5.2 Lokalna regresija $(p=1)$ [ESL - Poglavlje 1]

5.2.1 Jezgre (engl. kernel)

kNN regresija: za dani $k \leq n$, $\forall x \in \mathbb{R}$,

$$\hat{f}(x) := \frac{1}{k} \sum_{x^{(i)} \in N_k(x)} y_i .$$

skup od k najbližih susjeda od x u $\{x^{(1)}, \dots, x^{(n)}\}$

"weight"

Za $w_i(x) := \frac{1}{k} \mathbb{1}_{\{x^{(i)} \in N_k(x)\}}$, $x \in \mathbb{R}$, imamo

$$\hat{f}(x) = \sum_{i=1}^{n} w_i(x) y_i . \qquad (5.20)$$

"linear smoother"

Uočimo, $x \mapsto w_i(x)$ nije neprekidna $\Rightarrow$ $\hat{f}$ nije neprekidna.

Za danu f-ju $K: \mathbb{R}_+ \to \mathbb{R}_+$ ("jezgra") i $\lambda > 0$ ("bandwidth") defin.

$$K_\lambda(x, x^{(i)}) := K\left(\frac{|x - x^{(i)}|}{\lambda}\right), \quad \forall x, x^{(i)}. \qquad (5.21)$$

tzv. Nadaraya - Watson procjenitelj $\hat{f}^{(NW)}(x)$ dobijemo ako u (5.20) stavimo

$$w_i(x) := \frac{K_\lambda(x, x^{(i)})}{\sum_{j=1}^{n} K_\lambda(x, x^{(j)})}, \quad i = 1, \dots, n \qquad (5.22)$$

Nap.] Zbog normalizacije u (5.22) imamo

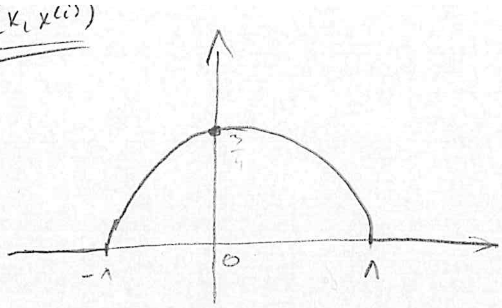
$$y_i = c, \ \forall i = 1, \dots, n \Rightarrow \hat{f}(x) = c, \ \forall x \in \mathbb{R}.$$

$\blacksquare$

Tipične jezgre: ← nije bitno za $K_\lambda(x, x^{(i)})$

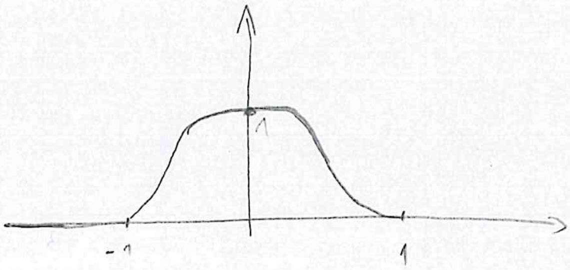
(i) $K(t) = \frac{3}{4}(1-t^2)\,\mathbb{1}_{\{|t|\le 1\}}$

(Epanechnikova jezgra)



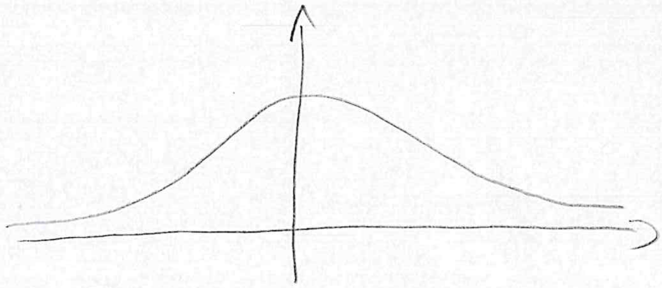
(ii) $K(t) = \left(1-|t|^3\right)^3 \mathbb{1}_{\{|t|\le 1\}}$

("Tricube" jezgra)



(iii) $K(t) = f_{N(0,1)}(t)$

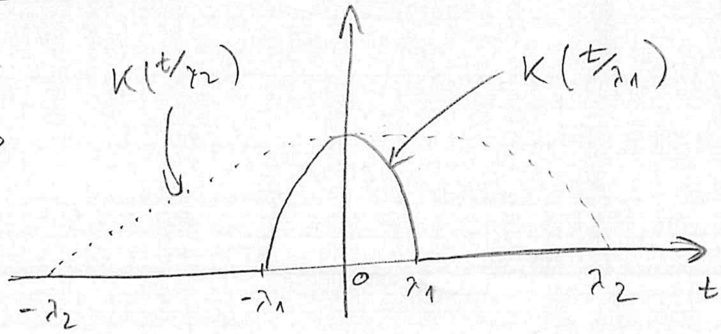
(Gaussova jezgra)



Hiperparametar λ kontrolira "veličinu okoline" tj. za $\lambda > 0$,

$K_\lambda(x_0, x^{(i)}) = K\left(\frac{|x_0 - x^{(i)}|}{\lambda}\right)$, a

npr. (i)



Nap.] • težina $w_i(x)$ veća za $x^{(i)}$-eve bliže x, a $w_i(x) \to 0$ kako $|x^{(i)} - x| \to +\infty$ (u (i) i (ii), $w_i(x) = 0$ ako $|x^{(i)} - x| \geq \lambda$)

• $x \mapsto w_i(x)$ neprekidne $\Rightarrow$ NW procjenitelj $\hat{f}$ je neprekidan.

