

1) Uvod

1.1 Problem (mehaničnog) učenja

Pretp. da imamo podatke

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

uvazak ili "skup za učenje"  
[training set]

gdje je npr.

- $y_i$  = cijena  $i$ -te nekretnine

odgovor ili zavisna varijabla  
[response]

$$x^{(i)} = (x_{i1}, x_{i2}, \dots, x_{ip})$$

↑  
veličina  
(u m<sup>2</sup>)

↑  
pogled  
na  
mjestu  
(DA/NE)

↑  
karakteristike ili  
nezavisne varijable  
(p ∈ N je broj karakteristika)

Najčešće, cilj učenja je predikcija:

$$T \rightsquigarrow \hat{f} = \hat{f}(T),$$

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

Ako je

- $y$  kvantitativna (neprekidna) varijabla

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

• y kategorijalna varijabla

→ problem klasifikacije

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

1.2 Teorija statističkog učenja

[ statistički pristup problemu učenja  
→ TSU ( sondno ) ]

Pretpostavke

- odgovor Y je stohastična varijabla

- koordinata  $X = (X_1, \dots, X_p)$  su stohastični vektor  
u  $\mathbb{R}^p$

( X i Y su zavisni ! npr.  $Y = f(X) + \epsilon$  )

- T je jedna realizacija od

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

gdje su, tipično,  $(X^{(i)}, Y_i), i = 1, \dots, m$ , mjdl

te  $(X^{(i)}, Y_i) \sim (X, Y)$

Cilj: Za dani f gubitka  $L: \mathbb{R}^2 \rightarrow \mathbb{R}$  [ loss function ]  
tražimo hat  $\hat{f} = \hat{f}(x): \mathbb{R}^p \rightarrow \mathbb{R}$  sa sv manjom bestnom

gneskom

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

[  $L(y, \hat{f}(x))$  je "trošček" predikcije  $\hat{f}(x)$  ako je strama unjednost  $y$  ] 3

Nap. 1 Ako imamo realizaciju

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

testni skup

jes jechwy mjal uwooku iz  $(X, Y)$ , za veliki  $m$ ,

$$L_{\tilde{Z}}(\hat{f}) := \frac{1}{m} \sum_{i=1}^m L(\tilde{y}_i, \hat{f}(\tilde{x}^{(i)})) \approx L(\hat{f}). \quad (1.2)$$

greshka na  $\tilde{Z}$

Primer tabari za  $L$  za mpr.

•  $y \in \mathbb{R} \rightarrow L(y, \hat{f}(x)) = (y - \hat{f}(x))^2 \quad (1.3)$

•  $y \in \{0, 1\} \rightarrow L(y, \hat{f}(x)) = \begin{cases} 1, & y \neq \hat{f}(x) \\ 0, & y = \hat{f}(x) \end{cases} \quad (1.4)$

uodimo,  $\forall \hat{f}: \mathbb{R}^P \rightarrow \mathbb{R}$ , ako 2 za  $x \in \mathbb{R}^P$  defin.

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

unjed-

testna greshka od  $\hat{f}$  u  $x$

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

Nap. 1

(i)  $\mathbb{P}_X(B) := \mathbb{P}(X \in B)$ ,  $\forall B \in \mathbb{R}^P$  Borelov  
[distribucija od  $X$ ]

100)

$$\mathbb{E}[L(Y, f(x)) | X=x] := h(x),$$

pri čemu je  $h: \mathbb{R}^p \rightarrow \mathbb{R}$  t.o.

$$h(X) = \mathbb{E}[L(Y, f(X)) | X] \quad \text{g.o.}$$

[memo]

$$\mathbb{E}[L(Y, f(X))] = \mathbb{E}[\mathbb{E}[L(Y, f(X)) | X]]$$

$$= \mathbb{E}[h(X)] = \int_{\mathbb{R}^p} h(x) \mathbb{P}_X(dx)$$

Iz (1.6) slijedi da  $f^*: \mathbb{R}^p \rightarrow \mathbb{R}$  defin. sa

$$f^*(x) := \underset{c \in \mathbb{R}}{\text{argmin}} \mathbb{E}[L(Y, c) | X=x], \quad x \in \mathbb{R}^p \quad (1.7)$$

minimizira (1.1), tj.  $L_x(f^*) \leq L_x(f)$  i  $L(f^*) \leq L(f)$ ,  
za sve  $f$  i sve  $x \in \mathbb{R}^p$ .

optimalna ("reducirana") greška

Pr. 1

(i) Ako je  $L$  iz (1.3),

$$f^*(x) = \mathbb{E}[Y | X=x], \quad (1.8)$$

regresijska f-ija

(ii) Ako je  $L$  iz (1.4),

$$f^*(x) = \begin{cases} 1, & \text{ako } \mathbb{P}(Y=1 | X=x) \geq \frac{1}{2}, \\ 0, & \text{inače.} \end{cases} \quad (1.9)$$

