1 Uvod

1.1 Problem (nadziranog) učenja

Pretp. da imamo podatke

$$\mathcal{I} = \{ (x^{(i)}, y_i) : i = 1, \dots, n \}$$

← uzorak ili "skup za učenje" [training set]

gdje je npr.

- y_i = cijena i-te nekretnine ← odziv ili zavisna varijabla [response]
- $x^{(i)} = (x_{i1}, x_{i2}, \dots, x_{ip})$ ← kovarijate ili nezavisne varijable ($p \in \mathbb{N}$ je broj kovarijata)
  - x_{i1}: veličina (u m^2)
  - x_{i2}: pogled na more (DA/NE)

Najčešće, cilj učenja je predikcija:

$$\mathcal{I} \rightsquigarrow \hat{f} = \hat{f}(\mathcal{I}),$$

t.d. je $\hat{f}(\tilde{x})$ "dobra" procjena za $\tilde{y}$, pri čemu je $(\tilde{x}, \tilde{y})$ "novi" par.

Ako je

- y kvantitativna (neprekidna) varijabla

  → problem regresije (oznaka $y \in \mathbb{R}$)

- y kategorijalna varijabla

  → problem klasifikacije

  (BSHOP $y \in S = \{0, 1, \dots, K-1\} \subseteq \mathbb{R}$)

(1.2) Teorija statističkog učenja

[statistički pristup problemu učenja → TSU (Sandrić)]

Pretpostavke

- očekivanje Y je skalarna varijabla
- koreanjete $X = (X_1, \dots, X_p)$ su skalarni vektor u $\mathbb{R}^p$

  (X i Y su slučajni ! npr. $Y = f(X) + \varepsilon$)

- T je jedna realizacija od

$$T = \{ (X^{(i)}, Y_i) : i = 1, 2, \dots, n \},$$

gdje su, tipično, $(X^{(i)}, Y_i)$, $i = 1, \dots, n$, nezavisni te $(X^{(i)}, Y_i) \sim (X, Y)$.

Cilj: Za danu funkciju gubitka $L : \mathbb{R}^2 \to \mathbb{R}$ [loss function] tražimo $\hat{f} = \hat{f}(T) : \mathbb{R}^p \to \mathbb{R}$ za koju minimum postiže najmanjom pogreškom

$$L(\hat{f}) := \mathbb{E}[L(Y, \hat{f}(X))]. \quad (1.1)$$

[$L(y, \hat{f}(x))$ je "trošak" predikcije $\hat{f}(x)$ ako je stvarna vrijednost y]

Nap.1 Ako imamo realizaciju (testni skup)

$$\tilde{T} = \{(\tilde{x}^{(i)}, \tilde{y}_i) : i = 1, \dots, m\}$$

koji je jednak [mjeri uzorke iz (X, Y)], za velike m,

$$L_{\tilde{T}}(\hat{f}) := \frac{1}{m} \sum_{i=1}^{m} L(\tilde{y}_i, \hat{f}(\tilde{x}^{(i)})) \overset{\text{zvuči}}{\simeq} L(\hat{f}). \quad (1.2)$$

$L_{\tilde{T}}(\hat{f})$: greška na $\tilde{T}$

Primjeri izbori za L su npr.

- $Y \in \mathbb{R} \longrightarrow L(y, \hat{f}(x)) = (y - \hat{f}(x))^2$ (1.3)
- $Y \in \{0,1\} \longrightarrow L(y, \hat{f}(x)) = \begin{cases} 1, & y \neq \hat{f}(x) \\ 0, & y = \hat{f}(x) \end{cases}$ (1.4)

Uočimo, $\forall f : \mathbb{R}^p \to \mathbb{R}$, ako $\exists x \in \mathbb{R}^p$ definiramo

$$L_x(f) := \mathbb{E}[L(Y, f(x)) \mid X = x], \quad (1.5)$$

$L_x(f)$: testna greška od f u x

onda

$$L(f) = \int_{\mathbb{R}^p} L_x(f) \, \mathbb{P}_X(dx). \quad (1.6)$$

Nap.1

(i) $\mathbb{P}_X(B) := \mathbb{P}(X \in B)$, $\forall B \subseteq \mathbb{R}^p$ Borelov

