

Strojno učenje – domaća zadaća 1

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

Zadano: 9. 10. 2016. Rok: 14. 10. 2016.

Napomena: Zadatke možete rješavati samostalno ili u grupi. Ako zadatke rješavate u grupi, pobrinite se da svi članovi grupe pridonose rješenju i da ga naposlijetu 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: *Na stvarim problemima razlikovati klasifikaciju od regresije.*] Objasnите razliku između klasifikacije i regresije. Koji je od ta dva pristupa prikladan za: (a) filtriranje neželjene e-pošte (*spam*), (b) predviđanje kretanja dionica, (c) rangiranje rezultata tražilice? Kako biste u ovim slučajevima definirali ciljne oznake y ?
2. [Svrha: *Razumjeti što je hipoteza, što je model i koja je veza između njih. Razumjeti što je prostor inačica.*]
 - (a) Dopunite praznine:

Hipoteza je funkcija koja _____ preslikava u _____, definirana do na _____. Model je _____ hipoteza, indeksiranih _____. Model također nazivamo prostorom _____, a dimenzija tog prostora jednaka je _____. Učenje modela odgovara pretraživanju _____ u potrazi za _____ hipotezom. To je ona hipoteza koja _____ klasificira označene primjere, što procjenjujemo pomoću _____ mjerene na _____. Drugim riječima, učenje modela svodi se na _____ parametara modela s _____ kao kriterijskom funkcijom.
 - (b) Rješavamo problem binarne klasifikacije u prostoru primjera $\mathcal{X} = \{0, 1\}^2$. Definirajte linearan model koji će primjere odvajati pravcem.
 - (c) Koja je dimenzija prostora parametra? Koliko različitih hipoteza postoji u \mathcal{H} ?
 - (d) Neka je skup označenih primjera sljedeći:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((0, 0), 0), ((1, 1), 0), ((1, 0), 1), ((0, 1), 1)\}.$$

Odredite konkretnu hipotezu $h \in \mathcal{H}$ koja ima najmanju empirijsku pogrešku.

- (e) Definirajte prostor inačica $VS_{\mathcal{H}, \mathcal{D}}$. Odredite taj prostor za ovaj konkretni problem.
3. [Svrha: *Shvatiti što je to induktivna pristranost i kako ona određuje klasifikaciju nevidjenih primjera.*] Pročitajte poglavlje 2.3 u skripti (tu temu nismo obradili na predavanju).
 - (a) Definirajte induktivnu pristranost (neformalno i formalno). Koje su dvije vrste induktivne pristranosti?

- (b) Raspolažemo skupom označenih primjera:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((0, 0, 0), 0), ((1, 0, 0), 1), ((1, 0, 1), 1), ((0, 1, 0), 1), ((0, 1, 1), 1)\}.$$

Koja je klasifikacija neviđenih primjera?

- (c) Definirajte linearan model \mathcal{H} za ulazni prostor $\mathcal{X} = \{0, 1\}^3$. Koja je to vrsta induktivne pristranosti?
- (d) Odredite klasifikaciju neviđenih primjera uz model \mathcal{H} . Je li induktivna pristranost dovoljna za jednoznačnu klasifikaciju primjera iz \mathcal{D} ? Odredite skup prostora inačica $VS_{\mathcal{H}, \mathcal{D}}$.
- (e) Definirajte (neformalno) neku dodatnu induktivnu pristranost takvu da klasifikacija svakog primjera slijedi jednoznačno na temelju skupa primjera \mathcal{D} . Koje je vrste ta dodatna induktivna pristranost?

4. [Svrha: Znati nabrojati osnovne komponente algoritma strojnog učenja i povezati ih s induktivnom pristranošću.]

- (a) Nabrojite tri osnovne komponente algoritma strojnog učenja.
 (b) Identificirajte uz koje se komponente veže koja vrsta induktivne pristranosti.

5. [Svrha: Razumjeti vezu između funkcije gubitka i empirijske pogreške te mogućnost njihove prilagodbe konkretnom problemu.]

