Strojno učenje – domaća zadaća 3

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

Zadano: 24. 10. 2016. Rok: 28. 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 naposlijetku 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: Znati definirati poopćeni linearan model i razumjeti dvije vrste nelinearnosti koje se kod njega javljaju. Razumjeti geometriju linearnog modela.***]

   (a) Definirajte poopćeni linearan model i navedite varijante modela koje su (i) nelinearne u parametrima i (ii) nelinearne u granici između klasa.

   (b) Dokažite da je $\mathbf{w}$ normala (hiper)ravnine.

   (c) Izvedite izraz za predznačenu udaljenost primjera $\mathbf{x}$ od hiperravnine.

2. [***Svrha: Isprobati na konkretnom kako se regresija može upotrijebiti za klasifikaciju. Razumjeti kako ostvariti višeklasnu klasifikaciju pomoću više binarnih modela. Razumjeti zašto je korištenje linearne regresije za klasifikaciju loša ideja.***] Na predavanjimo smo pokazali kako se linearan model regresije može koristiti za klasifikaciju. Pokažite to na sljedećim primjerima iz triju ($K = 3$) klasa:

$$\begin{aligned}\mathcal{D} &= \big\{(\mathbf{x}^{(i)}, y^{(i)})\big\}_{i=1}^{6}\\ &= \big\{\big((-3,1),0\big), \big((-3,3),0\big), \big((1,2),1\big), \big((2,1),1\big), \big((1,-2),2\big), \big((2,-3),2\big)\big\}.\end{aligned}$$

   (a) U gornjem primjeru za oznake klasa koristili smo cijele brojeve, no u praksi kod linearnih modela to ne radimo tako. Kako to radimo u praksi i zašto ne koristimo cijele brojeve kao oznake klasa?

   (b) Primijenite pristup *jedan-naspram-ostali* (OVR), definirajte dizajn-matricu i vektor oznaka $\mathbf{y}$ za svaki od triju modela, te izračunajte hipoteze $h_j(\mathbf{x})$ za svaku od triju klasa. Izračun možete napraviti ručno ili u nekom alatu.

   (c) Izračunajte diskriminacijske funkcije $h_{12}(\mathbf{x})$, $h_{23}(\mathbf{x})$ i $h_{13}(\mathbf{x})$ između svih parova hipoteza. Skicirajte primjere i dobivene granice u prostoru $\mathbb{R}^2$.

   (d) U koju bi klasu bio klasificiran primjer $\mathbf{x} = (-1, 3)$? Obrazložite odgovor.

   (e) Objasnite koja je prednost pristupa OVR nad pristupom *jedan-naspram-jedan* (OVO), a što je nedostatak.

   (f) Možete li reći koja je vjerojatnost da primjer pripada toj klasi? Obrazložite odgovor.

   (g) U praksi linearnu regresiju ne bismo željeli koristiti za klasifikaciju. Zašto? Pokažite na gornjem primjeru u čemu je problem (možete modificirati primjer).

3. [***Svrha: Razumjeti kriterij perceptrona i ograničenja koja proizlaze iz toga što ta funkcija nije derivabilna.***]

   Algoritam perceptrona minimizira pogrešku $E_p(\mathbf{w}|\mathcal{D})$ koju nazivamo *kriterij perceptrona*. Ta je funkcija aproksimacija udjela pogrešnih klasifikacija (engl. *misclassification ratio*), odnosno očekivanja gubitka 0-1, $E_m(\mathbf{w}|D)$, koju bismo idealno htjeli minimizirati, ali to ne možemo. Pogledajte (u slajdovima) kako izgleda pogreška perceptrona u prostoru parametara.

   (a) Objasnite zašto ne možemo izravno minimizirati $E_m(\mathbf{w}|\mathcal{D})$.

   (b) Je li pogreška perceptrona $E_p(\mathbf{w}|\mathcal{D})$ gornja ograda za empirijsku pogrešku $E(\mathbf{w}|\mathcal{D})$? Objasnite.

   (c) Jedan nedostatak perceptrona jest da rješenje $\mathbf{w}^*$ (a time i položaj granice) ovisi o početnim težinama i redoslijedu predočavanja primjera. Pozivajući se na sliku površine pogreške u prostoru parametara, objasnite zbog čega je to tako.

