i - \mathbf{X}_i w^* \right\|^2 \\ &= \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n/k} \left\| y_i - \mathbf{X}_i (\mathbf{X}_{-i}^T \mathbf{X}_{-i})^{-1} \mathbf{X}_{-i}^T y_{-i} \right\|^2 \end{aligned}$$

Calcul de la complexité :

- On sait déjà que la complexité de calcul de $(\mathbf{X}_{-i}^T \mathbf{X}_{-i})^{-1} \mathbf{X}_{-i}^T y_{-i}$ est égale à $O(d^3) + O((n - n/k)d^2)$.
- Calcul de $\mathbf{X}_i (\mathbf{X}_{-i}^T \mathbf{X}_{-i})^{-1} \mathbf{X}_{-i}^T y_{-i}$: Multiplication d'une matrice de taille $n/k \times d$ par une matrice de taille $d \times 1$: $O((n/k)d^2)$.

- Calcul de $y_i - \mathbf{X}_i(\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^T y_{-i}$: différence entre deux matrices de tailles $n/k \times 1$: $O(n/k)$.
- Calcul de $\left\| y_i - \mathbf{X}_i(\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^T y_{-i} \right\|^2$: on fait n/k multiplications et n/k) additions : $O((n/k))^2)$.

La complexité total est donc $O(kd^3)+O(k(n-n/k)d^2)+O(k(n/k))^2)$ (en considérant la somme allant de 1 à k). En majorant k par n, on obtient :

$$O(nd^3) + O(n(n-1)d^2) + O(n) = O(nd^3) + O(n(n-1)d^2)$$

(e) Complexité de calcul de cette expression :

$$\text{error}_{k-fold} = \frac{1}{k}\sum_{i=1}^{k} \left\| \frac{\mathbf{y}_i - \mathbf{X}_i\mathbf{w}^*}{\mathbf{I} - \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top} \right\|^2$$

- Calcul de $\mathbf{X}_i\mathbf{w}^*$: multiplication entre deux matrices de tailles $n/k \times d$ et $d \times 1$: $O((n/k)d)$.
- Calcul de $\mathbf{y}_i - \mathbf{X}_i\mathbf{w}^*$: différence entre deux matrices de taille $n/k \times 1$: $O(n/k)$.
- Calcul de X^TX : multiplication entre deux matrices de taille $d \times n$ et $n \times d$: $O(nd^2)$.
- Calcul de $(X^TX)^{-1}$ inversion d'une matrice de taille $d \times d$: $O(d^3)$.
- Calcul de $\mathbf{X}_i(X^TX)^{-1}$: multiplication de deux matrices de taille $n/k \times d$ et $d \times d$: $O((n/k)d^2)$.
- Calcul de $\mathbf{X}_i(X^TX)^{-1}\mathbf{X}_i^T$: multiplication de deux matrices de taille $n/k \times d$ et $d \times n/k$: $O((n/k)^2d)$.
- Calcul de $I - \mathbf{X}_i(X^TX)^{-1}\mathbf{X}_i^T$: différence entre 2 matrices de tailles $n/k \times n/k$: $O((n/k)^2)$.
- Calcul de $(I - \mathbf{X}_i(X^TX)^{-1}\mathbf{X}_i^T)^{-1}$: inversion d'une matrice de taille $n/k \times n/k$: $O((n/k)^3)$.
- Calcul de $(\mathbf{y}_i - \mathbf{X}_i\mathbf{w}^*)(I - \mathbf{X}_i(X^TX)^{-1}\mathbf{X}_i^T)^{-1}$: $O((n/k)^3)$.
- Calcul de $\left\| \frac{\mathbf{y}_i - \mathbf{X}_i\mathbf{w}^*}{\mathbf{I} - \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top} \right\|^2$: on fait n/k multiplications et n/k additions : $O((n/k)^2)$.