- (a) Definirajte empirijsku pogrešku preko funkcije gubitka L .
 (b) Kod asimetričnih gubitaka funkciju L možemo definirati preko matrice gubitka (v. skriptu: poglavlje 2.7 i primjer 2.6). Definirajte takvu matricu za problem klasifikacije neželjene e-pošte te izračunajte funkciju pogreške za slučaj pet pogrešno negativnih i dvije pogrešno pozitivne klasifikacije.

6. [Svrha: Razviti ispravnu intuiciju za odabir modela temeljem unakrsne provjere.]

- (a) Skicirajte krivulje pogreške učenje i ispitne pogreške u ovisnosti o složenosti modela. Naznačite područje prenaučenosti i podnaučenosti.
 (b) Objasnite zašto pogreška učenja s povećanjem složenosti modela teži k nuli.
 (c) Raspolažemo modelom \mathcal{H}_α koji ima hiperparametar α kojim se može ugađati složenost modela. Za odabrani α naučili smo hipotezu koja minimizira empirijsku pogrešku. Unakrsnom provjerom ustanovali smo da je ispitna pogreška znatno veća od pogreške učenja. Je li naš odabir hiperparametra α suboptimalan? Obrazložite odgovor.
 (d) Raspolažemo modelom \mathcal{H}_α s hiperparametrom α (veći α daje složeniji model). Raspolažemo dvama optimizacijskim algoritmima: L_1 i L_2 . Algoritam L_2 lošiji je od algoritma L_1 , u smislu da L_2 pronalazi parametre $\boldsymbol{\theta}_2$ koji su lošiji od parametara $\boldsymbol{\theta}_1$ koje pronalazi L_1 , tj. $E(\boldsymbol{\theta}_2 | \mathcal{D}) > E(\boldsymbol{\theta}_1 | \mathcal{D})$. Neka α_1^* označava optimalnu vrijednost hiperparametra za \mathcal{H}_α učenog algoritmom L_1 , a α_2^* optimalnu vrijednost za \mathcal{H}_α učenog algoritmom L_2 . Načinite skicu analognu onoj iz zadatka (a) i naznačite vrijednosti pogrešaka koje odgovaraju modelima $\mathcal{H}_{\alpha_1^*}$ i $\mathcal{H}_{\alpha_2^*}$.
 (e) Može li model učen lošijim algoritmom L_2 imati manju ispitnu pogrešku od modela koji je učen boljim algoritmom L_1 , ali nije optimalan? Skicirajte takvu situaciju na prethodnoj skici.

1. DOMAĆA ZADAĆA

14.10.2016

1. NA STARNIM PROBLEMIIMA RAZLIKOVATI KLASIFIKACIJU OD REGRESIJE

⇒ RAZLIKA između KLASIFIKACIJE i REGRESIJE:

- Kod KLASIFIKACIJE primjeru pridružujemo KLASU (RAZRED).
Koji je taj primjer pripada. Ciljna vrijednost je DISKRETNÄ.
- Kod REGRESIJE primjeru pridružujemo nizu KONTINUIRANU VR.

→ Koji od dva pristupa je primoran za:

- a) PREDVODANJE KRETANJA DIONICA : REGRESIJA
- b) FILTRIRANJE NEŽEYENE e-pošte : KLASIFIKACIJA
- c) RANGIRANJE REZULTATA TRAJELICE : BINARNA KLASIFIKACIJA
M PAROVIMA

2. HIPOTEZA, MODEL, VEZA između njih. PROSTOR INACICA.

- a) Hipoteza je funkcija koja PROSTOR PRIMJERA preslikava u SKUP OZNAKA KLASI; definirana do na PARAMETAR Θ . Model također nazivamo prostorom PARAMETARA, a dimenzija tog prostora jednaka je Broju PARAMETARA d . Model je SKUP MOGUĆIH hipoteza indeksiranih s PARAMETRIMA Θ . Učenje modela odgovara prenosićujući PROSTORA PARAMETARA u zonu za NAJBOLJOM hipotezom. To je ona hipoteza koja NAJTOČNije klasificira označene primjeri, što procijevamo pomoći EMPIRIJSKE POGREŠKE mjerue na SKUPU D (na skupu primjera). Drugim riječima, učenje modela svodi se na OPTIMIZACIJU parametara modela s EMPIRIJSKOM POGREŠKOM kao kriterijskom funkciju.

