SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 5179

Primjena stroja s potpornim vektorima za analizu sentimenta korisničkih recenzija

Dominik Stanojević

Zagreb, svibanj 2017.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Kako biste uklonili ovu stranicu, obrišite naredbu `\izvornik`.

Hvala Kurtzu.

SADRŽAJ

1. Uvod

Klasifikacijski i regresijski problemi jedni su od najvažnijih problema strojnog učenja. Modeli poput linearne i logističke regresije pogodni su za jednostavnije probleme. Zahvaljujući sve većoj dostupnosti podataka i povećanju procesorske moći današnjih računala, pojavljuju se složeniji zadaci za koje navedene metode nisu efikasne.

Pojava složenijih zadataka rezultirala je i pojavom složenijih metoda koje mogu doskočiti istima. Modeli poput slučajnih šuma i modeli iz skupine dubokog učenja u mogućnosti su rješavati i složenije, nelinearne proble me.

Osim navedenih modela, još jedan model koji je sposoban efikasno obraditi nelinearne podatke je **stroj s potpornim vektorima** (engl. *Support Vector Machine*, u nastavku SVM). Koristeći jezgreni trik, stroj s potpornim vektorima uspješno razdvaja linearno nerazdvojive podatke. Iako su temeljne ideje modela predstavljene prije više od pola stoljeća, stroj s potpornim vektorima i danas je jedan od najrobusnijih modela za klasifikaciju i regresiju.

Jedan od zanimljivih problema koji dobro prikazuje robusnost SVM-a je **analiza sentimenta** (engl. *Sentiment Analysis*). Subjektivnost emocija, kontekst te velika količina podataka svakako predstavljaju izazove u rješavanju problema. Koristeći SVM, uz uvjet kvalitetnog pretprocesiranja podataka, mogu se postići zavidni rezultati u polju analize sentimenta.

U radu je predstavljen model stroja s potpornim vektorima te problem analize sentimenta. U poglavlju 2 bit će predstavljen pregled područja, povijest modela stroja s potpornim vektorima te problem analize sentimenta. U poglavlju 3 detaljnije će se obraditi model SVM. Bit će opisana motivacija i interpretacija modela. Nadalje, detaljnije će se pojasniti algoritmi optimizacije modela. U poglavlju 4 formalizirat će se problem analize sentimenta. Prikazat će se postupak pretprocesiranja podataka koji će podatke pretvoriti u oblik razumljiv SVM-u. U petom poglavlju, provest će se eksperiment analize korisničkih recenzija uporabom opisanih metoda. Ukratko će se analizirati dobiveni rezultati. Poglavlje 6 sadrži zaključak i ideje za daljnje istraživanje.

2. Pregled područja

Započinje [4]

