Présentation finale double descente

Emett Haddad
Encadrants: Nicolas Vayatis, Samuel Gruffaz

24/06/2024

Plan

Rappel: Contexte

Fonction cible

On pose $\mathcal{X} = \mathbb{R}^D$ espace de départ et $\mathcal{Y} = \mathbb{R}$ espace d'arrivée. $\mathbf{X}$ et $\mathbf{Y}$ des variables aléatoires tel que $(\mathbf{X}, \mathbf{Y}) \hookrightarrow \mathcal{P}$.

$$y^\star : x \in \mathcal{X} \to \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathcal{P}}(\mathbf{Y}|\mathbf{X} = x) \in \mathcal{Y}$$

Échantillons

Échantillon d'apprentissage $\mathcal{D} := \{(x_n, y_n) \in \mathcal{X} \times \mathcal{Y}\}_{n=1}^N$ où
$y_n := y^\star(x_n) + \epsilon_n$, et les $x_n \hookrightarrow \mathbf{X}$ iid.
On modélise ici : $\mathbf{Y} = y^\star(\mathbf{X}) + \epsilon$ où ϵ représente le **bruit** tel que $\mathbb{E}(\epsilon|\mathbf{X}) = 0$, et $\mathbb{V}(\epsilon) = \sigma_\epsilon^2$.

Contexte

Estimateur

- Trouver déterminer un **estimateur** $\hat{y} : \mathcal{X} \to \mathcal{Y}$ tel que $\hat{y}(\mathbf{X}) \approx \mathbf{Y}$.
- $\hat{y} \in \mathcal{H}$ un espace de fonctions.

Vrai risque, risque empirique et excès de risque

Vrai risque et risque empirique: $\forall\, \hat{y} \in \mathcal{H}$

$$\mathcal{R}(\hat{y}) = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathcal{P}}((\mathbf{Y} - \hat{y}(\mathbf{X}))^2), \quad \hat{\mathcal{R}}_{\mathcal{D}}(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}(x_i))^2$$

Excès de risque: $\forall\, \hat{y} \in \mathcal{H}$

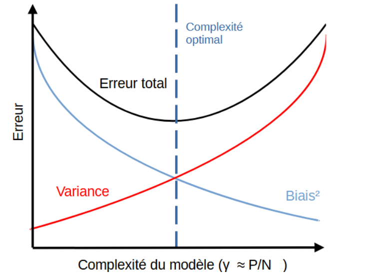
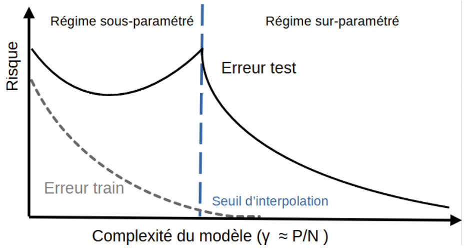
$$\mathcal{E}(\hat{y}) = \mathcal{R}(\hat{y}) - \mathcal{R}(y^\star)$$

Notion de Double Descente

Régime sous-paramétré: $P < N$ et **régime sur-paramétré**: $P > N$
Seuil d'interpolation: $P = N$

Notion de Double Descente:

On dit qu'il y a **double descente** quand l'erreur global minimal dans le régime sur-paramétré est inférieur à celle dans le régime sous-paramétré et qu'on observe un maximum au seuil d'interpolation.

Compromis biais-variance	Double descente
	

Modèles linéaire et régression pénalisée

Modèle linéaire et régression pénalisée

- $\mathcal{F} = \{f_i : \mathcal{X} \to \mathbb{R}, i \in \mathbb{N}\}$ ensemble de features.
- $X = [x_1, \cdots, x_N]^T \in \mathcal{M}_{N,D}(\mathbb{R})$, $Y = [y_1, \cdots, y_N]^T$
- $\forall x \in \mathcal{X},\ \Phi_P(x) = [f_1(x), \cdots, f_P(x)]^T$ et $Z = [\Phi_P(x_1), \cdots, \Phi_P(x_N)]^T = [f_j(x_i)] \in \mathcal{M}_{N,P}(\mathbb{R})$
- Risque empirique $\hat{\mathcal{R}}_{\mathcal{D}}(\hat{y}_\beta) = \frac{1}{N}||Y - Z\beta||_2^2$
- $\forall x \in \mathcal{X},\ \hat{y}_{\hat{\beta}}(x) = \Phi_P^T(x)\hat{\beta}$, pour $\hat{\beta} = Z^{\dagger}Y$.
- Modèle simple $f_i(x) = e_i(x)$ et X = Z

Pénalisation:

- Risque empirique pénalisé $\mathring{\mathcal{R}}_{\mathcal{D},\lambda}(\hat{y}_\beta) = ||Y - Z\beta||_2^2 + \lambda||\beta||_2^2$ où $\lambda > 0$.
- $\forall x \in \mathcal{X},\ \hat{y}_{\hat{\beta}_\lambda}(x) = \Phi_P^T(x)\hat{\beta}_\lambda$, pour $\hat{\beta}_\lambda = (Z^T Z + \lambda I_P)^{-1} Z^T Y$.

