

Strojno učenje – domaća zadaća 11

UNIZG FER, ak. god. 2016./2017.

Zadano: 20. 1. 2017.

Napomena: Zadatke možete rješavati samostalno ili u grupi. Ako zadatke rješavate u grupi, pobrinite se da svi članovi grupe pridonose rješenju i da ga napisljetu svi razumiju. Po potrebi konzultirajte sve dostupne izvore informacija. Rješenja zadataka ponesite na iduće auditorne vježbe. Zabilježite sve nejasnoće i nedoumice, kako bismo ih prodiskutirali.

1. [Svrha: Izvježbati izračun mjera uspješnosti modela na konkretnom primjeru.]

Raspolažemo skupom od 11 ispitnih primjera koje želimo klasificirati u tri klase. Oznaka $y^{(i)}$ i izlaz modela $h(\mathbf{x}^{(i)})$ za svaki od 11 primjera su sljedeći:

$$\{(y^{(i)}, h(\mathbf{x}^{(i)}))\}_{i=1}^{11} = \{(1, 1), (0, 2), (2, 2), (1, 2), (1, 1), (0, 0), (1, 1), (2, 1), (0, 1), (2, 0), (2, 1)\}.$$

Izračunajte preciznost, odziv i F_1 mjeru, i to *mikro* i *makro* varijante.

2. [Svrha: Razumjeti na koji se način provodi ugniježđena unakrsna provjera, kako se razdjeljuju primjeri kroz iteracije petlji te kako ugraditi dodatne predobradbe značajki, a pritom ne kompromitirati podjelu na skup za učenje i skup za ispitivanje.] Raspolažemo sa 1000 označenih primjera. Za vrednovanje SVM-a s hipерparametrima C i γ koristimo ugniježđenu unakrsnu provjeru sa po 5 ponavljanja u obje petlje. Hiperparametre optimiramo rešetkastim pretraživanjem u rasponima $C \in \{2^{-5}, 2^{-4}, \dots, 2^{15}\}$ i $\gamma \in \{2^{-15}, 2^{-14}, \dots, 2^3\}$.

- Koliko ćemo ukupno puta trenirati model?
- Koliko ćemo primjera u svakoj od iteracija koristiti za treniranje, koliko za provjeru, a koliko za ispitivanje?
- Kako glase odgovori na prethodna dva pitanja, ako bismo u vanjskoj petlji umjesto pterostrukе unakrsne provjere koristili unakrsnu provjeru *izdvoji jednoga* (engl. *leave one out, LOOCV*)?
- Klasifikator SVM posebno je osjetljiv na razlike u rasponima između značajki (zašto?), pa se preporuča standardizirati značajke. Što to točno znači i kako biste standardizaciju značajki ugradili u ugniježđenu unakrsnu provjeru?

3. [Svrha: Izvježbati izračun intervala pouzdanosti. Razumjeti kako broj primjera utječe na širinu intervala. Razumjeti u kojim slučajevima ne možemo izračunati interval pouzdanosti statistike srednje vrijednosti i kako je to povezano s metodom vrednovanja koju odlučimo primijeniti.]

F_1 -mjeru klasifikatora procjenjujemo desetorostrukom unakrsnom provjerom (*10-fold CV*), pri čemu smo dobili sljedeće vrijednosti:

$$0.68, 0.74, 0.71, 0.66, 0.58, 0.75, 0.76, 0.62, 0.78, 0.68.$$

- (a) Izračunajte 95%-tni i 99%-tni interval pouzdanosti za točnost ovog klasifikatora. Što ste pritom morali pretpostaviti i zašto?
- (b) Bi li interval bio širi ili uži da samo iste ove procjene za točnost i standardnu devijaciju dobili na temelju peterostrukre, a ne desetorostrukre unakrsne provjere?
- (c) Bismo li na isti način mogli izračunati interval pouzdanosti za unakrsnu provjeru *izdvoji jednoga* (engl. *leave one out*)? Zašto?
- (d) Bismo li na isti način mogli izračunati interval pouzdanosti procjene F_1 -mjere da niste radili unakrsnu provjeru, već samo ispitivali klasifikator na jednom ispitnom skupu (engl. *holdout method*). Zašto?

