(2) Procjena greške i odabir modela

[Model selection]

Za dani $\mathcal{T}$, zanima nas procjena greške

$$L(\hat{f}_\alpha) = \mathbb{E}\left[L(Y, \hat{f}_\alpha(X))\right]. \quad (2.1)$$

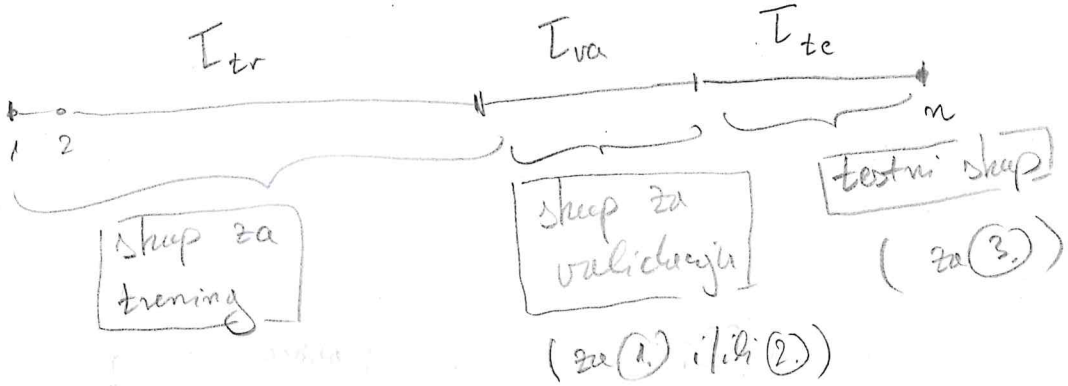
$\hat{f}_\alpha = \hat{f}_\alpha(\mathcal{T})$

↳ hiperparametar koji kontrolira kompleksnost modela (npr. $\alpha = \frac{1}{k}$ ili k u KNN metodi).

Ciljevi:

1. Odabir hiperparametra α; [Model selection]
2. Odabir između različitih modela; [Model selection]
3. Procjena testne greške odabranog modela. [Model assesment]

Ako je n velik, najbolji pristup je (na slučajan način) podijeliti $\mathcal{T}$:



1. $\mathcal{T}_{tr}$ → prilagodba mnogih modela i/ili istog modela za različite α [?]
2. $\mathcal{T}_{va}$ → za svaki model ili α procjenjujemo testnu grešku (2.1) te npr. izaberemo onaj s najmanjom procjenjenom greškom

3. T_{te} → procjenjujemo testnu grešku (2.1) odabranog modela.

Napr. 2.11 Zašto trebamo T_{te}?

npr. ako je T_{va} realizacija od njih nekoliko

$$T_{va} = \{(\tilde{X}^{(i)}, \tilde{Y}_i) : i = 1, \dots, m\}$$

za (X, Y), za svaku fiksnu f-ju $f: \mathbb{R}^p \to \mathbb{R}$,

$$L_{T_{va}}(f) = \frac{1}{m} \sum_{i=1}^{m} L(\tilde{Y}_i, f(\tilde{X}^{(i)})) \overset{\text{CGT}}{\approx} N\left(\underbrace{\mathbb{E}[L(Y, f(X))]}_{= L(f)}, \frac{\mathrm{Var}(L(Y, f(X)))}{m} \right) \quad (2.2)$$

[slučajna varijabla] $\forall f$!

ako je m dovoljno velik.

