Primjena stroja s potpornim vektorima za analizu sentimenta korisničkih recenzija

Završni rad br. 5179

Dominik Stanojević

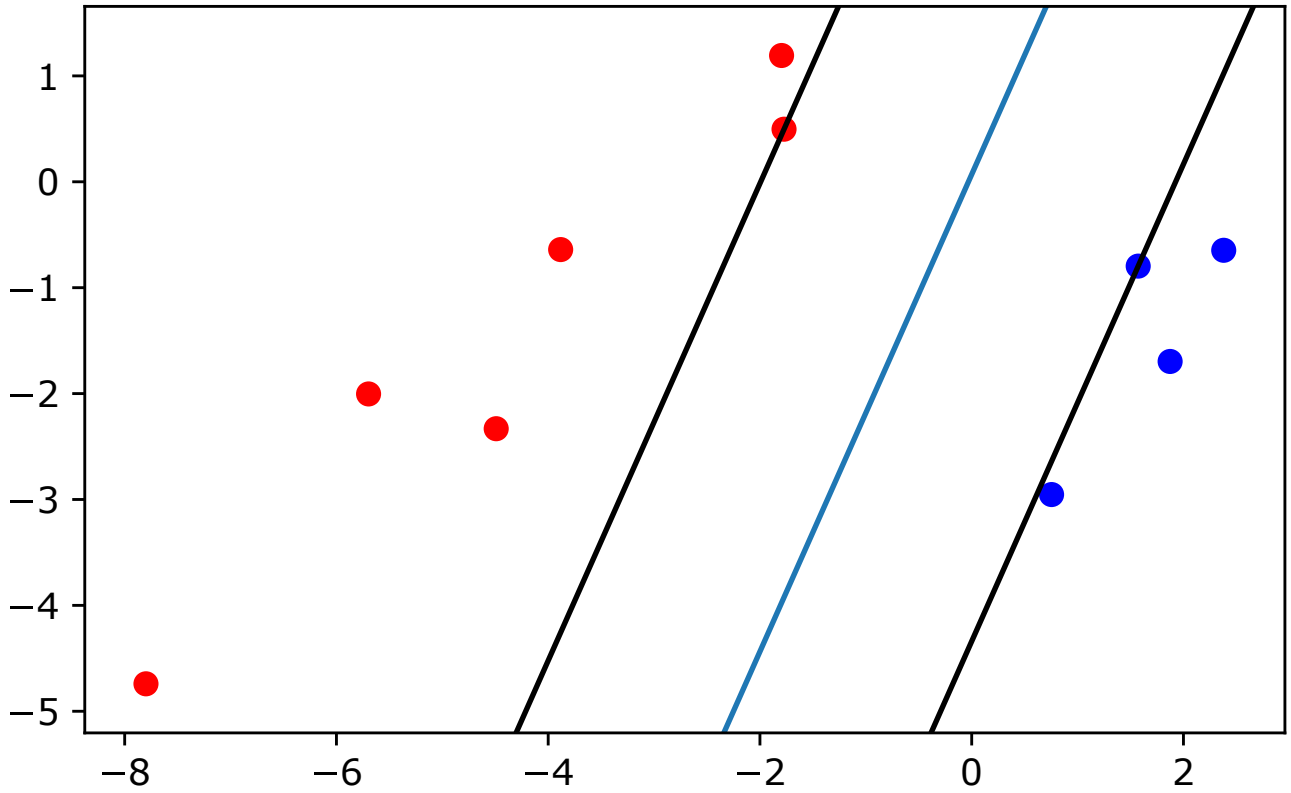
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva

3. srpnja 2017.

Sažetak

Stroj s potpornim vektorima

- Binarni klasifikator - primjerku $\mathbf{x}$ pridružena oznaka razreda $y \in \{-1, 1\}$
- Linearni klasifikator



Slika: SVM kao klasifikator

Hiperravnina

Neka je zadan vektorski prostor X dimenzije n, primjerice $\mathbb{R}^n$. Tada je **hiperravnina** definirana kao potprostor dimenzije $n-1$ unutar prostora X.

Jednadžba hiperravnine:

$$f(\mathbf{x}) = b + \mathbf{w}^T\mathbf{x} = 0$$

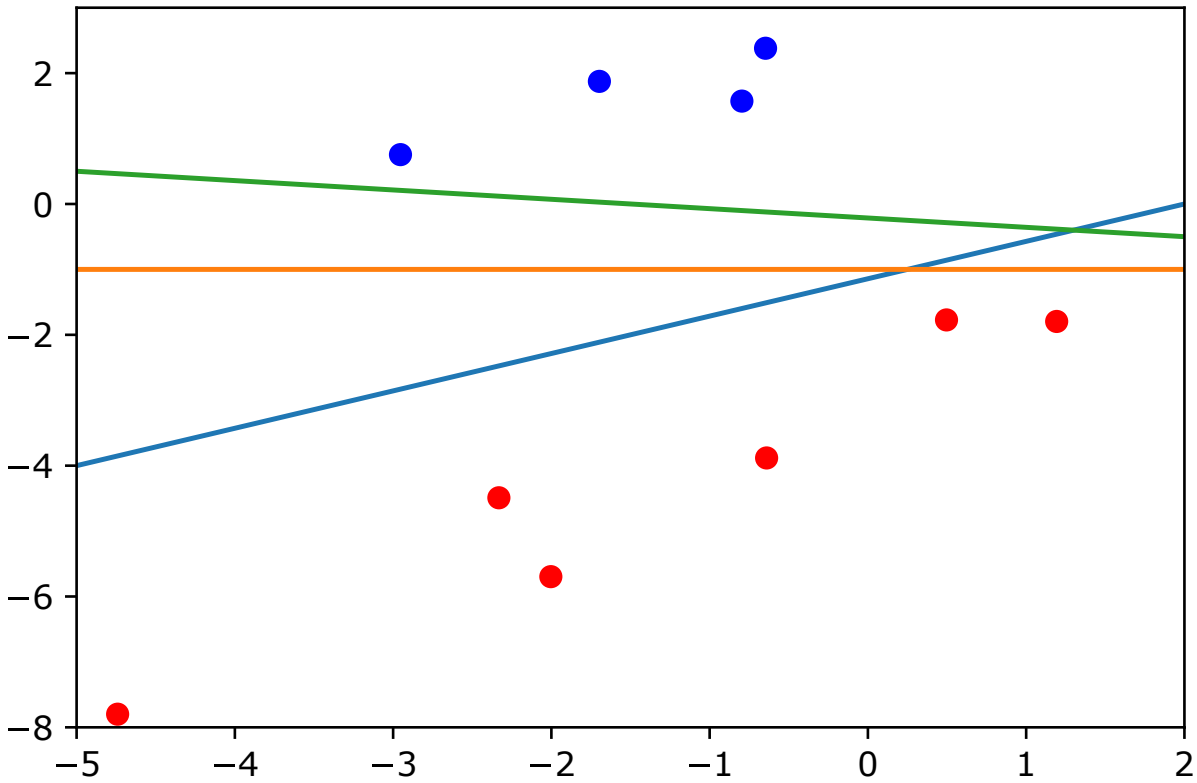
Svojstva hiperravnine:

1. za svaku točku T na hiperravnini vrijedi: $b = -\mathbf{w}^T\mathbf{x}$,
2. jedinični vektor normale je zadan izrazom: $\mathbf{n} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$,
3. udaljenost točke P od hiperravnine iznosi: $d = \frac{|f(\mathbf{x})|}{\|\mathbf{w}\|}$.

Razdvajajuća hiperravnina

Uvjet razdvajajuće hiperravnine:

$$y^{(i)} = sgn(b + \mathbf{w}^T \mathbf{x}^{(i)}), \forall i$$

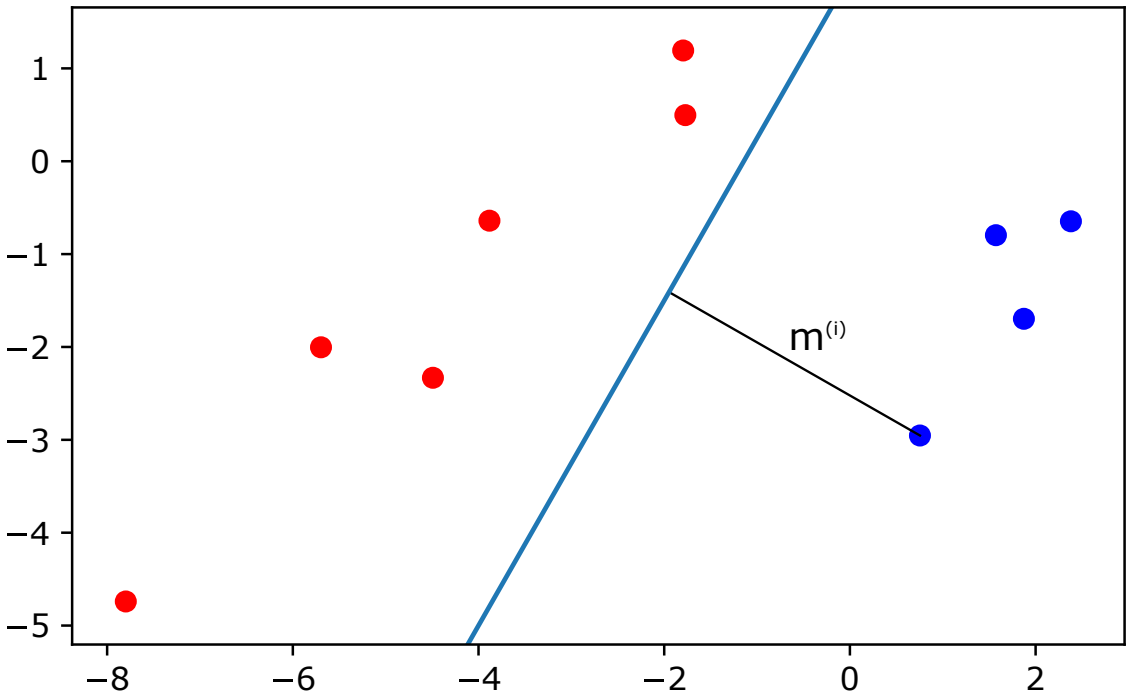


Slika: Nekoliko razdvajajućih hiperravnina

Margina razdvajajuće hiperravnine:

Margina hiperravnine:

$$M = \min_i \frac{y^{(i)}(b + \mathbf{w}^T \mathbf{x}^{(i)})}{\|\mathbf{w}\|}, y \in \{-1, 1\}.$$