Théorie modèle linéaire

Théorème: Expressions de l'estimateur

Dans le cas linéaire càd tel que $\hat{y}(x) = \Phi_P^T(x)\hat{\beta}$ nous avons:

$$\forall\, x \in \mathcal{X},\ \hat{y}(x) = [f_1(x), \cdots, f_P(x)][(f_i, f_j)_X]^{\dagger} \begin{bmatrix} (f_1, y)_X \\ \vdots \\ (f_P, y)_X \end{bmatrix} \tag{1}$$

$$\forall\, x \in \mathcal{X},\ \hat{y}(x) = ([\{G_P^{x_i}, G_P^{x_j}\}]^{\dagger}[\{G_P^{x}, G_P^{x_j}\}])^T Y \tag{2}$$

Théorème: Décroissance du paramètre $\hat{\beta}$

Dans le régime sur-paramétré, $||\hat{\beta}_P||_2$ est décroissante à partir du moment où le rang de la matrice Z_P devient maximale i.e $rg(Z_P) = N$.

Théorème modèle linéaire quasi-isotropique

Conjecture: Limite du quotient de matrices aléatoires suivant une distribution de Marchenko-Pastur

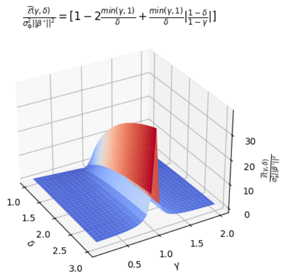
$$\lim_{N \to +\infty, \frac{P}{N} \to \gamma, \frac{E}{N} \to \delta} \frac{1}{min(P,N)} \mathrm{tr}[\mathrm{Z_E Z_E^T (Z_P Z_P^T)^{\dagger}}] = \left| \frac{1-\delta}{1-\gamma} \right|$$

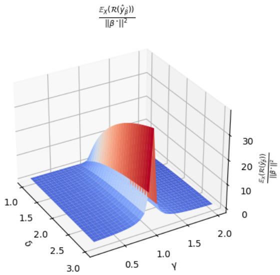
Théorème: Expression de l'excès de risque moyen asymptotique, cas du rang maximal

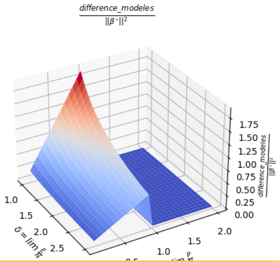
En prenant pour hypothèse la conjecture précédente.

En supposant : $\mathbb{E}(\beta^\star \beta^{\star T}) = \frac{||\beta^\star||_2^2}{E} I_E$, $rg(Z_P) = min(P,N)$, et $\mathcal{F}$ features orthonormées par rapport à $\mathbf{X}$.

$$\overline{\mathcal{E}}(\gamma,\delta) := \lim_{N \to +\infty, \frac{P}{N} \to \gamma, \frac{E}{N} \to \delta} \mathcal{E}_{moyen}(\hat{y}_{\hat{\beta}}) = \sigma_\Phi^2 \left[1 + \frac{min(\gamma,1)}{\delta}\left(-2 + \left|\frac{1-\delta}{1-\gamma}\right|\right)\right] \cdot ||\beta^\star||_2^2 \quad (3)$$


$\frac{\bar{\mathcal{E}}(\gamma,\delta)}{\sigma_\Phi^2||\beta^*||^2} = [1-2\frac{min(\gamma,1)}{\delta} + \frac{min(\gamma,1)}{\delta}|\frac{1-\delta}{1-\gamma}|]$
δ
γ
$\frac{\bar{\mathcal{E}}(\gamma,\delta)}{\sigma_\Phi^2||\beta^*||^2}$


$\frac{\mathcal{E}_X(\mathcal{R}(\hat{y}_{\hat{\beta}}))}{||\beta^*||^2}$
δ
γ
$\frac{\mathcal{E}_X(\mathcal{R}(\hat{y}_{\hat{\beta}}))}{||\beta^*||^2}$


$\frac{difference\_modeles}{||\beta^*||^2}$
$\frac{difference\_modeles}{||\beta^*||^2}$

Descente de gradient

Remarque:

Décomposition de l'erreur:

$$\sigma_{\Phi}^{-2}\mathcal{E}(\hat{y}_{\hat{\beta}_t},\beta^\star) = ||\hat{\beta}_t - \beta^\star||_2^2 = ||\hat{\beta}_t - \hat{\beta}||_2^2 + 2\mathrm{tr}[(\hat{\beta}_{\mathrm{t}} - \hat{\beta})^{\mathrm{T}}(\hat{\beta} - \beta^\star)] + ||\hat{\beta} - \beta^\star||_2^2$$

