

3. (N) Na skupu označenih primjera \mathcal{D} trenirali smo model logističke regresije. Dobili smo neki vektor težina \mathbf{w} i pomak $w_0 = 0.15$. Tako naučenom modelu neki primjer \mathbf{x} , čija je oznaka u skupu primjera $y = 0$, nanosi gubitak unakrsne entropije od $L(0, h(\mathbf{x})) = 0.274$. **Koliki gubitak unakrsne entropije bi nanosio primjer \mathbf{x} kada bismo njegove značajke pomnožili sa dva i promijenili mu oznaku?**

- A 4.03 B 2.54 C 7.11 D 1.19

$$L(\phi, h(\mathbf{x})) = 0.274$$

$$\stackrel{\downarrow}{y} = \phi$$

$$L(y, h(\mathbf{x})) = -y \ln h(\mathbf{x}) - (1-y) \ln (1-h(\mathbf{x}))$$

$$0.274 = -\ln (1-h(\mathbf{x})) / e$$

$$e^{0.274} = -1 + h(\mathbf{x})$$

$$h(\mathbf{x}) = \frac{1}{1 + e^{-0.274}} = 0.315$$

$$h(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \phi(\mathbf{x}))}$$

$$0.315 = \frac{1}{1 + \exp(-\mathbf{w}^\top \phi(\mathbf{x}))} \quad -\mathbf{w}^\top \phi(\mathbf{x}) = 0.5667$$

$$0.315(1 + \exp(-\mathbf{w}^\top \phi(\mathbf{x})) = 1$$

$$\exp(-\mathbf{w}^\top \phi(\mathbf{x})) = \frac{1 - 0.315}{0.315} = \frac{0.685}{0.315} = 2.155$$

$$\begin{array}{r} \text{bez } w_0 \\ \hline 0.5667 \\ -0.15 \\ \hline 0.4167x^2 \\ \hline 0.8374 \\ +0.15 \\ \hline 0.9874 \end{array}$$

$$h(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \phi(\mathbf{x}))} = 0.2714$$

$$\underline{L}(1, h(\mathbf{x})) = -\ln h(\mathbf{x}) = 1.31$$

5. (N) Model logističke regresije treniramo stohastičkim gradijentnim spustom. Primjere iz dvodimenziskog ulaznog prostora preslikali smo u prostor značajki funkcijom

$$\phi(\mathbf{x}) = (1, x_1, x_2, x_1x_2)$$

U jednoj iteraciji treniranja modela vektor parametara jednak je

$$\mathbf{w} = (0.2, 0.5, -1.1, 2.7)$$

Koliko u toj iteraciji iznosi L_2 -norma gradijenta gubitka za primjer $(\mathbf{x}, y) = ((-0.5, 2), 1)$?

- A 0.70 B 2.48 C 1.28 D 4.00

$$h(x) = \frac{1}{1+e^{-w^T\phi(x)}}$$

$$\nabla_w L(y, h(x)) = \left(-\frac{y}{h(x)} + \frac{1-y}{1-h(x)} \right) h(x) (1-h(x))^2 \phi(x)$$

$$= (h(x) - y) \phi(x)$$

$$= (w^T \phi(x) - y) \phi(x)$$

$$= \left([0, 2, 0, 5, -1, 1, 2, 7] \begin{bmatrix} 1 \\ -0,5 \\ 2 \\ -1 \\ 1,1 \end{bmatrix} - 1 \right) \phi(x)$$

0,25

2,2

2,7

1

$$\underline{\underline{6,15}}$$

$$\underline{\underline{-0,2}}$$

$$= (0,2 - 0,25 - 2,2 - 2,7 - 1) \phi(x)$$

$$= -5,95 [1 \ -0,5 \ 2 \ -1]$$

$$= [-5,95 + 2,975 - 1,9 + 5,95]$$

$$\|w\|_2 = \sqrt{w^T w} = 14,875$$

3. (N) Raspolažemo sljedećim skupom primjera u dvodimenziskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((1, 0), 1), ((2, -3), 2), ((3, 5), -1), ((5, 0), -4)\}$$

Na ovom skupu gradijentnim spustom trenirali smo L_1 -regularizirani model linearne regresije sa $\lambda = 1$. Dobili smo težine $\mathbf{w} = (2.12, -0.94, -0.08)$. **Koliko iznosi L_1 -regularizirana pogreška $E(\mathbf{w}|\mathcal{D})$?**

- A 7.10 B 2.69 C 1.58 D 0.29

$$E_R = \frac{1}{2} \sum_{i=1}^m (\mathbf{w}^\top \phi(\mathbf{x}) - y^{(i)})^2 + \frac{\lambda}{2} \|\mathbf{w}\|_1$$

$$\|\mathbf{w}\|_1 = \sum_{i=1}^m |w_i| = 0,94 + 0,08 = 1,02$$

$$E_R = \frac{1}{2} \left(\begin{bmatrix} 2,12 & -0,94 & -0,08 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - 1 \right)^2$$

$$+ \left(\begin{bmatrix} 2,12 & -0,94 & -0,08 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix} - 2 \right)^2$$

$$+ \left(\begin{bmatrix} 2,12 & -0,94 & -0,08 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} + 1 \right)^2$$

$$+ \left(\begin{bmatrix} 2,12 & -0,94 & -0,08 \end{bmatrix} \begin{bmatrix} 1 \\ 5 \\ 0 \end{bmatrix} + 4 \right)^2 + 0,51$$

$$= \frac{1}{2} (0,0394 + 2,3104 + 0,01 + 2,0164) + 0,51$$

$$= 2,6964$$

Osnovni koncepti

1. (T) Model \mathcal{H} je skup svih parametriziranih funkcija $h(\mathbf{x}; \boldsymbol{\theta})$ indeksiran parametrima $\boldsymbol{\theta}$. To jest:

$$\mathcal{H} = \underbrace{\{h(\mathbf{x}; \boldsymbol{\theta})\}}_{\boldsymbol{\theta}}$$

Što to zapravo znači?

- A Da različite funkcije h imaju različite parametre $\boldsymbol{\theta}$, da su sve one sadržane u modelu, to jest za sve njih vrijedi $h \in \mathcal{H}$
- B Da za različite parametre $\boldsymbol{\theta}$ dobivamo različite funkcije h , i da su sve one sadržane u modelu, to jest za sve njih vrijedi $h \in \mathcal{H}$
- C Da model sadrži beskonačno mnogo funkcija h čija konkretna definicija ovisi o vrijednostima parametara $\boldsymbol{\theta}$
- D Da su funkcije h definirane sa slobodnim parametrima $\boldsymbol{\theta}$ i da broj različitih funkcija odgovara broju parametara

2. (P) U ulaznom prostoru $\mathcal{X} = \{0, 1\}^3$ definiramo sljedeći klasifikacijski model:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{1}\{\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \geq 0\}$$

Koja je dimenzija prostora parametara te koliko različitih hipoteza postoji u ovom modelu?

- A Dimenzija prostora parametara je 4, a hipoteza ima beskonačno mnogo
- B Dimenzija prostora parametara je 4, a hipoteza ima manje od 256
- C Dimenzija prostora parametara i broj hipoteza su beskonačni
- D Dimenzija prostora parametara je 256, a hipoteza ima 14

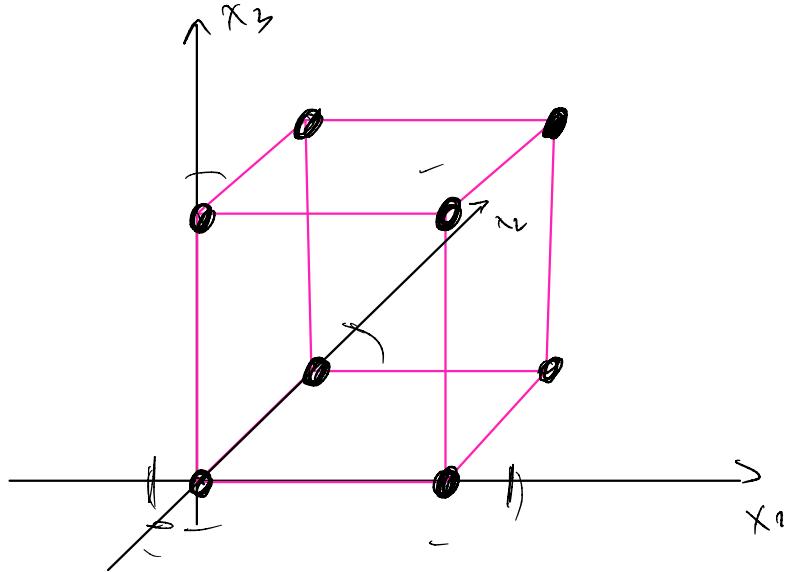
$$\theta_{0-3} \Rightarrow 3+1 = 4 \text{ dimenzija}$$
$$4 \cdot 4 \cdot 4 \cdot 4 = 256$$

3. (P) Za ulazni prostor $\mathcal{X} = \{0, 1\}^3$ definiramo klasifikacijski model \mathcal{H} kao skup parametriziranih funkcija definiranih na sljedeći način:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{1}\{(\theta_{1,1} \leq x_1 \leq \theta_{1,2}) \wedge (\theta_{2,1} \leq x_2 \leq \theta_{2,2}) \wedge (\theta_{3,1} \leq x_3 \leq \theta_{3,2})\}$$

Parametri su trodimenzijski vektori realnih brojeva, tj. prostor parametara definiran je kao $\boldsymbol{\theta} \in \mathbb{R}^6$.
Koliko iznosi $|\mathcal{H}|$?

- A 42 B ∞ C 56 D 28



1 vršak
1 vršak
8 vrhova
12 stranica
6 ploha

2 B

4. (P) Skup označenih primjera u dvodimensijskome ulaznom prostoru je:

$$\mathcal{D} = \{((0, 0), 0), ((0, 1), 0), ((1, 1), 1)\}$$

Koja je veličina prostora inačica, $|VS_{\mathcal{H}, \mathcal{D}}|$?

- A 16 B Pitanje nema smisla jer nije definiran model C Beskonačno mnogo D 14

5. (P) Za linearan klasifikator u $\mathcal{X} = \{0, 1\}^3$ zadan je sljedeći skup primjera za učenje:

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

Razmatramo dva modela:

$$\mathcal{H}_a : h_a(\mathbf{x}|\boldsymbol{\theta}) = \mathbf{1}\{\theta_0 + x_1\theta_1 + x_2\theta_2 + x_3\theta_3 \geq 0\}$$

$$\mathcal{H}_b : h_b(\mathbf{x}|\boldsymbol{\theta}) = h_a(\mathbf{x}; \boldsymbol{\theta}_1) \cdot h_a(\mathbf{x}; \boldsymbol{\theta}_2)$$

bilo i bilo ili nevačka

Uočite da svaka hipoteza iz modela \mathcal{H}_b kombinira dvije hipoteze iz modela \mathcal{H}_a (operacijom množenja). Neka:

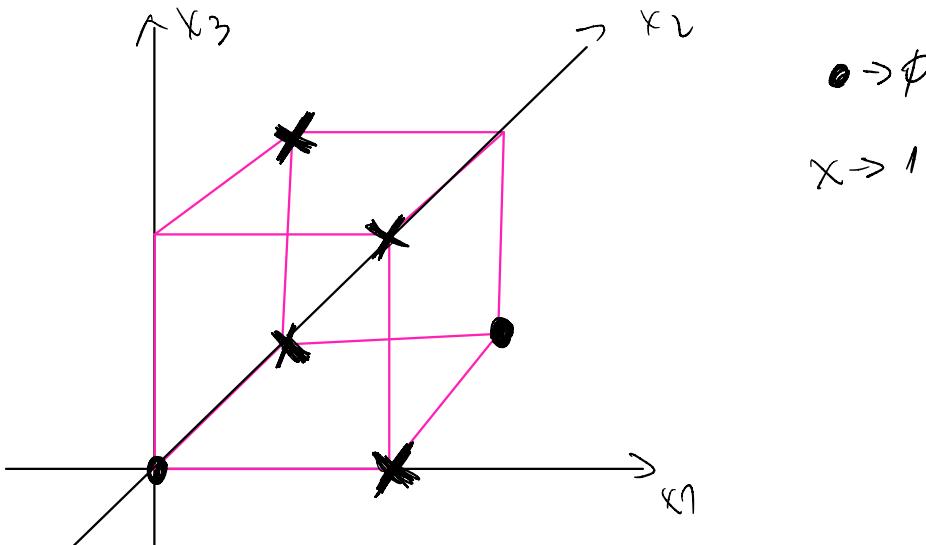
$$h_a^* = \operatorname{argmin}_{h \in \mathcal{H}_a} E(h|\mathcal{D})$$

$$h_b^* = \operatorname{argmin}_{h \in \mathcal{H}_b} E(h|\mathcal{D})$$

Koja je od navedenih tvrdnji točna?

- [A] $E(h_a^*|\mathcal{D}) \neq E(h_b^*|\mathcal{D}) > 0$
- [B] $E(h_a^*|\mathcal{D}) > E(h_b^*|\mathcal{D}) = 0$
- [C] $0 < (E(h_a^*|\mathcal{D}) - E(h_b^*|\mathcal{D})) < 1$
- [D] $E(h_a^*|\mathcal{D}) \neq E(h_b^*|\mathcal{D}) \neq 0$

*$E_{H_a} > \emptyset$
nije lin. oduzim*

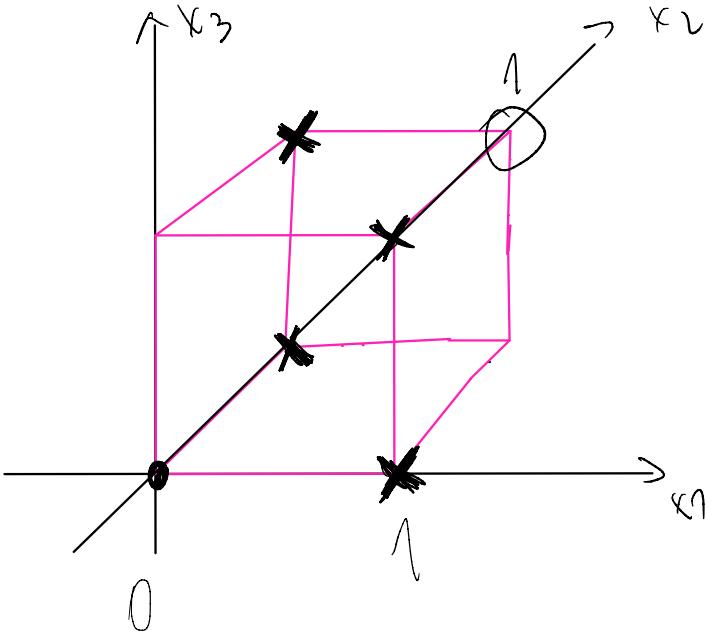


6. (P) Za linearan model u $\mathcal{X} = \{0, 1\}^3$ zadan je sljedeći skup primjera za učenje:

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

Optimizacijski postupak klasifikatora funkcioniра tako da minimizira empirijsku pogrešku, definiranu kao očekivanje funkcije gubitka 0-1, i postupak u tome uvijek uspijeva. Želimo znati koju bi klasu ovaj klasifikator dodijelio primjeru $\mathbf{x} = (1, 1, 1)$. Možemo li, na temelju iznesenih informacija, odrediti klasifikaciju dotičnog primjera i što nam to govori o induktivnoj pristranosti ovog algoritma?