Slika: Margina i-tog primjerka

Optimalna razdvajajuća hiperravnina

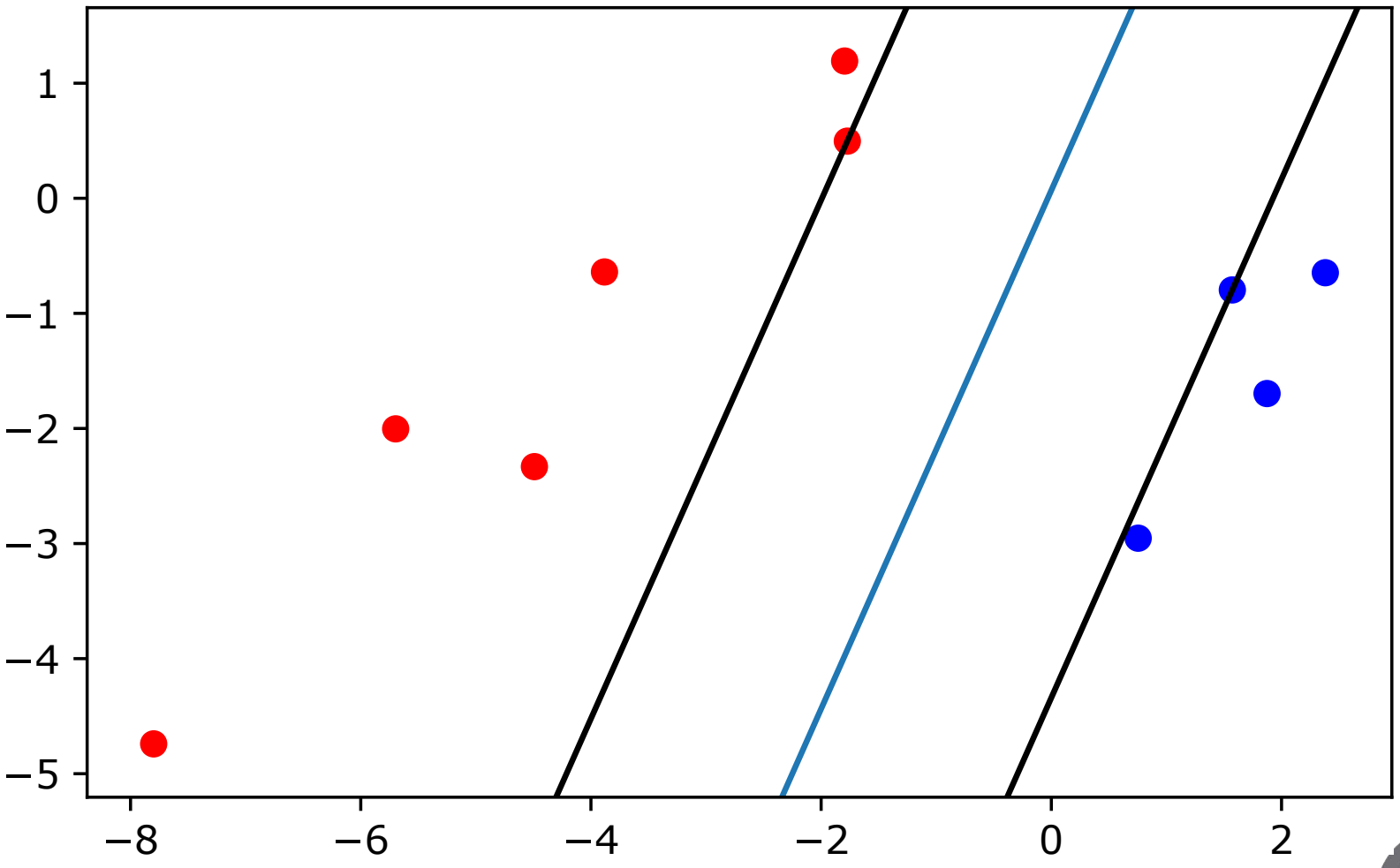
Inicijalni problem:

$$\begin{aligned} \max_{\mathbf{w},b} \quad & M \\ \text{s obzirom na} \quad & m^{(i)} \geq M,\ i = 1, \ldots, N \end{aligned}$$

Nakon sređivanja:

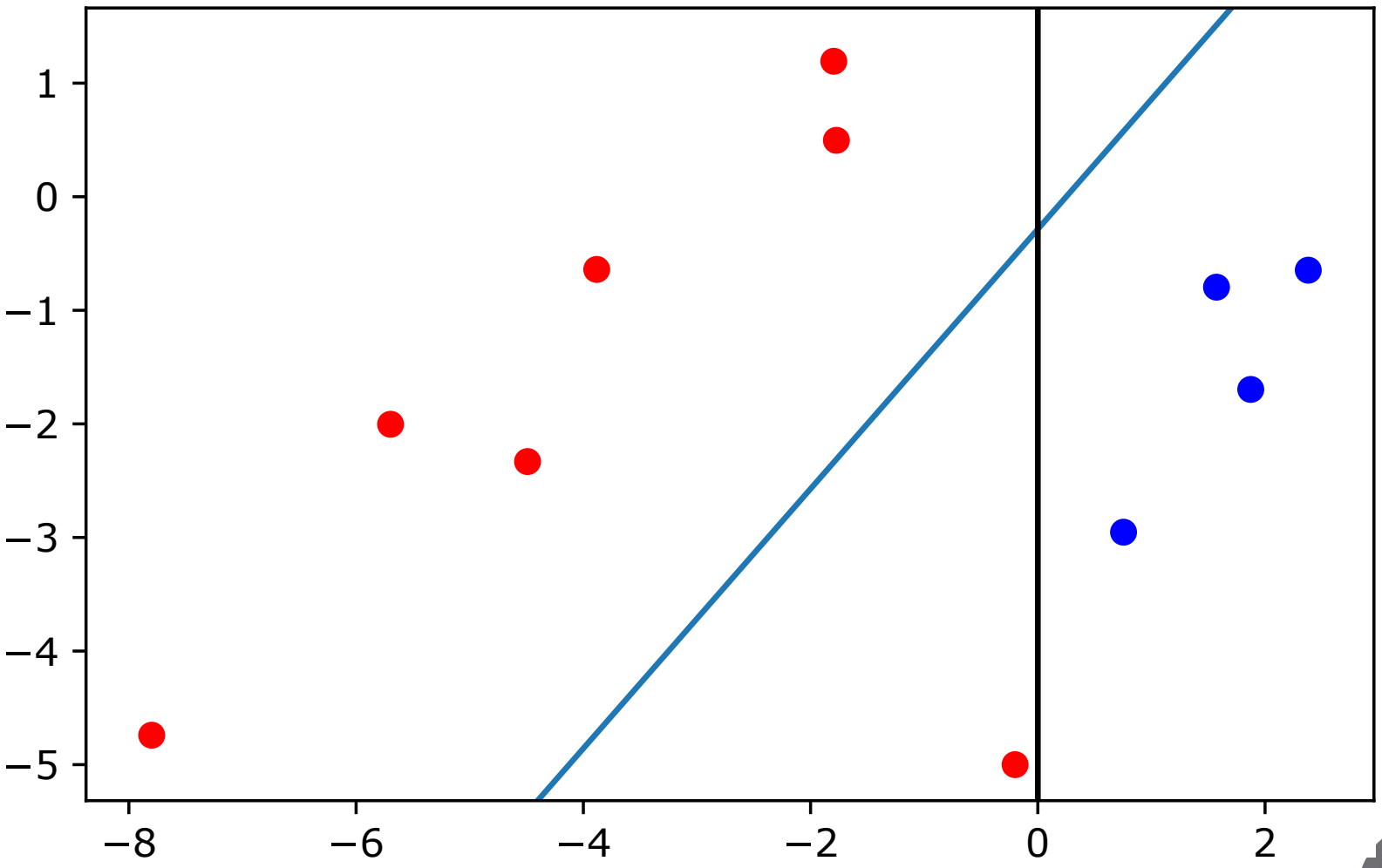
$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s obzirom na} \quad & y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1,\ i = 1, \ldots, N. \end{aligned}$$

Optimalna razdvajajuća hiperravnina



Slika: Maksimalna razdvajajuća hiperravnina

Regularizacija



Slika: Utjecaj stršeće vrijednosti na odabir hiperravnine

Regularizacija

Regularizacija je metoda kojom se sprječava pretreniranost modela koristeći funkciju kazne.