3. Stroj s potpornim vektorima

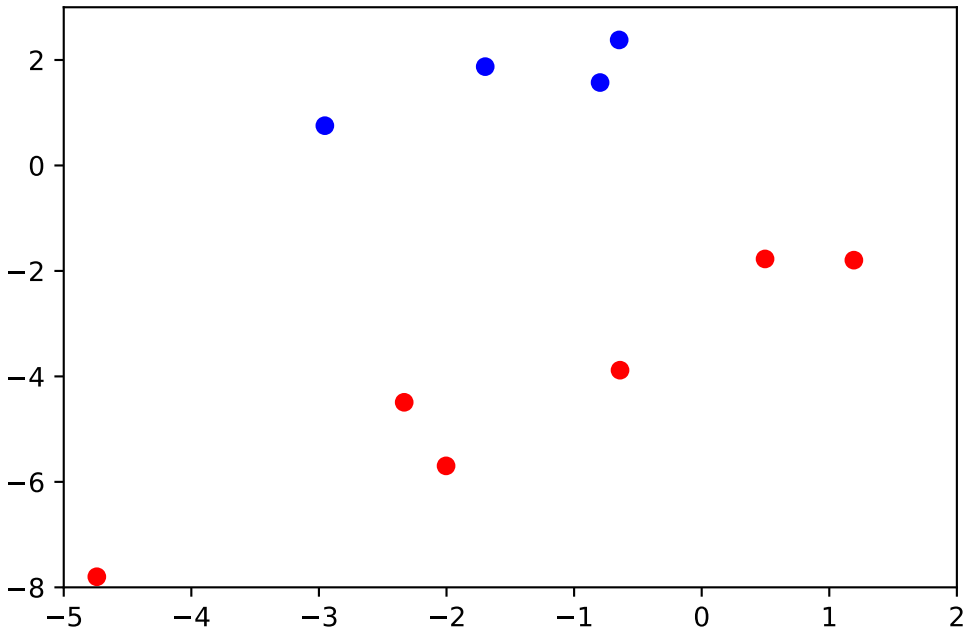
U ovom poglavlju bit će predstavljen model stroja s potpornim vektorima. Potpoglavlje 3.1 definira pojam klasifikacije i pojašnjava razliku između klasifikacije i regresije. U potpoglavlju 3.2 pojasnit će se ideja razdvajajuće hiperravnine. Potpogljavlje 3.3 predstavit će pojam margine razdvajajuće hiperravnine i njenu važnost u izgradnji klasifikatora. Koristeći ideje iz prethodnih potpoglavlja, potpoglavlje 3.4 definira optimalnu razdvajajuću hiperravninu, metodu koju koristi SVM prilikom klasifikacije podataka. Potpoglavlje 3.8 opisuje transformaciju prostora značajki koristeći jezgrene trikove. Jezgreni trikovi su efikasne metode koje omogućuju razdvajanje originalno linearno nerazdvojivih podataka. U potpoglavlju 3.5 daje se ideja regularizacije. Ova metoda omogućuje da stroj s potpornim vektorima pronađe optimalnu hiperravninu u slučaju linearno nerazdvojivih podataka.

3.1. Klasifikacija

Problemi nadziranog učenja uobičajno se dijele na dvije podskupine - klasifikaciju i regresiju. Neka vektorski prostor *X* dimenzije n, primjerice $\mathbb{R}^n$, predstavlja skup primjeraka. Pojedini primjerak može se zadati vektorom: $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Ako pojedinom primjerku $\mathbf{x}$ pridružimo oznaku razreda y, tada se govori o **klasifikaciji**. Pojednostavljeno, postupkom klasifikacije određuje se razred kojoj određen primjerak $\mathbf{x}$ pripada. Skup svih razreda *C* je konačan, a broj razreda dan je kardinalitetom $|C|$.

Primjer klasifikacijskog problema prikazan je slikom 3.1. Prostor primjeraka je $\mathbb{R}^2$, a skup razreda je dvočlani skup tj. $|C| = 2$. Klasifikacija podataka u dvočlane skupove naziva se **binarna klasifikacija**. Upravo je stroj s potpornim vektorima primjer binarnog klasifikatora, no postoje metode koje pružaju mogućnost višerazredne klasifikacije. Osim gore navedenog primjera, još neki primjeri klasifikacije su okrivanje neželjene pošte, prepoznavanje rukopisa, prepoznavanje prometnih znakova, itd..

Za razliku od problema klasifikacije u kojem je varijabli y pridružena vrijednost iz konačnog skupa, kod problema regresije primjerku pridružujemo vrijednost iz nekog beskonačnog skupa, primjerice $\mathbb{R}$. Postoji modifikacija stroja s potpornim vektorima koji omogućuje rješavanje regresijskih problema. Primjeri regresije su izračun plaće u ovisnosti o spolu,



Slika 3.1: Primjer klasifikacijskog problema

obrazovanju i sl., određivanje postotka glasova kandidata na izborima, itd.