Complexité totale en temps, avec la somme allant de 1 à k :

$$O(kd^3) + O(knd^2) + O(k(n/k)^2d) + O(k(n/k)^3)$$

Lorsqu'on majore k par n :

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

4. **Optimisation** [9 points] Soit la régression logistique à une dimension suivante:

$$\sigma(wx) = \frac{1}{1 + e^{-wx}}$$

où $x, w \in \mathbb{R}$, et la fonction de perte associée:

$$L(w) = -y \log \sigma(wx) - (1 - y) \log (1 - \sigma(wx))$$

(a) Montrer que la fonction de perte associée avec la régression logisitique est convexe en utilisant une des définitions de convexité suivante

- $L(w)$ est convexe si et seulement si

$$\forall w_1, w_2, t \in [0, 1] : L(tw_1 + (1 - t)w_2) \leq tL(w_1) + (1-t)L(w_2)$$

- $L(w)$ est convexe si et seulement si $\frac{d^2L}{dw^2}(w) \geq 0$ for all w

Vous pouvez aussi utiliser une autre définition, mais vous devez alors la donner explicitement

Réponse :

Soit la régression logistique à une dimension suivante:

$$\sigma(wx) = \frac{1}{1 + e^{-wx}}$$

où $x, w \in \mathbb{R}$, et la fonction de perte associée:

$$L(w) = -y \log \sigma(wx) - (1 - y) \log (1 - \sigma(wx))$$

Montrons que $L(w)$ est convexe en utilisant la deuxième définition.

Montrons que : $\frac{\partial^2 L}{\partial w^2}(w) \geq 0 \ \forall \ w$.

$$\begin{aligned}
\frac{\partial L}{\partial w}(w) &= \frac{\partial}{\partial w}(-y \log \sigma(wx) - (1-y) \log (1 - \sigma(wx))) \\
&= \frac{\partial}{\partial w}(-y \log \frac{1}{1+e^{-wx}} - (1-y) \log \left(1 - \frac{1}{1+e^{-wx}}\right)) \\
&= \frac{\partial}{\partial w}(-y \log \frac{1}{1+e^{-wx}} - (1-y) \log \left(\frac{e^{-wx}}{1+e^{-wx}}\right)) \\
&= -y \frac{xe^{-wx}}{1+e^{-wx}} + (1-y) \frac{xe^{-wx}}{(1-\frac{1}{1+e^{-wx}})(1+e^{-wx})^2} \\
&= -y \frac{xe^{-wx}}{1+e^{-wx}} + (1-y) \frac{xe^{-wx}}{e^{-wx}(1+e^{-wx})} \\
&= -y \frac{xe^{-wx}}{1+e^{-wx}} + (1+y) \frac{1}{1+e^{-wx}} \\
&= \frac{-yxe^{-wx} + x - xy}{1+e^{-wx}}
\end{aligned}$$

On a donc $\frac{\partial L}{\partial w}(w) = \frac{-yxe^{-wx}+x-xy}{1+e^{-wx}}$.
Calcul de $\frac{\partial^2 L}{\partial w^2}(w)$:

$$\begin{aligned}
\frac{\partial^2 L}{\partial w^2}(w) &= \frac{\partial^2}{\partial w^2}(\frac{-yxe^{-wx} + x - xy}{1+e^{-wx}}) \\
&= \frac{(yx^2e^{-wx})(1+e^{-wx}) + (-yxe^{-wx} + x - xy)(xe^{-wx})}{(1+e^{-wx})^2} \\
&= \frac{yx^2e^{-wx} + yx^2e^{-2wx} - x^2ye^{-2wx} + x^2e^{-wx} - x^2ye^{-wx}}{(1+e^{-wx})^2} \\
&= \frac{x^2e^{-wx}}{(1+e^{-wx})^2}
\end{aligned}$$

On sait que $x^2 \geq 0$, $e^{-wx} > 0$ et $(1+e^{-wx})^2 > 0$, $\forall \ x, w \in \mathbb{R}$.

(b) Donnez le gradient de $\sigma(wx)$ au point w. Quelles sont ses dimensions?

Réponse :

Le gradient de $\sigma(wx)$ au point w s'écrit :

$$\nabla_w \sigma(wx) = \frac{xe^{-wx}}{(1+e^{-wx})^2}$$

La dimension du gradient est 1×1.