- A Ne možemo, jer nije definirana induktivna pristranost preferencijom, pa činjenica da je model linearan nije dovoljan skup pretpostavki da bismo jednoznačno odredili klasifikaciju svih novih primjera
- B Možemo, klasifikacija je $y = 1$, i ovaj klasifikator ima definiranu induktivnu pristranost pomoću koje može jednoznačno odrediti klasifikaciju svakog primjera
- C Možemo, klasifikacija je $y = 1$, premda dane informacije nisu dovoljne za definiciju induktivne pristranosti, što se vidi iz toga da za ovaj skup primjera prostor inačica veći od jedan
- D Možemo, $y = 1$, jer klasifikator ima induktivnu pristranost jezikom (linearan model) i preferencijom (primjeri za koje je $h(x) \geq 0$ klasificiraju se pozitivno)



7. (P) Optimizacija parametara modela temelji se na funkciji gubitka $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_0^+$, gdje je $L(y, h(\mathbf{x}))$ gubitak na primjeru (\mathbf{x}, y) . U većini primjena koristimo simetričan gubitak 0-1. Međutim, u nekim primjenama ima više smisla definirati asimetričan gubitak. Jedan takav primjer je zadatak detekcije karcinoma iz medicinskih slika. Taj zadatak možemo formalizirati kao problem binarne klasifikacije s oznakama $\mathcal{Y} = \{0, 1\}$, gdje $y = 1$ označava postojanje karcinoma, a $y = 0$ nepostojanje karcinoma. **Koje od sljedećih svojstava bi trebala zadovoljiti asimetrična funkcija gubitka za takav zadatak?**

- A $L(0, 1) = 1$ i $L(1, 0) = L(1, 1) = L(0, 0) = 0$
- B $L(0, 1) > L(1, 0)$ i $L(1, 1) = L(0, 0) > 0$
- C $L(1, 0) > L(0, 1)$ i $L(1, 1) = L(0, 0) = 0$
- D $L(0, 1) = L(1, 0) > 0$ i $L(1, 1) = L(0, 0) = 0$

8. (T) Pogreška modela definirana je kao očekivanje funkcije gubitka na primjerima iz $\mathcal{X} \times \mathcal{Y}$. Međutim, u praksi tu pogrešku aproksimiramo empirijskom pogreškom, koju računamo kao srednju vrijednost funkcije gubitka na skupu označenih primjera $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$. Zašto pogrešku modela aproksimiramo empirijskom pogreškom i na kojoj se prepostavci temelji ta aproksimacija?

- A Očekivanje gubitka ne možemo izračunati jer primjera iz $\mathcal{X} \times \mathcal{Y}$ ima potencijalno beskonačno, stoga pogrešku računamo na temelju skupa \mathcal{D} za koji prepostavljamo da je konačan
- B Različitih primjera iz $\mathcal{X} \times \mathcal{Y}$ potencijalno ima beskonačno mnogo, pa pogrešku računamo na uzorku \mathcal{D} za koji prepostavljamo da je reprezentativan
- C Funkciju gubitka jednostavnije je definirati nego funkciju pogreške, a aproksimacija je točna uz pretpostavku i.i.d.
- D Ne možemo izračunati očekivanje gubitka jer nam nije poznata distribucija primjera iz $\mathcal{X} \times \mathcal{Y}$, no prepostavljamo da je \mathcal{D} reprezentativan uzorak iz te distribucije

9. (P) Zadan je sljedeći skup sa $N = 6$ označenih primjera iz \mathbb{R}^3 :

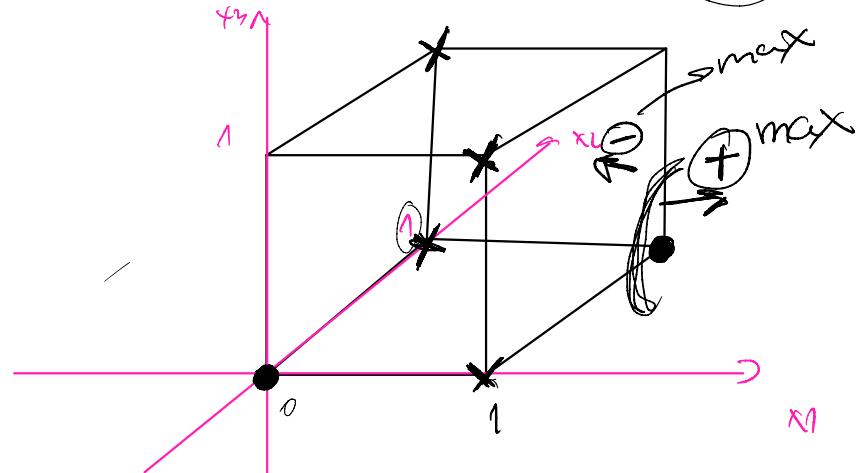
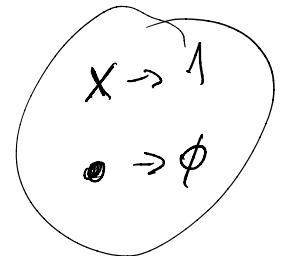
$$\begin{aligned}\mathcal{D} &= \{(\mathbf{x}^{(i)}, y^{(i)})\} \\ &= \{((0, 0, 0), 0), ((1, 1, 0), 0), ((1, 0, 0), 1), ((1, 0, 1), 1), ((0, 1, 0), 1), ((0, 1, 1), 1)\}\end{aligned}$$

Razmatramo linearan model i računamo empirijsku pogrešku $E(h|\mathcal{D})$ hipoteza iz tog modela definiranu kao očekivanje asimetričnog gubitka. Gubitak je definiran tako da lažno negativne primjere kažnjava sa 1, a lažno pozitivne primjere sa 0.5. **Koliko iznosi najmanja a koliko najveća moguća vrijednost tako definirane empirijske pogreške $E(h|\mathcal{D})$?**

- A $0 \leq E(h|\mathcal{D}) \leq 1/4$
- B $1/4 \leq E(h|\mathcal{D}) \leq 2/3$
- C $\frac{1}{48} \leq E(h|\mathcal{D}) \leq 2/3$
- D $1/12 \leq E(h|\mathcal{D}) \leq 3/4$

$$FP \Rightarrow 0,5$$

$$FN \Rightarrow 1$$

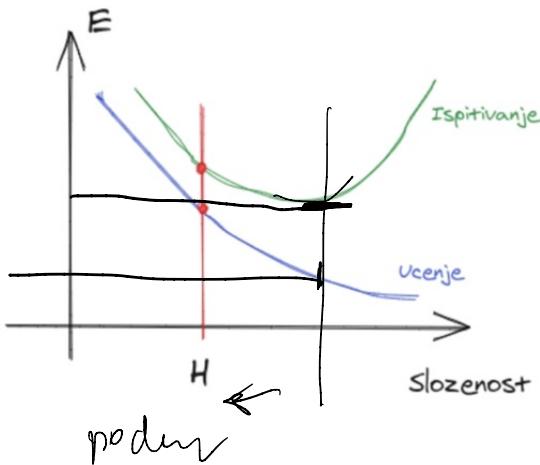


$$\text{max} = \frac{4 + 0,5}{6} = \frac{\frac{8+1}{2}}{6} = \frac{3}{4}$$

domet \downarrow

$$\text{min} = \frac{0,5}{6} = \frac{1}{12}$$

10. (P) Na slici ispod prikazan je graf funkcije pogreške učenje i pogreške ispitivanja za neku familiju modela i neki označeni skup primjera:



Crvenom linijom označena je složenost nekog modela \mathcal{H} . Crvene točke odgovaraju ispitnoj empirijskoj pogrešci i pogrešci učenja za hipotezu $h \in \mathcal{H}$ iz tog modela, dobivenoj nekim optimizacijskim algoritmom. **Što možemo reći o modelu \mathcal{H} i o hipotezi h ?**

- A Model \mathcal{H} nije optimalne složenosti, a čak ni hipoteza h ne mora biti optimalna na skupu za učenje, ako je optimizacijski algoritam loš
- B Model H je podnaučen, ali je barem hipoteza h hipoteza s najmanjom ispitnom pogreškom unutar takvog suboptimalnog modela
- C Model \mathcal{H} je nedovoljne složenosti, ali je barem hipoteza h optimalna u smislu najmanje moguće pogreške na skupu za učenje
- D Model \mathcal{H} je prenaučen, a hipoteza h će loše generalizirati na neviđene primjere

11. (T) Modeli strojnog učenja tipično imaju i parametre i hiperparametre. **Koja je razlika između parametara i hiperparametara?**

- A Algoritam strojnog učenja minimizira parametre te istovremeno maksimizira hiperparametre
- B Hiperparametri mogu biti diskretni ili kontinuirani, dok su parametri uvijek kontinuirani
- C Parametre optimira algoritam strojnog učenja, dok optimizacija hiperparametara nije u nadležnosti tog algoritma
- D Parametri određuju iznos empirijske pogreške na skupu za učenje, a hiperparametri iznos te pogreške na skupu za provjeru

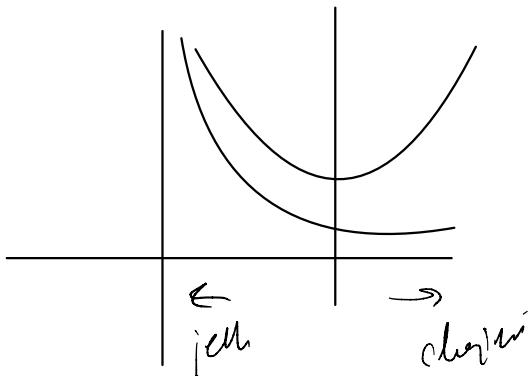
12. (P) Raspolažemo modelom \mathcal{H}_α koji ima hiperparametar α kojim se može ugađati složenost modela. Isprobavamo dvije vrijednosti hiperparametra: α_1 i α_2 . Treniramo modele \mathcal{H}_{α_1} i \mathcal{H}_{α_2} te dobivamo hipoteze h_{α_1} i h_{α_2} . Zatim računamo empirijske pogreške tih hipoteza na skupu za učenje \mathcal{D}_u i na skupu za ispitivanje \mathcal{D}_i . Utvrđujemo da vrijedi:

$$E(h_{\alpha_1} | \mathcal{D}_i) - E(h_{\alpha_1} | \mathcal{D}_u) < E(h_{\alpha_2} | \mathcal{D}_i) - E(h_{\alpha_2} | \mathcal{D}_u)$$

Što iz toga možemo zaključiti?

↳ (odgovor)

- A Model \mathcal{H}_{α_2} je prenaučen
- B Optimalan model je onaj s hiperparametrom iz intervala $[\alpha_1, \alpha_2]$
- C Model \mathcal{H}_{α_1} je podnaučen
- D Model \mathcal{H}_{α_1} je manje složenosti od modela \mathcal{H}_{α_2}



Regresija

1. (P) Pogreška najmanjih kvadrata definirana je kao:

$$E(\mathbf{w}|\mathcal{D}) = \frac{1}{2} \sum_{i=1}^N (\mathbf{w}^T \mathbf{x}^{(i)} - y^{(i)})^2$$

Izvedite matrični zapis ove funkcije. **Kako glasi matrični zapis ove funkcije, nakon sređivanja izraza, a prije deriviranja?**

- A $\frac{1}{2}(\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{y}^T \mathbf{X} + \mathbf{w}^T \mathbf{w})$
- B $\frac{1}{2}(\mathbf{w} \mathbf{X}^T \mathbf{X} \mathbf{w}^T - 2\mathbf{y}^T \mathbf{w} + \mathbf{y}^T \mathbf{y})$
- C $\frac{1}{2}(\mathbf{w}^T \mathbf{X}^T - 2\mathbf{w}^T \mathbf{X} \mathbf{y} + \mathbf{y}^T \mathbf{y})$
- D $\frac{1}{2}(\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{y}^T \mathbf{X} \mathbf{w} + \mathbf{y}^T \mathbf{y})$

$$\begin{aligned} E(\mathbf{w}|\mathcal{D}) &= \frac{1}{2} \sum_{i=1}^N (\mathbf{w}^T \mathbf{x}^{(i)} - y^{(i)})^2 \\ &= \frac{1}{2} \sum_{i=1}^N (\mathbf{w}^T \mathbf{x}^{(i)})^2 - 2 \sum_{i=1}^N (\mathbf{w}^T \mathbf{x}^{(i)}) \cdot y^{(i)} + \sum_{i=1}^N (y^{(i)})^2 \\ &\quad \text{(circled)} \quad \text{(circled)} \quad \text{(circled)} \\ &= \frac{1}{2} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{w}^T \mathbf{X} \mathbf{y} + \mathbf{y}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} &(\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w})^T - 2(\mathbf{y}^T \mathbf{X} \mathbf{w})^T + \mathbf{y}^T \mathbf{y} \\ &= \mathbf{y}^T \mathbf{X} \mathbf{w} \end{aligned}$$

2. (T) Rješenje najmanjih kvadrata za vektor težina \mathbf{w} jest:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Pod kojim uvjetima ćemo težine moći izračunati na ovaj način, i o čemu dominantno ovisi složenost tog postupka?

- A Ako je matrica $\mathbf{X}^T \mathbf{X}$ kvadratna i punog ranga, a složenost izračuna dominantno ovisi o N
- B Ako je rang matrice \mathbf{X} jednak $N + 1$, a složenost izračuna dominantno ovisi o N
- C Ako je rang matrice $\mathbf{X}^T \mathbf{X}$ jednak N , a složenost izračuna dominantno ovisi o n
- D Ako je rang matrice \mathbf{X} jednak $n + 1$, a složenost izračuna dominantno ovisi o n

3. (P) Razmatramo model jednostavne regresije:

$$h(x; w_0, w_1) = w_0 + w_1 x$$

Model linearne regresije inače koristi funkciju kvadratnog gubitka:

$$L(y, h(\mathbf{x})) = (y - h(\mathbf{x}))^2$$

Međutim, u našoj implementaciji greškom smo funkciju gubitka definirali ovako:

$$L(y, h(\mathbf{x})) = (y + h(\mathbf{x}))^2$$

S tako pogrešno definiranom funkcijom gubitka postupkom najmanjih kvadrata treniramo naš model na skupu primjera čije su oznake uzorkovane iz distribucije $\mathcal{N}(-1 + 2x, \sigma^2)$, gdje je varijanca σ^2 razmjerne malena (tj. nema mnogo šuma). **Koji vektor težina (w_0, w_1) očekujemo približno dobiti kao rezultat najmanjih kvadrata?**