[distribucija od X]

(iii)
$$\mathbb{E}[L(Y, t(x)) \mid X = x] := h(x),$$

pri čemu je $h : \mathbb{R}^p \to \mathbb{R}$ t.d.

$$h(X) = \mathbb{E}[L(Y, t(X)) \mid X] \quad \text{g.s.}$$

Imamo
$$\mathbb{E}[L(Y, t(X))] = \mathbb{E}[\mathbb{E}[L(Y, t(X)) \mid X]]$$
$$= \mathbb{E}[h(X)] = \int_{\mathbb{R}^p} h(x)\, \mathbb{P}_X(dx)$$

□

Iz (1.6) slijedi da $t^* : \mathbb{R}^p \to \mathbb{R}$ definira

$$t^*(x) := \underset{c \in \mathbb{R}}{\operatorname{argmin}} \mathbb{E}[L(Y, c) \mid X = x], \quad x \in \mathbb{R}^p \qquad (1.7)$$

minimizira (1.1), tj. $L_X(t^*) \le L_X(t)$ i $L(t^*) \le L(t)$,

za sve t i sve $x \in \mathbb{R}^p$.

optimalna ("ireducibilna") greška

Pr. 1

(i) Ako je L iz (1.3),

regresijska funkcija

$$t^*(x) = \mathbb{E}[Y \mid X = x], \qquad (1.8)$$

(ii) Ako je L iz (1.4),

Bayesov klasifikator

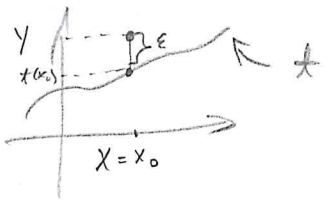
$$t^*(x) = \begin{cases} 1, & \text{ako } \mathbb{P}(Y = 1 \mid X = x) > \frac{1}{2}, \\ 0, & \text{inače.} \end{cases} \qquad (1.9)$$

(Dz - tretirajte $\mathbb{P}(\cdot \mid X=x)$ i $\mathbb{E}[\cdot \mid X=x]$ kao standardne $\mathbb{P}(\cdot)$ i $\mathbb{E}[\cdot]$)

[f^* nemamo ne znamo jer ne znamo distribuciju od (X,Y)]

[$\hat{f}$ biramo iz neke familije f-ja & konteci nekim algoritmom → PAC učenje, VC dimenzija... (TSU)]

(1.3) Primjer



Pretpostavimo model

$$Y = f(X) + \varepsilon, \qquad (1.10)$$

gdje (i) f neka f-ja, (ii) ε sl. varijabla nezavisna od X, (iii) $\mathbb{E}[\varepsilon]=0$, $\mathrm{Var}(\varepsilon)=\sigma^2<\infty$.

Uočimo, nužno je

$$f^*(x) = \mathbb{E}[Y \mid X=x] = \mathbb{E}[f(x)+\varepsilon \mid X=x]$$

$$= f(x) + \mathbb{E}[\varepsilon \mid X=x] = f(x) + \underbrace{\mathbb{E}[\varepsilon]}_{=0} = f(x).$$

Nadalje, $\forall x \in \mathbb{R}^p$ ireducibilna greška je

$$L_x(f) = \mathbb{E}[\underbrace{(Y-f(X))^2}_{\varepsilon^2} \mid X=x] \underset{(1.10)}{=} \sigma^2,$$

te

$$L(f) = \mathbb{E}[(Y-f(X))^2] = \sigma^2.$$

na to ne možemo utjecati

[ε predstavlja onaj dio info od Y koji nije sadržan u X.]