- b) RJEŠAVANJE PROBLEMA BINARNOG KLASIFIKATORA u PROSTORU PRIMJERA $X = \{0, 1\}^2$. POTREBNO JE DEF. LINEARAN MODEL koji će primjere odvajati pravcem.

$$X = \{0, 1\}^2 \quad h(x) = 1 \{ w_0 + w_1 x_1 + w_2 x_2 > 0 \}$$

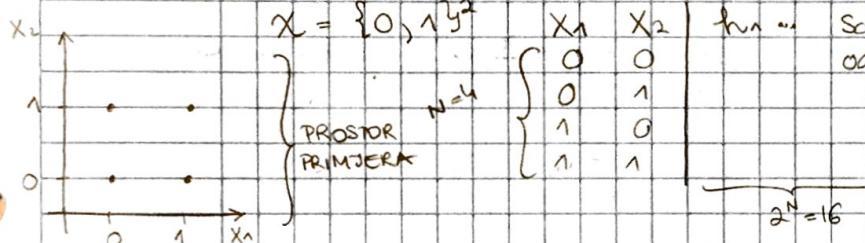
$$h(x_1, x_2 | w_0, w_1, w_2) = 1 \{ w_0 + w_1 x_1 + w_2 x_2 \}$$

(2) c) $X = \{0, 1\}^2$ i) LINEARAN MODEL KODI ĆE PR.
 ODVAJATI PRAVCEM ii) DIMENZIJA PROSTORA PARAM.
 iii) BR. RAZL. HIPOTEZA U \mathbb{H}^2 ?

i) $h(x_1, x_2; w_0, w_1, w_2) = 1 \{w_0 + w_1 x_1 + w_2 x_2 > 0\}$

iii) DIM. PROSTORA PARAMETARA = 3 ; w_0, w_1, w_2

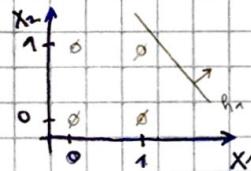
iii) BR. RAZL. HIPOTEZA U \mathbb{H}^2 ? $\rightarrow \max \text{ broj} = 2^N$ už $N = \text{broj vr. pr.}$
 \leftarrow naš slučaj $2^4 = 16$ max



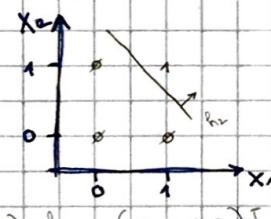
Sad razmatramo sve i gledamo kako odgovaraju nazna

14 HIPOTEZA

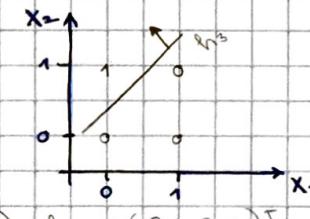
1) $h_1 = (0 \ 0 \ 0 \ 0)^T$



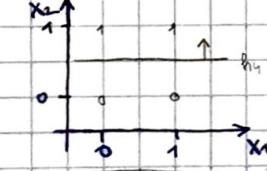
2) $h_2 = (0 \ 0 \ 0 \ 1)^T$



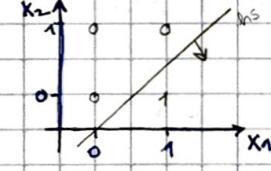
3) $h_3 = (0 \ 0 \ 1 \ 0)^T$



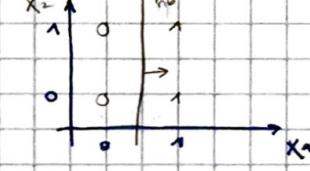
4) $h_4 = (0 \ 0 \ 1 \ 1)^T$



5) $h_5 = (0 \ 1 \ 0 \ 0)^T$

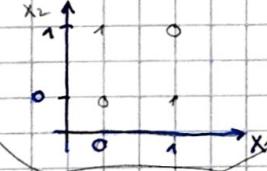


6) $h_6 = (0 \ 1 \ 0 \ 1)^T$

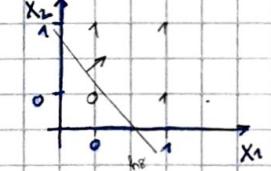


remake razliku postavi!