(c) Donnez une expression analytique pour tous les points stationnaires de $L(w)$ w.r.t w, en justifiant votre réponse.
Réponse :
Pour déterminer une expression analytique pour tous les points stationaires de $L(w)$ w.r.t w, il suffit de résoudre l'équation $\frac{\partial L}{\partial w}(w) = 0$.
On sait que : $\frac{\partial L}{\partial w}(w) = \frac{-yxe^{-wx}+x-xy}{1+e^{-wx}}$.
Résolvons alors l'équation :

$$\begin{aligned} \frac{-yxe^{-wx} + x - xy}{1 + e^{-wx}} &= 0 \\ -yxe^{-wx} + x - xy &= 0 \\ x(1 - y - ye^{-wx}) &= 0 \end{aligned}$$

Donc, soit $x = 0$, soit :

$$\begin{aligned} 1 - y - ye^{-wx} &= 0 \\ e^{-wx} &= \frac{1-y}{y} \\ w &= \frac{-\log\left(\frac{1}{y} - 1\right)}{x} \qquad \forall\ x, y \neq 0 \end{aligned}$$

L'expression analytique recherchée est donc :

$$\boxed{w = \frac{-\log\left(\frac{1}{y} - 1\right)}{x} \qquad \forall\ x, y \neq 0 \ \text{ et} y \neq 1}$$

(d) Donnez l'expression pour un étape de descente de gradient d'un point initial w_0 à point w_1 après, en utilisant le gradient du fonction de perte
Réponse :
En posant η le learning rate, on a :

$$w_{t+1} = w_t - \eta \frac{\partial}{\partial w_t} L(w_t)$$

On a donc :

$$\boxed{w_1 = w_0 - \eta \frac{-yxe^{-w_0 x} + x - xy}{1 + e^{-w_0 x}}}$$

5. **Dérivées et gradients** [10 points]

Dans cette question, nous examinerons les dérivées et les gradients. Le Chapitre 5 de "Mathematics for Machine Learning" peut être utilisé comme référence pour cette question.

(a) Calculer la dérivée $f^{'}(x)$ pour:

$$f(x) = -5log\left(x^5\right) sin\left(x^2\right)$$

Réponse :
On a $f(x) = -5log\left(x^5\right) sin\left(x^2\right)$. Calcul de la dérivée :

$$\begin{aligned} f'(x) &= -5\frac{d}{dx}(log\left(x^5\right) sin\left(x^2\right)) \\ &= -5(sin(x^2)\frac{d}{dx}log(x^5) + log(x^5)\frac{d}{dx}sin(x^2)) \\ &= -5(\frac{5x^4}{x^5}sin(x^2) + log(x^5)2xcos(x^2)) \\ &= \frac{-25}{x}sin(x^2) - 10xlog(x^5)cos(x^2) \end{aligned}$$

Donc :

$$\boxed{f'(x) = \frac{-25}{x}sin(x^2) - 10xlog(x^5)cos(x^2)}$$

(b) Calculer la dérivée $f^{'}(x)$ pour:

$$f(x) = 3exp\left(\frac{-5}{3\sigma}(x-\mu)^2\right)$$

où $\sigma, \mu \in \mathbb{R}$.

Réponse :
Calcul la dérivée $f^{'}(x)$ pour:

$$\begin{aligned} f(x) &= 3exp\left(\frac{-5}{3\sigma}(x-\mu)^2\right) \\ &= 3exp\left(\frac{-5}{3\sigma}\left(x^2 - 2x\mu + \mu^2\right)\right) \\ &= 3exp(\frac{-5}{3\sigma}x^2 + \frac{10\mu}{3\sigma}x - \frac{5\mu^2}{3\sigma}) \end{aligned}$$

où $\sigma, \mu \in \mathbb{R}$.