   (d) Drugi nedostatak perceptrona jest da postupak ne konvergira ako primjeri nisu linearno odvojivi. Pozivajući se opet na sliku površine pogreške u prostoru parametara, objasnite zašto je to tako.

4. [***Svrha: Razumjeti odnose između funkcija gubitaka različitih modela. Razumjeti kako funkcija gubitka određuje dobra i loša svojstva modela.***]

   (a) Skicirajte na jednome grafikonu sljedeće tri funkcije gubitka: (1) kvadratni gubitak regresije, (2) gubitak perceptrona i (3) gubitak 0-1.

   (b) Odgovorite čemu odgovara desna strana grafikona (x-os veća od nule), a čemu lijeva (x-os manja od nule).

   (c) Pozivajući se na skicu, odgovorite zašto kvadratni gubitak nije prikladan gubitak u slučajevima kada želimo minimizirati broj pogrešnih klasifikacija.

   (d) Pozivajući se na skicu, odgovorite za koje će modele očekivanje gubitka (empirijska pogreška) biti veće od broja pogrešnih klasifikacija.

③ DOMAĆA ZADAĆA

① a) DEF. POOPĆENI LIN. MODEL + NAVEDI VAR. MODELA KOJE SU: i) NELIN. U PARAM. ii) NELIN U GRANICI IZMEĐU KLASA

⇒ LINEARAN MODEL (Granica je linearna, hiperravnina) $h(x) = \vec{w}^T \tilde{x}$

⇒ POOPĆENI LINEARNI MODEL ⇒ $\boxed{h(x) = f(\vec{w}^T \tilde{x})}$ → granica u ul. prostoru je linearna premda je f nelin.

gdje je f aktivacijska fja koja spljošćuje linearnu funkciju na jedinični interval $[0,1]$ ili na interval $[-1,1]$.
$\vec{w}$ su parametri modela, dok je $\tilde{x}$ vektor značajki.

i) Model je NELINEARAN u PARAM. kada je f NELINEARNA fja!

ii) Mod. je NELIN. u GRANICAMA IZMEĐU KLASA kada je ulazni prostor X preslikan u prostor značajki $\phi(x)$ NELINEARNIM PRESLIKAVANJEM $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^m$ ⇒ $\boxed{h(x) = f(\vec{w}^T \phi(x))}$

b) DOKAŽI DA JE $\vec{w}$ NORMALA (HIPER)RAVNINE:

→ Razmatramo sljedeći model $h(\vec{x}; \vec{w}) = \omega_0 + \omega_1 x_1 + \omega_2 x_2$
→ Granica između klasa je $\omega_0 + \omega_1 x_1 + \omega_2 x_2 = \emptyset$
Primjer X klasificiramo u C_1 ako je $h(x) \geq 0$, inače u C_2.

⇒ Granica između klasa je $(n-1)$ dimenzijska hiperravnina definirana jednadžbom $h(x) = \emptyset$. Za dvije točke X_1 i X_2 KOJE LEŽE NA HIPERRAVNINI vrijedi:

$$h(x_1) = h(x_2)$$
$$w^T x_1 + w_0 = w^T x_2 + w_0$$
$$w^T (x_1 - x_2) = \emptyset$$ → $\vec{w}$ JE NORMALA HIPERRAVN.

c) IZVEDI IZRAZ ZA PREDZNAČNU UDALJENOST PR. $\vec{x}$ OD HIPERR.

d – PREDZNAČNA UDALJENOST točke X od hiperravnine
X – proizvoljno odabrana točka $X_\perp$ – ortog. projekcija na hiperravnini