- A (1, -2) B (2, -1) C (-1, 2) D (0, 0)

$$h(x) = -1 + 2x$$

$$\min L(y, h(x))$$

$$\begin{aligned} L(y, h(x)) &= (y - (-h(x))) \\ &= (y - (-1 - 2x)) \end{aligned}$$

$$\min L(y, h(x)) = (y - (-h(x))^2$$

$$1 - 2x$$

3. (P) Razmatramo model jednostavne regresije:

$$h(x; w_0, w_1) = w_0 + w_1 x$$

Model linearne regresije inače koristi funkciju kvadratnog gubitka:

$$L(y, h(\mathbf{x})) = (y - h(\mathbf{x}))^2$$

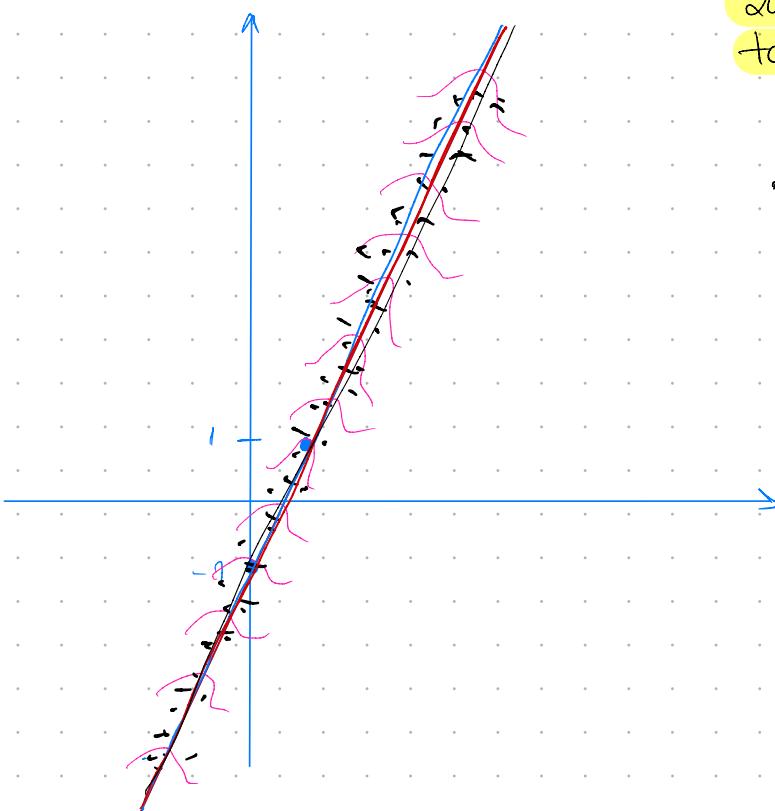
Međutim, u našoj implementaciji greškom smo funkciju gubitka definirali ovako:

$$L(y, h(\mathbf{x})) = (y + h(\mathbf{x}))^2$$

S tako pogrešno definiranom funkcijom gubitka postupkom najmanjih kvadrata treniramo naš model na skupu primjera čije su oznake uzorkovane iz distribucije $\mathcal{N}(-1 + 2x, \sigma^2)$, gdje je varijanca σ^2 razmjerne malena (tj. nema mnogo šuma). **Koji vektor težina (w_0, w_1) očekujemo približno dobiti kao rezultat najmanjih kvadrata?**

- A (1, -2) B (2, -1) C (-1, 2) D (0, 0)

$$W(-1 + 2x, \sigma^2)$$



zanimaju nas one težine w koje će nam dati
talušu hipotezu koja će primjere učiniti najviše
idealno: svaku bi prošao kroz vrh Gaussa

• želimo pronaći:

$$h(\vec{x}; \vec{w}) = j(\vec{x}) \rightarrow \text{pretp. podaci koje imamo su najvjerojatniji}$$

$$\Rightarrow -h(x) = \mu \quad \text{za } (1, -2)$$

$$L(y, h(x)) = (y + h(x))^2$$

$$\text{kvadrat reziduala} = (y - (-h(x)))^2$$

teže smješeno
g odsupnjuće

gleda udalj od $-h(x)$

$$\begin{pmatrix} (w_0, w_1) \\ (1, -2) \\ (-1, 2) \\ (2, -1) \end{pmatrix}$$

$$h(x) = -w_0 - w_1 x$$

$$-h(x) = -1 + 2x \rightarrow \text{isto je isto } W$$

4. (T) Model linearne regresije je poopćeni linearni model i ima probabilističku interpretaciju. Prijetite se, tu smo interpretaciju upotrijebili smo kako bismo opravdali empirijska funkcija pogreške definiranu na temelju kvadratnog gubitka. **Kako formalno glasi probabilistička pretpostavka modela linearne regresije?**

A $p(\mathbf{x}|y) = \mathcal{N}(f(\mathbf{x}), \sigma^2)$

B $p(y|\mathbf{x}) = \mathcal{N}(0, \sigma^2)$

C $p(y|\mathbf{x}) = \mathcal{N}(h(\mathbf{x}), \sigma^2)$

D $p(y) = \mathcal{N}(0, \sigma^2)$

$$y = h(x)$$

5. (T) Svaki algoritam strojnog učenja ima neku induktivnu pristranost. Induktivna pristranost sastoji se od pristranosti jezika i pristranosti preferencije. **Kako glasi induktivna pristranost preferencije (neregulariziranog) modela linearne regresije?**

- A Težine \mathbf{w} maksimiziraju iznos $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ \times
- B Hipoteza h je linearna kombinacija težina \mathbf{w} i značajki \mathbf{x} *nije pristranost*
- C Težine \mathbf{w} minimiziraju iznos $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$
- D Hipoteza h je funkcija iz \mathbb{R}^n u \mathbb{R} *nije pristranost*

Regresija II.

1. (T) Kao regularizacijski faktor kod modela linearne regresije tipično se koristi neka p-norma vektora težina, $\|\mathbf{w}\|_p$. Na kojoj se činjenici temelji korištenje norme kao regularizacijskog izraza?

- A Ako je model prenaučen, hipoteza će imati velike magnitude težina
- B Ako je model optimalne složenosti, hipoteza će imati male magnitude težina
- C Ako su težine hipoteze velike magnitude, model je prenaučen
- D Ako su težine hipoteze male magnitude, model je podnaučen

2. (T) L_1 -regularizacija ili LASSO kao regularizacijski izraz koristi prvu normu vektora težina, $\|\mathbf{w}\|_1$.
Što je prednost a što nedostatak L_1 -regularizacije?

- A Prednost je da zadržava sve značajke u modelu, a nedostatak je da Gramova matrica može biti blizu singularne ako u podatcima postoji multikolinearnost
 - B Prednost je da L_1 -regulariziranu pogrešku možemo minimizirati gradijentnim spustom, a nedostatak je da rezultira rijetkim modelima
 - C Prednost je da postoji rješenje u zatvorenoj formi (pseudoinverz), a nedostatak da izračun L_1 -regulariziranog pseudoinverza ovisi o broju značajki ali i o broju primjera
 - D Prednost je da izbacuje značajke iz modela, a nedostatak je da L_1 -regularizirana pogreška nema minimizator u zatvorenoj formi
-

3. (N) Raspolažemo sljedećim skupom primjera u dvodimenzijskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((1, 0), 1), ((2, -3), 2), ((3, 5), -1), ((5, 0), -4)\}$$

Na ovom skupu gradijentnim spustom trenirali smo L_1 -regularizirani model linearne regresije sa $\lambda = 1$. Dobili smo težine $\mathbf{w} = (2.12, -0.94, -0.08)$. Koliko iznosi L_1 -regularizirana pogreška $E(\mathbf{w}|\mathcal{D})$?

- A 7.10 B 2.69 C 1.58 D 0.29

$$\Phi[\mathbf{x}] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & -3 \\ 1 & 3 & 5 \\ 1 & 5 & 0 \end{bmatrix}$$

$$E_{\mathbf{w}}(\mathbf{w}|D) = E(\mathbf{w}|D) + \frac{\lambda}{2} \|\mathbf{w}\|_1$$

$\nabla_{\mathbf{w}} L \quad (h(x) - y) \Phi(x)$

$$h(x) = \mathbf{w}^T \mathbf{x} = [2, 12, -0.94, -0.08] \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 5 \\ 1 & 3 & 5 & 0 \\ 1 & 5 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 3 \\ 4 \end{bmatrix}$$

$$= [1, 18, 0.48, -1.1, -2.58]$$

$$\sum \nabla_{\mathbf{w}} L = ([1, 18, 0.48, -1.1, -2.58] - [1, 2, -1, -4]). \Phi(x)$$

$$= [0.18, -1.52, 0.1, 1.42] \underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & -3 \\ 1 & 3 & 5 \\ 1 & 5 & 0 \end{bmatrix}}$$

$$= [0.18, 4.54, 4.06]$$

$$E(w|D) = \frac{1}{2} \sum_{i=0}^n \nabla_w L = \frac{\lambda}{2} \|w\|_1 - 0.5$$

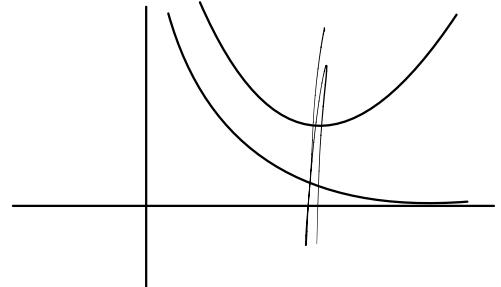
$$\epsilon_{b-1} =$$

4. (T) Svaki algoritam strojnog učenja ima neku induktivnu pristranost. Induktivna pristranost sastoji se od pristranosti jezika i pristranosti preferencije. **Koja je razlika između induktivnih pristranosti regularizirane i neregularizirane linearne regresije?**

- A Oba algoritma imaju isti model, definiran kao linearu kombinaciju značajki i težina, pa dakle imaju istu pristranost jezika, ali se razlikuju u pristranosti preferencije jer imaju različito definiranu empirijsku pogrešku (osim ako je regularizacijski faktor jednak nuli)
- B Algoritmi imaju različite pristranosti, i to različitu pristranost preferencije jer regularizirana regresija preferira jednostavnije hipoteze, a onda i različitu pristranost jezika jer je model neregularizirane regresije nadskup modela regularizirane regresije
- C Za razliku od neregularizirane regresije, regularizirana regresija preferira jednostavnije hipoteze, međutim pristranosti su im identične jer su oba algoritma definirana kao linearna kombinacija značajki i težina te oba koriste identičan optimizacijski postupak (pseudo-inverz matrice dizajna)
- D Algoritmi se ne razlikuju po pristranosti preferencijom budući da koriste istu funkciju gubitka (kvadratni gubitak), međutim regularizirana regresija ima jaču induktivnu pristranost jezika od regularizirane regresije budući da prvi model uključuje drugi model

5. (P) Raspolažemo skupom označenih primjera $\mathcal{D} \subset \mathbb{R}^n \times \mathbb{R}$ koji su u stvarnosti generirani funkcijom koja je polinom trećeg stupnja. Podataka imamo razmjerno malo, a šum u podatcima je velik. Skup \mathcal{D} dijelimo na skup za učenje i skup za ispitivanje. Neka je $\mathcal{H}_{d,\lambda}$ familija modela polinomialne regresije stupnja d s L2-regularizacijskim faktorom λ . Na skupu za učenje postupkom najmanjih kvadrata treniramo četiri modela iz te familije: $\mathcal{H}_{2,0}$, $\mathcal{H}_{5,0}$, $\mathcal{H}_{5,100}$ i $\mathcal{H}_{5,1000}$. Zatim izračunavamo empirijsku pogrešku (očekivanje kvadratnog gubitka) ovih modela na skupu za ispitivanje. **Što možemo zaključiti o ponašanju hipoteza naučenih iz ovih modela na skupu primjera \mathcal{D} ?**

- A Najbolje će generalizirati hipoteza iz $\mathcal{H}_{5,100}$ ili hipoteza iz $\mathcal{H}_{5,1000}$, ovisno o količini šuma u podatcima
- B Hipoteza iz $\mathcal{H}_{2,0}$ imati će veću pogrešku na skupu za učenje od hipoteze $\mathcal{H}_{5,0}$, ali mogu podjednako loše generalizirati
- C Hipoteza iz $\mathcal{H}_{5,1000}$ će generalizirati bolje od hipoteze iz $\mathcal{H}_{5,0}$, ali će imati veću pogrešku na skupu za učenje
- D Hipoteza iz $\mathcal{H}_{5,100}$ će bolje generalizirati od hipoteze iz $\mathcal{H}_{2,0}$ i imat će manju pogrešku na skupu za učenje



6. (T) Rješenje najmanjih kvadrata s L2-regularizacijom (hrbatna regresija) je:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

gdje $\lambda \mathbf{I} = \text{diag}(0, \lambda, \dots, \lambda)$. **Koji je efekt regularizacije na Gramovu matricu?**

- A Dodavanje vrijednosti λ na dijagonale Gramove matrice povećava normu težina $\|\mathbf{w}\|$ ✗
- B Minimizacija norme težina $\|\mathbf{w}\|$ povećava multikolinearnost Gramove matrice i smanjuje složenost modela ✗
- C Minimizacija norme težina $\|\mathbf{w}\|$ čini Gramovu matricu kvadratnom i singularnom
- D Dodavanje vrijednosti λ na dijagonale Gramove matrice povećava njezin rang

7. (P) Koristimo regresiju za predviđanje uspjeha na studiju. Kao značajke možemo koristiti ocjene u četiri razreda srednje škole (značajke x_1-x_4), prosjek ocjena sva četiri razreda (x_5) te uspjeh iz matematike (x_6) i fizike (x_7) na državnoj maturi (ukupno 7 značajki). Ne moramo iskoristiti sve značajke, ali ih želimo iskoristiti što više. Za preslikavanje u prostor značajki koristimo preslikavanje s kvadratnim, interakcijskim i linearnim značajkama. Od interakcijskih značajki uzimamo samo interakcije parova značajki (npr. x_1x_2) i interakcije trojki (npr. $x_1x_2x_3$) između svih značajki koje koristimo. **Koliko minimalno primjera za učenje trebamo imati, a da bi rješenje bilo stabilno i bez regularizacije?**

- A 75 B 38 C 48 D 63

$x_5 \rightarrow$ nepotrebno

broj. $\rightarrow 6$

$$\text{ku. : } 6 \quad \binom{6}{2} = \frac{3}{2} \cdot 5 = 15$$

dejlike

$$\text{trojke} \quad \binom{6}{3} = \frac{2}{2} \cdot 5 \cdot 4 = 20$$

+ dummy

48

Linearni diskriminativni modeli

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 predavanjima smo pokazali kako se linearan model regresije može (pokušati) koristiti za klasifikaciju. Pokažite to na sljedećim primjerima iz triju ($K = 3$) klasa:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^6$$

$$= \{((-3, 1), 0), ((-3, 3), 0), ((1, 2), 1), ((2, 1), 1), ((1, -2), 2), ((2, -3), 2)\}.$$

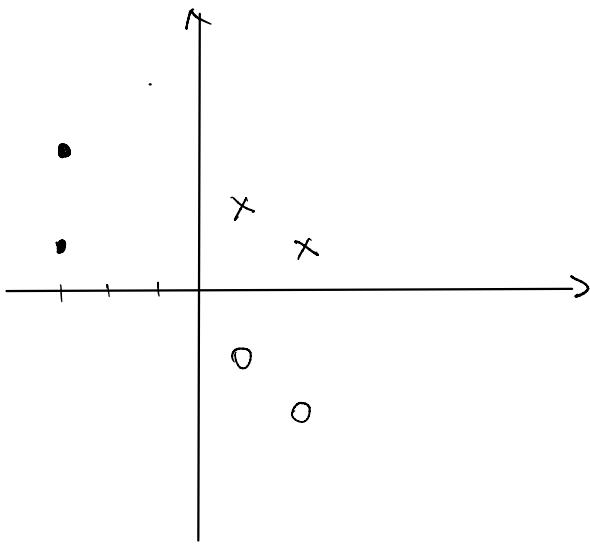
- (a) Primijenite pristup *jedan-naspram-ostali* (OVR), definirajte matricu dizajna i vektor oznaka \mathbf{y} za svaki od triju modela te izračunajte hipoteze $h_j(\mathbf{x})$ za svaku od triju klasa. Izračun možete napraviti ručno ili u nekom alatu.
- (b) Izračunajte diskriminacijske funkcije $h_{01}(\mathbf{x})$, $h_{12}(\mathbf{x})$ i $h_{02}(\mathbf{x})$ između parova susjednih klasa. Skicirajte primjere i dobivene granice u prostoru \mathbb{R}^2 .
- (c) U koju bi klasu bio klasificiran primjer $\mathbf{x} = (-1, 3)$? Obrazložite odgovor.
- (d) Možete li reći koja je vjerojatnost da primjer pripada toj klasi? Obrazložite odgovor.
- (e) Objasnite koja je prednost pristupa OVR nad pristupom *jedan-naspram-jedan* (OVO), a što je nedostatak.
- (f) U praksi linearu regresiju ne bismo željeli koristiti za klasifikaciju. Zašto? Pokažite na gornjem primjeru u čemu je problem (možete modificirati primjer).

a) OVR

$$\Phi = \begin{bmatrix} 1 & -3 & 1 \\ 1 & -3 & 3 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & -2 \\ 1 & 2 & -3 \end{bmatrix}$$

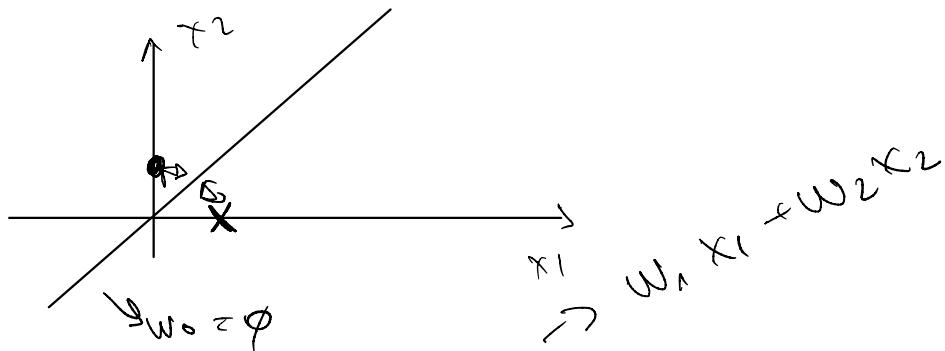
$$\mathbf{y}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{y}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{y}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

$$\mathbf{w} = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}$$



1. (P) Treniramo linearni diskriminativni model u dvodimenziskome prostoru primjera. Skup za učenje čine samo dva primjera, $(\mathbf{x}_1, y_1) = ((1, 0), +1)$ i $(\mathbf{x}_2, y_2) = ((0, 1), -1)$. Na tom skupu treniramo model koji ima induktivnu pristranost takvu da rješenje maksimizira minimalnu udaljenost primjera od hiperravnine. Naučen model ispravno klasificira oba primjera, pri čemu za oba primjera vrijedi $y \cdot h(\mathbf{x}) = 5$. Koliko iznosi težina w_2 tako naučenog modela?