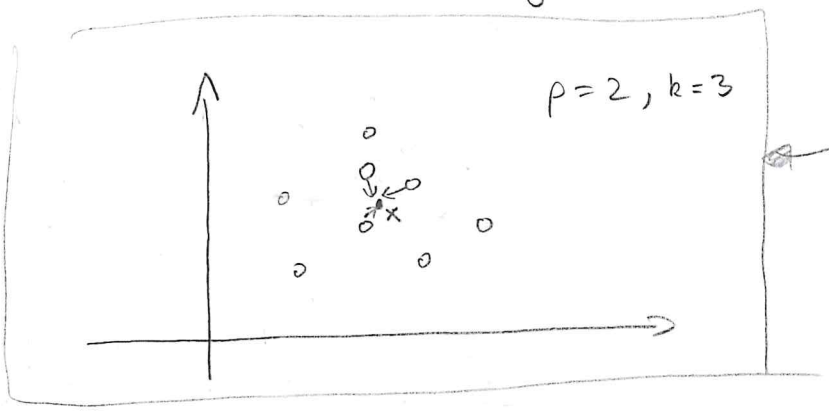
(1.3.1) Metoda k najbližih susjeda (k-nearest neighbours) → (kNN)

Za $k \in \{1, \dots, n\}$ fiksan,

$$\hat{f}_k(x) = \frac{1}{k} \sum_{x^{(i)} \in N_k(x)} y_i, \quad x \in \mathbb{R}^p, \qquad (1.11)$$

pri čemu je

$N_k(x)$ = skup od k točaka iz $T_X := \{x^{(1)}, \dots, x^{(n)}\}$ koje su najbliže točki x



[$\hat{f}_3(x)$ je prosjek izlaznih y-a za 3 najbliža susjeda od x u T]

Nap.

i) neparametarski pristup — $\hat{f}_k(x)$ direktno procjenjuje

$$f^*(x) = f(x) = \mathbb{E}[Y \mid X = x]$$

[jedini zahtjev je da je f dovoljno glatka.]

ii) [hiperparametar] parametar k kontrolira "kompleksnost" metode (tj. "fleksibilnost" od $\hat{f}_k$), npr.

- $k = 1 \Rightarrow \hat{f}_1(x^{(i)}) = y_i, \ \forall i = 1, \dots, n$!
  ("overfitting" jer $L_T(\hat{f}_1) = 0$)
- $k = n \Rightarrow \hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}, \ \forall x \in \mathbb{R}^p$!
  ($N_n(x) = T_X$) ("underfitting")

(iii) U praksi, k odlučujemo unakrsnom validacijom.
(kasnije)

1.3.2 Linearna regresija

Pretpostavljamo da je za neki $\beta = (\beta_0, \dots, \beta_p) \in \mathbb{R}^{p+1}$,

$$f(x) = \mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p, \quad (1.12)$$
$$=: f_\beta(x)$$

za sve $x = (x_1, \dots, x_p) \in \mathbb{R}^p$, a β procjenjujemo npr. metodom najmanjih kvadrata $\longrightarrow \hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$, te dobivamo

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p, \quad x \in \mathbb{R}^p.$$

Nap. 1

(i) parametarski pristup - pretpostavili smo da je $f \in \mathcal{F}_\beta = \{ f_\beta : \beta \in \mathbb{R}^{p+1} \}$ pri čemu je broj parametara $(p+1)$ konačan.

(ii) metoda ima smisla; kada (1.12) ni približno ne vrijedi $\longrightarrow$ aproksimiramo f da f-jom iz $\mathcal{F}_\beta$.

$\longrightarrow$ uvodno-predavanje.pdf

(kNN vs. lin. regresija)

(1.4) Odnos između pristranosti i varijance (!)

(bias - variance tradeoff) ⟶ (BVT)

[najvažnija stvar kada govorimo o predikciji!]

[terminologija koju ćemo koristiti stalno]

Pretp. da je kao u (1.10),

$$Y = f(X) + \varepsilon.$$

Ako promatramo $\hat{f} = \hat{f}(T)$ gdje je T slučajan uzorak, $\hat{f}$ postaje slučajna f-ja, a $L_x(\hat{f})$ i $L(\hat{f})$ slučajne varijable.