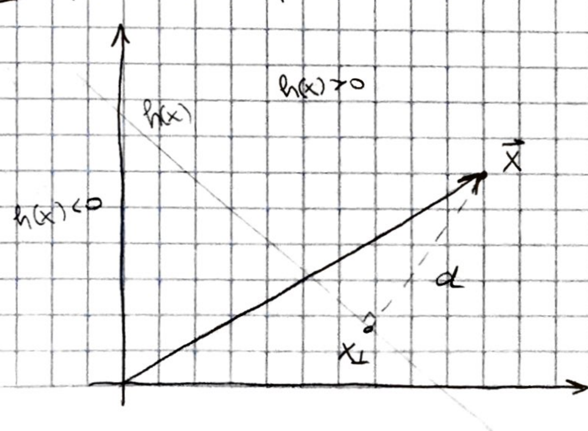
Vrijedi: $$x = x_\perp + d \frac{w}{\|w\|} \quad / \cdot w^T / + w_0$$

$$w^T x + w_0 = w^T x_\perp + w_0 + d \frac{w^T w}{\|w\|} \quad (w^T w = \|w\|^2)$$

$$h(x) = h(x_\perp) + d \|w\|$$

$h(x_\perp) = \emptyset$ jer leži na hiperravnini

$$\boxed{d = \frac{h(x)}{\|w\|}}$$



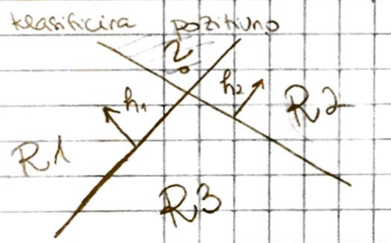
② $K=3$; $D=\{(x^{(i)}, y^{(i)})\}_{i=1}^{6} = \{((-3,1),0), ((-3,3),0), ((1,2),1), ((2,1),1), ((1,-2),2),$
$((2,-3),2)\}$

a) KAKVE OZNAKE ZA KLASE SE KORISTE U PRAKSI? ZAŠTO NE KORISTIMO CIJELE BROJEVE?

U praksi, umjesto višeklasne klasifikacije s cijelim brojevima radimo više binarnih klasifikacija. 3 su pristupa:

① OVR (one-vs-rest)

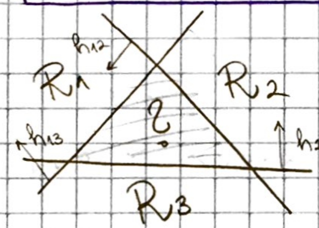
→ (K−1) binarnih klasifikatora, svaki bin. klas. h_j odjeljuje pr. klase C_j od pr. svih ostalih klasa. Pr. klas. u C_j ako je $\boxed{h_j(x) \geq 0}$
→ PROBLEM: Kad više bin. klas. pr. klasificira pozitivno



② OVO (one-vs-one)

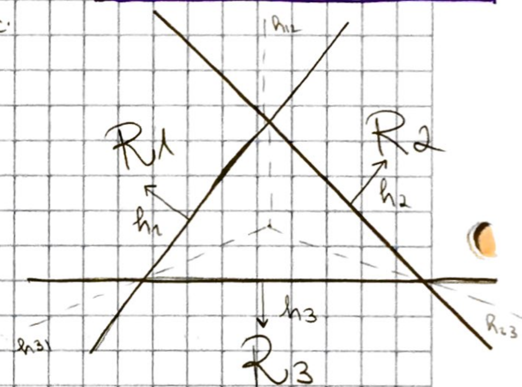
→ $\binom{K}{2}$ bin. klasif., po jedan za svaki par klasa. Klasif. h_{ij} odjeljuje pr. kl. C_i od pr. klase C_j. U sluč. da više klasa nekih pr. klasif. pozitivno odredtuje x:

$$\boxed{h(x) = \underset{C_i}{\operatorname{argmax}} \sum_{i \neq j} h_{ij}(x)}$$



③ OVR* − K bin. klasif. OVR i zatim klasifikacija u klasu s najvećom pouzdanošću (ovako u praksi radimo)

$$\boxed{h(x) = \underset{C_j}{\operatorname{argmax}} (h_j(x))}$$



U praksi se cijeli brojevi ne koriste jer prilikom regresije zna doći do problema, jednim pravcem ne mogu se odvojiti tri klase.