4. [Svrha: Izvježbati statističko testiranje hipoteze uparenim t-testom u svrhu ispitivanja razlike uspješnosti modela. Razumjeti što možemo zaključiti na temelju rezultata testa. Razumjeti pretpostavke testa.]

Trenirali smo model h_2 i želimo provjeriti je li njegova točnost bolja od referentnog (engl. *baseline*) modela h_1 koji sve primjere klasificira u većinsku klasu. Oba modela vrednjujemo desetorostrukom unakrsnom provjerom na ukupno $N = 1000$ primjera te računamo točnosti oba modela na svakom od deset preklopa. Rezultati su sljedeći:

i	$Acc(h_1)_i$	$Acc(h_2)_i$
1	0.52	0.60
2	0.56	0.54
3	0.44	0.58
4	0.58	0.58
5	0.49	0.46
6	0.39	0.58
7	0.47	0.50
8	0.57	0.67
9	0.55	0.61
10	0.43	0.55

- (a) Primijenite upareni t-test i provjerite hipotezu da je razlika u točnosti statistički značajna na razini značajnosti $\alpha = 5\%$ (dvostrani test). Iskažite zaključak.
- (b) Je li razlika statistički značajna na $\alpha = 1\%$? Iskažite zaključak.
- (c) Testirajte je li točnost modela h_2 statistički značajno *bolja* od točnosti modela h_1 , na razini značajnosti $\alpha = 5\%$ (jednostrani test). Iskažite zaključak. Vrijedi li isti zaključak za razinu značajnosti $\alpha = 1\%$?
- (d) Koje pretpostavke moraju vrijediti da biste uopće mogli primijeniti t-test? Vrijede li te pretpostavke u gornjim slučajevima? Igra li ukupan broj primjera N ikakvu ulogu pri statističkom testiranju?

11. DOMAĆA ZADĀCA

①

$$N=11, \quad K=3$$

$$y = [1 \ 0 \ 2 \ 1 \ 1 \ 0 \ 1 \ 2 \ 0 \ 2 \ 2] \text{ STVARNO}$$

$$h(x) = [1 \ 2 \ 2 \ 2 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1] \text{ KLASIFIKACIJA}$$

KLASA 0:

$$\overline{TP} : \{ (0,0) \} \rightarrow TP_0 = 1$$

$$TN: \{ (1,1), (2,2), (1,2), (1,1), (1,1), (2,1), (2,1) \} \rightarrow TN_0 = 7$$

$$FP: \{ (2,0) \} \rightarrow FP_0 = 1$$

$$FN: \{ (0,2), (0,1) \} \rightarrow FN_0 = 2$$

KLASA 1:

$$TP_1 = 3, \quad TN_1 = 4, \quad FP_1 = 3, \quad FN_1 = 1$$

KLASA 2:

$$TP_2 = 1, \quad TN_2 = 5, \quad FP_2 = 2, \quad FN_2 = 3$$

MIKRO VARIJANTE:

$$\text{PRECIZNOST} : P = \frac{TP_0 + TP_1 + TP_2}{TP_0 + TP_1 + TP_2 + FP_0 + FP_1 + FP_2}$$

$$P_{\text{MIKRO}} = \frac{5}{11} = 0.4545\%,$$

$$ODZIV: R = \frac{TP_0 + TP_1 + TP_2}{TP_0 + TP_1 + TP_2 + FN_0 + FN_1 + FN_2}$$

$$R_{\text{MIKRO}} = \frac{5}{11} = 0.4545 //$$

F1-MJERA:

$$F1_{\text{MIKRO}} = \frac{2P_{\text{MIKRO}} \cdot R_{\text{MIKRO}}}{P_{\text{MIKRO}} + R_{\text{MIKRO}}} = 0.4545 //$$