$$\begin{aligned} f'(x) &= \frac{\partial}{\partial x}(3exp(\frac{-5}{3\sigma}x^2 + \frac{10\mu}{3\sigma}x - \frac{5\mu^2}{3\sigma})) \\ &= 3(\frac{-10}{3\sigma}x + \frac{10\mu}{3\sigma})3exp\left(\frac{-5}{3\sigma}(x-\mu)^2\right) \\ &= \frac{10}{\sigma}(\mu - x)exp\left(\frac{-5}{3\sigma}(x-\mu)^2\right) \end{aligned}$$

Conclusion :

$$\boxed{f'(x) = \frac{10}{\sigma}(\mu - x)exp\left(\frac{-5}{3\sigma}(x-\mu)^2\right)}$$

(c) Considérez les fonctions suivantes:

$$f_1(x) = sin(2x_1)\, cos(3x_2)$$

où $x \in \mathbb{R}^2$, et

$$f_2(x, y) = 3x^T y$$
$$f_3(x) = -4xx^T$$

où $x, y \in \mathbb{R}^n$.

i. Quelles sont les dimensions de $\frac{\partial f_1}{\partial x}$, $\frac{\partial f_2}{\partial x}$ et $\frac{\partial f_3}{\partial x}$?

ii. Calculer la matrice Jacobienne pour chacune de ces fonctions.

Réponse :

(i) Les dimensions :
$\frac{\partial f_1}{\partial x} \in \mathbb{R}^{1\times 2}$.
$\frac{\partial f_2}{\partial x} \in \mathbb{R}^{1\times n}$.
$\frac{\partial f_3}{\partial x} \in \mathbb{R}^{n\times n\times n}$.

(ii) La matrice Jacobienne pour chacune de ces fonctions :

$$\frac{\partial f_1}{\partial x} = \begin{pmatrix} 2\ cos(2x_1)\ cos(3x_2) \\ -3\ sin(2x_1)\ sin(3x_2) \end{pmatrix}^T$$

$$\frac{\partial f_2}{\partial x} = 3y^T$$

$$\frac{\partial f_3}{\partial x} = -4 \begin{pmatrix} \frac{\partial x_1^2}{\partial x} & \frac{\partial x_1 x_2}{\partial x} & \dots & \frac{\partial x_1 x_n}{\partial x} \\ \frac{\partial x_2 x_1}{\partial x} & \frac{\partial x_2^2}{\partial x} & \dots & \frac{\partial x_2 x_n}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n x_1}{\partial x} & \frac{\partial x_n x_2}{\partial x} & \dots & \frac{\partial x_n^2}{\partial x} \end{pmatrix}$$

avec $\dim(\frac{\partial x_1^2}{\partial x}) = \dim(\frac{\partial x_1 x_2}{\partial x}) = ... = 1 \times n$

(d) Calculer les dérivées $\frac{df}{dx}$ des fonctions suivantes:

i. Utilisez la règle de dérivée en chaîne. Donnez les dimensions de chaque dérivée.

$$f(z) = 2exp\left(-\frac{1}{2}z^2\right)$$

$$z = g(y) = y^T S^{-1} y$$

$$y = h(x) = x - \mu$$

où $x, \mu \in \mathbb{R}^D, S \in \mathbb{R}^{D \times D}$.

ii.

$$f(x) = tr\left(xx^T + \sigma I\right)$$

où $x \in \mathbb{R}^D$ and $tr(A)$ est la trace de A.

iii. Utilisez la règle de la chaîne pour calculer les dérivées et fournir également les dimensions de la dérivée partielle pour

$$f = tanh^2(z), \quad z = Ax + b$$

où $f \in \mathbb{R}^M$, $x \in \mathbb{R}^N$, $A \in \mathbb{R}^{M \times N}$, $b \in \mathbb{R}^M$.