Prop. 1.1 (BVT)

Ako je $L(y, \hat{y}) = (y - \hat{y})^2$, $\forall x \in \mathbb{R}^p$ vrijedi

$$\mathbb{E}_T[L_x(\hat{f})] = \sigma^2 + \mathrm{Var}_T(\hat{f}(x)) + (\mathbb{E}_T[\hat{f}(x)] - f(x))^2 \qquad (1.13)$$

$\mathbb{E}_T[L_x(\hat{f})]$: "očekivana testna greška u x"

σ^2: ireducibilna greška

$\mathrm{Var}_T(\hat{f}(x)) + (\mathbb{E}_T[\hat{f}(x)] - f(x))^2$: "reducibilna greška" [cilj je to smanjiti]

$\mathrm{Var}_T(\hat{f}(x))$: varijanca

$(\mathbb{E}_T[\hat{f}(x)] - f(x))^2$: pristranost^2 (bias)

Napomena

(i) $\mathbb{E}_T$ predstavlja očekivanje u kojem slučajnost dolazi od T.

(ii) (1.13) $\Rightarrow$ $\mathbb{E}_T[L(\hat{f})] \overset{(1.6)+\text{FUBINIJEV TH.}}{=} \int_{\mathbb{R}^p} \mathbb{E}_T[L_x(\hat{f})]\, \mathbb{P}_X(dx)$ ∎

Dokaz!

(1°) $\forall g: \mathbb{R}^p \to \mathbb{R}$ i $\forall x \in \mathbb{R}^p$,

$$L_x(g) = \mathbb{E}[(Y - g(X))^2 \mid X = x] = \mathbb{E}[(f(x) - g(x) + \varepsilon)^2 \mid X = x]$$

$$= \mathbb{E}[(f(x) - g(x))^2 + (f(x) - g(x))\varepsilon + \varepsilon^2 \mid X = x]$$

$$= (f(x) - g(x))^2 + (f(x) - g(x)) \underbrace{\mathbb{E}[\varepsilon \mid X = x]}_{= \mathbb{E}[\varepsilon] = 0} + \underbrace{\mathbb{E}[\varepsilon^2 \mid X = x]}_{= \mathbb{E}[\varepsilon^2] = \sigma^2}$$

$$= (f(x) - g(x))^2 + \sigma^2$$

(2°) $\forall x \in \mathbb{R}^p$,

ovisi o T

$$\mathbb{E}_T[L_x(\hat{f})] \overset{(1°)}{=} \sigma^2 + \mathbb{E}_T[(f(x) - \hat{f}(x))^2]$$

$\pm \mathbb{E}_T[\hat{f}(x)]$

$$= \sigma^2 + \mathbb{E}_T[(f(x) - \mathbb{E}_T[\hat{f}(x)])^2]$$

$$+ \underbrace{(f(x) - \mathbb{E}_T[\hat{f}(x)])}_{\text{konstanta}} \underbrace{\mathbb{E}_T[\mathbb{E}_T[\hat{f}(x)] - \hat{f}(x)]}_{=0}$$

$$+ \mathbb{E}_T[(\mathbb{E}_T[\hat{f}(x)] - \hat{f}(x))^2]$$

$$= \sigma^2 + \underbrace{(f(x) - \mathbb{E}_T[\hat{f}(x)])^2}_{\text{pristranost}^2} + Var_T(\hat{f}(x))$$

∎

Tipično, kako kompleksnost metode raste, pristranost se smanjuje, a varijanca povećava.

