IFT 6390A Fondements de l'apprentissage machine
Prof. Guillaume RabusseauIFT 6390B Fundamentals of Machine Learning
Ioannis Mitliagkas

Devoir 2 - Partie Théorique

- Ce devoir doit être fait et envoyé sur Gradescope individuellement. 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:

[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) : $\mathbb{E}[(h_D(x')] = \mathbb{E}_{(x_1,y_1)\sim p} \cdots \mathbb{E}_{(x_n,y_n)\sim p} \mathbb{E}[(h_{\{(x_1,y_1),\ldots,(x_n,y_n)\}}(x')]$.

- 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.

Answer. On utilise x en place de x' et y en place de y', $h(x)$ en place de $h_D(x)$

$\mathbb{E}[(h(x) - y)^2]$

expanding the ^{2}

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

using linearity of expectations

$= \mathbb{E}[h(x)^2] - 2\,\mathbb{E}[yh(x)] + \mathbb{E}[y^2]$

add and subtract the term $E[h(x)]^2$

$= \mathbb{E}[h(x)^2] - 2\,\mathbb{E}[yh(x)] + \mathbb{E}[y^2] + E[h(x)]^2 - E[h(x)]^2$

using $Var(h(x)) = \mathbb{E}[(h(x) - \mathbb{E}[h(x)])^2]$

$= Var(h(x)) - 2\,\mathbb{E}[yh(x)] + \mathbb{E}[y^2] + E[h(x)]^2$

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

$= Var(h(x)) - 2\,\mathbb{E}[(f(x) + \epsilon)h(x)] + \mathbb{E}[(f(x) + \epsilon)^2] + E[h(x)]^2$

expand $(f(x)+\epsilon)^2$

$= Var(h(x)) - 2\,\mathbb{E}[(f(x)+\epsilon)h(x)] + \mathbb{E}[f(x)^2 + 2f(x)\epsilon + \epsilon^2] + E[h(x)]^2$

by linearity of expectations

$= Var(h(x)) - 2\,\mathbb{E}[(f(x)+\epsilon)h(x)] + \mathbb{E}[f(x)^2] + \mathbb{E}[2f(x)\epsilon] + \mathbb{E}[\epsilon^2] + E[h(x)]^2$

tricky part here $\mathbb{E}[2f(x)\epsilon] = 2 * (\mathbb{E}[f(x)]\,\mathbb{E}[\epsilon] + Covariance(f(x), \epsilon))$
and we can assume that $Covariance(f(x), \epsilon) = 0$ since the noise ϵ is random, and we know that $\mathbb{E}[\epsilon] = \mathbb{E}[\mathcal{N}(0, \sigma^2)] = 0$ so $\mathbb{E}[2f(x)\epsilon] = 0$

$= Var(h(x)) - 2\,\mathbb{E}[(f(x)+\epsilon)h(x)] + \mathbb{E}[f(x)^2] + \mathbb{E}[\epsilon^2] + E[h(x)]^2$

expand $(f(x)+\epsilon)h(x)$ and again use linearity of expectations

$= Var(h(x)) - 2\,\mathbb{E}[f(x)h(x)] + 2\,\mathbb{E}[\epsilon h(x)] + \mathbb{E}[f(x)^2] + \mathbb{E}[\epsilon^2] + E[h(x)]^2$

use the same trick as before to show that $2\,\mathbb{E}[\epsilon h(x)] = 0$

$= Var(h(x)) - 2\,\mathbb{E}[f(x)h(x)] + \mathbb{E}[f(x)^2] + \mathbb{E}[\epsilon^2] + E[h(x)]^2$

using $\mathbb{E}[f(x)^2] = f(x)^2$

$= Var(h(x)) - 2\,\mathbb{E}[f(x)h(x)] + f(x)^2 + \mathbb{E}[\epsilon^2] + E[h(x)]^2$

again tricky $\mathbb{E}[f(x)h(x)] = \mathbb{E}[f(x)]\,\mathbb{E}[h(x)] + Covariance(f(x), h(x))$
we can again assume $Covariance = 0$ so $\mathbb{E}[f(x)h(x)] = f(x)\,\mathbb{E}[h(x)]$

