NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION

Prof. Schmidt-Hieber (2020)

Literatur

Schmidt-Hieber, Johannes (Aug. 2020). "Nonparametric regression using deep neural networks with ReLU activation function". In: *Annals of Statistics* 48.4, S. 1875–1897. ISSN: 0090-5364. DOI: `10.1214/19-aos1875`. URL: `http://dx.doi.org/10.1214/19-AOS1875`.

1 Modell

Beobachte n i.i.d. Datenpunkte aus dem nichtparametrisches multivariaten Regressionsmodell

$$Y_i = f_0(\boldsymbol{X}_i) + \varepsilon_i \qquad i = 1, \dots, n \tag{1}$$

- $Y_i \in \mathbb{R}$, $\boldsymbol{X}_i \in [0,1]^d$ zufällig
- $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ unabhängig von den $(\boldsymbol{X}_i)_i$

Ziel: Rekonstruktion der unbekannten Funktion $f_0 : [0,1]^d \to \mathbb{R}$ anhand der Daten $(\boldsymbol{X}_i, Y_i)_i$.

2 Notation

- Vektoren werden dick gedruckt $\boldsymbol{X}_i$
- $|\boldsymbol{x}|_p = (\sum_i |x_i|^p)^{1/p}$
- $|\boldsymbol{x}|_\infty = \max_{i=1,\dots,d} |x_i|$
- $|\boldsymbol{x}|_0 = \sum_i \mathbf{1}(x_i \neq 0)$

3 Neuronale Netzwerke

- Rectified Linear Unit (ReLU) als *Aktivierungsfunktion*:

$$\sigma : \mathbb{R} \to \mathbb{R} \qquad \sigma(x) := \max(0, x)$$

- Für $\mathbf{v} \in \mathbb{R}^r$ *verschobene Aktivierungsfunktion*:

$$\sigma_{\mathbf{v}} : \mathbb{R}^r \to \mathbb{R}^r \qquad \sigma_{\mathbf{v}}(\begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix}) := \begin{pmatrix} \sigma(y_1 - v_1) \\ \vdots \\ \sigma(y_r - v_r) \end{pmatrix}$$

- Ein Tupel $(L, \mathbf{p})$ mit einer *Tiefe* $L \in \mathbb{N}_{>0}$ und *Breitenvektor* $\mathbf{p} = (p_0, \dots, p_{L+1}) \in \mathbb{N}^{L+2}$ heißt *Netzwerk Architektur.*

- Eine Funktion $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ heißt *Neuronales Netzwerk* mit Netzwerk Architektur $(L, \mathbf{p})$, wenn *Gewichtsmatrizen* $W_i \in \mathbb{R}^{p_{i+1} \times p_i}$ $(i = 0, \dots, L)$ und *Verschiebungsvektoren* $\mathbf{v}_i \in \mathbb{R}^{p_i}$ $(i = 1, \dots, L)$ existieren, sodass für f gilt

$$f(\mathbf{x}) := W_L \sigma_{v_L} W_L \sigma_{\mathbf{v}_{L-1}} \dots W_1 \sigma_{\mathbf{v}_1} W_0 \mathbf{x} \tag{2}$$

  Die Gewichtsmatrizen W_i und Verschiebungsvektoren v_i heißen *Parameter* des Neuronalen Netzwerkes.

- $\mathcal{F}(L, \mathbf{p}) := \left\{ f \text{ von Gestalt (2)} : \max_{j=0\dots,L} ||W_j||_\infty \vee |\mathbf{v}_j|_\infty \leq 1 \right\}$

- $\mathcal{F}(L, \mathbf{p}, s, F) := \left\{ f \in \mathcal{F}(L, \mathbf{p}) : \sum_{j=0}^{L} ||W_j||_0 + |v_j|_0 \leq s, |||f|_\infty||_\infty \leq F \right\}$

Im Folgende bezeichne $\hat{f}_n$ ein Schätzer für die Regressionsfunktion f_0, welcher stets eine Funktion aus der Klasse $\mathcal{F}(L, \mathbf{p}, s, F)$ liefert

- $\Delta_n(\hat{f}_n, f_0) := \mathbb{E}_{f_0} \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_n(\mathbf{X}_i))^2 - \inf_{f \in \mathcal{F}(L, \mathbf{p}, s, F)} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i))^2 \right]$

- $\hat{f}_n^{\text{ERM}}$ mit $\Delta_n(\hat{f}_n^{\text{ERM}}, f_0) = 0$ heißt *Empirical Risk Minimizer*

- *Prediction error*

$$R(\hat{f}_n, f_0) = \mathbb{E}_{f_0}[(\hat{f}_n(\mathbf{X}) - f_0(\mathbf{X}))^2]$$

  mit $\mathbf{X} \overset{D}{=} \mathbf{X}_i$ unabhängig von den Daten $(\mathbf{X}_i, Y_i)_i$

4 Annahmen

4.1 Klassische Nichtparametrische Statistik

- *Annahme*: f_0 ist β-glatt
- *Minimax-Rate* für Prediction error $R(\hat{f}_n, f_0)$: $n^{\frac{-2\beta}{2\beta+d}}$
- **Problem:** d kann sehr groß werden $\implies$ *Curse of Dimensionality*

4.2 Annahmen von Prof. Schmidt-Hieber

- *Ziel:* Minimax-Rate unabhängig von der Input-Dimension d
- Die Regressionsfunktion f_0 lässt sich als Komposition von unterschiedlichen Funktionen schreiben

$$f_0 = g_q \circ g_{q-1} \circ \cdots \circ g_1 \circ g_0$$

mit $g_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$.