3.2. Razdvajajuća hiperravnina

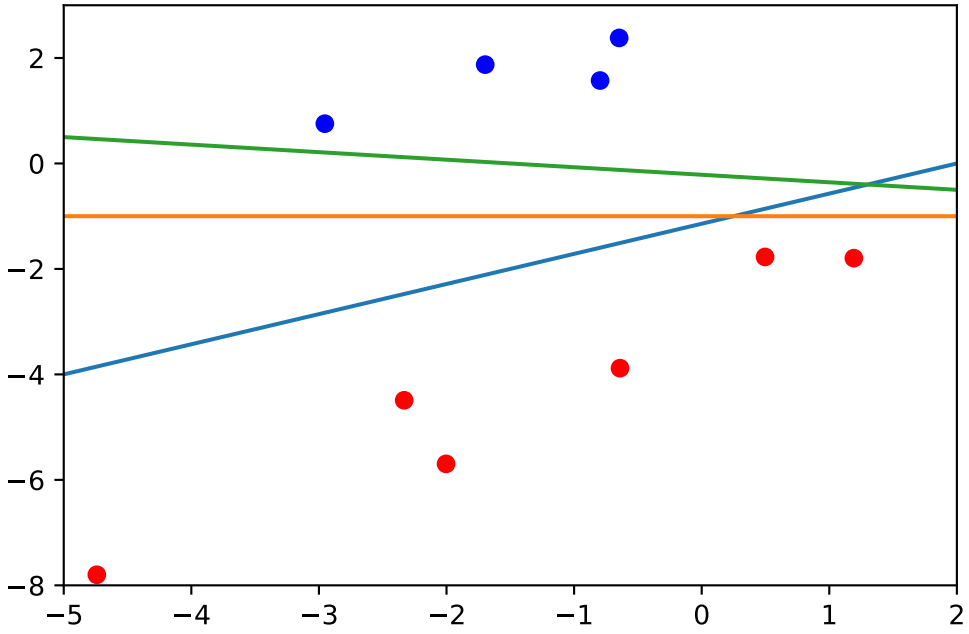
Interpretaciju modela stroja s potpornim vektorima potrebno je započeti s pojmom koji nije strogo vezan uz sam model. Primjerice model logističke regresije, iako temeljen na vjerojatnosti, u konačnici pronalazi hiperravninu kojom razdvaja podatke.

Neka je zadan vektorski prostor X dimenzije n, primjerice $\mathbb{R}^n$. Tada je **hiperravnina** definirana kao potprostor dimenzije $n-1$ unutar prostora X. Primjerice, u jednodimenzijskom prostoru hiperravnina je točka, u dvodimenzijskom prostoru hiperravninu predstavlja bilo koji pravac koji leži u ravnini, a u trodimenzijskom prostoru hiperravnina je predstavljena ravninom. Analogno, pojam hiperravnine vrijedi i za prostore većih dimenzija.

Za hiperravninu zadanom jednadžbom $f(\mathbf{x}) = b + \mathbf{w}^T\mathbf{x} = 0$ vrijede sljedeća svojstva:

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: $d = \frac{f(\mathbf{x})}{\|\mathbf{w}\|}$.

Hiperravnina ovisi o vektoru težina $\mathbf{w}$ i slobodnom članu b. U daljnjem tekstu hiperravnina bit će zadana svojim parametrima, $\langle \mathbf{w}, b \rangle$.