$= Var(h(x)) - 2f(x)\,\mathbb{E}[h(x)] + f(x)^2 + \mathbb{E}[\epsilon^2] + E[h(x)]^2$

using $Bias(h(x))^2 = \mathbb{E}[h(x)]^2 - 2f(x)\,\mathbb{E}[h(x)] + f(x)^2$

$= Var(h(x)) + Bias(h(x))^2 + \mathbb{E}[\epsilon^2]$

we can subtract $\mathbb{E}[\epsilon]^2$ since it is 0 because $\mathbb{E}[\epsilon] = 0$

$= Var(h(x)) + Bias(h(x))^2 + \mathbb{E}[\epsilon^2] - \mathbb{E}[\epsilon]^2$

using $Var(\epsilon) = \mathbb{E}[\epsilon^2] - \mathbb{E}[\epsilon]^2$

$= Var(h(x)) + Bias(h(x))^2 + Var(\epsilon)$

since $\epsilon \sim \mathcal{N}(0, \sigma^2)$ we know its variance is σ^2

$= Var(h(x)) + Bias(h(x))^2 + \sigma^2$

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.

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

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

Answer.

(a) Oui

$\boxed{x' = (x-1.5)^2}$

$\boxed{x' = (x-2)(x-1)}$

(b) Oui

$\boxed{x = X_1 * X_2}$

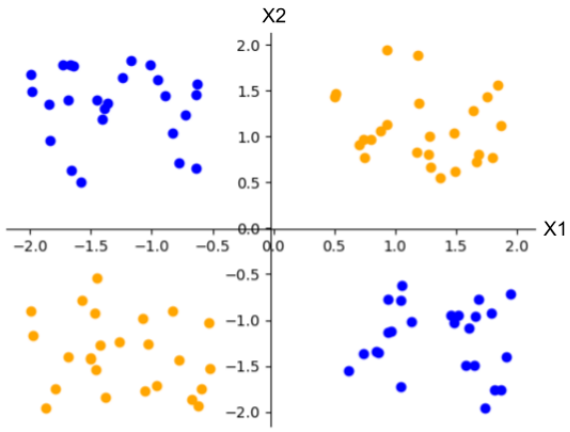

Figure 2: Jeu de données 2D.

(c) Oui $X' = (X_1 - 2) * (X_2 - 3)$ Le noyau correspondent

$$\begin{aligned} K(X,Y) =& \phi(X)^T \phi(Y) \\ =& [(X_1 - 2) * (X_2 - 3)] * [(Y_1 - 2) * (Y_2 - 3)] \\ =& (X_1X_2 - 3X_1 - 2X_2 + 6) * (Y_1Y_2 - 3Y_1 - 2Y_2 + 6) \\ =& X_1X_2Y_1Y_2 - 3X_1X_2Y_1 - 2X_1X_2Y_2 + 6X_1X_2 - 3X_1Y_1Y_2 \\ & + 9X_1Y_1 + 6X_1Y_2 - 18X_1 - 2X_2Y_1Y_2 + 6X_2Y_1 + 4X_2Y_2 \\ & - 12X_2 + 6Y_1Y_2 - 18Y1 - 12Y_2 + 36 \end{aligned}$$

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 : $D_{\setminus i} = \{(x_j, y_j) \mid j \notin \text{ind}[i]\}$.

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

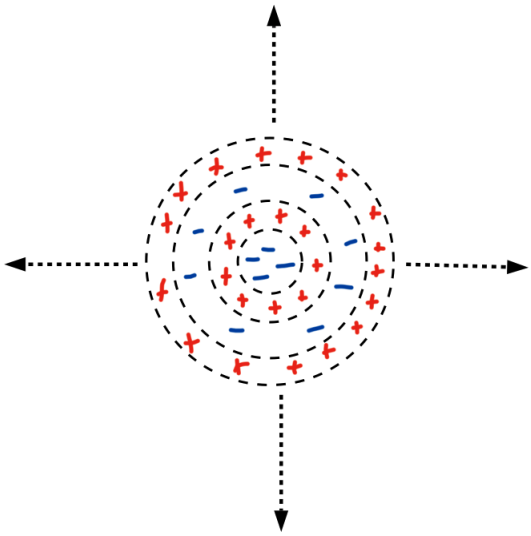

Figure 3: Un autre jeu de données 2D. Les points entre les aires de rayon $2k$ et $2k+1$ sont étiquetés par -. Les points entre les aires de rayon $2k+1$ et $2k+2$ sont étiquetés par $+$.