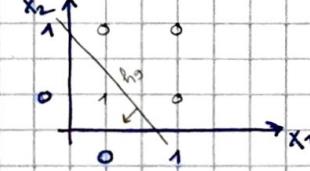
7) $h_7 = (0 \ 1 \ 1 \ 0)^T$



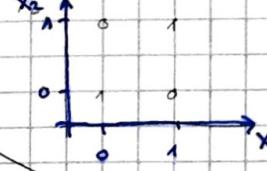
8) $h_8 = (0 \ 1 \ 1 \ 1)^T$



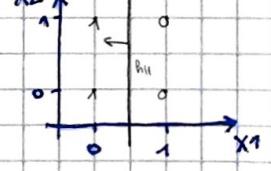
9) $h_9 = (1 \ 0 \ 0 \ 0)^T$



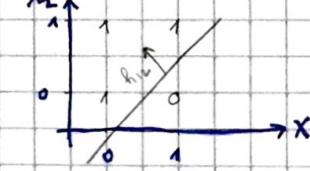
10) $h_{10} = (1 \ 0 \ 0 \ 1)^T$



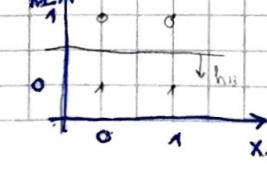
11) $h_{11} = (1 \ 0 \ 1 \ 0)^T$



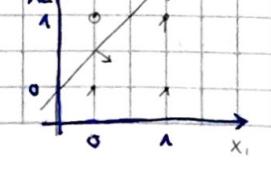
12) $h_{12} = (1 \ 0 \ 1 \ 1)^T$



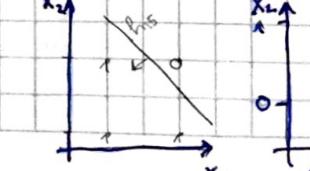
13) $h_{13} = (1 \ 1 \ 0 \ 0)^T$



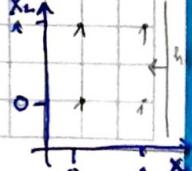
14) $h_{14} = (1 \ 1 \ 0 \ 1)^T$



15) $h_{15} = (1 \ 1 \ 1 \ 0)^T$

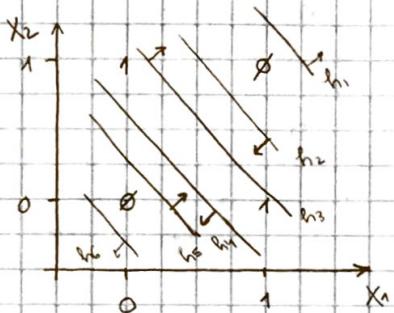


16) $h_{16} = (1 \ 1 \ 1 \ 1)^T$

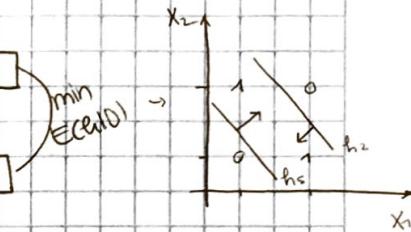


$$d^*) D = \{x^{(i)}, y^{(i)}\} \quad y = \{(0,0), 0, (1,1), 0, (1,0), 1, (0,1), 1\}$$

HIPOTEZA S NAJMANJOM EMPIRIJSKOM POGREŠKOM?



$E(h_1 D) = 0.5$
$E(h_2 D) = 0.25$
$E(h_3 D) = 0.75$
$E(h_4 D) = 0.75$
$E(h_5 D) = 0.25$
$E(h_6 D) = 0.5$



e*) PROSTOR INACICA DEF. + PROSTOR INACICA ZA OVAJ KONKRETAN PROBLEM.