Théorème: Descente de gradient à pas constant par morceaux

On suppose que $\hat{\beta}_0 = 0$ et avec $t_1, \cdots, t_r$ changements de pas de descente de gradient. Alors, avec $t_{r+1} := t > t_r$:

$$\hat{\beta} - \hat{\beta}_t = V \begin{bmatrix} \Sigma_R^{-1}\prod_{j=0}^{r}(I_R - \underline{\alpha_j'}\Sigma_R^2)^{t_{j+1}-t_j} & 0 \\ 0 & 0 \end{bmatrix} U^T Y \tag{4}$$

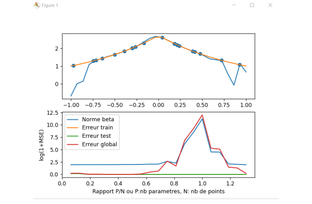
Corollaire: Approximation par descente de gradient à pas variable

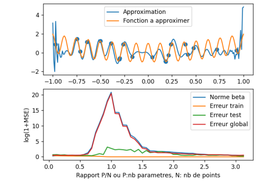
Dans le cas où $\hat{\beta}_0 = 0$ et $\max(\underline{\alpha_{\mathrm{j}}}') < \sigma_{\max}{}^{-2}$, en notant $t_{r+1} := t > t_r$ on a:

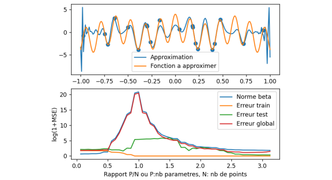
$$\sigma_{max}^{-1}\prod_{j=0}^{r}(1 - \underline{\alpha_j}'\sigma_{max}^2)^{t_{j+1}-t_j}||Y||_2 \leq ||\hat{\beta}_t - \hat{\beta}||_2 \leq \sigma_{min}^{-1}\prod_{j=0}^{r}(1 - \underline{\alpha_j}'\sigma_{min}^2)^{t_{j+1}-t_j}||Y||_2$$

Régression polynomial

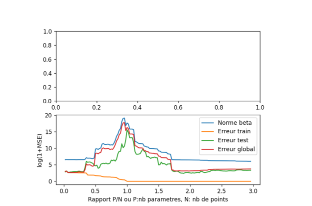
D = 1, N = 16 [1]	D = 1, N = 16	D = 1, N = 16
see plot 1	see plot 2	see plot 3
$y^\star(x) = exp(1 - \lvert x \rvert)$	$y^\star(x) = cos(50x) + x^2$	$y^\star(x) = 3cos(40x) + cos(10x) - x^2$

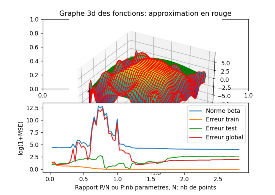






D = 3, N = 48	D = 2, N = 32	D = 2, N= 40
see plot 4	see plot 5	see plot 6
$y^\star(x) = exp(3 - (x_0^2 + x_1^2 + x_2^2))$	$y^\star(x) = exp(2 - (x_0^2 + x_1^2))$	$y^\star(x) = 2exp(2 - (x_0^2 + x_1^2))$





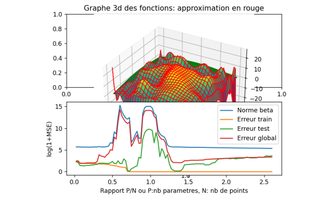


Figure: Phénomène de double descente en dimension 1, 2 et 3 pour des features polynomiales

MLP

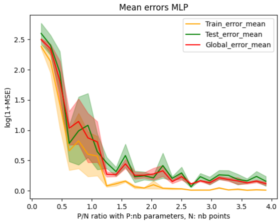
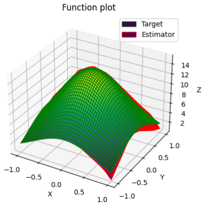
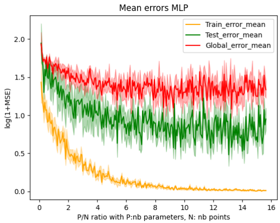
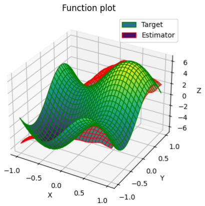
	
Risque moyen en fonction de $P \backslash N$, MLP	$y^{\star}(x) = 2exp(2 - (x_0^2 + x_1^2))$
	
Risque moyen en fonction de $P \backslash N$, MLP	$y^{\star}(x) = 20 \cdot [sin(3x_0) + sin(3x_1)] + \cdots$

Figure: Graphe de la MSE d'un algorithme Multilayer Perceptron.

Bibliographie I

[1] Emett Haddad. *Github Double Descente.* https://github.com/EmettGabrielH/Double-descente---Emett-Haddad. [Online]. 2024.