Najčešće funkcije kazne:

- L1-SVM: $\max(1 - y^{(i)}\mathbf{w}^T\mathbf{x}^{(\mathbf{i})}, 0)$
- L2-SVM: $\max(1 - y^{(i)}\mathbf{w}^T\mathbf{x}^{(\mathbf{i})}, 0)^2$

Redefiniranje optimizacijskog problema:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s obzirom na} \quad & y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi_i, \ i = 1, \ldots, N \\ & \xi_i \geq 0, \ i = 1, \ldots, N. \end{aligned}$$

Primalni i dualni optimizacijski problem

Optimizacijski problem:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s obzirom na} \quad & g_i(x) \leq 0, \ i = 1, \ldots, m \\ & h_i(x) = 0, \ i = 1, \ldots, n. \end{aligned}$$

Lagrangeova funkcija:

$$\mathcal{L}(x, \alpha, \beta) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{i=1}^{n} \beta_i h_i(x)$$

α_i, β_i nazivamo Lagrangeovim multiplikatorima.

Primalni i dualni optimizacijski problem

Primarni problem:

$$p = \min_{x} \max_{\alpha,\beta,\alpha_i \geq 0} \mathcal{L}(x, \alpha, \beta).$$

Dualni problem:

$$d = \max_{\alpha,\beta,\alpha_i \geq 0} \min_{x} \mathcal{L}(x, \alpha, \beta).$$

Interesantni slučaj:

$$p = d = \mathcal{L}(x, \alpha, \beta).$$

Karush-Kuhn-Tucker uvjeti

$$\frac{\partial \mathcal{L}(x, \alpha, \beta)}{\partial x_i} = 0, \ i = 1, \ldots, p$$

$$\frac{\partial \mathcal{L}(x, \alpha, \beta)}{\partial \beta_i} = 0, \ i = 1, \ldots, n$$

$$\alpha_i g_i(x) = 0, \ i = 1, \ldots, m$$

$$g_i(x) \leq 0, \ i = 1, \ldots, m$$

$$\alpha_i \geq 0, \ i = 1, \ldots, m$$

Optimizacija stroja s potpornim vektorima

Lagrangeova funkcija:

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y^{(i)}(\mathbf{w}^T\mathbf{x^{(i)}} + b) - (1-\xi_i)) - \sum_{i=1}^{N}\mu_i\xi_i.$$

Deriviranjem po $\mathbf{w}, b, \xi$ i sređivanjem dolazi se do dualnog problema:

$$\max_{\alpha} \quad \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^{(i)}y^{(j)}x^{(i)T}x^{(j)}$$

$$\text{s obzirom na} \quad 0 \leq \alpha_i \leq C,\ i = 1, \ldots, N$$

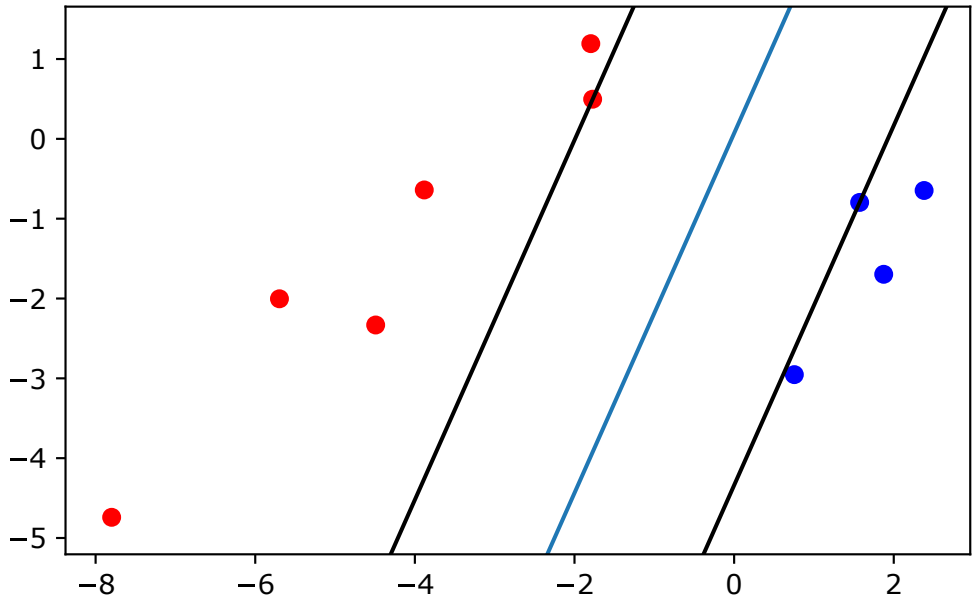
$$\sum_{i=1}^{N}\alpha_i y^{(i)} = 0.$$

Potporni vektori

Uvjet nejednakosti:

$$g_i(\mathbf{w}, b) = -(y^{(i)}(\mathbf{w}^T\mathbf{x}^{(\mathbf{i})} + b) - (1 - \xi_i)) \leq 0.$$