PROSTOR INACICA VS_{H,D} modela H je skup hipoteza koje su konzistentne s primjerenim za učenje D:

$$VS_{H,D} = \{h \in H \mid \text{Consistent}(h, D)\}$$

\Rightarrow Za ovaj konkretan problem $VS_{H,D} = \emptyset$, jer nema skup, što je i vidljivo iz d)*. Ne postoji hipoteza koja je $E(h|D) = 0$, odnosno za koju vrijedi da je $h(x) = y$.

3) a) INDUKTIVNA PRISTRANOST DEF. (formalno, neformalno).
DUJE VRSTE INDUKTIVNE PRISTRANOSTI?

INDUKTIVNA PRISTRANOST je skup apriornih pretpostavki koje omogućavaju induktivno učenje

DUJE VRSTE INDUKTIVNE PRISTRANOSTI:

- 1) PRISTRANOST OGRANIČAVANjem ili PRISTRANOST JEZICA
(odabiremo model H i time ogranicimo skup hipoteza koje se mogu prikazati tim modelom)
- 2) PRISTRANOST PREFERENCIJOM ili PRISTRANOST PRETRAVLJAVANjem
(definiramo način pretravljanja hipoteza unutar H i na taj način damo prednost jednemu hipotezama u odnosu na druge).

Vidimo se da 2) čini bolje nego 1), budući da ne ograničava u skup pretravljivih hipoteza, u praksi se obično voli biti 1) zbog jednostavnosti modela.

INDUKTIVNU PRISTRANOST možemo shvatiti kao dodatnu inf. koja nam omogućava da na temelju nepoznate inf. iz problema možemo zaključiti o kojoj je točno hipotezi vrijed.

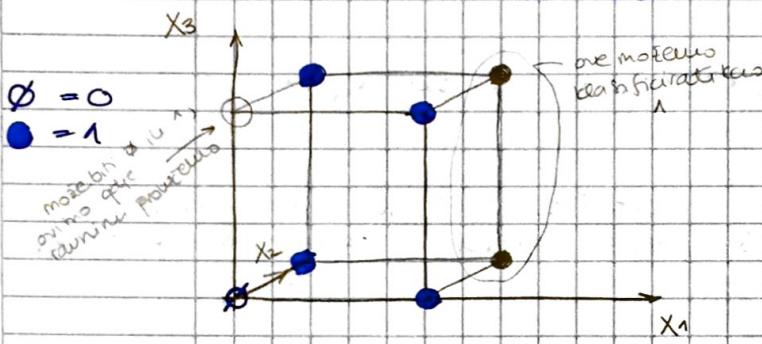
FORMALNA DEF. Nella je \mathcal{L} algoritam za učenje, h_x hipoteza inducirana pomoći \mathcal{X} na skupu za učenje \mathcal{D} i nella je $\text{h}_x(x)$ klasifikacija primjera $x \in \mathcal{X}$ temeljena te hipoteze. Induktivna pristranost od \mathcal{X} je, bilo koji skup minimalnih pretpostavki \mathcal{B} takođe da

$$\forall D, \forall x \in \mathcal{X}. ((\mathcal{B} \wedge D \wedge x) \vdash \text{h}_x(x))$$

\Rightarrow UKRATKO INDUKTIVNA PRISTRANOST je skup pretpostavki temeljem kojih klasifikacija primjera slijedi DEDUKTIVNO.

b) $D = \{(x^{(i)}, y^{(i)})\}^7 = \{((0,0,0), 0), ((1,0,0), 1), ((1,0,1), 1), ((0,1,0), 1), ((0,1,1), 1)\}$

KLASIFIKACIJA NEVIDENIH PRIMJERA?



$$|D| = 5 \quad |\mathcal{X}| = 8$$

KLASIFIKACIJA NEVIDENIH PRIMJERA? NEMA

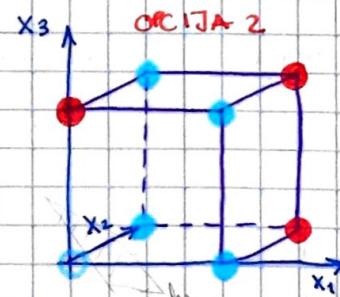
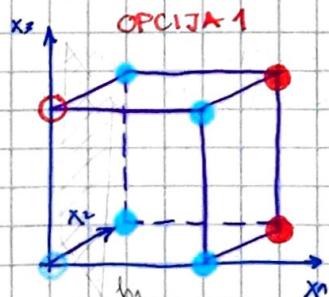
jer nema dovoljno prisutnih vještača

Potrebno je uvesti induktivnu pr.

c) $\mathcal{X} = \{0, 1\}^3 \rightarrow$ DEF. LIN. MOD. TLE ŽE UL. PROSTOR. KOJA JE OVO VRŠNA INDUKTIVNE PRISTRANOSTI?