- A -1 B 5 C -5 D 1



$$y \cdot h(\mathbf{x}) = 5$$

$$y \cdot (w_1 x_1 + w_2 x_2) = 5$$

$$(1) \quad 1 \cdot (w_1 \cdot 1 + w_2 \cdot 0) = 5$$

$$(2) \quad -1 \cdot (w_1 \cdot 0 + w_2 \cdot 1) = 5$$

$$w_1 = 5$$

$$w_2 = -5$$

1. (P) Treniramo linearni diskriminativni model u dvodimenzijском простору примјера. Скуп за учење чине само два примјера, $(\mathbf{x}_1, y_1) = ((1, 0), +1)$ и $(\mathbf{x}_2, y_2) = ((0, 1), -1)$. На том скупу тренирамо модел који има индуктивну пристраност такву да решење максимизира минималну удаљеност примјера од хиперправне. Наведен модел исправно класифицира оба примјера, при чему за оба примјера vrijedi $y \cdot h(\mathbf{x}) = 5$. Колико изнosi тежина w_2 тако наведеног модела?

- A -1 B 5 C -5 D 1

$y \cdot h(\mathbf{x}) = 5 \rightarrow$ дато нам је
точност класификације

- да објективира $\rightarrow 0$
 \rightarrow точно класифициран

$$n=2$$

$$\frac{y \cdot h(\mathbf{x}) = 5}{w_2 = ?}$$

$$\vec{w} = (\phi, 1, -1)$$

$$h(1, 0) = 1 - 0 = 1$$

$$h(0, 1) = 0 - 1 = -1 //$$

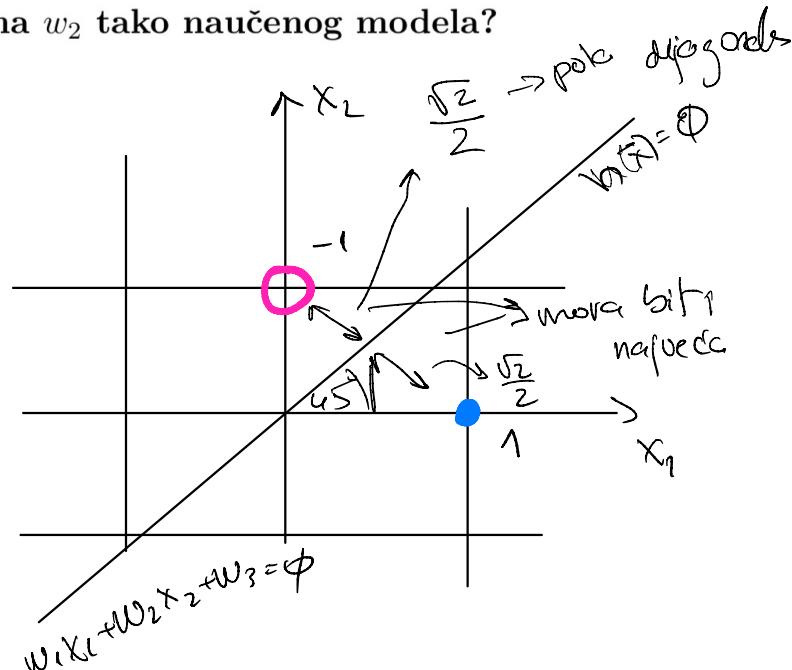
$$d = \frac{|h(\vec{\mathbf{x}})|}{\|\vec{w}\|} = \frac{|y \cdot h(\vec{\mathbf{x}})|}{\sqrt{w^T w}} = \frac{1}{\sqrt{2}} \cdot \frac{\sqrt{2}}{\sqrt{2}} = \frac{\sqrt{2}}{2}$$

\downarrow
бсја w_0

$$5 \cdot \vec{w} = (\phi, 5, -5)$$

$$\begin{aligned} h(1, 0) &= 5 - 0 = 5 & y \cdot h(\mathbf{x}) &= 5 \\ h(0, 1) &= 0 - 5 = -5 & & -5 \end{aligned}$$

$$d = \frac{|y \cdot h(\vec{\mathbf{x}})|}{\|\vec{w}\|} = \frac{5}{\sqrt{50}} = \frac{\sqrt{2}}{\sqrt{50}} = \frac{\sqrt{2}}{5\sqrt{2}} = \frac{1}{5}$$



$$w_2 x_2 = -w_1 x_1 - w_3$$

$$x_2 = \left(-\frac{w_1}{w_2} x_1 \right) - \left(\frac{w_3}{w_2} \right) = \frac{w_1}{w_2} x_1 - \frac{w_3}{w_2} = \phi$$

$$-\frac{w_1}{w_2} = 1$$

$$w_1 = w_2$$

$$x_1 - x_2 + 0 = 0$$

$$h(\vec{\mathbf{x}}) = x_1 - x_2$$

модел је с

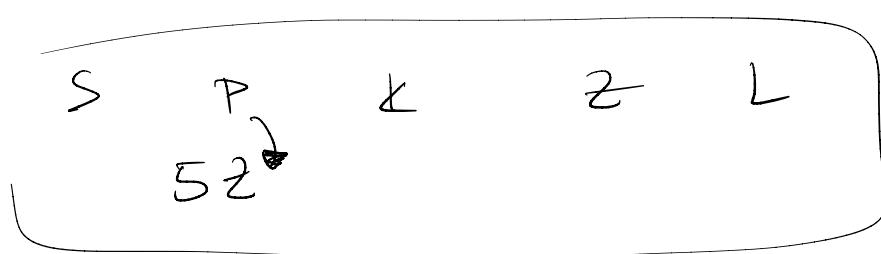
вело димензија

(вело већа константа)

\Rightarrow већи - исте

2. (P) Razvijamo sustav za automatsku klasifikaciju novinskih članaka u jednu od pet kategorija. Tih pet kategorija su "sport", "politika", "kriminal", "znanost" i "lifestyle". Najveća razlika u veličini klase je između kategorija "politika" i "znanost". Očekivano, u kategoriji "politika" ima najviše članaka, dok ih u kategoriji "znanost" ima $5 \times$ manje, što je u redu jer to ionako nitko ne čita. Svaki novinski članak prikazujemo kao vektor riječi, gdje su komponente vektora broj pojavljivanja pojedine riječi. Problem rješavamo algoritmom perceptronra. Budući da je perceptron binaran klasifikator, odlučili smo primijeniti shemu OVR ili shemu OVO za dekompoziciju višeklasnog klasifikacijskog problema u skup binarnih klasifikacijskih problema. **Što možemo očekivati?**

- A OVO će imati $2 \times$ puta manje značajki od OVR, ali bi mogao raditi bolje na člancima iz kategorije "znanost"
- B OVR će imati $2 \times$ manje značajki od OVO, ali bi mogao raditi lošije na člancima iz kategorije "znanost"
- C OVO će imati $5 \times$ puta manje značajki od OVR, ali bi mogao raditi lošije na člancima iz kategorije "znanost"
- D OVR će imati $5 \times$ manje značajki od OVO, ali bi mogao raditi bolje na člancima iz kategorije "znanost"



$$\text{OVO } \binom{K}{2} = 10$$
$$\text{OVR } K = 5$$

OVO \rightarrow br. glasova 12! \rightarrow nejednakost

3. (T) Na predavanjima smo za klasifikaciju pokušali upotrijebiti algoritam regresije. Zaključili smo da to ne funkcionira, tj. da algoritam linearne regresije jednostavno nije klasifikacijski algoritam. **Koje bismo minimalne preinake trebale učiniti u algoritmu linearne regresije, a da on dobro funkcionira kao klasifikacijski algoritam?**

- A Promijeniti model, funkciju gubitka i optimizacijski postupak
- B Promijeniti funkciju gubitka i optimizacijski postupak
- C Promijeniti funkciju gubitka
- D Promijeniti model i funkciju gubitka

4. (N) Raspolažemo sljedećim skupom za učenje u dvodimensijskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}(i), y(i))\} = \{((1, 0), +1), ((2, -3), -1), ((2, 5), -1)\}$$

Na ovom skupu treniramo perceptron. Pritom koristimo funkciju preslikavanja u peterodimensijski prostor značajki, koja je definirana na sljedeći način:

$$\phi(\mathbf{x}) = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$$

Početne težine perceptronu neka su sljedeće:

$$\mathbf{w} = (1, 0, -1, 2, -2, 0)$$

Koliko iznosi empirijska pogreška perceptrona na skupu za učenje prije početka treninga (dakle, s početnim težinama)?

- A 8 B 9 C 16 D 25

$$e_p(\mathbf{w}) = \sum_{y_i \neq h(\mathbf{x}_i)} y_i$$

$$\Phi_1 = [1 \ 1 \ 0 \ 0 \ 1 \ 0]$$

$$= -(-1 \cdot 1 + 8 \cdot (-1))$$

$$\Phi_2 = [1 \ 2 \ -3 \ -6 \ 4 \ 9]$$

$$= -(-1 - 8)$$

$$\Phi_3 = [1 \ 2 \ 5 \ 10 \ 4 \ 25]$$

$$= 9$$

$$h(\mathbf{x}^1) = [1 \ 0 \ -1 \ 2 \ -2 \ 0] \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = 1 - 2 = -1 \rightarrow \textcircled{-1}$$

$$h(\mathbf{x}^2) = [1 \ 0 \ -1 \ 2 \ -2 \ 0] \begin{bmatrix} 1 \\ 2 \\ -3 \\ -6 \\ 4 \\ 9 \end{bmatrix} = 1 + 3 - 12 - 8 = -16 \rightarrow \textcircled{-1}$$

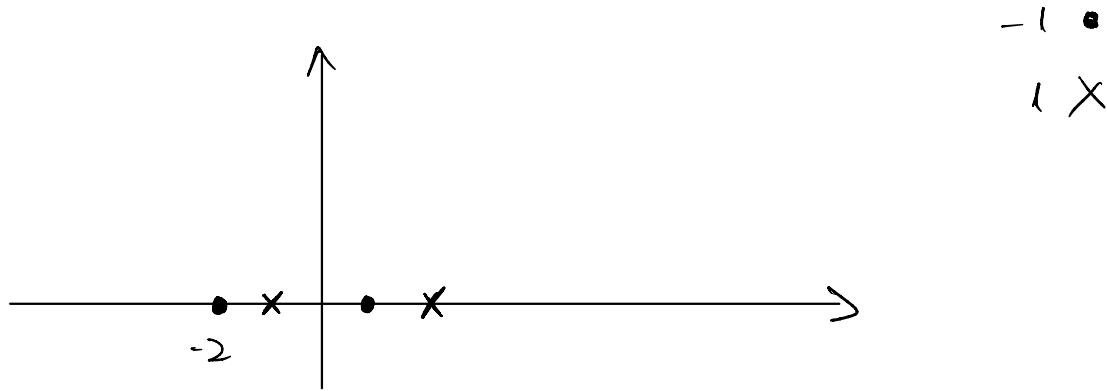
$$h(\mathbf{x}^3) = [1 \ 0 \ -1 \ 2 \ -2 \ 0] \begin{bmatrix} 1 \\ 2 \\ 5 \\ 10 \\ 4 \\ 25 \end{bmatrix} = 1 - 5 + 20 - 8 = 8 \rightarrow \textcircled{1}$$

5. (P) Razmotrimo sljedeći skup označenih primjera:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((-2, 0), -1), ((-1, 0), +1), ((1, 0), +1), ((2, 0), -1)\}$$

Ovaj skup nije linearno odvojiv i algoritam perceptronu neće konvergirati. Linearna neodvojivost podataka je konceptualni razlog zašto algoritam ne konvergira. **Koji je tehnički razlog zašto algoritam perceptronu na ovom skupu primjera neće konvergirati?**

- A U svakoj točki prostora parametara postoji barem jedan primjer za koji je gradijent gubitka veći od nule
- B Premda je empirijska pogreška na ovom skupu primjera derivabilna, ona je uglavnom konstantna
- C U prostoru parametara ne postoji točka u kojoj je gradijent empirijske pogreške jednak nuli
- D U prostoru parametara postoji više točaka za koje je empirijska pogreška jednaka nuli



Logistička regresija

1. (T) Poopćeni linearни model definirali smo kao $h(\mathbf{x}) = f(\mathbf{w}^T \phi(\mathbf{x}))$, gdje je f neka (moguće nelinearna) aktivacijska funkcija, a ϕ je (moguće nelinearna) funkcija preslikavanja u prostor značajki. **Koji od navedenih uvjeta je dovoljan uvjet da granica između klasa u ulaznom prostoru bude linearna?**

- A f je afina funkcija i $\phi(\mathbf{x}) = (1, \mathbf{x})$ B f je afina funkcija C $f(\mathbf{x}) = \mathbf{x}$ D $\phi(\mathbf{x}) = (1, \mathbf{x})$

2. (T) Kod logističke regresije, pogrešku unakrsne entropije izveli smo modelirajući distribuciju vjerojatnosti oznaka y u skupu označenih primjera. **Na koji smo način modelirali distribuciju vjerojatnosti pojedinačnog primjera y ?**

A $P(y|\mathbf{x}) = h(\mathbf{x})^y(1 - h(\mathbf{x}))^{1-y}$

B $P(y|\mathbf{x}) = h(\mathbf{x})(1 - h(\mathbf{x}))$

C $P(y|\mathbf{x}) = (y - h(\mathbf{x}))\mathbf{x}$

D $P(y|\mathbf{x}) = y^{h(\mathbf{x})}(1 - y)^{h(\mathbf{x})}$

3. (N) Na skupu označenih primjera \mathcal{D} trenirali smo model logističke regresije. Dobili smo neki vektor težina w i pomak $w_0 = 0.15$. Tako naučenom modelu neki primjer x , čija je oznaka u skupu primjera $y = 0$, nanosi gubitak unakrsne entropije od $L(0, h(x)) = 0.274$. Koliki gubitak unakrsne entropije bi nanosio primjer x kada bismo njegove značajke pomnožili sa dva i promijenili mu oznaku?