Réponse :

(i) En utilisant la règle de dérivée en chaine :

$$\begin{aligned} \frac{df}{dx} &= \frac{dz}{dx}\frac{df}{dz} \\ &= \frac{df}{dz}\frac{dz}{dy}\frac{dy}{dx} \end{aligned}$$

Or, avec I la matrice identité, on a que :

$$\begin{aligned}
\frac{dy}{dx} &= I \\
\frac{dz}{dy} &= y^T(S^{-1} + S^{-1^T}) \\
&= (x-\mu)^T(S^{-1} + S^{-1^T}) \\
\frac{df}{dz} &= -2\ z\ exp(-\frac{1}{2}z^2) \\
&= -2(x-\mu)^T S^{-1}(x-\mu)\ \ exp(-\frac{1}{2}((x-\mu)^T S^{-1}(x-\mu))^2)
\end{aligned}$$

$\frac{df}{dz}$ est un scalaire $\in \mathbb{R}$,
$\frac{dz}{dy} \in \mathbb{R}^{1\times D}$ et
$\frac{dy}{dx} \in \mathbb{R}^{D\times D}$.
Ainsi, $\frac{df}{dx} \in \mathbb{R}^{1\times D}$, et :

$$\begin{aligned}
\frac{df}{dx} =&(x-\mu)^T(S^{-1} + S^{-1^T}) \\
&\times (-2(x-\mu)^T S^{-1}(x-\mu)\ \ exp(-\frac{1}{2}((x-\mu)^T S^{-1}(x-\mu))^2))
\end{aligned}$$

(ii) On considère f tel que :

$$f(x) = tr\left(xx^T + \sigma I\right)$$

où $x \in \mathbb{R}^D$.
On a :

$$xx^T + \sigma I = \begin{pmatrix} x_1^2+\sigma & x_1x_2 & \cdots & x_1x_D \\ x_2x_1 & x_2^2+\sigma & \cdots & x_2x_D \\ \vdots & \vdots & \ddots & \vdots \\ x_Dx_1 & x_Dx_2 & \cdots & x_D^2+\sigma \end{pmatrix}$$

Donc :

$$tr\left(xx^T + \sigma I\right) = \sum_{i=1}^{D}(x_i^2+\sigma) = D\sigma + \sum_{i=1}^{D} x_i^2$$

D'où :

$$\boxed{f'(x) = 2x^T}$$

(iii) Calcul de la dérivée de f tel que :

$$f = tanh^2(z), \quad z = Ax + b$$

où $f \in \mathbb{R}^M$, $x \in \mathbb{R}^N$, $A \in \mathbb{R}^{M \times N}$, $b \in \mathbb{R}^M$.
En utilisant la règle de dérivée en chaine :

$$\frac{df}{dx} = \frac{df}{dz}\frac{dz}{dx}$$

Avec :

$$\frac{df}{dz} = \begin{pmatrix} \frac{d\, tanh^2(z_1)}{dz_1} & \frac{d\, tanh^2(z_1)}{dz_2} & \cdots & \frac{d\, tanh^2(z_1)}{dz_M} \\ \frac{d\, tanh^2(z_2)}{dz_1} & \frac{d\, tanh^2(z_2)}{dz_2} & \cdots & \frac{d\, tanh^2(z_2)}{dz_M} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d\, tanh^2(z_M)}{dz_1} & \frac{d\, tanh^2(z_M)}{dz_2} & \cdots & \frac{d\, tanh^2(z_M)}{dz_M} \end{pmatrix}$$

$$= \begin{pmatrix} 2\ tanh(z_1)\frac{d\, tanh(z_1)}{dz_1} & 0 & \cdots & 0 \\ 0 & 2\ tanh(z_2)\frac{d\, tanh(z_2)}{dz_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2\ tanh(z_M)\frac{d\, tanh(z_M)}{dz_M} \end{pmatrix}$$

$$= \begin{pmatrix} 2\ tanh(z_1)(1 - tanh^2(z_1)) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2\ tanh(z_M)(1 - tanh^2(z_M)) \end{pmatrix}$$

et

$$\frac{dz}{dx} = A$$

$\frac{df}{dz}$ est de dimension : $M \times M$, et $\frac{dz}{dx}$ est de dimension : $M \times N$.
Donc $\frac{df}{dx}$ est de dimension $M \times N$, et :

$$\boxed{\frac{df}{dx} = \begin{pmatrix} 2\ tanh(z_1)(1 - tanh^2(z_1)) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2\ tanh(z_M)(1 - tanh^2(z_M)) \end{pmatrix} A}$$