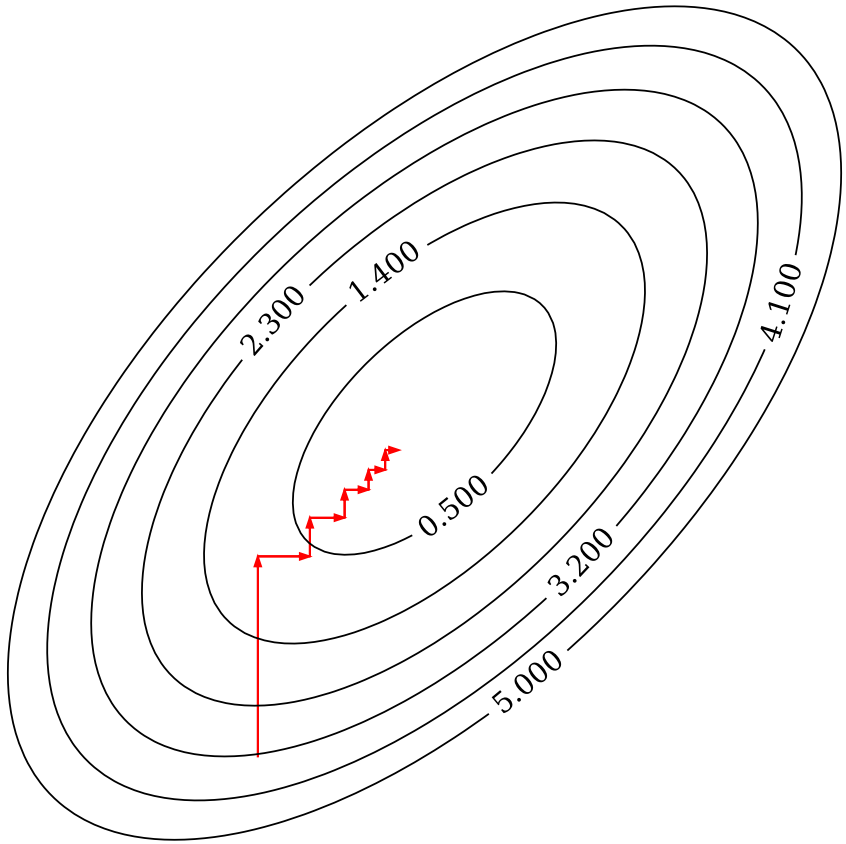
U slučaju da vrijedi $g_i(\mathbf{w}, b) = 0$ za neki primjerak $(x^{(i)}, y^{(i)})$ tada se taj primjerak naziva potpornim vektorom. Za te primjerke pripadajući Lagrangeov multiplikator je različit od nule.



Slika: Potporni vektori

Dualni koordinatni spust

- Rješavanje dualnog optimizacijskog problema
- U jednoj iteraciji osvježava se vrijednost samo jednog multiplikatora



Slika: Koordinatni spust

Dualni koordinatni spust

Optimizacijski problem:

$$\begin{aligned} \min_{\alpha} \quad & f(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T \bar{\mathbf{Q}} \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \\ \text{s obzirom na} \quad & 0 \leq \alpha_i \leq U, \ i = 1, \ldots, N. \end{aligned}$$

Za L1-SVM vrijedi $U = C$ dok za L2-SVM vrijedi $U = \infty$.
Za matricu $\bar{\mathbf{Q}}$ vrijedi izraz $\bar{\mathbf{Q}} = \mathbf{Q} + \mathbf{D}$.
Elementi matrice $\mathbf{Q}$ su $Q_{ij} = \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)}$, dok je $\mathbf{D}$ dijagonalna matrica. Za L1-SVM $\mathbf{D}_{ii} = 0$, a za L2-SVM $D_{ii} = \frac{1}{2C}$.

Dualni koordinatni spust

Vanjska iteracija osvježava sve vrijednosti Lagrangeovih multiplikatora.
Unutarnja iteracija osvježava vrijednost jednog multiplikatora:

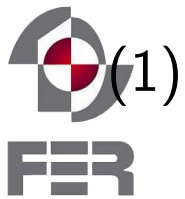
$$\begin{aligned} \min_{d} \quad & f(\boldsymbol{\alpha}_i^k + d\mathbf{e}_i) = \frac{1}{2}\bar{Q}_{ii}d^2 + \nabla_i f(\boldsymbol{\alpha}_i^k)d + C \\ \text{s obzirom na} \quad & 0 \leq \alpha_i + d \leq U,\ i = 1, \ldots, N \end{aligned}$$

Osvježavanje vrijednosti multiplikatora:

$$\alpha_i^{k+1} = \min(\max(0, \alpha_i^k - \frac{\nabla_i f(\boldsymbol{\alpha}_i^k)}{\bar{Q}_{ii}}), U)$$

Direktno osvježavanje težina:

$$\mathbf{w} = \mathbf{w} + (\alpha_i^{k+1} - \alpha_i^k)y^{(i)}\mathbf{x}^{(i)} \tag{1}$$



Jezgreni trikovi

Klasifikacija primjerka koristeći Lagrangeove multiplikatore:

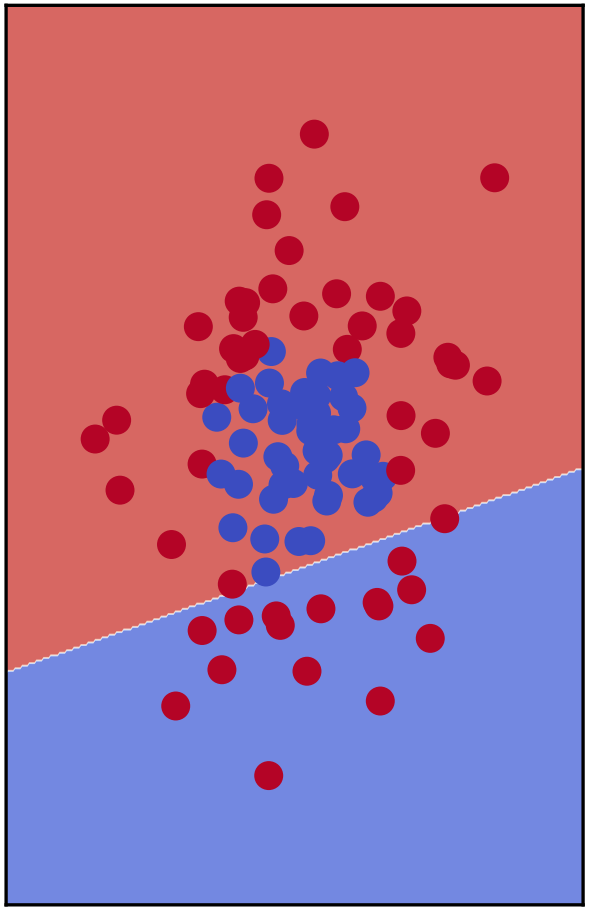
$$\hat{y} = sgn(\sum_{i=1}^{N} \alpha_i y^{(i)} \langle \phi(x), \phi(x^{(i)}) \rangle + b).$$