Slika 3.2: Razdvajajuće hiperravnine

Hiperravnine same po sebi nisu pretjerano interesantne. No, za klasifikaciju interesantan je određen podskup hiperravnina. Hiperravnina koja razdvaja dva razreda podataka naziva se **razdvajajuća hiperravnina.** Uz pretpostavku $y \in \{-1, 1\}$, za razdvajajuću hiperravninu vrijedi:

$$y^{(i)} = sgn(b + \mathbf{w}^T \mathbf{x}^{(i)}), \forall i \tag{3.1}$$

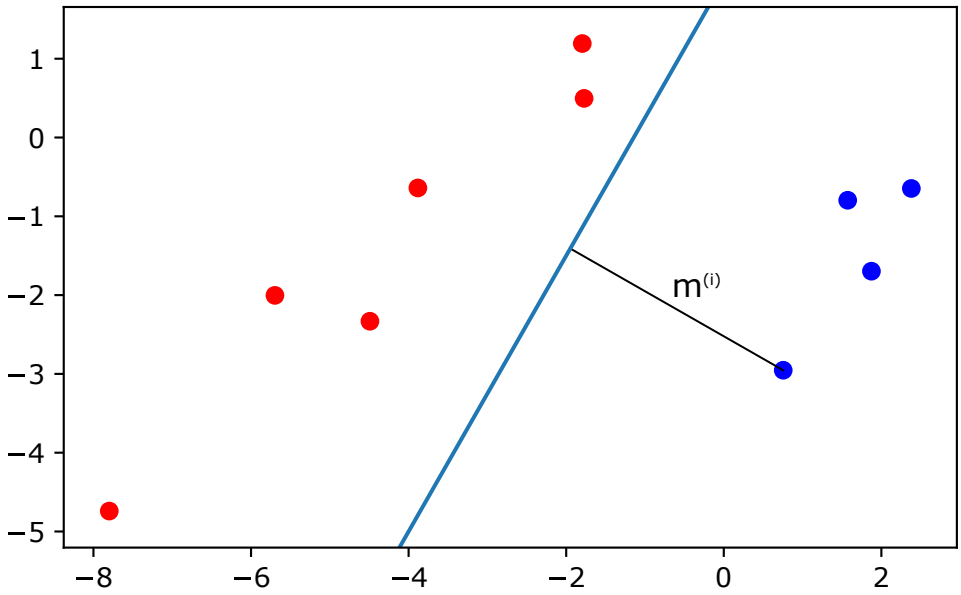
gdje je $\mathbf{x}^{(i)}$ primjer iz skupa podataka, a $y^{(i)}$ je oznaka razreda pridružena primjeru $\mathbf{x}^{(i)}$.

Na slici 3.2 prikazani su primjerci jednaki onima sa slike 3.1. Također, prikazane su i tri razdvajajuće hiperravnine. Valja primijetiti kako je moguće konstruirati beskonačno mnogo razdvajajućih hiperravnina.

3.3. Margina razdvajajuće hiperravnine

Za razliku od drugih klasifikatora koji traže bilo koju razdvajajuću hiperravninu kako bi klasificirali podatke, stroj s potpornim vektorima uzima u obzir i udaljenosti primjeraka od hiperravnine. Intuitivno se može zaključiti kako je sigurnije odrediti razred za one primjerke koji su udaljeniji od hiperravnine. Udaljenost primjerka od hiperravnine nazivamo **margina**.

Neka je zadan i-ti primjerak $(\mathbf{x}^{(i)}, y^{(i)})$ gdje je $\mathbf{x}^{(i)}$ vektor značajki, a $y^{(i)}$ pripadajuća oznaka razreda. Također, neka je $\mathbf{r^{(i)}}$ radij-vektor točke T koja se nalazi na hiperravnini i najmanje je udaljena od primjerka. Udaljenost i-tog primjerka od hiperravnine iznosi $m^{(i)}$.