MAKRO VARIJANTE

PRECIZNOST: $P_0 = \frac{TP_0}{TP_0 + FP_0} = 0.5$

$$P_1 = \frac{TP_1}{TP_1 + FP_1} = 0.5 \quad P_2 = \frac{TP_2}{TP_2 + FP_2} = \frac{1}{3} = 0.33$$

$$P_{\text{MAKRO}} = \frac{1}{3} [0.5 + 0.5 + 0.33] = \frac{4}{9} = 0.4444 //$$

ODZIV: $R_0 = \frac{TP_0}{TP_0 + FN_0} = \frac{1}{3}$

$$R_1 = \frac{3}{4}, \quad R_2 = \frac{1}{4}$$

$$R_{\text{MAKRO}} = \frac{1}{3} \left[\frac{1}{3} + \frac{3}{4} + \frac{1}{4} \right] = \frac{4}{9} = 0.4444 //$$

F1-MJERA:

$$F_{1,1} = \frac{2P_1 R_1}{P_1 + R_1} = 0.6$$

$$F_{1,2} = 0.287$$

$$F_{1,0} = \frac{2P_0 R_0}{P_0 + R_0} = 0.4$$

$$F_{1,\text{MAKRO}} = \frac{1}{3} (0.4 + 0.6 + 0.287)$$

$$F_{1,\text{MAKRO}} = 0.429$$

(2)

$$N = 1000, \quad \underbrace{N_u = 5, \quad N_v = 5}_{\text{PONAVLJANJA U VANJSKOJ}}$$

PONAVLJANJA U UNUTARNJOJ)

$N_c = 16$ (16 RAZLIČITIH PARAMETARA c)

$N_p = 19$ (19 RAZLIČITIH PARAMETARA p)

a)

$$N_{uk} = N_v \cdot N_u \cdot N_c \cdot N_p = \underbrace{5 \cdot 5 \cdot 16 \cdot 19}_{\text{GRID SEARCH}} + 5 = 7605$$

b) $N = 1000$

GRIDS
SEARCH VJEĆENJE S
OPTIMALNIM PARAM.
U VANJSKOJ PETLJI

IMAMO 5 PONAVLJANJA U VANJSKOJ PETLJI.

TO ZNAČI DA SKUP OD 1000 PODATAKA

DIJELIMO NA SKUP ZA VJEĆENJE I ISPITNI SKUP U OMJERU 4:1.

$$N(\text{ISPITNI SKUP}) = \frac{4}{5} \cdot 1000 = 800,$$

$$N(\text{VJEĆENJE}) = \frac{4}{5} \cdot 1000 = 800,$$

U UNUTARNJOJ PETLJI SKUP ZA VJEĆENJE DODATNO DIJELIMO NA SKUP ZA TREĆIRANJE I SKUP ZA PROVJERU, PONOVO U OMJERU 4:1 (JER U UNUTARNA PETLJA IMA 5 ITERACIJA).

$$N(\text{TREĆIRANJE}) = \frac{4}{5} \cdot 800 = 640,,$$

$$N(\text{PROVJERA}) = \frac{1}{5} \cdot 800 = 160,,$$

ZAKLJUČNO, U SVAKOJ ITERACIJI:

$$N(\text{TREĆIRANJE}) = 640, \quad N(\text{PROVJERA}) = 160, \quad N(\text{ISPITIVANJE}) = 200$$

c)

i) UMJESTO 5, SADA VANJSKA PETLJA. IMA $N=1000$ PONAVLJANJA.