Schreibe $g_i = (g_{ij})^T_{j=1,\ldots,d_{i+1}}$ mit Komponenten g_{ij}. Es bezeichne t_i die kleinste Zahl, sodass alle g_{ij} von nicht mehr als t_i Variablen abhängen.

- β-**Hölder Funktionen** (mit Radius K)

$$\mathcal{C}_r^\beta(D, K) = \left\{ f : D \subseteq \mathbb{R}^r \to \mathbb{R} : \sum_{\boldsymbol{\alpha}:|\alpha|<\beta} ||\partial^{\boldsymbol{\alpha}} f||_\infty + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in D\\ x\neq y}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|_\infty^{\beta-\lfloor\beta\rfloor}} \leq K \right\}$$

- Definiere

$$\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K) := \Big\{ f = g_q \circ \cdots \circ g_0 : g_i = (g_{ij})_j : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}} \\ \text{mit } g_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([a_i, b_i]^{t_i}, K), \text{für } |a_i|, |b_i| \leq K \Big\}$$

mit $\boldsymbol{d} := (d_0, \ldots, d_{q+1}), \boldsymbol{t} := (t_0, \ldots, t_q), \boldsymbol{\beta} := (\beta_0, \ldots, \beta_q)$

- **Effektive Glattheits-Indizes**

$$\beta_i^* := \beta_i \prod_{\ell=i+1}^{q} (\beta_\ell \wedge 1)$$

- **Rate**

$$\phi_n := \max_{i=0,\ldots,q} n^{-\frac{2\beta_i^*}{2\beta_i^*+t_i}}$$

5 Hauptaussagen

Theorem 1. Betrachte das d-variate Regressionsmodell (1) mit unbekannter Regressionsfunktion f_0 aus der Klasse $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$. Es sei $\hat{f}_n$ ein Schätzer mit Werten in $\mathcal{F}(L, \boldsymbol{p}, s, F)$ sodass

(i) $F \geq \max(K, 1)$,

(ii) $\sum_{i=0}^{q} \log_2(4t_i \vee 4\beta_i) \log_2 n \leq L \lesssim n\phi_n$,

(iii) $n\phi_n \lesssim \min_{i=1,\ldots,L} p_i$,

(iv) $s \asymp n\phi_n \log n$

Dann existieren Konstanten C, C', die nur von $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, F$ abhängen, sodass

$$\Delta_n(\hat{f}_n, f_0) \leq C\phi_n L \log^2 n \quad \Longrightarrow \quad R(\hat{f}_n, f_0) \leq C'\phi_n L \log^2 n$$
$$\Delta_n(\hat{f}_n, f_0) \geq C'\phi_n L \log^2 n \quad \Longrightarrow \quad \frac{1}{C'}\Delta_n(\hat{f}_n, f_0) \leq R(\hat{f}_n, f_0) \leq C'\Delta_n(\hat{f}_n, f_0)$$

Korollar 1. (Obere Schranke) Sei $\hat{f}_n^{\text{ERM}} \in \arg\min f \in \mathcal{F}(L, \boldsymbol{p}, s, F) \sum_{i=1}^{n}(Y_i - f(\boldsymbol{X}_i))^2$ ein empirischer Risiko-Minimierer. Dann existiert unter den Annahmen von Theorem 1 eine Konstante C', die nur von $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, F$ abhängt, sodass

$$R(\hat{f}_n^{\text{ERM}}, f_0) \leq C'\phi_n L \log^2 n$$

Theorem 3. Betrachte das d-variate Regressionsmodell (1). Weiter habe die Verteilung der $\boldsymbol{X}_i$ auf $[0,1]^d$ eine Dichte bezüglich des Lebesgue-Maßes, welche von oben und unten durch positive Konstanten beschränkt ist. Dann existiert für beliebiges nicht-negatives q, beliebigen Dimensionsvektoren $\boldsymbol{d}$ und $\boldsymbol{t}$ mit $t_i \leq \min(d_0, \ldots, d_{i-1})$ für alle i, beliebige Regularitätsvektoren $\boldsymbol{\beta}$ und alle hinreichend große Konstanten $K > 0$, eine positive Konstante c, sodass

$$\inf_{\hat{f}_n} \sup_{f_0 \in \mathcal{G}(q,\boldsymbol{d},\boldsymbol{t},\boldsymbol{\beta},K)} R(\hat{f}_n, f_0) \geq c\phi_n$$

wobei das erste Infimum über alle Schätzer $\hat{f}_n$ gebildet wird.

Theorem 2. Betrachte das d-variate Regressionsmodell (1) mit unbekannter Regressionsfunktion f_0 mit $||f_0||_\infty \leq F$ für ein $F \geq 1$. Sei $\hat{f}_n$ ein beliebiger Schätzer mit Werten in $\mathcal{F}(L, \boldsymbol{p}, s, F)$. Dann existiert für jedes $\varepsilon \in (0, 1]$ eine Konstante C_ε, die nur von ε abhängt, sodass mit

$$\tau_{\varepsilon,n} := C_\varepsilon F^2 \frac{(s+1)\log(n(s+1)^L p_0 p_{L+1})}{n}$$

gilt

$$(1-\varepsilon)^2 \Delta_n(\hat{f}_n, f_0) - \tau_{\varepsilon,n} \leq R(\hat{f}_n, f_0) \leq (1+\varepsilon^2)\left(\inf_{f \in \mathcal{F}(L,\boldsymbol{p},s,F)} ||f - f_0||_\infty^2 + \Delta_n(\hat{f}_n, f_0)\right) + \tau_{\varepsilon,n}$$