Slika 3.3: Margina hiperravnine s obzirom na primjerak $(\mathbf{x}^{(i)}, y^{(i)})$

Vrijede dvije jednadžbe:

$$\mathbf{r}^{(i)} = \mathbf{x}^{(i)} - y^{(i)} m^{(i)} \frac{\mathbf{w}}{\|\mathbf{w}\|},$$

$$b + \mathbf{w}^T \mathbf{r}^{(\mathbf{i})} = 0.$$

Riješavanjem ovog sustava po $m^{(i)}$ dobiva se margina hiperravnine $\langle \mathbf{w}, b \rangle$ s obzirom na primjerak $(\mathbf{x}^{(i)}, y^{(i)})$:

$$m^{(i)} = \frac{y^{(i)}(b + \mathbf{w}^T \mathbf{x}^{(i)})}{\|\mathbf{w}\|}. \tag{3.2}$$

Na slici 3.3 prikazana je udaljenost primjerka od razdvajajuće hiperravnine. Valja uočiti kako za pozitivne oznake razreda, $y^{(i)} = 1$, vrijednost $b + \mathbf{w}^T \mathbf{x}^{(i)}$ je pozitivna. Analogno, za $y^{(i)} = -1$ vrijednost $b + \mathbf{w}^T \mathbf{x}^{(i)}$ je negativna. Može se zaključiti kako je vrijednost margine za svaki primjerak strogo pozitivna. U slučaju hiperravnine koja ne razdvaja podatke to svojstvo ne vrijedi.

Osim svojstva pozitivnosti, za marginu je zanimljiva i otpornost na skaliranje. Neka je hiperravnina $\langle \mathbf{w}, b \rangle$ skalirana nekim faktorom k. Za marginu $m'^{(i)}$ vrijedi:

$$m'^{(i)} = \frac{y^{(i)}(kb + k\mathbf{w}^T \mathbf{x}^{(i)})}{\|k\mathbf{w}\|} = \frac{y^{(i)} k (b + \mathbf{w}^T \mathbf{x}^{(i)})}{k\|\mathbf{w}\|} = m^{(i)}.$$

Ovo svojstvo omogućuje da duljina vektora težina bude proizvoljna što će se pokazati veoma korisnim kod postavljanja optimizacijskog problema.

Nakon definiranja margine hiperravnine za pojedini primjerak, potrebno je definirati i marginu hiperravnine uzimajući u obzir cijeli skup podataka za učenje. **Margina hiperravnine** u odnosu na skup podataka za učenje je margina onog primjerka koji je najbliži hiperravnini tj.

$$M = \min_i m^{(i)}.$$

3.4. Optimalna razdvajajuća hiperravnina

Nakon definicije margine, sljedeći cilj je pronaći razdvajajuću hiperravninu koja najbolje razdvaja podatke. Uzimajući u obzir činjenicu da veća udaljenost primjerka od hiperravnine pruža veću sigurnost za ispravnu klasifikaciju, intuitivno se može zaključiti kako će optimalna razdvajajuća hiperravnina biti ona koja maksimizira marginu hiperravnine s obzirom na skup primjeraka za učenje. **Optimalna razdvajajuća hiperravnina** zadana je sljedećim optimizacijskim problemom:

$$\begin{aligned} \max_{\mathbf{w},b} \quad & M \\ \text{s obzirom na} \quad & y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq M,\ i = 1, \ldots, N \\ & \|\mathbf{w}\| = 1. \end{aligned} \tag{3.3}$$