$$N_{\text{ne}} = 1000 \cdot 5 \cdot 16 \cdot 19 = 1520000,$$

ii) LOOCV U VANJSKOJ PETLJI. DJELI PODATKE NA $(N-n)$ U SKUPU ZA UČENJE i n U SKUPU ZA ISPITIVANJE:

$$N(\text{UČENJE}) = 999$$

$$N(\text{ISPITIVANJE}) = n$$

UNUTARNJA PETLJA JE I DALJE, ISTE VELIČINE

$$N(\text{TRENING}) = \frac{4}{5} \cdot 999 = \{\text{ZAOKRUŽENO}\} = 799,$$

$$N(\text{PROVERA}) = \frac{1}{5} \cdot 999 = \{\text{ZAOKRUŽENO}\} = 200,$$

ZAKLJUČNO:

$$N(\text{TRENING}) = 799, \quad N(\text{PROVERA}) = 200, \quad N(\text{ISPITIVANJE}) = 1$$

d) SVM U DUALNOJ FORMI PROMATRA SLIČNOST DVAJU PRIMJERA (NJIHOV SKALARNI PRODUKT), A NA SKALARNI PRODUKT VIŠE VJEĆU ZNAČAJKE VELIKOG ZNOSA. NPR., IMAMO TRI PRIMJERA:

$$x_1 = \begin{bmatrix} 100 \\ 0.1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 100 \\ 0.3 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 98 \\ 0.1 \end{bmatrix}$$

x_1 i x_3 SU SLIČNIJII NEGO x_1 i x_2 JER SE DOSTA DOBRO POPULARAJU U OBJE ZNAČAJKE, DOK SE x_1 i x_2 JAKO RAZLIKUJU U DRUGOJ ZNAČAJKI.

NO, RAČUNAJUĆI SKALARNI PRODUKT BEZ SKALIRANJA:

$$x_1^T \cdot x_2 = 10000.03, \quad x_1^T \cdot x_3 = 9800.01$$

ISPADA DA SU x_1 I x_2 SLIČNIJI, ŠTO NIJE ISTINA. STOGLA SE ODLUČUJEMO SKALIRATI SVE ZNAČAJKE NA RASPON $[0, 1]$.

STANDARDIZACIJU BIH UGRADIO U UGNJEŽDENU UNAKRSNU PROVJERU KORISTEĆI GJEVOROD (KAO G. LABOS), TO SKALIRANJE BI SE IZRVAČAVALO SVAKI PUT NEPOSREDNO PRIJE IZRVAŠAVANJA SVRHA.

(2)

a) $N < 30$, NE ZNAMO $\sigma \rightarrow$ NORAMO PROCIJENITI $\hat{\sigma}$

t - STATISTIKA

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = \frac{1}{10} (0.68 + 0.74 + 0.71 + 0.66 + 0.58 + 0.75 + 0.76 + 0.62 + 0.78 + 0.68)$$

$$\boxed{\bar{x} = 0.696}$$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - 0.696)^2}$$

$$\boxed{\hat{\sigma} = 0.06432}$$

$$\mu = \bar{x} + k \cdot \frac{\hat{\sigma}}{\sqrt{N}}$$

IMAMO $N < 30$ PRIMJERA PA KORISTIMO STUDENTOVU (t) DISTRIBUCIJU I KOEFICIJENT ŠIRINE INTERVALA NE ZNAMO

→ TREBA NAĆI NA INTERNETU k ZA 95% I 99%

$$\left. \begin{array}{l} k_{0.95\%} = 2.262 \\ k_{0.99\%} = 3.2498 \end{array} \right\} \text{ISČITANO IZ TABLICE}$$

S INTERNETA (NA ISPITU
VALJA DAJU PODATKE)

$$\mu = 0.696 + 2.262 \cdot \frac{0.06432}{\sqrt{10}}, \quad 95\%$$

$$\boxed{\mu = 0.696 + 0.046 \quad \text{UZ } 95\% \text{ SIGURNOST}}$$

$$\mu = 0.696 + 3.2498 \cdot \frac{0.06432}{\sqrt{10}}, \quad 99\%$$

$$\boxed{\mu = 0.696 + 0.0661 \quad \text{UZ } 99\% \text{ SIGURNOST}}$$

MORALI SMO PRETPOSTAVITI DA SE PODACI POKORAVAJU STUDENTOVU (t) DISTRIBUCIJI JER IMAMO SAMO $N=10$ PODATAKA.

lr) BIO BI ŠIRI JER BISMU IMALI MANJE PRIMJERA, $N=5$, A ŠIRINA INTERVALA JE OBRSNUTO PROPORCIJALNA S \sqrt{N} (MANJI N, VEĆA ŠIRINA).

c) DA, IMAMO 10 PRIMJERA PA JE UNAKRSNA PROVJERA "IZDVOJI JEDNOGA" ISTO KTO I DESETEROSTRUKA UNAKRSNA PROVJERA.

d) NE, JER SE U TOM SLUČAJU NAŠA PROCJENA POGREŠKE TEMELJI SAMO NA JEDNOM UZORKU.