Pak, ako računamo $L_{T_{va}}(\hat{f}_\alpha)$ za sve α te $\hat{f}_\alpha = \hat{f}_\alpha(T_{tr})$

$$\hat{\alpha} = \arg\min_{\alpha} L_{T_{va}}(\hat{f}_\alpha), \quad (2.3)$$

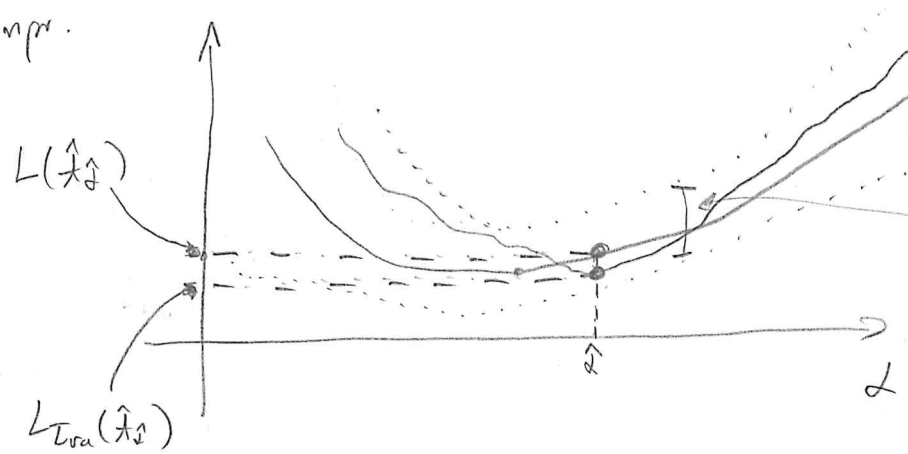
imamo

$$\mathbb{E}\left[\underbrace{L_{T_{va}}(\hat{f}_{\hat{\alpha}})}_{\leq L_{T_{va}}(\hat{f}_\alpha), \forall \alpha \text{ po } (2.3)} \right] \leq \min_{\alpha} \underbrace{\mathbb{E}[L_{T_{va}}(\hat{f}_\alpha)]}_{= L(\hat{f}_\alpha)} \leq L(\hat{f}_{\hat{\alpha}})$$

[slučajna varijabla]

$$\mathbb{E}\left[L(\hat{f}_{\hat{\alpha}}) - L_{T_{va}}(\hat{f}_{\hat{\alpha}}) \right] \geq 0 \text{ ! [pristranost]}$$

npr.



Ovisi o $\boxed{m}$
i $Var(L(Y, \hat{f}_\alpha(x)))$
(vidi (2.2))

[veći m, manja greška]

Slično, ako imamo više različitih modela $\rightarrow \hat{f}^{(1)}, \dots, \hat{f}^{(k)}$ te

$$\hat{i} := \underset{i=1,\dots,k}{\operatorname{argmin}} L_{Tva}(\hat{f}^{(i)}),$$

imamo

$$\mathbb{E}\left[L_{Tva}(\hat{f}^{(\hat{i})})\right] \leq L(\hat{f}^{(\hat{i})}).$$

[Malo je kompliciranije u slučaju kada biramo i $\hat{f}^{(i)}$ i $\hat{i}$.]

(2.1) Unakrsna validacija (CV)

Kada (m) nije velik, tj. T_{tr}, T_{va} i T_{te} su premaleni, "izbacujemo" T_{va} te odabir modela i/ili hiperparametara radimo na $\underline{T_{tr}}$.

↳ u tom slučaju CV predstavlja najjednostavniju i najpopularniju metodu.

"k-fold CV": (neka je $\mathcal{L} := \mathcal{L}_{tr}$)

- za dani $k \in \{2, 3, \dots, n\}$ (na slučajan način) podijelimo $\mathcal{L}$ na k približno jednakih disjunktnih "dijelova" ("folds")

$\mathcal{L}_1, \dots, \mathcal{L}_k$

→ neka je $n_j := |\mathcal{L}_j| = \#$ elemenata u $\mathcal{L}_j$, $j = 1, \dots, k$

→ pretp. $r_k := \frac{n}{k} \in \mathbb{N} \Rightarrow$ $\boxed{n_j = r_k}$, $\forall j = 1, \dots, k$ ← veličina svakog bloka