- A 4.03 B 2.54 C 7.11 D 1.19

$$L(y, h(x)) = -y \ln(h(x)) - (1-y) \ln(1-h(x))$$

$$L(0, h(x)) = -1 \ln(1 - h(x))$$

$$-0.274 = +\ln(1 - h(x)) \quad \left| e^{\textcolor{pink}{n}} \right.$$

$$1 - h(x) = e^{-0.274}$$

$$-h(x) = e^{-0.274} - 1$$

$$h(x) = 1 - e^{-0.274} = 0.239$$

$$h(x) = \frac{1}{1 + \exp^{-w^\top \phi(x)}}$$

$$0.239 = \frac{1}{1 + \exp^{-w^\top \phi(x)}}$$

$$0.239 (1 + \exp^{-w^\top \phi(x)}) = 1$$

$$1 + \exp^{-w^\top \phi(x)} = \frac{1}{0.239}$$

$$\exp^{-w^\top \phi(x)} = \frac{1}{0.239} - 1 = 3.18 \quad \left| \text{h} \right.$$

$$(-w^\top \phi(x)) = 1.16$$

$$-\phi(x) = 2(-w^\top \phi(x) - w_0) + w_0$$

$$= -2w^\top \phi(x) - w_0 = -2 \cdot 0.156 - 0.15 = -2.48$$

$$h(x) = \frac{1}{1 + \exp^{-w^\top \phi(x)}} =$$

$$L(1, h(x)) = -\ln(h(x)) = -\ln\left(\frac{1}{1 + \exp^{-w^\top \phi(x)}}\right)$$

$$= -\ln\left(\frac{1}{1 + \exp^{-1.16}}\right) = 2.54 \text{ ln}$$

4. (N) Na skupu \mathcal{D} označenih primjera trenirali smo model binarne logističke regresije. Naknadno smo uočili da jedan primjer iz skupa \mathcal{D} modelu nanosi razmjerno velik gubitak. Konkretno, iznos gubitka za dotični primjer je $L(y, h(\mathbf{x})) = 1.20$. Ispostavilo se da je taj primjer pogrešno označen. **Koliko bi iznosio gubitak modela na istom ovom primjeru, ako bismo sada naknadno promijenili njegovu oznaku, ali model ostavili nepromijenjenim?**

- A 0.70 B 0.28 C 0.36 D 0.52

$$L(y_1, h(x)) = 1.2$$

$$\overbrace{1.2 = -\ln(1-h(x))}^{\text{PRETPOSTAVIMO } y=0}$$

$$1-h(x) = e^{-1.2}$$

$$h(x) = 1 - e^{-1.2} = 0.1699$$

$$\overbrace{L(1, h(x)) = -\ln(h(x))}^{y=1} = 0.136 \text{ m}$$

5. (N) Model logističke regresije treniramo stohastičkim gradijentnim spustom. Primjere iz dvodimenzionskog ulaznog prostora preslikali smo u prostor značajki funkcijom

$$\phi(x) = (1, x_1, x_2, x_1x_2)$$

U jednoj iteraciji treniranja modela vektor parametara jednak je

$$w = (0.2, 0.5, -1.1, 2.7)$$

Koliko u toj iteraciji iznosi L_2 -norma gradijenta gubitka za primjer $(x, y) = ((-0.5, 2), 1)$?

- A 0.70 B 2.48 C 1.28 D 4.00

$$\phi(x) = (1 \quad -0.5 \quad 2 \quad -1)$$

$$w^\top \phi(x) = [0.2 \ 0.5 \ -1.1 \ 2.7] \begin{bmatrix} 1 \\ -0.5 \\ 2 \\ -1 \end{bmatrix} = 0.2 - 0.25 - 1.1 - 2.7 \\ = -0.05 - 3.18 \\ = -4.23$$

$$h(x) = \frac{1}{1 + e^{-w^\top \phi(x)}} = 0.007$$

$$\nabla_w L(y, h(x)) = (h(x) - y) \phi(x) \\ = -0.993 [1 \quad -0.5 \quad 2 \quad -1] \\ = [-0.993 \quad +0.4965 \quad -1.986 \quad 0.993]$$

$$\|\nabla_w L\| = \sqrt{(-0.993)^2 + (0.4965)^2 + (-1.986)^2 + (0.993)^2}$$

$$= 2.48$$

6. (P) Na skupu označenih primjera treniramo tri modela: (1) model neregularizirane logističke regresije (NR), (2) model L2-regularizirane logističke regresije (L2R) i (3) perceptron s funkcijom preslikavanja. Sva tri modela koriste iste značajke. Za sva tri algoritma promatramo iznos empirijske pogreške učenja kroz iteracije optimizacijskog postupka. Nakon određenog broja iteracija, algoritam perceptrona uspješno se zaustavlja s rješenjem. **Kako se u ovom slučaju ponaša empirijska pogreška učenja kroz iteracije za dva spomenuta modela logističke regresije, NR i L2R?**

→lin odgovor

- A Pogreška učenja modela NR nakon određenog broja iteracije doseže nulu, dok pogreška učenja modela L2R ~~najprije pada pa raste~~
- B Pogreške učenja modela NR i modela L2R ~~dosežu nulu~~, ali modelu L2R za to treba više iteracija
- C Pogreške učenja modela NR i modela L2R obje stagniraju nakon određenog broja iteracija, ali ~~modelu NR za to treba više iteracija~~
- D Pogreška učenja modela NR asymptotski teži nuli, dok pogreška učenja modela L2R nakon određenog broja iteracija stagnira

7. (P) Razmotrimo sljedeći skup označenih primjera:

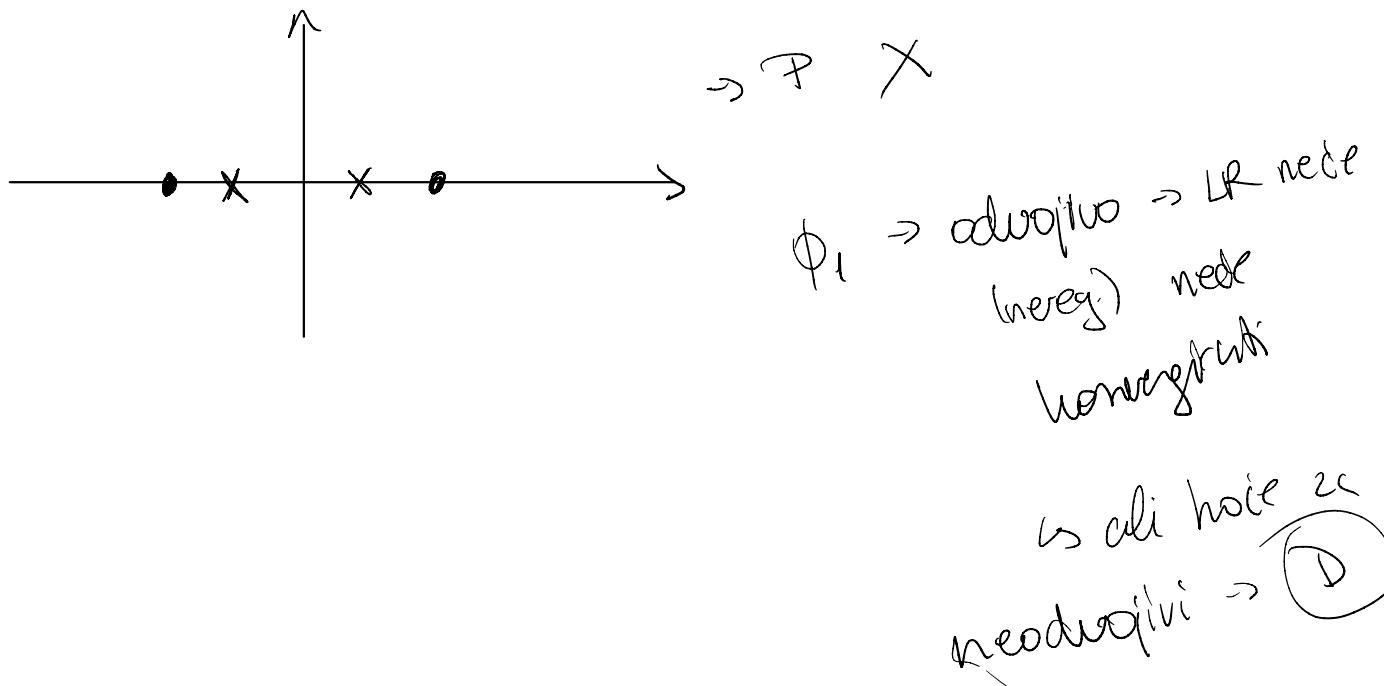
$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((-2, 0), -1), ((-1, 0), +1), ((1, 0), +1), ((2, 0), -1)\}$$

Nad ovim skupom treniramo dva modela: perceptron (P) i neregulariziranu logističku regresiju (LR). Pored toga, razmatramo tri funkcije preslikavanja:

$$\begin{aligned}\phi_0(\mathbf{x}) &= (1, x_1, x_2) \\ \phi_1(\mathbf{x}) &= (1, x_1, x_2, x_1^2, x_2^2) \\ \phi_2(\mathbf{x}) &= (1, x_1, x_2, x_1 x_2)\end{aligned}$$

Ukupno, dakle, isprobavamo šest kombinacija modela i funkcije preslikavanja. **Za koje će algoritme (model+preslikavanje) empirijska pogreška na skupu za učenje konvergirati?**

- A P+ ϕ_0 B P+ ϕ_2 C LR+ ϕ_1 D LR+ ϕ_2



Logistička regresija

II.

1. (T) Kod logističke regresije optimizaciju tipično provodimo gradijentnim spustom. Međutim, kod linearne regresije optimizaciju smo provodili izračunom pseudoinverza matrice dizajna. **Zašto optimizaciju kod logističke regresije također ne provodimo izračunom pseudoinverza matrice dizajna?**

- A Optimizaciju parametara linearne regresije također možemo napraviti gradijentnim spustom po empirijskoj pogrešci, ali to ne radimo jer izračun pseudoinverza ima manju računalnu složenost
- B Maksimizacija log-izglednosti oznaka logističke regresije kao rješenje za parametre ne daje izraz u zatvorenoj formi koji sadržava pseudoinverz matrice dizajna 
- C Zbog nelinearnosti logističke funkcije, kod logističke regresije izračun pseudoinverza matrice dizajna nije moguće napraviti u zatvorenoj formi
- D Optimizaciju možemo provesti izračunom pseudoinverza matrice dizajna, međutim, za razliku od gradijentnog spusta, taj postupak ne funkcioniра kada je matrica dizajna singularna

2. (T) Kod logističke regresije za optimizaciju tipično koristimo gradijentni spust ili Newtonov optimizacijski postupak. **Što su prednosti, a što nedostatci gradijentnog spusta u odnosu na Newtonov postupak, i to konkretno kod logističke regresije?**

- A Za razliku od Newtonovog postupka, gradijentni spust može se koristiti za "online" (pojedinačno) učenje, no može krivudati i zato sporije konvergirati od Newtonovog postupka
- B Za razliku od Newtonovog postupka, gradijentni spust može se koristiti za L2-regulariziranu logističku regresiju, no ako je stopa učenja prevelika, postupak može divergirati, dok Newtonov postupak nema taj problem
- C Newtonov postupak brže konvergira, ali se može koristiti samo za konveksnu funkciju pogreške, dok gradijentni spust nema tog ograničenja, ali može zaglaviti u lokalnom optimumu
- D Gradijentni spust znatno je računalno jednostavniji od Newtonovog postupka, no za razliku od Newtonovog postupka kod L2-regularizirane regresije ne konvergira ako primjeri nisu linearno odvojivi

3. (T) Kod Newtonovog postupka optimizacije za logističku regresiju ažuriranje težina provodi se prema sljedećem pravilu:

$$\mathbf{w} \leftarrow \mathbf{w} - \mathbf{H}^{-1} \nabla E(\mathbf{w} | \mathcal{D})$$

Očito, za provođenje ovog postupka potrebno je računati inverz Hesseove matrice, tj. matrice parcijalnih derivacija \mathbf{H} . Općenito, operacija invertiranja matrice nije uvijek izvediva, a čak i kada jest izvediva, rješenje nije uvijek numerički stabilno. **Kod logističke regresije, koji je nužan i dovoljan uvjet za izvedivost i numeričku stabilnost Newtonovog optimizacijskog postupka?**

- A Značajke moraju biti linearno zavisne
- B U podatcima ne smije biti multikolinearnosti 
- C Broj primjera mora biti veći od broja značajki plus jedan
- D Funkcija pogreške mora biti konveksna

4. (T) Svi poopćeni linearni modeli mogu se trenirati u “online” (pojedinačnom) načinu, primjernom algoritma LMS. To vrijedi i za algoritam linearne regresije, za koji smo prvotno kao minimizaciju kvadrata provodili računajući pseudoinverz matrice dizajna. Jedna od prednosti algoritma LMS u odnosu na izračun pseudoinverza kod linearne regresije je manja računalna složenost LMS-a. Neka E označava broj epoha, N je broj primjera, a m broj značajki u prostoru značajki. **Koja je (vremenska) računalna složenost algoritma LMS, primijenjenog na linearnu regresiju?**

- [A] $\mathcal{O}(ENm)$ [B] $\mathcal{O}(E(N + m))$ [C] $\mathcal{O}(EN^2m)$ [D] $\mathcal{O}(ENm^2)$

5. (T) Problem višeklasne ($K > 2$) klasifikacije logističkom regresijom možemo riješiti na više načina. Možemo primijeniti (1) multinomijalnu logističku regresiju (MLR), (2) binarnu logističku regresiju sa shemom OVO (BLR-OVO) ili (3) binarnu logističku regresiju sa shemom OVR (BLR-OVR). **Koja je prednost MLR nad BLR-OVO i BLR-OVR?**

- A MLR ima više parametara od BLR-OVR, ali nije osjetljiva na neuravnoteženost broja primjera po klasama
- B Za razliku od BLR-OVR i BLR-OVO, kod MLR ne postoje područja u ulaznom prostoru za koje klasifikacijska odluka nije određena
- C MLR i BLR-OVR imaju manje parametara od BLR-OVO, no jedino za MLR vrijedi $\sum_k P(y = k | \mathbf{x}) = 1$
- D Za razliku od BLR-OVO i BLR-OVR, MLR koristi funkciju softmax, pa minimizacija L1-regularizirane pogreške ima rješenje u zatvorenoj formi

6. (N) Raspolažemo označenim skupom primjera iz triju klasa ($K = 3$) u trodimenzijskome ulaznom prostoru ($n = 3$). Na tom skupu treniramo model multinomijalne logističke regresije. Treniranje provodimo gradijentnim spustom. U nekoj od iteracija gradijentnog spusta matrica težina je sljedeća (stupci odgovaraju težinama za pojedine klase):

$$\mathbf{W} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & -2 \\ 3 & -4 & 6 \\ -3 & 0 & 2 \end{pmatrix}$$

Jedan od primjera u skupu za učenje je primjer $\mathbf{x} = (3, 2, -1)$ s oznakom $\mathbf{y} = (0, 1, 0)$. Koliko iznosi gubitak unakrsne entropije koji u ovoj iteraciji optimizacijskog postupka nanosi dotični primjer?