Zamjena umnoška s jezgrenom funkcijom $K(x, x')$. Primjer:

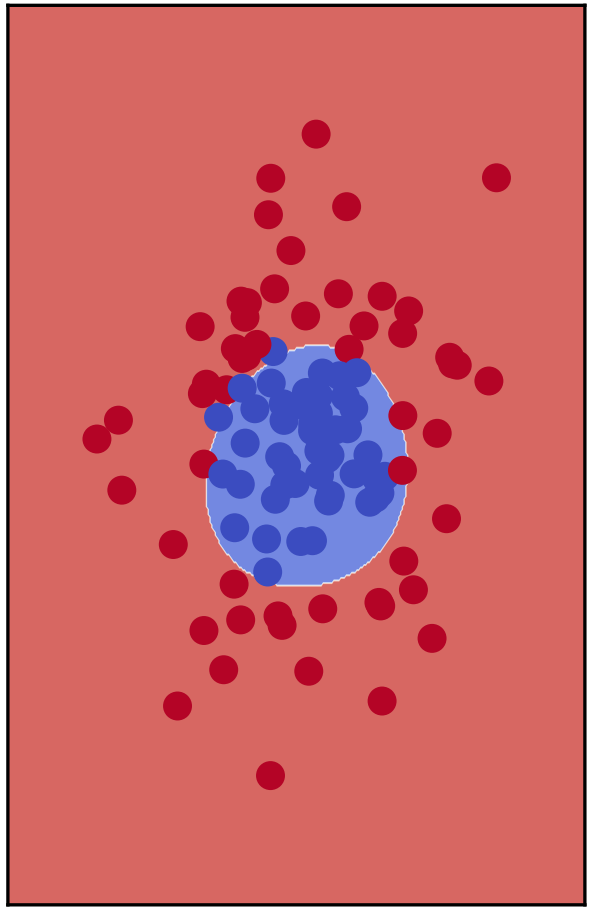
$$K(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{2\sigma^2})$$

Jezgreni trik

Linearna jezgra



Radijalna jezgra



Slika: Klasifikacija podataka bez jezgrenog trika i s RBF

Višerazredna klasifikacija

1. Jedan-naspram-ostalih - $N - 1$ klasifikatora
2. Jedan-naspram-jedan - $\frac{N(N-1)}{2}$ klasifikatora

Primjena SVM-a na regresijske probleme

Optimizacijska funkcija:

$$f(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i^+ + \xi_i^-)$$

Uz uvjete:

$$y^{(i)} - (\mathbf{w}^T\mathbf{x}^{(i)} + b) \leq \epsilon + \xi_i^+, \forall i$$

$$\mathbf{w}^T\mathbf{x}^{(i)} + b - y^{(i)} \leq \epsilon + \xi_i^-, \forall i$$

Postupak pretvorbe u dualni problem jednak kao i slučaju klasifikacije.

Definicija

Analiza sentimenta (engl. *Sentiment Analysis*) je područje posvećeno prikupljanju, obradi i analizi subjektivnih informacija, posebice stavova.

Objektivna rečenica: *"Danas je sunčano."*
Subjektivna rečenica: *"Volim sunčan dan."*

Uređena petorka: (*objekt*, *karakteristika*, *izvor*, *sentiment*, *vrijeme*)

Primjer analiziranog teksta

Užasan operacijski sustav. Osim osnovnih alata, upakiran je i s gomilu programa koji samo usporavaju rad sustava. Nadam se da korisnici nisu ljubitelji privatnosti jer ju s ovim operacijskim sustavom sigurno neće imati. Jedina pozitivna opcija je uvođenje radnih sati tako da me sustav rjeđe maltretira s ponovnim pokretanjem. Savjet: potražite alternativu. Marko, 2016.

(Operacijski sustav, "općenito", Marko, negativan, 2016.)
(Operacijski sustav, programski paket, Marko, negativan, 2016.)
(Operacijski sustav, razina privatnosti, Marko, negativan, 2016.)
(Operacijski sustav, vrijeme osvježavanja sustava, Marko, pozitivan, 2016.)

Tipovi mišljenja

- **Regularno mišljenje** - *Volim kolače.*
- **Komparativno mišljenje** - *"Iako volim voćne kolače, draži su mi čokoladni."*
- **Eksplicitno mišljenje** - *Volim kolače.*
- **Implicitno mišljenje** - *"Baterija na mobitelu ne traje dovoljno dugo."*.

Razine analize sentimenta

1. Analiza na razini dokumenta
2. Analiza sentimenta na razini rečenica
3. Analiza sentimenta na razini karakteristika

Analiza sentimenta na razini dokumenta

Cilj analize na razini dokumenta je pronaći sentiment osnovne informacijske jedinice, u ovom slučaju cijelog dokumenta.

Formalno: $(\_, \text{"općenito"}, \_, s, \_)$

Značajke:

- Subjektivne riječi i izrazi
- Frekvencija izraza
- Negacija

Problemi prilikom analize sentimenta

- Dvoznačnost
- Nositelji sentimenta i implicitne rečenice
- Sarkazam
- Spam

Vektorizacija teksta

Pretvorba dokumenta u vektor realnih brojeva - razumljivo SVM-u.

Koraci:

1. Leksička analiza
2. Lematizacija
3. Izbacivanje zaustavnih riječi i interpunkcijskih znakova
4. Izgradnja N-grama
5. Izvlačenje značajki teksta

Leksička analiza

Leksička analiza ili tokenizacija je postupak razdvajanja prepoznatljivih riječi, fraza i simbola. Jedna leksička jedinica zove se leksem. Leksičkom analizom tekst se pretvara u slijed leksema.