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. Quelle est la complexité du calcul de cette expression?

Answer.

(a) $\boxed{risk = \sum_D (y - h(x))^2}$

(b)

$$\begin{aligned}
\mathop{\mathbb{E}}_{D\sim p}[\text{error}_{k-fold}] &= \mathop{\mathbb{E}}_{D\sim p}\left[\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\sum_{j\in\text{ind}[i]}\ell(h_{D\setminus\text{ind}[i]}(x_j),y_j)\right] \\
&= \frac{1}{k}\sum_{i=1}^{k}\mathop{\mathbb{E}}_{D\sim p}\left[\frac{1}{n/k}\sum_{j\in\text{ind}[i]}\ell(h_{D\setminus\text{ind}[i]}(x_j),y_j)\right] \\
&\approx \frac{1}{k}\sum_{i=1}^{k}\mathop{\mathbb{E}}_{\substack{D'\sim p,\\(x,y)\sim p}}[\ell(h_{D'}(x),y)] \\
&= \mathop{\mathbb{E}}_{\substack{D'\sim p,\\(x,y)\sim p}}\left[(y-h_{D'}(x))^2\right]
\end{aligned}$$

(c) Here we assume the complexity of multiplying an $a\times b$ matrix by a $b\times c$ matrix is $\mathcal{O}(abc)$.
Then calculating $\boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ takes
$\mathcal{O}\left(dnd + d^3 + ddn + dn\right) = \boxed{\mathcal{O}\left(d^3 + d^2n\right)}$ time.

(d)

$$\text{error}_{k-fold} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\sum_{j\in\text{ind}[i]}\left(\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right]\right)_j^2$$

L'évaluation prends

$$\begin{aligned}
&(\mathbf{X}_{-i}^\top\mathbf{X}_{-i}) \implies \mathcal{O}\left(d^2n(1-\frac{1}{k})\right) \implies \mathcal{O}\left(d^2n\right) \\
&(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1} \implies \mathcal{O}\left(d^2n + d^3\right) \\
&(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top \implies \mathcal{O}\left(d^2n + d^3 + d^2n(1-\frac{1}{k})\right) \implies \mathcal{O}\left(d^2n + d^3\right) \\
&\mathbf{X}_i\left[(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right] \implies \mathcal{O}\left(d^2n + d^3 + dn(1-\frac{1}{k})\right) \implies \mathcal{O}\left(d^2n + d^3\right) \\
&\frac{1}{n/k}\sum_{j\in\text{ind}[i]}(\mathbf{y}_i - \cdots)_j^2 \implies \mathcal{O}\left(d^2n + d^3 + \frac{n}{k}\right) \implies \mathcal{O}\left(d^2n + d^3\right) \\
&\text{error}_{k-fold} \implies \mathcal{O}\left(k\cdot(d^2n + d^3)\right) \implies \mathcal{O}\left(d^2nk + d^3k\right)
\end{aligned}$$

(e)

$$\text{error}_{k-fold} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\sum_{j\in\text{ind}[i]}\left(\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right]\right)_j^2$$