↳ pr. Imamo $x \in R_3$, koji se nalazi blizu R_1, a daleko od R_2. Prilikom regresije x može biti bliže R_1, što nije točno.

b) PRIMIJENI OVR, DEF. DIZAJN MATRICU Φ I VEKTOR OZNAKA y ZA SVAKI OD 3 MODELA TE IZR. HIPOTEZU $h_j(x)$ ZA SVAKU OD 3 KLASE. → $K=3$ → $(K-1)=2$ binarne klasif. → h_1, h_2

$$\Phi(x) = \begin{bmatrix} 1 & -3 & 1 \\ 1 & -3 & 3 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & -2 \\ 1 & 2 & -3 \end{bmatrix} \quad y_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad y_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad y_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

$$\boxed{\vec{w}_j = (\Phi^T \Phi)^{-1} \Phi^T \vec{y}_j = \Phi^+ \vec{y}_j}$$

$$\boxed{h_j(x) = w_j^T \tilde{x} = w_j^T \Phi(x)}$$

$$w = (\overbrace{3\times6 \;\; 6\times3}^{3\times3})^{-1} \; \overbrace{3\times6 \;\; 6\times1}^{3\times1} = 3\times1 = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

$$h_j(x) = w_{0j} + w_{1j} x_1 + w_{2j} x_2$$

$h_1 = 0.3351 - 0.2173 x_1 - 0.0052 x_2$
$h_2 = 0.2592 + 0.2343 x_1 + 0.2225 x_2$
$h_3 = 0.4058 - 0.017 x_1 - 0.2173 x_2$

c) IZR. DISKR. F-JE $h_{12}(\vec{x})$, $h_{23}(\vec{x})$ I $h_{13}(\vec{x})$ IZMEĐU SVIH PAROVA HIPOTEZA. SKICIRAJ PR. I DOBIVENE GRANICE U PROSTORU $\mathbb{R}^2$.

$h_{12}(x) = ? = w_{0,12} + w_{1,12} x_1 + w_{2,12} x_2$; $\boxed{\vec{w}_{12} = \vec{w}_1 - \vec{w}_2}$

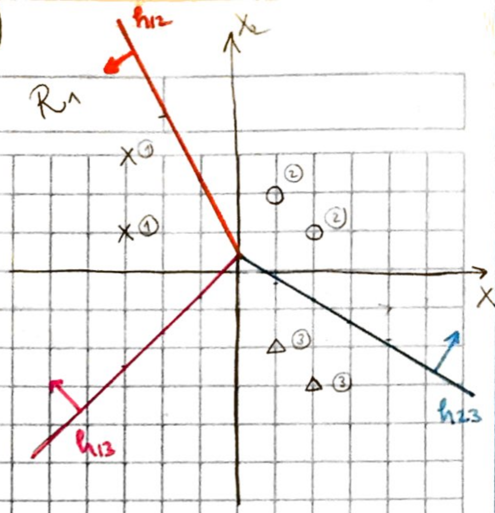
granica između h_1 i h_2

analogno za h_{23} i h_{13}:
$\vec{w}_{23} = \vec{w}_2 - \vec{w}_3$
$\vec{w}_{13} = \vec{w}_1 - \vec{w}_3$

$\Rightarrow$ $h_{12} = 0.0759 - 0.4516x_1 - 0.2277x_2 = \emptyset$ $\boxed{x_2 = -1.9833x_1 + \frac{1}{3}}$

$h_{23} = -0.1466 + 0.2513x_1 + 0.4398x_2 = \emptyset$ $\boxed{x_2 = -0.571x_1 + \frac{1}{3}}$

$h_{13} = -0.0707 - 0.2003x_1 + 0.212x_2 = \emptyset$ $\boxed{x_2 = 0.945x_1 + \frac{1}{3}}$



$X = \begin{bmatrix} -3 & 1 \\ -3 & 3 \\ 1 & 2 \\ 2 & 1 \\ 1 & -2 \\ 2 & -3 \end{bmatrix}$ $y = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 2 \\ 2 \end{bmatrix}$ $\begin{matrix} \} \text{KLASA 1} \\ \} \text{KLASA 2} \\ \} \text{KLASA 3} \end{matrix}$