PRISTRANOST FEZIKA! (PRISTRANOST OGRANIČAVANJA) jer smo odabrali model \mathcal{H} i time ogranicili skup hipoteza.

d) KLASIFIKACIJA NEVIDENIH PRIMJERA uz model \mathcal{H} . JE LI INDUKTIVNA PRISTRAK, DOVOLJNA ZA JEDNOZNAČNU KLASIFIKACIJU PR. IZ \mathcal{D} ? ODREDI SKUP PROSTORA INACICA VS_{H,D}.



$$VS_{H,D} = \{h_1, h_2\}$$

$$h_1(\vec{x}|\vec{w}) = \begin{cases} 0 & \text{za } \vec{x} = (0,0,*) \\ 1 & \text{nije} \end{cases}$$

$$h_2(\vec{x}|\vec{w}) = \begin{cases} 0 & \text{za } \vec{x} = (0,1,*) \\ 1 & \text{nije} \end{cases}$$

\rightarrow nije dovoljna induktivna pristranost za jednoznačnu klasifikaciju

e) DEF. DODATNE INDUKT. PRISTRANOSTI TAKVA DA KLASIF. SVAKOG PR. SLJEDI JEDNOZNAČNO NA TEMELJU SKUPA PRIMJERA D. KJE JE VRSTE TA DODATNA INDUKTIVNA PRISTRANOST?

Npr. ako pri trajećoj hipotezi odaberemo ravninu u blizini točki $(0,0,0)$.

Ova dodatna induktivna pristranost je PRISTRANOST PRETRAVANJA!

(4) a) TRI OSNOVNE KOMPONENTE ALGORITMA SNOJNOG UČENJA

- 1) MODEL (ili prostor hipoteza)
- 2) FUNKCIJA GUBITKA (kriterijska fja, empirijska pogreška)
- 3) OPTIMIZACIJSKI POSTUPAK

b) MEŽ KOJE SE KOMPONENTE VEŽE KOJA VRSTA INDUKTIVNE PRISTRANOSTI?

- 1) ODABIR MODELA \rightarrow PRISTRANOST JEŽIKA (ograničavanja)
- 2) fja GUBITKA i 3) OPTIMIZACIJSKI POSTUPAK \rightarrow PRISTRANOST PRETRAVANJA

(5) a) DEF. EMPIRIJSKE POGR. PREKO FUNKCIJE GUBITKA L

$$f_j \text{ GUBITKA: } L(y^{(i)}, h(x^{(i)})|\theta)) = 1_{\{h(x^{(i)})|\theta) \neq y^{(i)}\}} = |h(x^{(i)})|\theta) - y^{(i)}|$$

$$\text{EMP. POGR: } E(\Theta|D) = \frac{1}{N} \sum_{i=1}^N L(y^{(i)}, h(x^{(i)})|\theta)$$

Ocenjujući fje gubitka
rad primjerima iz skupa
za učenje

Nebitni faktor jer ne utiče
za kje vr. Θ fja f. doseže min.

b) DEF. MATRICU GUBITKA ZA PROBLEM KLASIFIKACIJE NEŽEYENE e-pošte. Izr. fja POGREŠKE ZA SLUČAJ 5 POGREŠNO NEGATIVNIH I DVIJE POGREŠNO POZITIVNE KLASIFIKACIJE.