$$\begin{aligned}
&\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right] \\
&= \frac{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top}\cdot\left(\mathbf{y}_i - \mathbf{X}_i(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right) \\
&= \boxed{\frac{\mathbf{y}_i - \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top\mathbf{y}_i - \mathbf{X}_i(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i} + \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top}} \\
&= \boxed{\frac{\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top\mathbf{y}_i + (\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right]}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top}} \\
&= \boxed{\frac{\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}^\top\mathbf{X})^{-1}\left(\mathbf{X}_i^\top\mathbf{y}_i + (\mathbf{X}^\top\mathbf{X})(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i} - \mathbf{X}_i^\top\mathbf{X}_i(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right)\right]}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top}} \\
&= \frac{\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}^\top\mathbf{X})^{-1}\left(\mathbf{X}_i^\top\mathbf{y}_i + (\mathbf{X}^\top\mathbf{X} - \mathbf{X}_i^\top\mathbf{X}_i)(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right)\right]}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top} \\
&= \frac{\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}^\top\mathbf{X})^{-1}\left(\mathbf{X}_i^\top\mathbf{y}_i + (\mathbf{X}_{-i}^\top\mathbf{X}_{-i})(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right)\right]}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top} \\
&= \frac{\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}^\top\mathbf{X})^{-1}\left(\mathbf{X}_i^\top\mathbf{y}_i + \mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right)\right]}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top} \\
&= \frac{\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{y})\right]}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top} \\
&= \frac{\mathbf{y}_i - \mathbf{X}_i\mathbf{w}^*}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top}
\end{aligned}$$

So,

$$\begin{aligned}\text{error}_{k-fold} &= \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\sum_{j\in\text{ind}[i]}\left(\mathbf{y}_i - \mathbf{X}_i\left[(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^\top\mathbf{y}_{-i}\right]\right)_j^2 \\ &= \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n/k}\sum_{j\in\text{ind}[i]}\left(\frac{\mathbf{y}_i - \mathbf{X}_i\mathbf{w}^*}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top}\right)_j^2\end{aligned}$$

L'évaluation prends

$$\begin{aligned}&(\mathbf{X}^\top\mathbf{X}) \implies \mathcal{O}\left(d^2n\right)\\ &(\mathbf{X}^\top\mathbf{X})^{-1} \implies \mathcal{O}\left(d^2n+d^3\right)\\ &\mathbf{w}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \implies \mathcal{O}\left(d^2n+d^3+d^2n+dn\right) \implies \mathcal{O}\left(d^2n+d^3\right)\\ &(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top \implies \mathcal{O}\left(d^2n+d^3+d^2\frac{n}{k}\right) \implies \mathcal{O}\left(d^2n+d^3\right)\\ &\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top \implies \mathcal{O}\left(d^2n+d^3+d\left(\frac{n}{k}\right)^2\right) \implies \mathcal{O}\left(d^2n^2+d^3\right)\\ &\left(1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top\right)^{-1} \implies \mathcal{O}\left(d^2n+d^3+\left(\frac{n}{k}\right)^3\right) \implies \mathcal{O}\left(d^2n+d^3+n^3\right)\\ &\frac{\mathbf{y}_i - \mathbf{X}_i\mathbf{w}^*}{1-\mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top} \implies \mathcal{O}\left(d^2n+d^3+\left(\frac{n}{k}\right)^3\right) \implies \mathcal{O}\left(d^2n+d^3+n^3\right)\end{aligned}$$

$\mathcal{O}\left(d^3+d^2k\right)$ pour calculer $(\mathbf{X}^\top\mathbf{X})^{-1}$ et $\mathbf{w}^*$, et $\mathcal{O}\left(d^2\right)$ par point pour un calcul supplémentaire, soit un temps d'exécution total de $\boxed{\mathcal{O}\left(d^3+d^2n+n^3\right)}$, ce qui est un facteur de n meilleur que dans (d).

k-fold

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:

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

(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

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

- $L(w)$ 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

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

(c) Donnez une expression analytique pour tous les points stationnaires de $L(w)$ w.r.t w, en justifiant votre réponse.

(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

Answer.

(a)

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

simplifie l'expression commençant avec le premier terme

$$\begin{aligned} -y \log \sigma(wx) &= -y \log \frac{1}{1+e^{-wx}} \\ &= -y[\log 1 - \log(1+e^{-wx})] \\ &= y \log(1+e^{-wx}) \end{aligned}$$

le deuxième terme

$$\begin{aligned} -(1-y)\log\left(1-\sigma(wx)\right) &= -(1-y)\log 1 - \frac{1}{1+e^{-wx}} \\ &= -(1-y)\log\frac{1+e^{-wx}}{1+e^{-wx}} - \frac{1}{1+e^{-wx}} \\ &= -(1-y)\log\frac{e^{-wx}}{1+e^{-wx}} \\ &= -(1-y)[\log e^{-wx} - \log\left(1+e^{-wx}\right)] \\ &= -(1-y)[-wx - \log\left(1+e^{-wx}\right)] \\ &= (1-y)[wx + \log\left(1+e^{-wx}\right)] \end{aligned}$$