(-3,1) $h_1 = 0.9817$ $h_2 = 0.2212$ $h_3 = 0.2395$ $\Rightarrow$ KLASA = 1

(-3,3) $h_1 = 0.9712$ $h_2 = 0.2238$ $h_3 = -0.1950$ $\Rightarrow$ KLASA 1

(1,2) $h_1 = 0.1073$ $h_2 = 0.9385$ $h_3 = -0.0458$ $\Rightarrow$ KLASA 2

(2,1) $h_1 = -0.1047$ $h_2 = 0.9503$ $h_3 = 0.1545$ $\Rightarrow$ KL. 2

(1,-2) $h_1 = -0.1283$ $h_2 = 0.0484$ $h_3 = 0.8233$ $\Rightarrow$ KL. 3

(2,-3) $h_1 = -0.0838$ $h_2 = 0.0602$ $h_3 = 1.0232$ $\Rightarrow$ KL. 3

d) U KOJU BI KLASU BIO KLASIFICIRAN PR. $x = (-1,3)$? OBRAZLOŽI

Primjer bi bio klasificiran u klasu 2 što je i vidljivo sa slike iz prethodnog podzadatka. Primjer se nalazi između hiperravnina h_{12} i h_{23}, dakle, klasa 1.
$h_{12} = 0.9215$

e) KOJA JE PREDNOST PRISTUPA OVR NAD OVO? ŠTO JE NEDOSTATAK?

PREDNOST: - Kod OVR pristupa NEMA JEDNOZNAČNOSTI! Jedan primjer pripada samo jednoj klasi!
- OVR ima manje modela od OVO! OVR - K, OVO - $\frac{K}{2}$

MANA: OVR lako rezultira neuravnoteženim brojem primjera kroz klase!

f) KOJA JE VJEROJATNOST DA PR. PRIPADA TOJ KLASI? OBRAZLOŽI

Ne možemo izr. vj. da primjer pripada određenoj klasi s obzirom da klasifikacija linearizacijom NIJE PREDIKTIVNI POSTUPAK.

Izlaz modela NEMA VJEROJATNOSNU INTERPRETACIJU! (VRIJEDNOSTI HIPOTEZE NISU OGRANIČENE NA INT. [0,1])

g) U PRAKSI LIN. REG. NE KORISTIMO ZA KLASIFIKACIJU. ZAŠTO? POKAŽI NA GORNJEM PR. U ČEMU JE PROBLEM

Lin. Reg. se ne koristi za klasifikaciju jer, uz to što izlazi nemaju vjerojatnosnu interpretaciju, model je jako NEROBUSAN NA VRIJEDNOSTI KOJE ODSKAČU! Uzrok tomu je što funkcija gubitka L kažnjava i dobro klasificirane primjere. Što su ti primjeri, jako na dobroj strani granice, dalje od granice, oni tim jače utječu na njen njezin položaj, te samim time i na klasifikaciju primjera!!

Ako npr. u prethodnom primjeru zamijenimo ((2,-3),2) s ((2,-11),2):

$h_1 = 0.3245 - 0.2243x_1 - 0.0088x_2$

$h_2 = 0.4198 + 0.2060x_1 + 0.0862x_2$

$h_3 = 0.2556 + 0.0183x_1 - 0.0777x_2$

!!! KRIVO KLASIF.!

KLASIF. $h = [1\ 1\ 2\ 2\ 2\ 3]$

$h_{12} = -0.0953 - 0.4304x_1 - 0.0953x_2 = \emptyset$

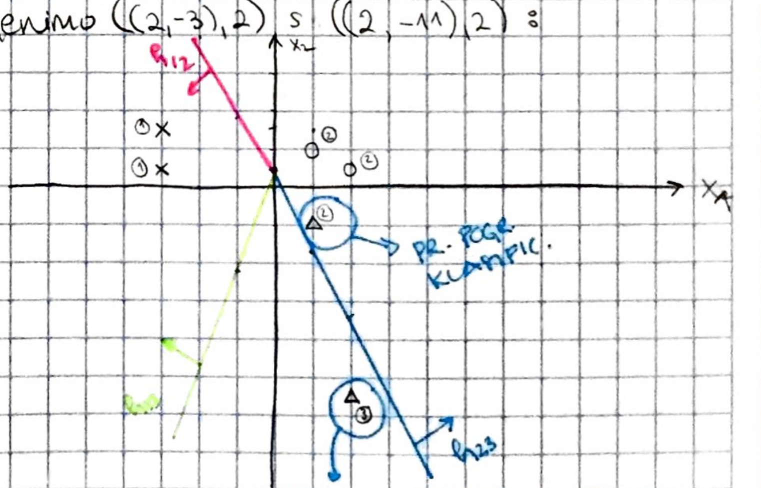
$h_{13} = 0.0689 - 0.2427x_1 + 0.0689x_2 = \emptyset$