- $\forall j = 1, \dots, k$, računamo [ovisi o algoritmu]

$$CV_j^{(k)} = CV_j^{(k)}(\hat{f}) := \frac{1}{r_k} \sum_{(x^{(i)}, y_i) \in \mathcal{L}_j} L\left(y_i, \hat{f}^{-j}(x^{(i)})\right), \quad (2.4)$$

pri čemu je

$$\boxed{\hat{f}^{-j} := \hat{f}(\mathcal{L}^{-j})}$$

za

$$\boxed{\mathcal{L}^{-j} := \bigcup_{i \neq j} \mathcal{L}_i}$$

svi osim $\mathcal{L}_j$

- k-CV procjena testne greške od $\hat{f} = \hat{f}(\mathcal{L})$ je

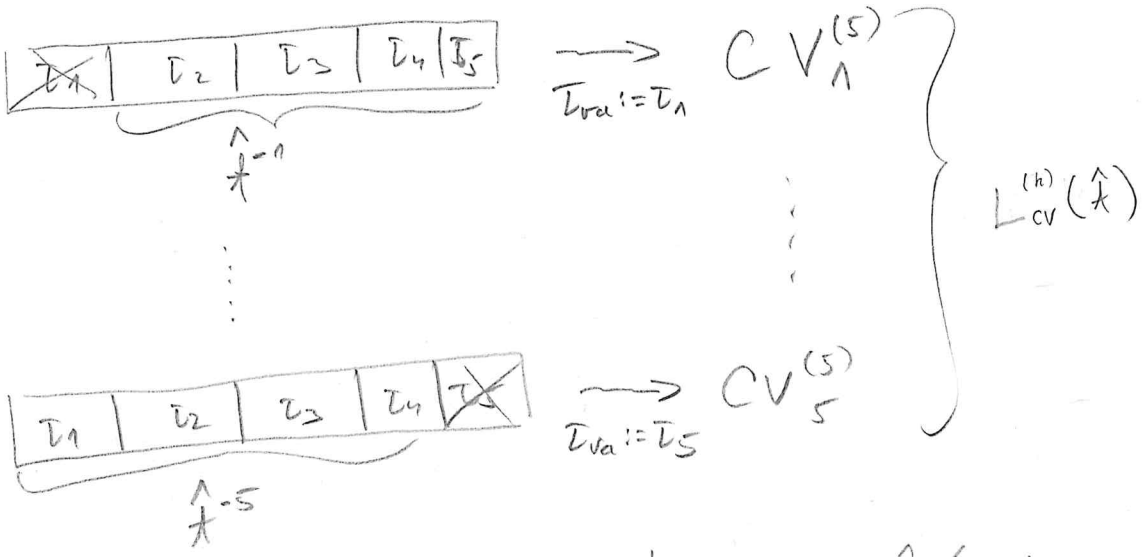
$$L_{CV}^{(k)}(\hat{f}) := \frac{1}{k} \sum_{j=1}^{k} CV_j^{(k)} \quad (2.5)$$

$$= \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, \hat{f}^{-j(i)}(x^{(i)})\right), \quad (2.6)$$

pri čemu je $j(i) \in \{1, \dots, k\}$ t.d. $(x^{(i)}, y_i) \in \mathcal{L}_j$.

[(2.6) je dobro def. i u slučaju kada $\frac{n}{k} \notin \mathbb{N}$, tj. blokovi su različite duljine.]

npr. $k=5$



↳ u j-tom koraku $\mathcal{I}_j$ igra ulogu skupa za validaciju (tj. nezavisnog testnog skupa). ▣

Nap. 2.2 Ispostavlja se da $L_{CV}^{(k)}(\hat{f})$ bolje procjenjuje $\mathbb{E}_{\mathcal{T}}[L(\hat{f})]$ nego $L(\hat{f})$ (za dani $T = \mathcal{I}$) (vidi ESL, poglavlje 7).
Ipak, ako $L_{CV}^{(k)}$ koristimo za odabir modela / hiperparametra typično izabrano model / hiperparameter koji ima manju testnu grešku $L(\hat{f})$. [CV je još uvijek teoretski nedovoljno istražena!] ▣

Kako odabrati k?