- [A] 7 [B] 11 [C] 23 [D] 35

$$\text{Jedan od primjera u skupu za učenje je primjer } \mathbf{x} = (3, 2, -1) \text{ s oznakom } \mathbf{y} = (0, 1, 0). \text{ Koliko iznosi gubitak unakrsne entropije koji u ovoj iteraciji optimizacijskog postupka nanosi dotični primjer?}$$

$$L(y_1, h_2(x)) = - \sum_{k=1}^K y_k \ln h_k(x; w) \xrightarrow{\text{sune}} k_1 = 1$$

$$h_2(x; w) = \text{softmax}(w^\top \phi(x)) = \frac{\exp(w_2^\top \phi(x))}{\sum_i \exp(w_i^\top \phi(x))} \quad \phi(x) = [1 \ 3 \ 2 \ -1]$$

$$\boxed{w^\top \vec{x} = \vec{x}^\top w}$$

$$\begin{bmatrix} 1 & 3 & 2 & -1 \end{bmatrix}_{1,4} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -2 \\ 3 & -4 & 6 \\ -3 & 0 & 2 \end{bmatrix}_{4,3} = \begin{bmatrix} k_1 & k_2 & k_3 \\ 16 & -7 & 5 \end{bmatrix}_{1,3}$$

$$h_2(x; w) = \frac{\exp(-7)}{\exp(16) + \exp(-7) + \exp(5)} = 1,026 \cdot 10^{-10}$$

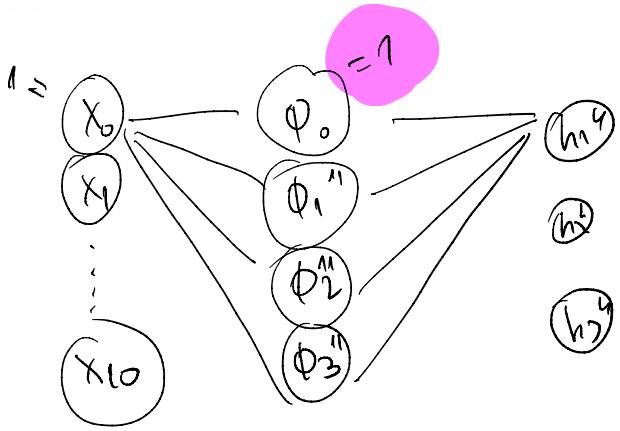
$$L(y_1, h_2(x)) = -\ln 1,026 \cdot 10^{-10} = 23 \frac{1}{2}$$

7. (P) Poopćeni linearni modeli mogu koristiti adaptivne bazne funkcije. Prednost toga je da ne moramo ručno definirati preslikavanje ϕ u prostor značajki, već se to preslikavanje može naučiti na temelju podataka. Rasplažemo podatcima iz $K = 3$ klase u 10-dimenzijском ulaznom prostoru. Za taj višeklasni problem koristimo multinomijalnu logističku regresiju, ali s adaptivnim baznim funkcijama. Svaka adaptivna bazna funkcija ϕ_j parametrizirana je kao skalarni produkt vektora značajki i vektora primjera, kao što smo radili na predavanjima. Naš model definirali smo ovako:

$$h_k(\mathbf{x}) = \text{softmax}_k \left(\sum_{j=0}^3 w_{j,k} \phi_j(\mathbf{x}) \right)$$

Ovime je definirana hipoteza za klasu k . Svaka klasa ima svoju hipotezu h_k . Svaka klasa ima i svoje težine $w_{j,k}$. Međutim, bazne funkcije ϕ_j zajedničke su za sve klase (dakle, ti parametri su dijeljeni između klasa). **Koliko ukupno parametara ima ovaj model?**

- A 45 B 49 C 136 D 142



$$\text{Parametara u } \phi : 11 + 11 + 11 = 33$$

$$\text{Parametara u } h_k : 4 + 4 + 4 = 12$$

45

Stroj potpornih vektora

1. (T) Kod izvoda algoritma SVM s tvrdom marginom, pretpostavili smo da za primjere $\mathbf{x} \in \mathbb{R}^n$ vrijedi sljedeći uvjet linearne odvojivosti:

$$\forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}. y^{(i)} h(\mathbf{x}^{(i)}) \geq 0$$

Koliko hipoteza zadovoljava ovaj uvjet, i kako algoritam SVM odabire jednu od njih?

- A Uvjet zadovoljava konačan broj hipoteza koje su linearno odvojive, a SVM između njih odabire onu jednu koja minimizira kvadrat vektora težina
- B Uvjet zadovoljava konačan broj hipoteza koje su linearno odvojive, no one se razlikuju samo po faktoru koji množi težine (\mathbf{w}, w_0) , pa SVM odabire onu jednu za koju vrijedi $y h(\mathbf{x}) \geq 1$ za sve primjere
- C Uvjet zadovoljava beskonačno mnogo hipoteza, a SVM odabire onu jednu koja minimizira kvadrat vektora težina te koja ispravno klasificira sve primjere, uz uvjet da $h(\mathbf{x})$ nije u intervalu $(-1, +1)$
- D Uvjet zadovoljava beskonačno mnogo hipoteza, međutim samo za jednu vrijedi $y h(\mathbf{x}) = 1$ za najbliže primjere, i to je hipoteza koju odabire SVM

2. (T) Kod SVM-a, problem maksimalne margine sveo se na problem minimizacije izraza $\frac{1}{2}\|\mathbf{w}\|^2$ uz određena ograničenja. **Zašto minimizacija ovog izraza daje maksimalnu marginu?**

- A Što je vektor \mathbf{w} kraći, to je manja udaljenost d primjera od hiperravnine, a to efektivno znači da je margina to šira jer je margina fiksna a udaljenosti d se smanjuju
- B Što je vektor \mathbf{w} kraći, to je manja vrijednost $h(\mathbf{x})$, ali je težina w_0 konstantna, pa se udaljenosti izmeđi hiperravnine i primjera povećavaju, što znači da se margina širi
- C Što je vektor \mathbf{w} kraći, to je manja vrijednost $h(\mathbf{x})$, pa primjeri moraju biti što dalje da bi vrijedilo $h(\mathbf{x}) = \pm 1$, a to znači da je margina to šira
- D Što je vektor \mathbf{w} kraći, to je veća udaljenost d primjera od hiperravnine, što znači da se potporni vektori udaljavaju od hiperravnine, a to znači da margina postaje šira

31. (T) Svaki algoritam strojnog učenja ima neku induktivnu pristranost. Bez induktivne pristranosti nije moguće naučiti model koji bi generalizirao. **Po čemu se induktivna pristranost algoritma SVM (tvrdi margini) razlikuje od induktivne pristranosti algoritma perceptron?**

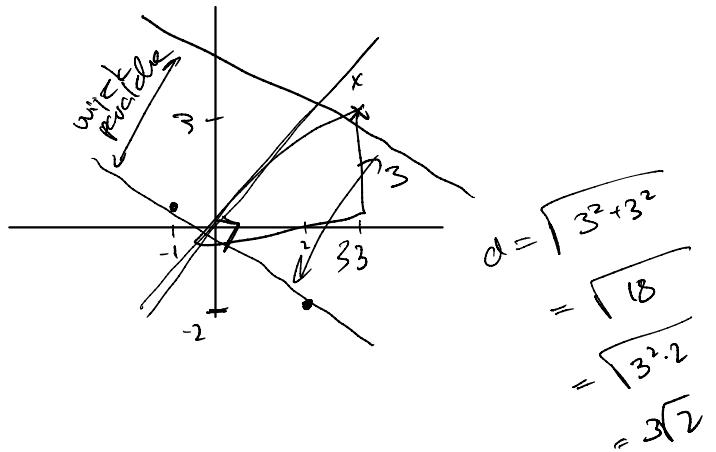
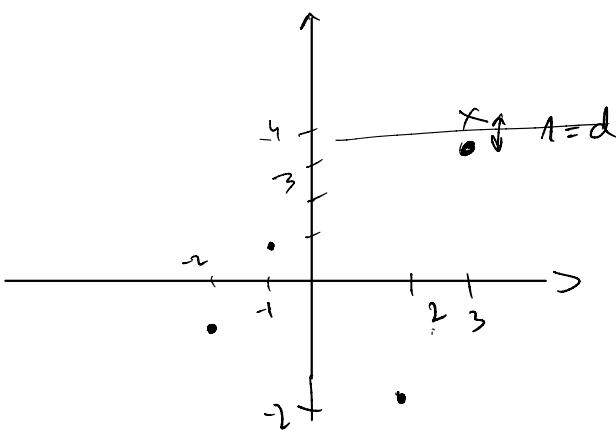
- A Razlikuju se po pristranost preferencijom, jer perceptron ne maksimizira marginu, premda se može dogoditi da pronađe rješenje koje maksimizira marginu
- B SVM ima pristranost preferencijom kojom maksimizira marginu, dok perceptron nema induktivnu pristranost preferencijom već samo pristranost jezika
- C ~~Imaju istu pristranost preferencijom~~, a to je da primjeri moraju biti linearno odvojivi, no SVM ima dodatnu pristranost ograničenjem u vidu optimizacijskih ograničenja
- D ~~Imaju istu pristranost jezika~~, a pristranost preferencijom također će biti ista ako se oba optimiraju gradijentnim spustom s istim početnim težinama i istom stopom učenja

4. (P) Raspolažemo sljedećim skupom označenih primjera u dvodimensijskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((-1, 1), -1), ((-2, -1), -1), ((2, -2), -1), ((3, 3), -1), ((3, 4), +1)\}$$

Na ovom skupu treniramo model SVM-a s tvrdom marginom. Međutim, naknadno smo utvrdili da je primjer $(3, 3)$ imao pogrešnu oznaku, pa smo to ispravili te ponovno trenirali SVM. Na ispravljenom skupu primjera dobili smo granicu između klasa sa znatno širom marginom nego na početnom skupu primjera. **Koliko je nova margina šira od stare?**

- A $3\sqrt{2}$ puta B $2\sqrt{5}$ puta C $\sqrt{26}$ puta D $\frac{5}{2}\sqrt{3}$ puta



5. (T) Kod optimizacije SVM-a iskoristili smo Lagrangeovu dualnost kako bismo se iz primarnog optimizacijskog problema prebacili u dualni optimizacijski problem. To smo učinili tako da smo na temelju Lagrangeove funkcije L definirali dualnu Lagrangeovu funkciju \tilde{L} i uveli nova ograničenja, što nam je opet dalo kvadratni program. **Kako onda u konačnici glasi optimizacijski problem tvrde margine u dualnoj formulaciji (ako zanemarimo ograničenja)?**

- A $\operatorname{argmax}_{\alpha} \min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \alpha)$
- B $\operatorname{argmin}_{\alpha} \max_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \alpha)$
- C $\operatorname{argmax}_{\mathbf{w}, w_0} \min_{\alpha} L(\mathbf{w}, w_0, \alpha)$
- D $\operatorname{argmin}_{\mathbf{w}, w_0} \max_{\alpha} L(\mathbf{w}, w_0, \alpha)$

6. (N) Rješavamo binarni klasifikacijski problem. Raspolažemo označenim skupom primjera. Odgovarajuća matrica dizajna je sljedeća:

$$\mathbf{X} = \begin{pmatrix} 1 & 3 & 16 & -8 & -11 \\ 1 & -5 & 4 & -8 & -7 \\ 1 & 7 & -4 & 11 & 9 \\ 1 & 15 & -20 & 25 & 25 \end{pmatrix}$$

Na ovom skupu treniramo model SVM-a s tvrdom marginom i linearom jezgrenom funkcijom (tj. bez preslikavanja u prostor značajki). Model treniramo u primarnoj formulaciji. Za rješenje maksimalne marge dobili smo ovaj vektor težina (uključivo s težinom w_0):

$$\mathbf{w} = (+0.1370, -0.0290, +0.0194, -0.0461, -0.0388)$$

Umjesto u primarnoj formulaciji, model smo mogli trenirati u dualnoj formulaciji, pa bismo umjesto vektora težina \mathbf{w} dobili vektor dualnih parametara α , odnosno Lagrangeove multiplikatore. Prisjetite se da su vektori čiji su Lagrangeovi multiplikatori veći od nule potporni vektori. Premda to nije uvijek moguće, u ovom konkretnom slučaju dualni parametri modela mogu se izvesti iz rješenja primarnog modela. Izvedite vektor dualnih parametara α . **Koliko iznosi najveća vrijednost parametra u vektoru dualnih parametara α ?** (Rezultate uspoređujte po prve tri decimale.)

- A 0.0013 B 0.0024 C 0.0045 D 0.0089

$$w = \sum_{i=1}^n \alpha_i y^{(i)} X^{(i)}$$

$$y = w^\top x$$

$$y = [0.137 \quad -0.029 \quad 0.0194 \quad -0.0461 \quad -0.0388] \begin{bmatrix} 1 & 1 & 1 & 1 \\ 3 & -5 & 7 & 15 \\ 16 & 4 & -4 & -20 \\ -8 & -8 & 11 & 25 \\ -11 & -7 & 9 & 25 \end{bmatrix}$$

potporni

$\alpha_2, \alpha_3 > 0$

$$= [1.156 \quad 1 \quad -1 \quad -2.81]$$

$$W_1 = \mathcal{L}_2 \gamma^{(2)} x_1^{(2)} + \mathcal{L}_3 \gamma^{(3)} x_4^{(3)}$$

$$W_2 = \mathcal{L}_2 \gamma^{(2)} x_2^{(2)} + \mathcal{L}_3 \gamma^{(3)} x_2^{(3)}$$

$$-0,029 = \mathcal{L}_2 (-5) + \mathcal{L}_3 (-1) \quad |+ \\$$

$$0,0194 = \mathcal{L}_2 \cdot 4 + \mathcal{L}_3 (-1) (-4)$$

$$-0,029 = -5\mathcal{L}_2 - \mathcal{L}_3$$

$$0,0194 = 4\mathcal{L}_2 + 4\mathcal{L}_3 /: 4$$

$$-0,029 = -5\mathcal{L}_2 - \mathcal{L}_3$$

$$4,85 \cdot 10^{-3} = \mathcal{L}_2 + \mathcal{L}_3 \rightarrow \mathcal{L}_2 = 4,85 \cdot 10^{-3} - \mathcal{L}_3$$

$$\boxed{\mathcal{L}_2 = 2,375 \cdot 10^{-3}}$$

$$-0,029 = -5(4,85 \cdot 10^{-3} - \mathcal{L}_3) - \mathcal{L}_3$$

$$-4,75 \cdot 10^{-3} = -2\mathcal{L}_3$$

$$\boxed{\mathcal{L}_3 = 2,375 \cdot 10^{-3}}$$

7. (T) Model SVM-a može se definirati i optimirati u primarnoj ili dualnoj formulaciji. **Koncepcionalno, kada će primjer \mathbf{x} u dualnoj formulaciji SVM-a biti klasificiran u pozitivnu klasu?**

- A Ako je linearna kombinacija značajki iz \mathbf{x} s pozitivnim težinama veća ili jednaka linearnoj kombinaciji značajki iz \mathbf{x} s negativnim težinama
- B Ako je vektor \mathbf{x} po skalarnom umnošku sličniji potpornim vektorima s pozitivnom oznakom nego potpornim vektorima s negativnom oznakom
- C Ako je skalarni umnožak vektora \mathbf{x} i vektora oznaka \mathbf{y} veća od praga definiranog parametrom w_0
- D Ako većina od ukupno α primjera iz skupa za učenje koji su po euklidskoj udaljenosti najbliži primjeru \mathbf{x} ima pozitivnu oznaku

Stroj potpornih vektora II.

1. (T) Problem meke margine SVM-a formulirali smo kao problema optimizacije uz ograničenja, preciznije kao problem kvadratnog programiranja. Neka je n broj značajki, a N broj primjera. **Koliko primarni optimizacijski problem meke margine ima ukupno ograničenja, a koliko varijabli po kojima optimiramo?**

- A N ograničenja i $2N + 1$ varijabli
- B N ograničenja i $n + 1$ varijabli
- C $2N$ ograničenja i $N + n + 1$ varijabli
- D $2N$ ograničenja i $2n$ varijabli

2. (T) Kod optimizacijskog problema meke margine jedan od uvjeta KKT koji vrijede u točki rješenja je sljedeći uvjet komplementarne labavosti:

$$\alpha_i(y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + w_0) - 1 + \xi_i) = 0$$

Što možemo zaključiti na temelju ovog uvjeta?