Bayesov klasifikator

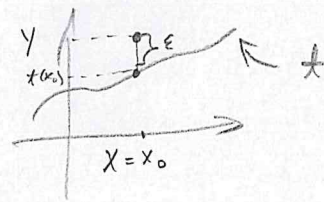
(DZ - tretinajte  $P(\cdot | X=x)$ ;  $E[\cdot | X=x]$  kao standardne  $P(\cdot)$ ;  $E[\cdot]$ ) [5]

[  $f^*$  manje ili više je mehanizam distribucija od  $(X, Y)$  ]

[  $f^*$  binomna iz meke familije  $f$ -ja  $\mathcal{F}$  koristeći neki algoritam  $\rightarrow$  PAC učenje, VC dimenzija... ] (TSU)

### 1.3 Primer

Pretpostavimo model



$$Y = f(X) + \epsilon, \quad (1.10)$$

gdje (i)  $f$  meka  $f$ -ja, (ii)  $\epsilon$  sl. varijabla meranizirana od  $X$ , (iii)  $E[\epsilon] = 0$ ,  $\text{Var}(\epsilon) = \sigma^2 < \infty$ .

Uočimo, nužno je

$$\begin{aligned} f^*(x) &= E[Y | X=x] = E[f(x) + \epsilon | X=x] \\ &= f(x) + E[\epsilon | X=x] = f(x) + \underbrace{E[\epsilon]}_{=0} = f(x). \end{aligned}$$

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

$$L_x(f) = E[\underbrace{(Y - f(x))^2}_{\epsilon^2} | X=x] \stackrel{(1.10)}{=} \sigma^2,$$

te

$$L(f) = E[(Y - f(x))^2] = \sigma^2.$$

na to me nužno utjecati.

$\epsilon$  predstavlja onaj dio info od  $Y$  koji nije sadržan u  $X$ .

(1.3.1) Metoda k najbližih susjeda (k-nearest neighbors) → kNN [6]

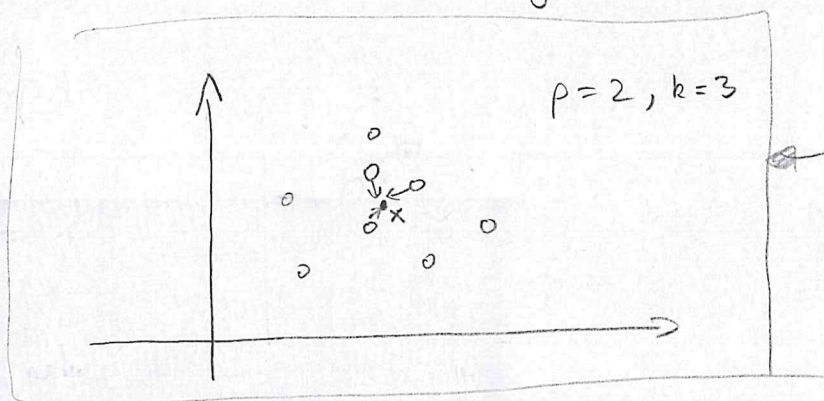
Za  $k \in \{1, \dots, m\}$  fiksno,

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

pri čemu je

$N_k(x)$  = skup od  $k$  točaka iz

$\mathcal{T}_x := \{x^{(1)}, \dots, x^{(m)}\}$  koje su najbliže točki  $x$



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

Uop:

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

$$f^*(x) = f(x) = \mathbb{E}[Y | X=x] \quad [\text{jedini zahtjev je da je } f \text{ dovoljno glatka}]$$

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

•  $k=1 \Rightarrow \hat{f}_1(x^{(i)}) = y_i, \forall i=1, \dots, m$ !  
("overfitting" jer  $L_{\mathcal{T}}(\hat{f}_1) = 0$ )

•  $k=m \Rightarrow \hat{f}_m(x) = \frac{1}{m} \sum_{i=1}^m y_i = \bar{y}, \forall x \in \mathbb{R}^p$ !  
 $N_m(x) = \mathcal{T}_x$  ("underfitting")

(iii) u praksi, k odabiremo unakrsnom validacijom.  
(kasnije)

7

### 1.3.2 Linearna regresija

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

$$f(x) = E[Y | X=x] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p, \quad (1.12)$$

$=: f_{\beta}(x)$

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

stavimo

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

Nap-1

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

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

$\rightarrow$  uvodno - predavanje. pdt  
(kNN vs. lin. regresija)

1.4 Odnos između pristranosti i varijance (!)  
 (bias - variance trade off) → (BVT)

[najveća stvar kada govorimo o predikciji!]

[terminologija koju ćemo koristiti stalno]

Pretp. da je kao u (1.10),

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

Ako posmatramo  $\hat{f} = \hat{f}(T)$  gdje je  $T$  slučajni uzorak,  $\hat{f}$  postaje slučajna  $f$ -ja,

a  $L_X(\hat{f})$  i  $L(\hat{f})$  slučajne varijable.