↳ $|\mathcal{I}^{-j}| = n - r_k = \frac{k-1}{k} \cdot n < n$

$\Rightarrow$ typično imamo $\mathbb{E}_{\mathcal{T}}[L_{CV}^{(k)}(\hat{f})] > \mathbb{E}_{\mathcal{T}}[L(\hat{f})]$ za male k (odnosno velike r_k) ← pristranost!

S druge strane, u slučaju $\boxed{k=n}$, tj. $T^{-j} = T \setminus \{(x^{(j)}, y_j)\}$, $j=1,\dots,n$, imamo manju pristranost, ali budući da su T^{-j}, $j=1,\dots,n$ vrlo slični, procjene $CV_j^{(n)}$, $j=1,\dots,n$ su pozitivno koralirane pa je $Var_T(L_{CV}^{(n)}(\hat{A}))$ veća.

[opet TBVT!]

→ tzv. "leave-one-out CV" (LOOCV)

[→ za velike n, $L_{CV}^{(n)}$ može biti zahtjevno za izračunati.]

U praksi najčešće $\boxed{k=5 \text{ ili } 10}$ te $\boxed{k=n}$.

[Također ne postoji puno teorijskih rezultata.]

Neka je $\hat{\lambda} := \hat{\lambda}_{min}$ za $\lambda_{min} := \arg\min_{\lambda} L_{CV}^{(k)}(\hat{\lambda}_\lambda)$ često se uzima $\hat{\lambda}$ koji odgovara najjednostavnijem modelu t.d.

$$L_{CV}^{(k)}(\hat{\lambda}) \leq L_{CV}^{(k)}(\hat{\lambda}_{\lambda_{min}}) + \widehat{SE}(\lambda_{min}),$$

gdje je

$$\widehat{SE}(\lambda) := \sqrt{\frac{Var(CV_1^{(k)}, \dots, CV_k^{(k)})}{k}} \qquad (2.7)$$

(uzoračka varijanca)

$SE(\lambda) = \sqrt{Var_T(L_{CV}^{(k)}(\hat{\lambda}_\lambda))}$

(nearivni) procjenitelj za standardnu grešku procjenitelja $L_{CV}^{(k)}(\hat{\lambda}_\lambda)$.

- → (2.7) ima smisla kada je k velik i $CV_1^{(k)}, \dots, CV_k^{(k)}$ približno nekorelirani! ((2.7) podcjenjuje stvarnu std. grešku)
- tzv. "one SE rule"

- analogne prenke možemo primijeniti kod odabira modela
- Općenit princip: Između jednako preciznih modela (do na jednu SE), biramo najjednostavniji.

→ CV_ilustracija.R

Pr. 2.4 Pretp. da je $p = 5000$, a $n = 100$ (dakle, $p >> n$), te $Y \in \{0,1\}$. Naš algoritam je

(1.) Na temelju T izračunaj uzoračke korelacije

$$\rho((x_{1j}, y_1), (x_{2j}, y_2), \dots, (x_{nj}, y_n)),$$

j-te kovarijate i odziva, $j = 1, \dots, p$.

→ Ostavi samo $m = 100$ kovarijata sa najvećom korelacijom

(2.) Koristeći samo tih m kovarijata i neku metodu generiraj $\hat{f}$.

↳ Kako izgleda CV algoritam za procjenu greške?

→ i (1.) mora biti uključen u svaki korak CV algoritma!