- A Da se primjeri koji nisu potporni vektori sigurno nalaze izvan margine
- B Da se potporni vektori ne nalaze izvan margine na pravoj strani granice
- C Da se potporni vektori nalaze na margini ili izvan nje, a na pravoj strani granice
- D Da se primjeri koji nisu potporni vektori nalaze na margini ili unutar margine

3. (N) Raspolažemo sljedećim skupom označenih primjera u trodimenzijskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((-1, 3, 6), -1), ((-4, 4, 4), -1), ((-2, 4, 1), +1)\}$$

Na ovom skupu primjera treniramo model SVM-a s linearom jezgrenom funkcijom i sa $C = 0.01$. Postupak treniranja algoritmom SMO završio je s vektorom Lagrangeovih koeficijenata $\alpha = (0, 0.01, 0.01)$. Iz ovoga se da izračunati da vrijedi $w_0 = -0.8$. Umjesto algoritma SMO, za optimizaciju smo mogli upotrijebiti gradijentni spust i optimirati težine u primarnoj formulaciji problema. U tom slučaju koristili bismo empirijsku pogrešku SVM-a definiranu kao L2-regularizirani gubitak zglobnice. Međutim, tu pogrešku možemo izračunati i naknadno, nakon što smo naučili model. **Koliko iznosi empirijska pogreška ovog SVM-a na skupu primjera \mathcal{D} ?**

- A 1.935 B 33.935 C 1.135 D 33.135

$$G_R(\vec{w}|\mathcal{D}) = \sum_{i=1}^n \max(0, 1 - y_i h(\mathbf{x}_i)) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

$$\vec{w} = \sum_{i=1}^n \alpha_i y_i^{(i)} \mathbf{x}^{(i)}$$

$$w_1 = -0.01 \cdot (-4)^2 \cdot x_1^2 - 0.01 \cdot 4^3 \cdot x_1^3 \quad \|\mathbf{w}\|_2^2 = 0.02^2 + 0.03^2 = 1.13 \cdot 10^{-3}$$

$$w_2 = -0.01 \cdot 4^2 \cdot x_2^2 - 0.01 \cdot 4^3 \cdot x_2^3$$

$$w_3 = -0.01 \cdot 4^2 \cdot x_3^2 - 0.01 \cdot 4^3 \cdot x_3^3$$

$$\frac{\lambda}{2} = \frac{1}{2C} \Rightarrow \frac{1}{2C} \cdot \|\mathbf{w}\|_2^2 = 0.065$$

$$w_1 = -0.01 \cdot (-4) + 0.01 \cdot (-2) = 0.01(-4 - 2) = 0.02$$

$$w_2 = -0.01 \cdot 4 + 0.01 \cdot 4 = 0.01(-4 + 4) = 0$$

$$w_3 = -0.01 \cdot 4 + 0.01 \cdot 1 = 0.01(-4 + 1) = -0.03$$

$$h(x) = \begin{bmatrix} 1 & -1 & 3 & 6 \\ 1 & -4 & 4 & 4 \\ 1 & -2 & 4 & 1 \end{bmatrix} \begin{bmatrix} -0.08 \\ 0.02 \\ 0 \\ -0.03 \end{bmatrix} = \begin{bmatrix} -1 & -1 & -0.087 \end{bmatrix}$$

svi neg. \rightarrow 1 kivo

$$(3, 4) \quad (4, 1) \\ \xrightarrow{-0.087}$$

$$G(w) = \max(0, 1 - h(x)) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

$$= 1.135 + 0.065 = 1.1935$$

$$\begin{bmatrix} 1 & -1 & 3 & 6 \\ 1 & -4 & 4 & 4 \\ 1 & -2 & 4 & 1 \end{bmatrix} \begin{pmatrix} 1 \\ -2 \\ 4 \\ 1 \end{pmatrix} = (4, 3)$$

Jezgrena funkcije

1. (P) Na 1000 primjera sa 100 značajki treniramo rijetki jezgreni stroj s Gaussovim jezgrama. Sve Gaussove jezgre imaju istu varijancu. Nakon treniranja, dobivamo model koji ima 28 prototipa. **Koliko ovaj model ima hiperparametara, koliko parametara moramo optimirati te koliko parametara ima naučeni model?**

- A Model nema hiperparametara, optimiramo 1001 parametara, a naučeni model ima 2857 parametara
- B Model ima 2800 hiperparametara, optimiramo 101 parametar, a naučeni model ima 29 parametara
- C Model ima 1 hiperparametar, optimiramo 1001 parametar, a naučeni model ima 2829 parametara
- D Model 100 hiperparametara, optimiramo 2800 parametara, a naučeni model ima 2801 parametar

$$28 \text{ prototipa} \times 100 \text{ značajki} + 28 \text{ alfa} + w_0 = 2829$$

$$\text{optimiramo novo } \alpha_{\text{fj}} + w_0$$

2. (T) Neke jezgre funkcije nazivamo Mercerove jezgre ili pozitivno definitne jezgre. Takve jezgre daju pozitivno definitnu Gramovu matricu. **Zašto je dobro da je jezgrena funkcija Mercerova jezgra?**

- A Zato što takva jezgra inducira Hilbertov prostor, tj. prostor beskonačnodimenzijskih značajki, što nam daje potencijalno vrlo složene modele
- B Zato što takva jezgra omogućava da, umjesto da vektoriziramo primjere, klasifikaciju određujemo na temelju sličnost između primjera i prototipnih primjera
- C Zato što takva jezgra nužno daje nenegativne vrijednosti sličnosti između parova primjera, što je nužno kako gubitak ne bi bio negativan
- D Zato što takva jezgra odgovara skalarnom produktu u nekom prostoru značajki, a to je nužno da bismo mogli primijeniti jezgreni trik

3. (N) Treniramo SVM s polinomijalnom jezgrom definiramo kao:

$$\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z})^2$$

Ova jezgra je Mercerova jezgra, što znači da postoji funkcija ϕ takva da $\kappa(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T \phi(\mathbf{z})$. Konkretno, u slučaju dvodimenziskog ulaznog prostora ($n = 2$), ova jezgra odgovara preslikavanju u šesterodimenzinski prostor. Međutim, postoji odstupanje u konkretnim koeficijentima polinoma. Razmotrite primjer $\mathbf{x} = (1, 0)$ te izračunajte $\phi_\kappa(\mathbf{x})$, koji dobivamo preslikavanjem definiranim implicitno preko jezgre, te $\phi_p(\mathbf{x})$, koji dobivamo preslikavanjem definiranim kao polinom drugog stupnja. **Koliko iznosi euklidska udaljenost između $\phi_\kappa(\mathbf{x})$ i $\phi_p(\mathbf{x})$?**

- A 0 B $2\sqrt{2}$ C 4 D $\sqrt{2}$

primjer 12 predavanja
sve ostalo $\phi : \phi_0, \phi_1, \phi_2, \dots$

$$\phi_p(\mathbf{x}) = \left(\frac{x_1^2}{2}, \frac{\sqrt{2}x_1 x_2}{1}, \frac{x_2^2}{3} \right)$$

$$\phi_p(\mathbf{x}) = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2) \rightarrow 6D$$

$$\phi_\kappa(x^{(1)}) = (\phi_0, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5)$$

$$\phi_p(x^{(1)}) = (1, 1, 0, 0, 1, 0)$$

$$d = \sqrt{1+1+0} = \sqrt{2}$$

3. (N) Treniramo SVM s polinomijalnom jezgrom definiramo kao:

$$\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z})^2$$

Ova jezgra je Mercerova jezgra, što znači da postoji funkcija ϕ takva da $\kappa(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T \phi(\mathbf{z})$. Konkretno, u slučaju dvodimenziskog ulaznog prostora ($n = 2$), ova jezgra odgovara preslikavanju u šesterodimenzinski prostor. Međutim, postoji odstupanje u konkretnim koeficijentima polinoma. Razmotrite primjer $\mathbf{x} = (1, 0)$ te izračunajte $\phi_\kappa(\mathbf{x})$, koji dobivamo preslikavanjem definiranim implicitno preko jezgre, te $\phi_p(\mathbf{x})$, koji dobivamo preslikavanjem definiranim kao polinom drugog stupnja. **Koliko iznosi euklidska udaljenost između $\phi_\kappa(\mathbf{x})$ i $\phi_p(\mathbf{x})$?**

- A 0 B $2\sqrt{2}$ C 4 D $\sqrt{2}$

$$k(x, z) = (x^T z)^2 \quad x = (x_1, x_2) \\ z = (z_1, z_2)$$

$$\begin{aligned} &= ((x_1, x_2)^T (z_1, z_2))^2 \\ &= (x_1 z_1 + x_2 z_2)^2 \\ &= x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\ &= x_1^2 z_1^2 + \sqrt{2} x_1 x_2 \sqrt{2} z_1 z_2 + x_2^2 z_2^2 \\ &= (x_1^2, \sqrt{2} x_1 x_2, x_2^2)^T (z_1^2, \sqrt{2} z_1 z_2, z_2^2) \\ &= \phi(x) \phi(z) \end{aligned}$$

$$\Phi_K = (\phi \ \phi \ \phi \ \sqrt{2} x_1 x_2 \ x_1^2 \ x_2^2)$$

$$\Phi_P = (1 \ x_1 \ x_2 \ x_1 x_2 \ x_1^2 \ x_2^2) \quad 60$$

$$\Phi_K = (0 \ 0 \ 0 \ 0 \ 1 \ 0)$$

$$\Phi_P = (1 \ 1 \ 0 \ 0 \ 1 \ 0)$$

$$d = \sqrt{1+1} = \sqrt{2}$$

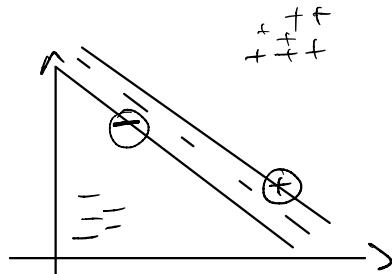
DZ
1. [Svrha: Razumjeti potrebu za mekom marginom. Znati izvesti problem meke margine SVM-a preko Lagrangeove dualnosti.]

- Objasnite motivaciju za uvođenje meke margine. Skicirajte primjer prenaučenosti kod tvrde margine, i to za linearno odvojiv i linearne neodvojiv slučaj.
- Formulirajte problem optimizacije meke margine.
- Definirajte dualni kvadratni problem za meku marginu.
- Krenuvši od uvjeta KKT, dokažite da potporni vektori za koje vrijedi $0 < \alpha_i < C$ leže na margini, a da vektori za koje $\alpha_i = C$ leže na margini ili se nalaze unutar nje.

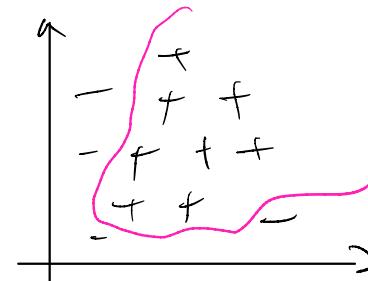
a) Za početak, nećemo uvijek imati linearne odvojiv slup; možemo ga naravno učiniti odvojivim preslikavanjem u prostor značajki, ali to bi nas moglo dovesti do prenaučenih modela. Oni će loše generalizirati jer u podacima često imamo šum, a naš će model na njega reagovati.

→ Želimo kompromis → dopuštamo da dio primjera ulijede u marginu te one postaju sira i poroznata.

LINEARNO ODVOJIV



UNIJEKNO NEODVOJIV



b) Problem optimizacije meke margine

→ uodimo rezervnu var. ξ_i , po jednoj za svaki primjer

$$y^{(i)} (w^\top x^{(i)} + w_0) \geq 1 - \xi_i \quad i=1, \dots, N$$

→ sada je ciljna funkcija koju želimo minimizirati:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

↳ hiperparan

kompromis između
sinine marge i kazne
 $C = \frac{1}{\lambda}$

$$\Rightarrow \underset{w, w_0, \xi}{\operatorname{argmin}} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \right\}$$

C veći
→ veća dozvoljena

u2 ograničenja

$$y^{(i)} (w^\top x^{(i)} + w_0) \geq 1 - \xi_i, \quad i = 1, \dots, N$$

$$\xi_i \geq 0, \quad i = 1, \dots, N$$

c) Dualni kvadratni problem mješane marge

$$\underset{\alpha}{\operatorname{argmax}} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^\top x^{(j)}$$

→ isto kao i za vrdu, ali uodimo novo ograničenje:

- budući da je $\beta_i \geq 0$

$$\text{iz } \frac{\partial L}{\partial \xi_i} = 0 \quad \alpha_i = C - \beta_i$$

$$\Rightarrow \beta_i = C - \alpha_i$$

$$\Rightarrow C - \alpha_i \geq \beta_i; \quad 0 \leq \beta_i \leq C$$

$$C - \alpha_i \geq 0 \Rightarrow \alpha_i \leq C \Rightarrow \alpha_i \leq C; \quad \alpha_i \geq 0$$

$$\Rightarrow \boxed{0 \leq \xi_i \leq C}$$

$$(\dagger) \quad \sum_{i=1}^n \xi_i y^{(i)} = \phi$$

d) za potporne vektore uvjedi: $\xi_i > 0$

\rightarrow kod male margene ξ_i dodatno omrežje $0 \leq \xi_i \leq C$

$$\hookrightarrow \text{potporu: } \boxed{0 < \xi_i \leq C}$$

$$\rightarrow \text{uvjedi } \boxed{\xi_i = C - \beta_i} \quad \rightarrow \text{kompl. ljestvost: } \beta_i \xi_i = \phi$$

$$\underline{\xi_i < C}$$

\rightarrow na margini

$$\text{ako je } \xi_i < C \\ \text{iz } \xi_i = C - \beta_i \quad \left. \begin{array}{l} \\ \end{array} \right\}$$

$$C - \beta_i < C$$

$$-\beta_i < 0$$

$$\beta_i > \phi \quad (\dagger) \quad \beta_i \xi_i = \phi \rightarrow \boxed{\xi_i = \phi}$$

$$y^{(i)} \left\{ w^T x^{(i)} + w_0 \right\} \geq 1 - \xi_i$$

ako leži

$$y^{(i)} (w^T x^{(i)} + w_0) = 1$$

$$\underline{\xi_i = C}$$

\rightarrow mogu ležati i unutar marge

$$\underline{\xi_i = C}$$

$$\underline{\xi_i = C - \beta_i}$$

$$\underline{C = C - \beta_i}$$

$$\underline{\beta_i = \phi}$$

$$\underline{\xi_i > \phi}$$

\downarrow
primjer ne margini
ili unutar nje

3. [Svrha: Razumjeti jezgreni trik kod SVM-a.]