- Obrada HTML i XML oznaka - brisanje, posebni slučajevi (**, *<b>*, ...)
- Odvajanje leksema - razmaci, emotikoni, telefonski brojevi...
- Negacija - prefiks NEG_

"Ovo nije pozitivna rečenica :(." $\Longrightarrow$ [*Ovo*, *NEG_pozitivna*, *NEG_rečenica*, :(, .]

Lematizacija

Postupak svođenja riječi na kanonski oblik.

Primjer: *najlošiji* $\Longrightarrow$ *loš*

Problemi: Agresivnost, gubitak intenziteta

N-grami

N-grami su izrazi koji se sastoje od *n* slijednih leksema.

Primjer: *"Njegov savjet uzeo sam sa zrnom soli."*

- Unigrami (1-gram): [*savjet*, *uzeti*, *zrno*, *sol*]
- Bigrami (2-gram): [*savjet uzeti*, *uzeti zrno*, *zrno soli*]

Izvlačenje značajki teksta

Metode temeljne na modelu zbirki značajki (engl. *Bag of words*).

Modeli:

- Binarni - pojavnost izraza
- Frekvencijski - frekvencija pojavljivanja
- Model TF-IDF - relevantnost izraza za pojedini dokument i odnos relevantnosti između dokumenata

Primjer izvlačenja značajki

1. *"Ovo je jako pozitivna rečenica."*
2. *"Ovo je jako, jako negativna rečenica."*

Primjer za drugu rečenicu:

Model/Izraz	Ovo	je	jako	poztivna	rečenica	negativna
Binarni	1	1	1	0	1	1
Frekvencijski	1	1	2	0	1	1
TF-IDF	0.3338	0.3338	0.6676	0	0.3338	0.4691

$$\text{tfidf}(t,d) = \text{tf}(t,d)\text{idf}(t,D) = (1+\log(1+f_{t,d}))\log(1+\frac{N}{n_t})$$

Analiza korisničkih recenzija

Large Movie Dataset Review:

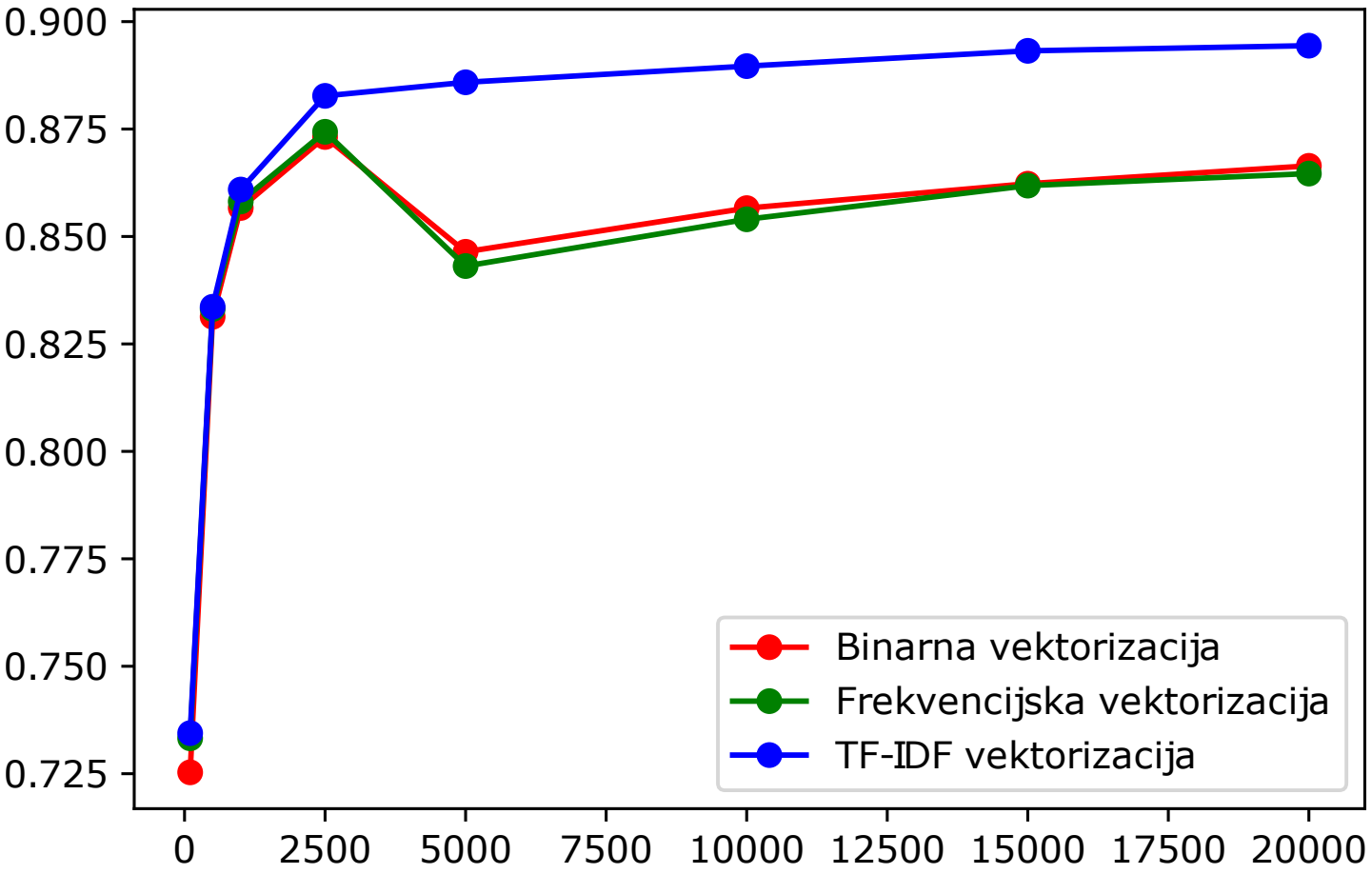
- Prikupljeno s internetske baze filmova IMDB, obrađeno od strane sveučilišta Stanford
- Više od 50000 korisničkih recenzija filmova
- Pozitivna/negativna klasifikacija
- 25000 pozitivnih recenzija; 25000 negativnih recenzija
- 25000 trening recenzija; 25000 test recenzija; dodatne neoznačene recenzije