(vidi ESL, poglavlje 7.10.2)

Npr., neka je

- $X_1, \dots, X_p$ nezavisne s $N(0,1)$ razdiobom
- $Y \sim \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$ nezavisno od $X_1, \dots, X_p$

$\Rightarrow \forall f: \mathbb{R}^p \to \{0,1\}$ vrijedi

$$L(t) = \mathbb{P}(Y \neq t(X))$$

prep. 0-1 gubitak

$$= \frac{1}{2}\mathbb{P}(t(X)=1 \mid Y=0) + \frac{1}{2}\mathbb{P}(t(X)=0 \mid Y=1)$$

$$= \frac{1}{2}\mathbb{P}(t(X)=1) + \frac{1}{2}\mathbb{P}(t(X)=0) = \boxed{\frac{1}{2}}$$

(nez. od Y i X)

[neraun!]

$\Rightarrow$ za bilo koju metodu $\hat{t}$, $\hat{t} \in T$,

$$L(\hat{t}) = \frac{1}{2} = \mathbb{E}_T[L(\hat{t})].$$

→ CV_primjer.R

Nap. Ukoliko radi kompletnosti radimo neku metodu nenadziranog učenja (npr. analiza glavnih komponenti), tj. ne koristimo y_i-eve, taj korak nije potrebno uvrstiti u CV algoritam!

(2.2) Stupnjevi slobode (degrees of freedom) → (df)

↳ Želimo usporediti kompleksnost različitih metoda

Pretp. da je model kao u (1.10), $Y = t(X) + \varepsilon$, te nadalje da je

$$Y_i = \underbrace{t(x^{(i)})}_{=:\mu_i} + \varepsilon_i, \quad i = 1, \dots, n$$

uz

- $\varepsilon_1, \dots, \varepsilon_n$ nez. varijable t.d. $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $\forall i \neq j$
- $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$ fiksni → $T = \{(x^{(i)}, Y_i) : i = 1, \dots, n\}$ ("fixed-design")

(alternativno, mogli smo u nastavku uvjetovati na $X^{(i)} = x^{(i)}$, $i = 1, \dots, n$

uz $Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^{n \times 1}$, $\mu = (\mu_1, \dots, \mu_n)^T \in \mathbb{R}^{n \times 1}$ te

$\mathcal{E} = (\varepsilon_1, \dots, \varepsilon_n)^T \in \mathbb{R}^{n \times 1}$, (2.8) postaje

$$Y = \mu + \mathcal{E}, \quad \mathbb{E}[\mathcal{E}] = 0, \quad \mathrm{Cov}(\mathcal{E}) = \sigma^2 I_n$$

↑ kovarijacijska matrica

Za metodu $A: T \to \hat{\mathcal{A}}(T)$ definiramo

$$dt(A) := \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{Y}_i, Y_i), \quad (2.9)$$

↑ jedina slučajnost dolazi od $\varepsilon_1, \dots, \varepsilon_n$

pri čemu je $\hat{Y}_i := \hat{f}(x^{(i)})$, $i = 1, \dots, n$.

[dakle, ako se "više" prilagođavamo podacima, Cov pa raste time i dt raste !]

Motivacija za (2.9): usporedit ćemo

- $\mathbb{E}\left[\underbrace{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{L_T(\hat{f})} \right]$ ← očekivanu grešku na skupu za treniranje

- $\mathbb{E}\left[\underbrace{\frac{1}{n} \sum_{i=1}^{n} (\tilde{Y}_i - \hat{Y}_i)^2}_{L_{\tilde{T}}(\hat{f})} \right]$ ← tzv. (očekivanu) "in-sample" testnu grešku

gdje $\tilde{T} = \{(x^{(i)}, \tilde{Y}_i) : i = 1, \dots, n\}$ testni skup nezavisan

od T ($\tilde{Y}_i = \mu_i + \tilde{\varepsilon}_i$, $i = 1, \dots, n$, $(\tilde{\varepsilon}_i)_{i=1,\dots,n}$ nez. od $(\varepsilon_i)_{i=1,\dots,n}$).