$h_{23} = 0.1642 + 0.1877x_1 + 0.1642x_2 = \emptyset$

$x_{2,12} = -4.52x_1 - 1$

$x_{2,13} = 3.5225x_1 - 1$

$x_{2,23} = -2.72x_1 - 1$



UDALJEN PRIMJER UTJEČE NA GRANICU, PRIVLAČI JU

3. ALG. PERCEPTRONA MINIMIZIRA POGR. $E_p(\vec{w}|D)$ KOJU NAZIVAMO KRITERIJ PERCEPTRONA. TA JE FJA APROX. UDJELA POGR. KLASIF. → OČEK. GUB. 0-1, E_m.

a) ZAŠTO NE MOŽEMO IZRAVNO MINIMIZIRATI $E_m(\vec{w}|D)$?

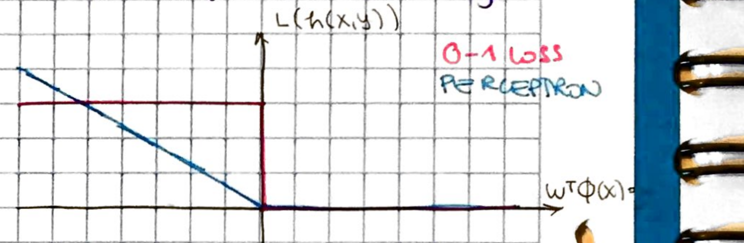
$E_m = E_m(\vec{w}|D) = \frac{1}{N}\sum \mathbb{1}\{f(w^T\phi(x^{(i)})) \neq y^{(i)}\}$ } OČEKIVANJE GUBITKA 0-1 KOJU BISMO IDEALNO HTJELI MINIMIZIRATI

NIJE DERIVABILNA

Ova funkcija je PO DIJELOVIMA KONSTANTNA te stoga ne postoji mogućnost korištenja GRADIJENTNOG SPUSTA! (Za taj alg. moramo biti u stanju izr. derivaciju pogreške)

b) DA LI JE $E_p(\vec{w}|D)$ GORNJA OGRADA ZA EMP. POGR. $E(\vec{w}|D)$? OBJASNI

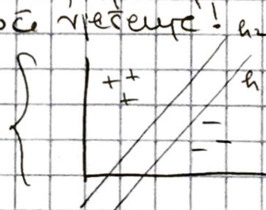
Ne nije, kao što je vidljivo i sa slike. 0-1 loss funkcija gubitka konstantnog je iznosa (1) za pogrešno klasificirane primjere, dok funkcija gubitka perceptrona linearno raste od 0, stoga postoji dio na kojem je $E_p < E_{0-1}$, samim time E_p nije gornja ograda za E.



c) JEDAN ⊖ PERC. JEST DA RJ. $\vec{w}^*$ (A SAMIM TIME I POLOŽAJ GRANICE) OVISI O POČ. TEŽINAMA I REDOSLIJEDU PREDOČAVANJA PR. POZIVAJUĆI SE NA SLIKU POVRŠINE POGREŠKE U PROSTORU PARAM., OBJASNI ZBOG ČEGA JE TO TAKO.

Perceptron će na temelju POČETNIH TEŽINA i REDOSLIJEDA PREDOČAVANJA PRIMJERA odabrati samo jednu hiperravninu od njih ∞. Zašto? Zato što se algoritam perceptrona zaustavlja čim ispravno klasificira primjere za učenje, odnosno u konačnom broju koraka za lin. odvojive primjere. U konačnici, rješenje neće nužno biti hipoteza koja najbolje klasificira (h_2), već bilo koje moguće rješenje!

Ne znamo koja će hipoteza biti izabrana kod perceptrona { h_1 lošija hipoteza jer se nalazi preblizu neg. pr., dok je h_2 bolja jer ima bolje svojstvo generalizacije.