Komponente leksičkog analizatora

Metoda	Negacija isljućena	Negacija uključena
Bez lematizacije i steminga	89.18%	89.44%
Lematizacija	89.01%	89.29%
Steming	88.9%	89.33%

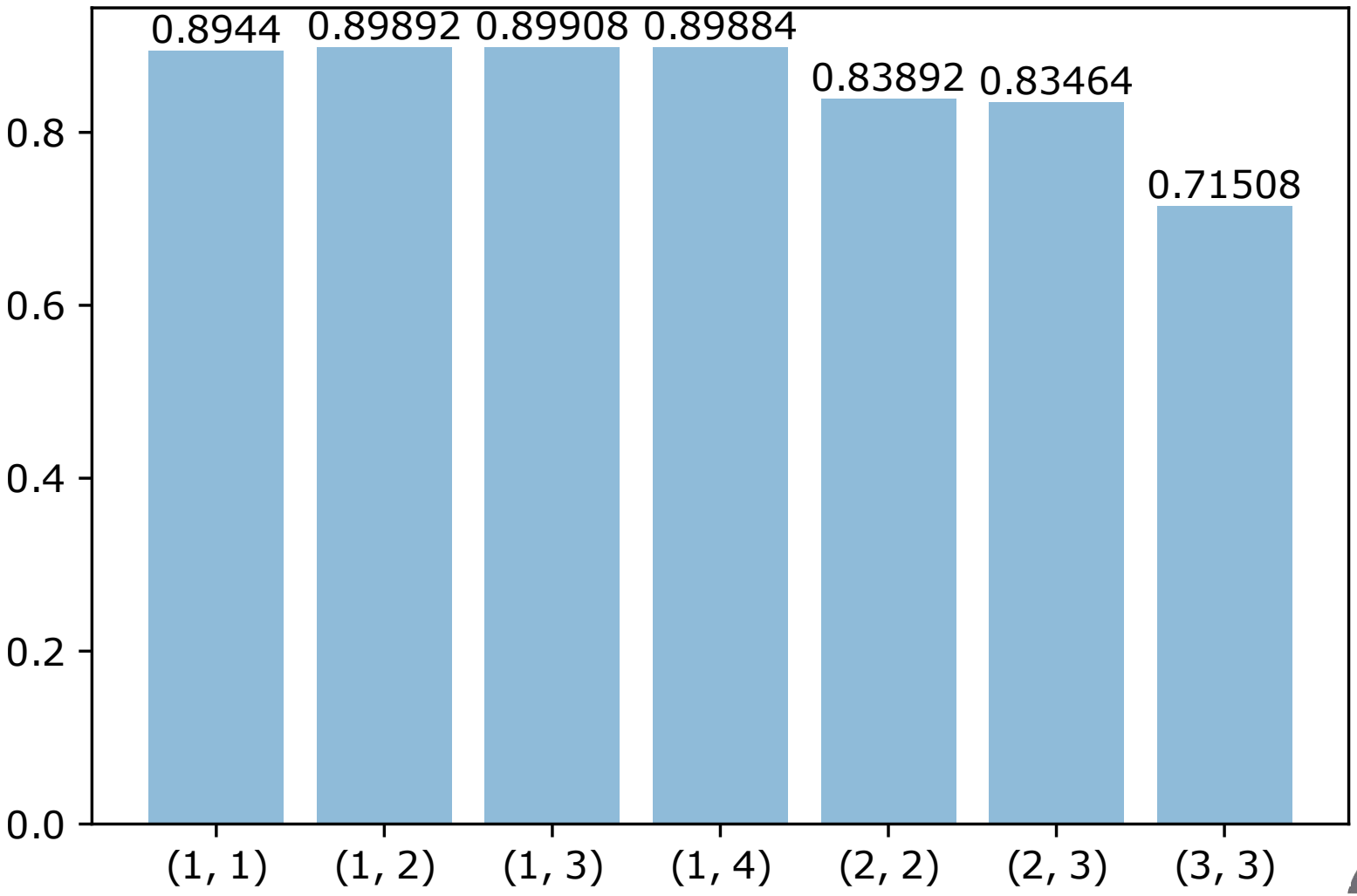
Tablica: Točnost klasifikacije recenzija u odnosu na komponente leksičkog analizatora

Model vektorizacije i dimenzionalnost



Slika: Točnost klasifikacije u odnosu na model vektorizacije i dimenziju vektora značajki

Izbor n-grama



Slika: Točnost klasifikatora u odnosu na izbor *n*-grama

Nositelji sentimenta

Rank	Leksem	Težina	**Rank**	Leksem	Težina
1	worst	-5.097	**11**	fails	-2.966
2	7/10	3.866	**12**	great	2.941
3	awful	-3.821	**13**	disappointment	-2.933
4	bad	-3.495	**14**	8/	2.857
5	excellent	3.456	**15**	disappointing	-2.805
6	dull	-3.418	**16**	poor	-2.788
7	waste	-3.413	**17**	waste_NEG	-2.779
8	4/10	-3.35	**18**	8	2.714
9	boring	-3.208	**19**	unfortunately	-2.686
10	terrible	-3.04	**20**	amazing	2.67

Tablica: Nositelji sentimenta koji najviše utječu na klasifikaciju

Primjeri pogrešno klasificiranih primjeraka

Isječak recenzije	Označeno	Klasificirano
...there's something compelling and memorable about it. Like another commenter on the film, I saw this in childhood. It's been thirty three years since 1952, but I have never forgotten the story or its ridiculously cumbersome title. See it if you have the opportunity.	0	1
...HOWEVER, understand that the self-indulgent director also had many "funny gags" that totally fell flat and hurt the movie. His "camera tricks" weren't so much tricky but annoying and stupid. IGNORE THESE AND KEEP WATCHING–it does get better. The film is fast paced, funny and worth seeing. In particular, I really liked watching the acting and mugging of Max Linder–he was so expressive and funny! Too bad he is virtually forgotten today.	1	0

Tablica: Isječci pogrešno klasificiranih recenzija

DEMO

Hvala na pažnji.