Prop. 1.11 (BVT)

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

$$E_T[L_X(\hat{f})] = \sigma^2 + \text{Var}_T(\hat{f}(x)) + (E_T[\hat{f}(x)] - f(x))^2 \quad (1.13)$$

irreducibilna greška + varijanca + pristranost<sup>2</sup> (bias)

"očekivana testna greška u  $x$ "

"reducibilna greška" [odj je to smanjiti]

Nap. 1  
 (i)  $E_T$  predstavlja očekivanje u kojem slučajnost dolazi od  $T$ .

(ii) (1.13)  $\Rightarrow E_T[L(\hat{f})] = \int_{\mathbb{R}^p} E_T[L_X(\hat{f})] P_X(dx)$  (1.6) + FUBINIJEV TR.



(10)  $\forall g: \mathbb{R}^p \rightarrow \mathbb{R} ; \forall x \in \mathbb{R}^p,$

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

(20)  $\forall x \in \mathbb{R}^p,$

$$\mathbb{E}_T[L_x(\hat{f})] \stackrel{(10)}{=} \sigma^2 + \mathbb{E}_T[(f(x) - \hat{f}(x))^2]$$

$\downarrow$   
 ozn. s T  
 $\downarrow$   
 $= \mathbb{E}_T[\hat{f}(x)]$

$$\begin{aligned}
 &= \sigma^2 + \mathbb{E}_T[(f(x) - \mathbb{E}_T[\hat{f}(x)])^2] \\
 &\quad + \underbrace{(f(x) - \mathbb{E}_T[\hat{f}(x)])}_{\text{konstanta}} \underbrace{\mathbb{E}_T[(\mathbb{E}_T[\hat{f}(x)] - \hat{f}(x))]}_{= 0} \\
 &\quad + \mathbb{E}_T[(\mathbb{E}_T[\hat{f}(x)] - \hat{f}(x))^2]
 \end{aligned}$$

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

Tipično, kako kompleksnost metode raste, pristranost  $\hat{f}$  se smanjuje, a varijanca poraste.

Pr. 1 u kNN metodi,

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

$$\hat{f}(x) = f(X^{(1)}) + \epsilon \quad (1.14)$$

→ za veliki  $n$  (pod nekim ujetima),

$$\hat{f}(X^{(1)}) \approx f(x), \text{ g.s.} \quad (1.15)$$

te

$$E_T[\hat{f}(x)] \approx f(x) + E[\epsilon] = f(x)$$

i

$$\text{Var}_T[\hat{f}(x)] \approx 0 + \text{Var}(\epsilon) = \sigma^2$$

⇒ ako je  $n$  velik i  $\sigma$  ne premalen, imamo

malu pristranost i veliku varijancu

(dodatni oči  $\epsilon$ -a u (1.14)

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

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

JZVB

→ za veliki  $n$ ,

$$\hat{f}(x) \approx E[f(X)] + E[\epsilon]$$

$$= E[f(X)] \quad \text{g.s.}$$

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

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

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

$\Rightarrow$  Za veliki  $n$ , ako  $f$  nije konstantna  $f$ -je, imamo L

velika pristranost i mala varijanca

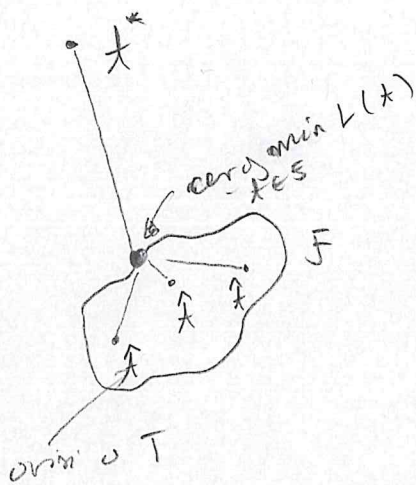
jer  $f(x)$  procjenjujemo  
konstanti: sk  $f(x^{(i)})$

Nop-1 BVT je fundamentalna stvar u stat. učenju  
(kada govorimo o predikciji), i javlja se  
bez obzira na izbor  $A$ -je  $L$ .

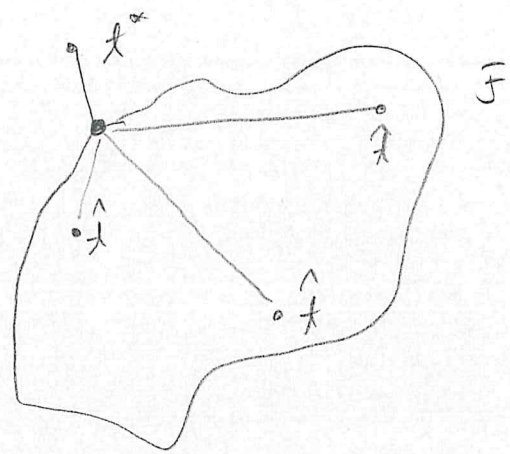
$$L(\hat{f}) = L(f^*) + (L(\hat{f}) - L(f^*))$$

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

Intuitivno,



velika pristranost  
mala varijanca



mala pristranost  
velika varijanca

$\rightarrow$  uvodno predavanje pet (Odnos između pristranosti i varijance.)