combine les termes

$$\begin{aligned} L(w) &= y\log(1+e^{-wx}) + (1-y)[wx + \log(1+e^{-wx})] \\ &= (1-y)wx + \log(1+e^{-wx}) \end{aligned}$$

prenons maintenant la deuxième dérivée pour montrer la convexité

commençons par la première dérivée

$$\begin{aligned} \frac{dL}{dw} &= (1-y)x - \frac{xe^{-wx}}{1+e^{-wx}} \\ &= (1-y)x - \frac{(1+e^{-wx})x - x}{1+e^{-wx}} \\ &= (1-y)x - x + \frac{x}{1+e^{-wx}} \\ &= -yx + \frac{x}{1+e^{-wx}} \\ &= -yx + x\sigma(wx) \end{aligned}$$

maintenant nous faisons le second dérivé en utilisant notre connaissance du dérivé d'un sigmoïde

$$\begin{aligned} \frac{d^2L}{dw^2} &= x\sigma(wx)(1-\sigma(wx))x \\ &= x^2\sigma(wx)(1-\sigma(wx)) \end{aligned}$$

Puisque $\sigma() \in (0,1)$, donc $\sigma(wx) \in (0,1)$ et $(1-\sigma(wx)) \in (0,1)$. Enfin, $x^2 > 0$ donc chacun de nos trois termes > 0 donc le produit $\frac{d^2L}{dw^2} > 0$ et cette perte est convexe.

(b)

$$\begin{aligned}
\frac{d}{dw}\sigma(wx) &= \frac{d}{dw}\frac{1}{1+e^{-wx}} \\
&= \frac{d}{dw}(1+e^{-wx})^{-1} \\
&= -(1+e^{-wx})^{-2} * \frac{d}{dw}(1+e^{-wx}) \\
&= -(1+e^{-wx})^{-2} * e^{-wx} * -x \\
&= \frac{xe^{-wx}}{(1+e^{-wx})^2} \\
&= x\frac{1}{1+e^{-wx}}\frac{e^{-wx}}{1+e^{-wx}} \\
&= x\frac{1}{1+e^{-wx}}(1-\frac{1}{1+e^{-wx}}) \\
&= x\sigma(wx)(1-\sigma(wx))
\end{aligned}$$

Cette dimension est la même que le nombre de points représentés par x. Pour un seul point x, il s'agit simplement d'un scalaire.

(c) Les points fixes sont où $\frac{dL(w)}{dw} = 0$. En utilisant notre formule de (a)

$$\begin{aligned}
\frac{dL(w)}{dw} &= 0 \\
-yx + x\sigma(wx) &= 0 \\
x\sigma(wx) &= yx
\end{aligned}$$

assume $x \neq 0$

$$\begin{aligned}
\sigma(wx) &= y \\
\frac{1}{1+e^{-wx}} &= y \\
e^{-wx} &= \frac{1}{y} - 1 \\
-wx &= \log(\frac{1-y}{y}) \\
w &= -\frac{1}{x}\log(\frac{1-y}{y})
\end{aligned}$$

Pour déterminer les valeurs possibles pour w, voyons nos valeurs possibles pour y. Si l'exemple est positif, $y = 1$ et $w = -\frac{1}{x} \log 0$. C'est indéfini, mais on peut voir que *in the limit*

$$\lim_{y \to 0+} \log y = -\infty$$

Par conséquent, *au limite*, il y a un point fixe

$$\begin{aligned} w &= -\frac{1}{x} * -\infty \\ &= \infty \end{aligned}$$

De même, pour un exemple négatif $y = 0$, on s'attendrait à ce que le point stationnaire dans la limite soit $-\infty$

(d) α est le taux d'apprentissage

$$\begin{aligned} w_1 &= w_0 - \alpha * \frac{d}{dw} L(w) \\ &= w_0 - \alpha * (-yx + x\sigma(wx)) \\ &= w_0 + \alpha(yx - x\sigma(wx)) \end{aligned}$$