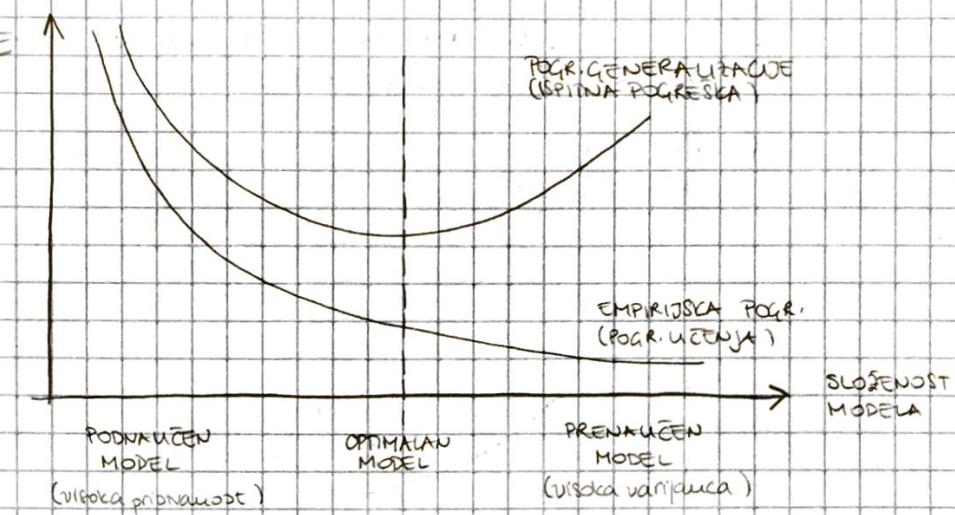
MATRICA GUBITKA \rightarrow RETCI - STVARNE KLASE
STUPCI - ODABRANE KLASE

		false positive		true positive	
		spam	not spam	spam	not spam
0 - SPAM	0	5	1	0	0
	1 - SPAM	0	1	0	5

legende: main
seku_dos a class
nje spaur a class
madel spaur
spaur a class

$$E = \frac{1}{N} \cdot (5 \cdot 5 + 1 \cdot 5) = \frac{30}{10} = 3$$

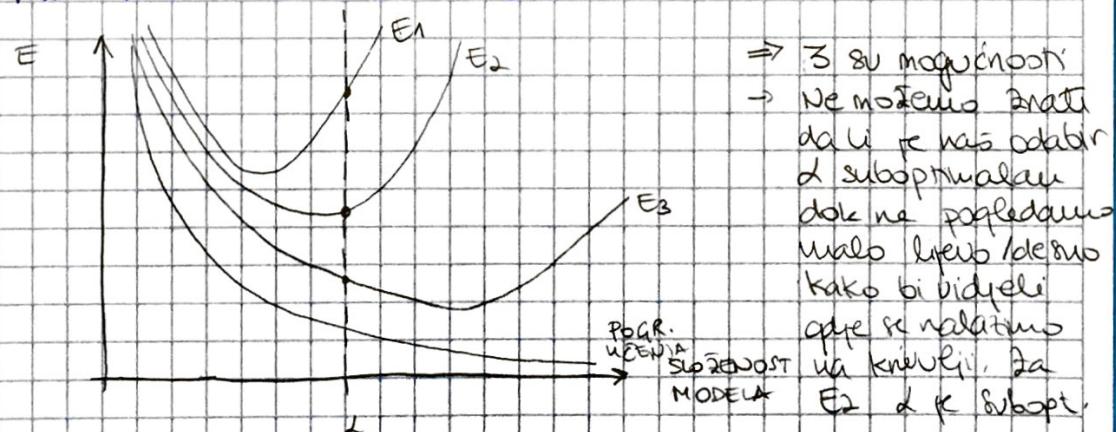
6) a) KRIVULJA POGR. UČENJA I ISPITNE POGR. U OVISNOSTI O SLOŽENOSTI MODELAA. NAZNAČI PODR. PRENAUČ. I PODNAUČ.