- Za klasifikaciju primjera u ulaznom prostoru $X = \mathbb{R}^2$ koristimo polinomijalnu jezgrenu funkciju $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 1)^2$. Pokažite da je za $n = 2$ jezgra κ Mercerova jezgra. Zašto je to bitno?
- Izvedite pripadno preslikavanje $\phi(\mathbf{x})$ za $n = 2$. U koji vektor će efektivno biti preslikan primjer $\mathbf{x} = (2, 3)$ primjenom jezgre $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 1)^2$?
- Kada će broj parametara neparametarske inačice ovog modela za $n = 2$ biti veći od broja parametara njegove parametarske inačice? (U oba slučaja, parametri su vektori realnih brojeva.)
- Provjerite je li u dobivenom prostoru značajki XOR-problem linearne odvojiv. Objasnite. Vrijedi li isti zaključak za jezgrenu funkciju $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z})^2$?

a) $\kappa(x_1, z_2) = (x^T z + 1)^2$
 npr. $x = (x_1, x_2)$ $z = (z_1, z_2)$

$$\begin{aligned}\kappa(x, z) &= (x^T z + 1)^2 = ((x_1, x_2) + 1)^T ((z_1, z_2) + 1) \\ &= ((x_1+1, x_2+1)^T (z_1+1, z_2+1))^2 \\ &= ((x_1+1)(z_1+1) + (x_2+1)(z_2+1))^2 \\ &= (x_1+1)^2(z_1+1)^2 + \sqrt{2}(x_1+1)(x_2+1)\sqrt{2}(z_1+1)(z_2+1) + (x_2+1)^2(z_2+1)^2 \\ &= ((x_1+1)^2, \sqrt{2}(x_1+1), (x_2+1)^2)^T ((z_1+1)^2, \sqrt{2}(z_1+1)(z_2+1), (z_2+1)^2) \\ &= \phi(x)^T \phi(z)\end{aligned}$$

MERCEROVA J

b) $\phi(x) = ((x_1+1)^2, \sqrt{2}(x_1+1)(x_2+1), (x_2+1)^2)$

$$x = (2, 3)$$

$$\phi(x) = (9, 12\sqrt{2}, 16)$$

4. (N) Na skupu primjera za učenje iz ulaznog prostora $n = 4$ trenirali smo SVM s polinomijalnom jezgrenom funkcijom $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 2)^3$. Potporni vektori i njihove oznake su sljedeći:

$$\begin{aligned}(\mathbf{x}^{(1)}, y^{(1)}) &= ((9, 30, 21), -1) \\(\mathbf{x}^{(2)}, y^{(2)}) &= ((-11, -26, -15), -1) \\(\mathbf{x}^{(3)}, y^{(3)}) &= ((-1, -7, -6), +1)\end{aligned}$$

Lagrangeovi koeficijenti su $\alpha_1 = 2.214 \cdot 10^{-8}$, $\alpha_2 = 3.803 \cdot 10^{-8}$ i $\alpha_3 = 6.017 \cdot 10^{-8}$. Upotrijebite jezgreni trik da biste odredili vrijednost hipoteze $h(\mathbf{x})$ za primjer $\mathbf{x} = (3, 0, -3)$.

- A -2.330 B -0.676 C +0.947 D +1.434

$$\begin{aligned}h(\mathbf{x}) &= \sum_{i=1}^n \alpha_i y^{(i)} \mathbf{x}^T \mathbf{x}^{(i)} + w_0 \\&= \sum_{i=1}^n \alpha_i y^{(i)} \kappa(\mathbf{x}, \mathbf{x}^{(i)}) + w_0 \\&= 2.214 \cdot 10^{-8} (-1) \cdot (-39304) \\&\quad + 3.803 \cdot 10^{-8} (-1) \cdot 2744 \\&\quad + 6.017 \cdot 10^{-8} (1) \cdot 4913 + w_0 = 1.06 \cdot 10^{-3} + 0.945 \\&= 0.946\end{aligned}$$

$$\begin{aligned}\kappa(\mathbf{x}_1, \mathbf{z}) &= (\mathbf{x}^T \mathbf{z} + 2)^3 \\&= ([3 \ 0 \ -3] \begin{bmatrix} 9 \\ 30 \\ 21 \end{bmatrix} + 2)^3 \\&= (27 - 63 + 2)^3 = (-34)^3 = -39304 \\[10pt]\kappa(\mathbf{x}_2, \mathbf{z}) &= ([3 \ 0 \ -3] \begin{bmatrix} -11 \\ -26 \\ -15 \end{bmatrix} + 2)^3 = (-33 + 65 + 2)^3 \\&= 14^3 = 2744 \\[10pt]\kappa(\mathbf{x}_3, \mathbf{z}) &= ([3 \ 0 \ -3] \begin{bmatrix} -1 \\ -7 \\ -6 \end{bmatrix} + 2)^3 = (-3 + 18 + 2)^3 \\&= 17^3 = 4913\end{aligned}$$

$$w_0 = \frac{1}{|S|} \left(\sum_{j \in S} y^{(j)} - \sum_{j \in S} y^{(j)} \underbrace{\left(\underbrace{(x^{(j)})^T}_{\kappa(x, z)} x^{(j)} \right)}_{\kappa(x, z)} \right)$$

(x1)

$$x_1: 2,216 \cdot 10^{-8} \cdot (-1) \left([9 \ 30 \ 21] \begin{bmatrix} 9 \\ 30 \\ 21 \end{bmatrix} + 2 \right)^3$$

$$= -2,216 \cdot 10^{-8} (81 + 900 + 441 + 2)^3 = -63,93$$

$$x_2 = 3,803 \cdot 10^{-8} (-1) \left([9 \ 30 \ 21] \begin{bmatrix} -11 \\ -26 \\ -15 \end{bmatrix} + 2 \right)^3 = 64,41$$

$$x_3 = 6,017 \cdot 10^{-8} \cdot 1 \left([9 \ 30 \ 21] \begin{bmatrix} -1 \\ -7 \\ -6 \end{bmatrix} + 2 \right)^3 = -2,43$$

$$-1 + 1,95 = \boxed{0,95}$$

$$\sum -1,95$$

(x2)

$$x_1: 2,216 \cdot 10^{-8} \cdot (-1) \left([-11 \ -26 \ -15] \begin{bmatrix} 9 \\ 30 \\ 21 \end{bmatrix} + 2 \right)^3 = 37,498$$

$$x_2 = 3,803 \cdot 10^{-8} (-1) \left([-11 \ -26 \ -15] \begin{bmatrix} -11 \\ -26 \\ -15 \end{bmatrix} + 2 \right)^3 = -40,83$$

$$x_3 = 6,017 \cdot 10^{-8} \cdot 1 \left([-11 \ -26 \ -15] \begin{bmatrix} -1 \\ -7 \\ -6 \end{bmatrix} + 2 \right)^3 = 1,393$$

$$-1 + 1,939 = \boxed{0,939}$$

$$\sum -1,939$$

$$x_3 = 2,214 \cdot 10^{-8} (-1) \begin{bmatrix} -1 & -7 & -6 \end{bmatrix} \begin{bmatrix} 9 \\ 30 \\ 21 \end{bmatrix} + 2)^3 = 0,893$$

$$3,803 \cdot 10^{-8} (-1) \begin{bmatrix} -1 & -7 & -6 \end{bmatrix} \begin{bmatrix} -11 \\ -26 \\ -15 \end{bmatrix} + 2)^2 = -0,88$$

$$6,017 \cdot 10^{-8} \cdot 1 \begin{bmatrix} -1 & -7 & -6 \end{bmatrix} \begin{bmatrix} -1 \\ -7 \\ -6 \end{bmatrix} + 2)^2 = 0,041$$

$$1 - 0,054 = \boxed{0,946}$$

$$\sum 0,054$$

$$W_0 = \frac{0,95 + 0,939 + 0,946}{3} = 0,945$$

5. (P) Neka je $\mathcal{H}_{C,\gamma}$ model SVM-a s Gaussovom jezgrom. Hiperparametri tog modela su regularizacijski faktor C i preciznost jezgre γ . Odabir modela provodimo unakrsnom provjerom i to pretraživanjem po rešetci za sljedeće vrijednosti hiperparametara:

$$C = \{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, \underbrace{1, 2^1, 2^2, 2^3, 2^4, 2^5}\} \xrightarrow{\text{PRENAUČEN}} \textcircled{5}$$
$$\gamma = \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, \underbrace{1, 10^1, 10^2, 10^3, 10^4, 10^5}\} \xrightarrow{\text{PRENAUČEN}} \textcircled{5}$$

Za model sa $C = 1, \gamma = 1$ utvrdili smo da je prenaučen. **Koliko modela od ovih koje ćemo još ispitati će sigurno također biti prenaučeni?**

- A 10 B 35 C 65 D 95

$$6 \cdot 6 - 1 = 35$$

8 mali
prenač.

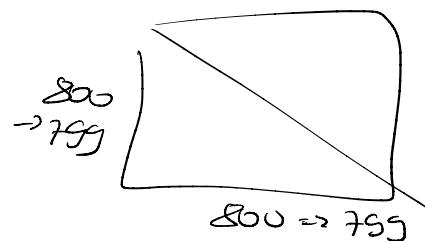
6. (T) Treniramo model SVM s nekom jezgrenom funkcijom. Nakon što smo naučili model na skupu primjera, za neki primjer x želimo izračunati udaljenost tog primjera od hiperravnine u prostoru značajki. **Je li moguće izračunati tu udaljenost?**

- A Ne, jer u dualnoj (neparametarskoj) formulaciji problema maksimalne margine nemamo vektor značajki
- B Da, ako nismo koristili Gaussovou jezgru ili neku složeniju jezgru koja koristi Gaussovou jezgru kao gradivni blok
- C Ne, jer granica između klase u prostoru značajki općenito ne mora biti hiperravnina, već može biti hiperpovršina
- D Da, ako smo koristili linearnu jezgru, odnosno ako je ulazni prostor jednak prostoru značajki

7. (P) Na skupu od $N = 1000$ primjera rješavamo problem višeklasne klasifikacije u $K = 4$ klase. Dvije klase imaju svaka po 400 primjera, a dvije svaka po 100 primjera. Razmatramo bismo li koristili SVM u shemi OVO ili SVM u shemi OVR. Model treniramo s jezgrenom funkcijom, no zbog ograničenja na raspoloživu računalnu memoriju moramo pripaziti da Gramova matrica ne postane prevelika. Prisjetite se da je Gramova matrična simetrična, pa je dovoljno pohraniti samo polovicu matrice (bez dijagonale). **Koji je u ovom slučaju najveći omjer veličine Gramove matrice za sheme OVO i OVR?**

- A OVO:OVR $\approx 1:3$ B OVO:OVR $\approx 1:405$ C OVO:OVR $\approx 4:5$ D OVO:OVR $\approx 32:50$

OVO: 400 : 400 $\Rightarrow 800$ primjera



$$100 + 100 + 400 + 400 = 1000$$

$$\frac{799 \times 799}{2} = 319200,5 \approx 32$$

OVR

$$\frac{999 \times 999}{2} = 499000,5 \approx 50$$

$$\frac{\cancel{799 \times 799}}{\cancel{999 \times 999}} = \frac{32}{50} //$$

Neparametarske metode

1. (T) Algoritam SVM može biti parametarski i neparametarski, ovisno o tome provodimo li optimizaciju u primarnoj ili dualnoj formulaciji. U oba slučaja preferiramo da je model rijedak, tj. da je nakon treniranja što više parametara postavljeno na nulu. **Kako rijetkost modela ovisi o hiperparametru C ?**

- A Što je C veći, to je neparametarski model manje rijedak, dok je parametarski to rjedi jer λ pada
- B Što je C veći, to je neparametarski model manje rijedak, dok parametarski model nije rijedak jer ima L_2 -regularizaciju a ne L_1 -regularizaciju
- C Što je C manji, to je neparametarski model rjedi, ali to nema utjecaja na rijetkost parametarskog modela jer on nema potporne vektore
- D Što je C manji, to je neparametarski model rjedi, a također je to rjedi i parametarski model jer λ raste

2. (N) Bavimo se zadatkom određivanja etimologije riječi. Zanima nas je li neka nama nepoznata riječ latinskog ili slavenskog porijeka. Zadatak rješavamo kao binarnu klasifikaciju. Prikupili smo označeni skup primjera, koji se sastoji od latinskih riječi i riječi iz svih dvanaest živućih slavenskih jezika. Npr., u našem skupu imamo: stroj 1 strues 0 tracto 0 trasa 1 gdje 1 označava da je to slavenska riječ, a 0 da je latinska. Na ovom skupu primjera treniramo algoritam k -NN (k najблиžih susjeda). Kao funkciju udaljenosti koristimo Levenshteinovu udaljenost. Levenshteinova udaljenost L između dviju riječi najmanji je broj umetanja, brisanja i zamjena jednog znaka potrebnih da se jedna riječ pretvori u drugu. Npr., $L(stroj, straja) = 2$. Razmatramo dva modela. Model h_1 je 3-NN. Model h_2 je težinski k -NN s jezgrenom funkcijom definiranom kao $\kappa(\mathbf{x}, \mathbf{x}') = 1/(1+L(\mathbf{x}, \mathbf{x}'))$. **Koja je klasifikacija riječi $x = straja$ prema modelu h_1 i h_2 ?**

- A $h_1 = h_2 = 0$ B $h_1 = h_2 = 1$ C $h_1 = 1, h_2 = 0$ D $h_1 = 0, h_2 = 1$

$$\underline{k=3}$$

$$L(\underline{stroj}, \underline{straj\dot{c}a}) = 2 \quad \textcircled{1}$$

$$h_1(x) = 1$$

$$L(\underline{strues}, \underline{straj\dot{c}a}) = 3 \quad \textcircled{2}$$

$$L(\underline{tracto}, \underline{straj\dot{c}a}) = 4 \quad \textcircled{3}$$

$$L(\underline{trasa}, \underline{straj\dot{c}a}) = 2 \quad \textcircled{4}$$

$$\kappa(\underline{stroj}, \underline{straj\dot{c}a}) = \frac{1}{1+2} = 0,33 \quad \textcircled{5}$$

$$h_2(x) = 1$$

$$= \frac{1}{1+3} = 0,25 \quad \textcircled{6}$$

$$= \frac{1}{1+4} = 0,2 \quad \textcircled{7}$$

$$= \frac{1}{1+2} = 0,33 \quad \textcircled{8}$$

3. (N) Algoritam k-NN koristimo za višeklasnu klasifikaciju riječi prema jeziku kojemu pripadaju. Skup za učenje sastoji se od sljedećih riječi i oznaka klasa:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{("water", 0), ("voda", 1), ("zrak", 1), ("luft", 2), ("feuer", 2)\}$$

Kao mjeru sličnosti između primjera koristimo jezgrenu funkciju nad znakovnim nizovima, definiranu kao $\kappa(\mathbf{x}_1, \mathbf{x}_2) = |\mathbf{x}_1 \cap \mathbf{x}_2| / |\mathbf{x}_1 \cup \mathbf{x}_2|$, gdje je su operacije unije i presjeka definirane nad skupovima slova od kojih se riječi sastoje. Npr., $\kappa("water", "voda") = 1/8 = 0.125$. Razmatramo dvije varijante algoritma: 3-NN i težinski k-NN. Kod potonjeg u obzir uzimamo sve primjere, tj. $k = N$. Odredite klasifikaciju primjera $\mathbf{x} = "zemlja"$ pomoću ova dva algoritma. U slučaju jednakih sličnosti između dva primjera, kao susjed se uzima onaj koji je u skupu \mathcal{D} naveden prvi. U slučaju izjednačenja glasova između klase, prednost se daje klasi s numerički manjom oznakom y . **U koju će klasu biti klasificiran primjer x algoritmom 3-NN, a u koju algoritmom težinski k -NN?**

- A $y = 0$ i $y = 0$ B $y = 0$ i $y = 1$ C $y = 0$ i $y = 2$ D $y = 1$ i $y = 1$

$$L("water", "zemlja") = \frac{2}{9} \quad \emptyset$$

$$L("voda", "zemlja") = \frac{1}{9} \quad \begin{matrix} 1 \\ 1 \\ \end{matrix} \quad \left. \begin{matrix} 1 \\ 1 \\ \end{matrix} \right\} ①$$

$$L("zrak", "zemlja") = \frac{2}{8}$$

$$L("luft", "zemlja") = \frac{1}{9} \quad 2$$

$$L("feuer", "zemlja") = \frac{1}{9} \quad 2$$