• λ kontrolira kompleksnost, tj. BVT

→ koristimo CV za odabir λ

→ izbor jezgre zapravo nebitan! ■

Nap. 5.7] Umjesto fiksne veličine okoline λ za sve $x \in \mathbb{R}$, možemo koristiti

$$K_\lambda(x, x^{(i)}) = K\left(\frac{|x - x^{(i)}|}{h_\lambda(x)}\right) \qquad (5.23)$$

gdje je $h_\lambda : \mathbb{R} \to \langle 0, \infty \rangle$ f-ja (koja može ovisiti o $\{x^{(1)}, \dots, x^{(n)}\}$).

npr. 1 • $\lambda = k \in \mathbb{N}$ ($k \le n$)

• $h_k(x)$ = udaljenost od x do k-tog najbližeg susjeda u $\{x^{(1)}, \dots, x^{(n)}\}$

• $K(t) := \mathbb{1}\{t \in [-1, 1]\}$ (uniformna jezgra)

za $x \in \mathbb{R}$,

$$K_k(x, x^{(i)}) = \begin{cases} 1, & x^{(i)} \in N_k(x) \\ 0, & \text{inače} \end{cases}$$

te $\sum_{i=1}^{n} K_k(x, x^{(i)}) = k$

$\Rightarrow \quad w_i(x) = \frac{1}{k} \mathbb{1}_{\{x^{(i)} \in N_k(x)\}}$, tj. $\boxed{\text{NW procjenitelj} = k\text{NN}}$. ⊠

Ipak, NW proc. $\hat{f}(x)$ ima veliku pristranost ako točke $x^{(1)}, \dots, x^{(n)}$ nisu "simetrično" raspoređene oko x [osim ako je f^* lokalno oko x konstantna.]

→ najveći problem za rubne x-eve!

5.2.2 Lokalna linearna regresija

DZ Pokažite da NW procjenitelj $\hat{f}(x)$ zadovoljava $\hat{f}(x) = \hat{\beta}_0(x)$ za

$$\hat{\beta}_0(x) := \underset{\beta_0 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \underbrace{K_\lambda(x, x^{(i)})}_{\text{ili } w_i(x) \text{ iz } (5.22)} [y_i - \beta_0]^2, \quad \forall x \in \mathbb{R} \qquad (5.24)$$

⊠

Lok. lin. regresija: za dani $\lambda > 0$, $\forall x \in \mathbb{R}$, $\boxed{\hat{f}(x) := \hat{\beta}_0(x) + \hat{\beta}_1(x) \cdot x}$ za

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \underset{\beta_0, \beta_1 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} K_\lambda(x, x^{(i)}) [y_i - \beta_0 - \beta_1 x^{(i)}]^2 \qquad (5.25)$$

koristi se samo za $\hat{f}(x)$!

"weighted LS" problem

Za $\forall x \in \mathbb{R}$, $\boxed{b(x)^T := (1, x) \in \mathbb{R}^2}$,

$$\mathbb{B} := \begin{bmatrix} b(x^{(1)})^T \\ \vdots \\ b(x^{(n)})^T \end{bmatrix} \in \mathbb{R}^{n \times 2}, \quad \mathbb{W}_\lambda(x) := \mathrm{diag}\left(K_\lambda(x, x^{(i)}),\ i = 1, \dots, n\right) \in \mathbb{R}^{n \times n} \tag{5.26}$$

$$(\hat\beta_0(x), \hat\beta_1(x)) = \underset{\beta = (\beta_0, \beta_1) \in \mathbb{R}^2}{\mathrm{argmin}} (y - \mathbb{B}\beta)^T \, \mathbb{W}_\lambda(x) \, (y - \mathbb{B}\beta)$$

$$= \left[y_\lambda(x) := \sqrt{\mathbb{W}_\lambda(x)} \cdot y, \ \mathbb{B}_\lambda(x) := \sqrt{\mathbb{W}_\lambda(x)} \cdot \mathbb{B} \right]$$

$$= \underset{\beta \in \mathbb{R}^2}{\mathrm{argmin}} \| y_\lambda(x) - \mathbb{B}_\lambda(x) \cdot \beta \|_2^2$$

LS problem!