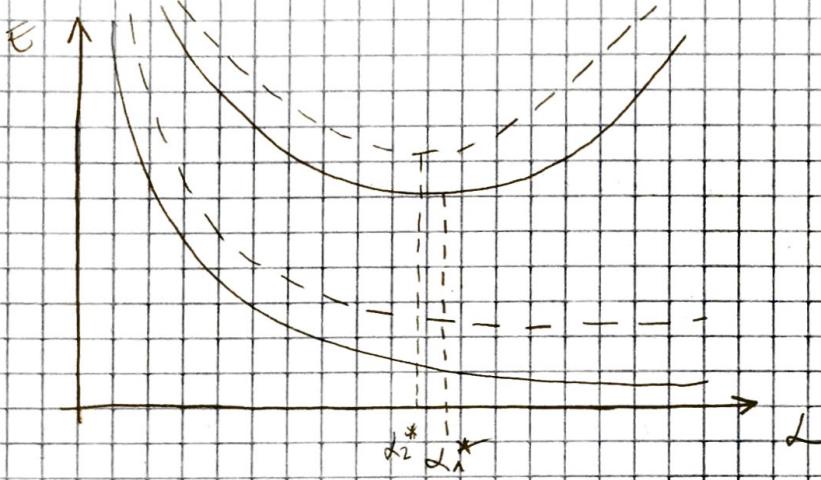
b) ŽAŠTO POGREŠKA UČENJA S POVEĆANJEM SLOŽENOSTI MODELAA TEŽI K NULI?

jer se hipoteza u potpunosti prilagođi primjerenim za učenje (uključujući i šum),

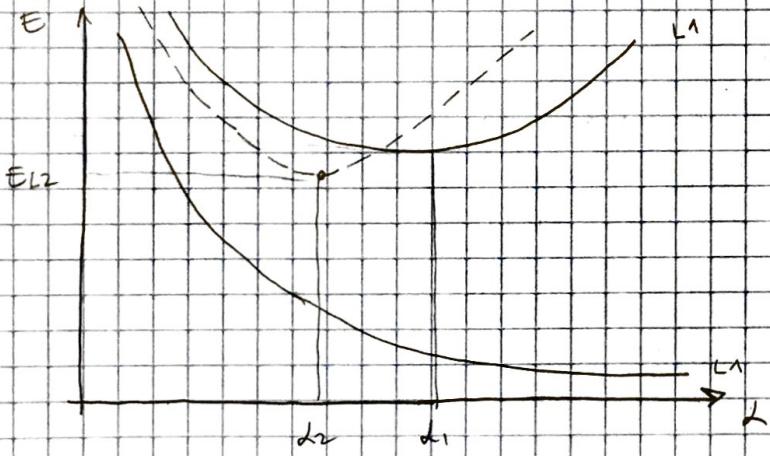
c) RASPOLAŽEŠ MODELOM H_λ KOJI IMA HIPERPARAMETR λ S KOJIM SE MOže UGADATI SLOŽENOST MODELAA.
* ZA λ HIPOTEZA MINIMIZIRA EMPIRIJSKU POGREŠKU, UNAKRINU PROVJEROM USTANOVILI smo DA JE ISPITNA POGR. \gg POGR. UČENJA. JE LI NAJ ODABIR HIPERPARAMETRA λ SUBOPTIMALAN?



d) H_1, H_2 (daju sečenjevi model). RASPOLAŽEMO S DVA OPTIMIZ. MODELA $\rightarrow L_1 \text{ i } L_2$. L_2 PROVALA ZI PARAMETRE Θ_2 KOP SU LOŠIJI OD PARAMETRA Θ_1 KOJE NALAZI L_1 . $E(\Theta_2|D) > E(\Theta_1|D)$, L_2 LOŠIJI OD L_1 . L_1^* OPTIMALNA VR. HIPERPARAMETRA ZA UČENJENO S L_1 , L_2^* U- L_2 . NAPRAVI SKICU KAO U a) (NAJNAČI VR. POGR. KOJE ODGOVARAJU H_1 i H_2)



e) MOŽE LI MODEL UČEN LOŠIJIM ALG. L_2 IMATI MANJU ISPITNU POGR. OD ONOG UČENJOG S L_1 , A NJEG OPTIMALAN? SKICAJ TAKU SITUACIJU



\Rightarrow JEDINO STO JE GRESKA UČENJA SA SIGURNOSĆU JE TO DA OSTALO JE VARIJABILNO.