Pr.] U kNN metodi,

(i) ako je $k=1$, $\forall x \in \mathbb{R}^p$,

$$\hat{f}(x) = f(X^{(i)}) + \varepsilon \qquad (1.14)$$

($X^{(i)}$: najbliži susjed od x u $\{X^{(1)}, \dots, X^{(n)}\}$)

$\longrightarrow$ za veliki n (pod nekim uvjetima),

$$f(X^{(i)}) \approx f(x), \quad \text{g.s.} \qquad (1.15)$$

te

$$\mathbb{E}_T\left[\hat{f}(x)\right] \approx f(x) + \mathbb{E}[\varepsilon] = f(x)$$

i

$$\mathrm{Var}_T\left[\hat{f}(x)\right] \approx 0 + \mathrm{Var}(\varepsilon) = \sigma^2$$

$\Rightarrow$ ako je n velik i σ ne premalen, imamo

malu pristranost i **veliku varijancu**

(oboje: od ε-a u (1.14))

(ii) ako je $k=n$, $\forall x \in \mathbb{R}^p$,

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(X^{(i)}) + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i ,$$

JZVB $\longrightarrow$ za veliki n,

$$\hat{f}(x) \approx \mathbb{E}[f(X)] + \mathbb{E}[\varepsilon]$$
$$= \mathbb{E}[f(X)], \quad \text{g.s.}$$

(ista konstanta $\forall x \in \mathbb{R}^p$!)

te $\mathbb{E}_T[\hat{f}(x)] \approx \mathbb{E}[f(X)]$,

i $\mathrm{Var}_T[\hat{f}(x)] \approx 0$ [sve heuristički!]

=> za veliki n, ako $\mathcal{A}$ nije konstantna $\mathcal{A}$-ja, imamo velike pristranost i male varijance

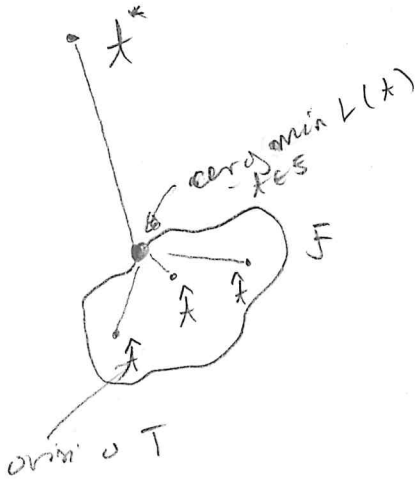
jer $\mathcal{A}(x)$ procjenjujemo koristeći se $\mathcal{A}(X^{(i)})$

Nap.1 BVT je fundamentalna stvar u stat. učenju (kada govorimo o _predikciji_), i javlja se bez obzira na izbor $\mathcal{A}$-je L.

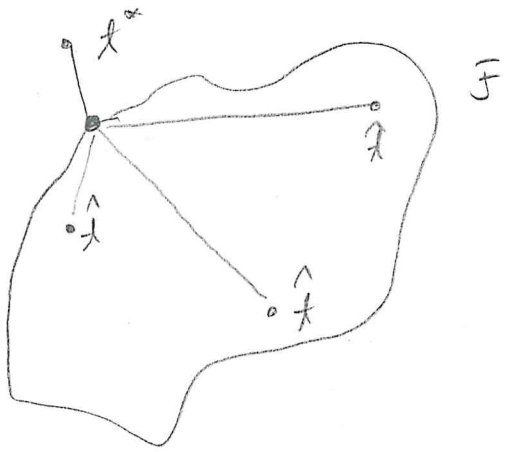
$$L(\hat{\mathcal{A}}) = L(\mathcal{A}^*) + (L(\hat{\mathcal{A}}) - L(\mathcal{A}^*))$$

$$= \underbrace{L(\mathcal{A}^*)}_{\text{ireducibilna greška}} + \underbrace{(L(\hat{\mathcal{A}}) - \inf_{\mathcal{A}\in\mathcal{F}} L(\mathcal{A}))}_{\text{greška procjene ("varijance")}} + \underbrace{(\inf_{\mathcal{A}\in\mathcal{F}} L(\mathcal{A}) - L(\mathcal{A}^*))}_{\text{greška aproksimacije ("pristranost")}}$$

Intuitivno,



velika pristranost
mala varijanca



mala pristranost
velika varijanca

→ [uvodno_predavanje.pdf] (Odnos između pristranosti i varijance.)