$$\Rightarrow \boxed{\hat\beta(x)} = (\hat\beta_0(x), \hat\beta_1(x)) = \left(\mathbb{B}_\lambda(x)^T \mathbb{B}_\lambda(x)\right)^{-1} \mathbb{B}_\lambda(x)^T y_\lambda(x)$$

$$= \boxed{\left(\mathbb{B}^T \mathbb{W}_\lambda(x) \mathbb{B}\right)^{-1} \mathbb{B}^T \mathbb{W}_\lambda(x) \, y} \tag{5.27}$$

ne ovisi o y! * tzv. "ekvivalentna jezgra"

Specijalno, $\forall x \in \mathbb{R}$,

$$\hat f(x) = b(x)^T \cdot \hat\beta(x) = \sum_{i=1}^n l_i(x) \cdot y_i, \tag{5.28}$$

za $l(x)^T = (l_1(x), \dots, l_n(x)) := b(x)^T (\mathbb{B}^T \mathbb{W}_\lambda(x) \mathbb{B})^{-1} \mathbb{B}^T \mathbb{W}_\lambda(x) \in \mathbb{R}^{1 \times n}$. (5.29)

<u>Pristranost</u>: ako je $Y_i = \overset{\text{točni}}{f}(x^{(i)}) + e_i$, (5.28) daje $\forall x \in \mathbb{R}$,

$$\mathbb{E}[\hat f(x)] \overset{*}{=} \sum_{i=1}^n l_i(x) \cdot f(x^{(i)})$$

Taylor oko x, $\forall i$

$$= f(x) \sum_{i=1}^n l_i(x) + f'(x_0) \sum_{i=1}^n l_i(x)(x^{(i)} - x) + \overset{\text{ostatak}}{R} \tag{5.30}$$

[vrijedi za svaki $\hat f$ oblika (5.28) (npr. NW procjen.)]

$$\overset{\text{Lema 5.18}}{=} f(x) + \boxed{0} + R$$

↳ u lok. lin. regr. $\mathbb{E}[\hat f(x)] - f(x) = R$ ← ne ovisi o nagibu od f! (Za razliku od NW procjenitelja)

Lema 5.8 $\sum_{i=1}^{n} l_i(x) = 1, \quad \sum_{i=1}^{n} l_i(x)(x^{(i)} - x) = 0, \quad \forall x \in \mathbb{R}.$

Dokaz Iz (5.27) i (5.28) za $y = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ te $y = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(n)} \end{bmatrix}$, imamo

$$\left[\sum_{i=1}^{n} l_i(x), \sum_{i=1}^{n} l_i(x) x^{(i)} \right] = b(x)^T (\mathbb{B}^T \mathbb{W}(x) \mathbb{B})^{-1} \mathbb{B}^T \mathbb{W}(x) \cdot \underbrace{\begin{bmatrix} 1 & x^{(1)} \\ \vdots & \vdots \\ 1 & x^{(n)} \end{bmatrix}}_{= \mathbb{B}}$$

$$= b(x)^T = [1, x]$$ ∎

(5.2.3) Lokalna polinomijalna regresija ("loess")

za $d \in \mathbb{N}_0$, $\hat{f}(x) = \hat{\beta}_0(x) + \sum_{j=1}^{d} \hat{\beta}_j(x) x^j$ za

$$\hat{\beta}(x) = (\hat{\beta}_0(x), \ldots, \hat{\beta}_d(x)) := \underset{\beta \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i=1}^{n} K_\lambda(x, x^{(i)}) \left[y_i - \sum_{j=0}^{d} \hat{\beta}_j(x) x^j \right]$$

- $d = 0 \Rightarrow$ NW procjenitelj
- $d = 1 \Rightarrow$ lok. lin. regresija

Opet, za $b(x)^T = (1, x, \ldots, x^d)$ te $\mathbb{B} \in \mathbb{R}^{n \times (d+1)}$ i $\mathbb{W}_\lambda(x)$ kao su u (5.26),

$$\hat{f}(x) = \sum_{i=1}^{n} l_i(x) y_i, \qquad (5.31)$$

za $l(x)^T = (l_1(x), \ldots, l_n(x))$ kao u (5.29).

Može se pokazati (DZ) da $\boxed{\mathbb{E}[\hat{f}(x)] - f(x)}$ u (5.30) ovisi samo o članovima uz $f^{(k)}(x)$ za $\boxed{k \geq d+1}$

$\rightarrow$ značajno smanjenje pristranosti za veći d ako je f "zakrivljena" oko x

$\rightarrow$ ipak, veća $\mathrm{Var}(\hat{f}(x))$, pogotovo na rubu područja (ako nas zanima ekstrapolacija, tipično $\boxed{d=1}$)

Nap. 5.10 (i) (5.31) $\Rightarrow \hat{y} = L_\lambda \cdot y$ za $L_\lambda = \begin{bmatrix} - l(x^{(1)})^T - \\ \vdots \\ - l(x^{(n)})^T - \end{bmatrix} \in \mathbb{R}^{n \times n}$ $\Rightarrow \boxed{d_\lambda = \mathrm{tr}(L_\lambda)}$

ovise o λ!

(ii) λ biramo CV metodom

(vrijede analogni od (5.18) i (5.19) za LOOCV i GCV!) ∎