4.) a)

HIPOTEZA:

$$H_0: \mu_1 = \mu_2, \quad \alpha = 0.05$$

$$H_1: \mu_1 \neq \mu_2$$

t-TEST

$$d_i = ACC_1^i - ACC_2^i$$

$$\bar{d} = \frac{1}{n} \sum_{i=1}^N (ACC_1^i - ACC_2^i) = -0.067,$$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (d_i - \bar{d})^2} = \sqrt{\frac{1}{9} \sum_{i=1}^9 (d_i + 0.067)^2}$$

$$\hat{\sigma} = 0.0726$$

$$t = \frac{\bar{d} - 0}{\hat{\sigma} / \sqrt{n}} = \frac{-0.067}{0.0726 / \sqrt{10}} = -2.918355$$

P - VRJEDNOST : (ONLINE KALKULATOR,

$$df = N - 1 = 9, \quad t = -2.918355$$

$$P = 0.0171$$

ZAKLJUČAK:

$p < \alpha \rightarrow$ S RAZINOM ZNAČAJNOSTI 5%

PODACI DAJU DOVOLJNO DOKAZA DA JE $\mu_1 \neq \mu_2$

ZNAČI DA smo 95% SIGURNI DA TOČNOSTI h_1 I h_2 Nisu iste

b) $\alpha = 0.01, \quad p = 0.0171$

$p > \alpha \rightarrow$ PODACI NE DAJU DOVOLJNO DOKAZA
DA SE OBORI NUL-HIPOTEZA H_0

c) JEDNOSTRANI TEST:

H₀: $\mu_1 = \mu_2$

H_a: $\mu_1 < \mu_2$

t - TEST: JEDNAKO KAO U a) ZADATKU,
DOBRIJE SE

$t = -2.918355$

p - VRIJEDNOST JE DVOSTRUKO MANJA NEGO
U a) ZADATKU (JER KORISTIMO JEDNOSTRANI
TEST):

$$p = \frac{0.0171}{2} = 0.00855 //$$

$\lambda = 5\%$:

$p < \lambda \rightarrow$ S RAZINOM ENAČAJNOSTI 5%
PODACI DAJU DOVOLJNO DOKAZA DA JE $\mu_1 < \mu_2$

ISTI ZAKLJUČAK VRIJEDI I ZA $\lambda = 0.01$ JER
I DALJE $p < \lambda$. INTERPRETACIJA: 99% SMO SIGURNI
DA JE TOČNOST MODELA b_1 MANJA OD TOČNOSTI MODELA b_2 .

d) MORAMO PRETPOSTAVITI DA SE PODACI POKORAVAJU
NORMALNOJ (ILI STUDENTOVoj za $N < 30$) DISTRIBUCIJI.
TA PRETPOSTAVKA VRIJEDI JER JE TOČNOST MODELA
SREDNJA VRIJEDNOST BROJA ISPRAVNO KLASIFICIRANIH
PRIMJERA, A PREMA CENTRALNOM GRANIČNOM TEOREMU
DISTRIBUCIJA UZORKOVANJA SREDNJE VRIJEDNOSTI POKORAVA
SE NORMALNOJ (ZA $N \rightarrow \infty$) NEOVISNO O DISTRIBUCIJI

POPULACIJE. BROJ PRIMJERA N IGLA VLOGU:
ZA VEĆI N VIŠE SMO SIGURNI U NAŠE
PODATKE I MOŽEMO POBIVATI SNAŽNJE ZAKLJUČKE.