Prop. 2.5 Vrijedi

$$\underbrace{\mathbb{E}\left[L_{\tilde{T}}(\hat{A})\right] - \mathbb{E}\left[L_T(\hat{A})\right]}_{\text{tzv. očekivani "optimizam" od } L_T(\hat{A})} = \frac{2\sigma^2}{n} \cdot df(A) \qquad (2.10)$$

Dokaz:

Lijeva strana u (2.10) $= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{Y}_i^2 - 2\tilde{Y}_i\hat{Y}_i + \hat{Y}_i^2 - Y_i^2 + 2Y_i\hat{Y}_i - \hat{Y}_i^2\right\}\right]$

$= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}[\tilde{Y}_i^2] - 2\,\mathbb{E}[\tilde{Y}_i\hat{Y}_i] - \mathbb{E}[Y_i^2] + 2\,\mathbb{E}[Y_i\hat{Y}_i]\right\}$

$\mathbb{E}[\tilde{Y}_i^2] = \mathrm{Var}(\tilde{Y}_i) + \mu_i^2 = \sigma^2 + \mu_i^2$

$\mathbb{E}[\tilde{Y}_i\hat{Y}_i] = \mathbb{E}[\tilde{Y}_i]\,\mathbb{E}[\hat{Y}_i] = \mu_i$

$\mathbb{E}[Y_i^2] = \mathrm{Var}(Y_i) + \mu_i^2 = \sigma^2 + \mu_i^2$

$= \frac{2}{n}\sum_{i=1}^{n}\left\{\mathbb{E}[Y_i\hat{Y}_i] - \mu_i \cdot \mathbb{E}[\hat{Y}_i]\right\} = \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(Y_i, \hat{Y}_i)$ ∎

($\mathbb{E}[Y_i] = \mu_i$)

Pr. 2.6

(i) Ako je $\hat{Y}_i := \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\forall i$ $\Rightarrow df(A) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Cov}(Y_i, \bar{Y}) = 1.$

($\mathrm{Cov}(Y_i, \bar{Y}) = \frac{1}{n}\mathrm{Var}(Y_i) = \frac{1}{n}\sigma^2$)

(ii) Ako je $\hat{Y}_i := Y_i$, $\forall i$ $\Rightarrow df(A) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = n.$

($\mathrm{Var}(Y_i) = \sigma^2$)

(iii) Kod lin. regresije metodom najmanjih kvadrata, $df(A) = p$ (hasrije)

(iv) Kod kNN regresije, $df(A) = \frac{n}{k}$ (DZ)

Svi primjeri iz Pr. 2.5 su oblika

$$\hat{Y}_i = \sum_{j=1}^{n} s_{ij} Y_j \,, \quad i = 1, \dots, n \qquad (2.11)$$

gdje je $S = (s_{ij} : i,j \in \{1,\dots,n\}) \in \mathbb{R}^{n \times n}$ (deterministička) matrica čiji elementi ovise samo o $x^{(1)}, \dots, x^{(n)}$.

Prop. 2.7 Ako je A t.d. vrijedi (2.11),

$$df(A) = tr(S) \qquad (2.12).$$

Dokaz:

$$Cov(Y_i, \hat{Y}_i) = Cov\left(Y_i, \sum_{j=1}^{n} s_{ij} Y_j\right)$$

$$= \sum_{j=1}^{n} s_{ij} \underbrace{Cov(Y_i, Y_j)}_{=0 \ \forall j \neq i} = s_{ii} \cdot \underbrace{Var(Y_i)}_{=\sigma^2}, \quad \forall i = 1, \dots, n$$

(s_{ij} nisu slučajni)

Kor. 2.7 Iz (2.11) slijedi da je statistika

$$T = L_T(\hat{A}) + \frac{2\sigma^2}{n} df(A) \qquad (2.13)$$

nepristran procjenitelj za $\mathbb{E}\left[L_T^2(\hat{A})\right]$

→ T možemo koristiti za odabir modela.

→ kod lin. regresije metodom najmanjih kvadrata,

$$T = L_T(\hat{A}) + \frac{2\sigma^2}{n} \cdot p \qquad (2.14)$$

← tzv. C_p statistika

(koristi se za odabir varijabli)