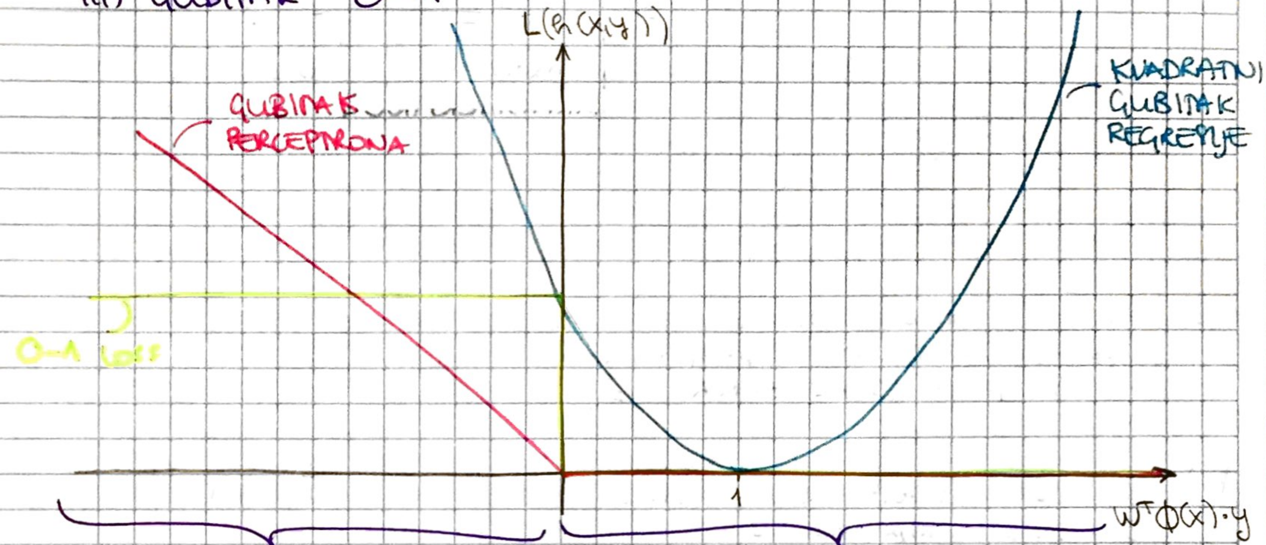


d) DRUGI NEDOSTATAK JE TAJ DA POSTUPAK NE KONVERGIRA AKO PR. NISU LINEARNO ODVOJIVI! OBJASNI ZAŠTO.

Algoritam perceptrona funkcionira tako da se postupak ponavlja dokle god svi primjeri nisu ispravno klasificirani, odnosno dok algoritam ne konvergira. Ukoliko primjeri NISU LIN. ODVOJIVI, ne postoji ispravna klasifikacija svih primjera stoga algoritam perceptrona NE KONVERGIRA!

(rez. neće biti hipoteza h s najmanjom mogućom pogreškom)

4) a) SKICIRAJ NA JEDNOM GRAFU SLJEDEĆE 3 FJE GUBITKA:
i) KVADRATNI GUBITAK REGRESIJE, ii) GUBITAK PERCEPTRONA,
iii) GUBITAK 0-1



b) → POGREŠNO KLASIFICIRANI PRIMJERI   ISPRAVNO KLASIFICIRANI PRIMJERI

c) ZAŠTO KVADRATNI GUBITAK NIJE PRIKLADAN GUBITAK U SLUČAJEVIMA KADA ŽELIMO MINIMIZIRATI BROJ POGREŠNIH KLASIFIKACIJA?

Kvadratni gubitak nije prikladan za min. broja pogr. klasif. zbog toga što JAKO KAŽNJAVA SUVIŠE TOČNO KLASIFICIRANE PRIMJERE, stoga te primjere broji kao neispravno klasificirane primjere.

d) ZA KOJE ĆE MODELE OČEKIVANJE GUBITKA BITI VEĆE OD BROJA POGR. KLASIF.?

Za klasifikaciju REGRESIJOM (kvadratni gubitak)