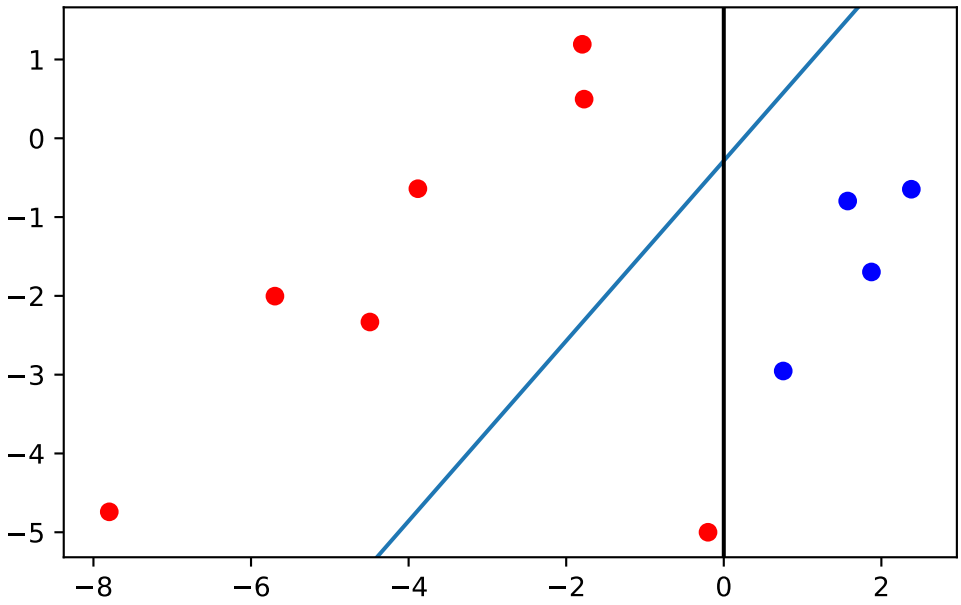
Gledajući gornji problem, može se uočiti kako su uvedena dva ograničenja. Ova ograničenja omogućuju da su svi primjerci udaljeni od hiperravnine za minimalno M. Drugo ograničenje koje normalizira vektor težina onemogućuje rješavanje problema u ovom obliku budući da uvjet $\|\mathbf{w}\| = 1$ nije konveksan. Koristeći jednadžbu 3.2, oba ograničenja mogu se svesti na jedno:

$$y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq M\|\mathbf{w}\|,\ i = 1, \ldots, N.$$

Nadalje, budući da duljina vektora težina ne utječe na marginu i klasifikaciju, moguće je proizvoljno odabrati njegovu duljinu. Za $\|\mathbf{w}\| = \frac{1}{M}$ cilj optimizacije se svodi na pronalazak maksimuma funkcije $f(\mathbf{w}) = \frac{1}{\|\mathbf{w}\|}$. Znajući kako se maksimizacija ove funkcije može svesti na minimizaciju kvadrata norme, može se pisati:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s obzirom na} \quad & y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1,\ i = 1, \ldots, N. \end{aligned} \tag{3.4}$$

Koristeći nekoliko različitih transformacija, problem je redefiniran na jednostavniji način. Budući da je dobivena konveksna kvadratna funckija cilja uz linearne uvjete ovaj problem je rješiv metodama kvadratnog programiranja. Kasnije u radu će se ovaj optimizacijski problem dodatno transformirati kako bi se iskoristili algoritmi koji rješavaju problem efikasnije od generičkog softvera za rješavanje ovakvih optimizacijskih problema.



Slika 3.4: Utjecaj stršeće vrijednosti na odabir hiperravnine

3.5. Regularizacija

U dosadašnjem radu klasificiralo se linearno razdvojive podatke. Međutim, podaci najčešće nisu linearno razdvojivi ili optimalna razdvajajuća hiperravnina nije najbolji klasifikator budući da nije otporna na stršeće vrijednosti. Slučaj kada stršeća vrijednost utječe na izbor hiperravnine prikazan je na slici 3.4. Vidljivo je kako hiperravnina nacrtana plavom bojom dobro razdvaja podatke osim stršeće vrijednosti označene točkom T. Druga hiperravnina označena crnom bojom, iako je razdvajajuća, nije najsretniji izbor za klasifikaciju.

Optimizacijski problem je napisan tako da bira razdvajajuću hiperravninu u bilo kojem slučaju, čak i kada ona nije dobar izbor. Kao rješenje ovog problema uvodi se metoda regularizacije. **Regularizacija** je metoda kojom se sprječava pretreniranost modela koristeći funkciju kazne. Iako je hiperravnina nacrtana crnom bojom na slici 3.4 dobro prilagođena danim primjercima, kod klasificiranja novih primjerka neće dati dobre rezultate kao hiperravnina označena plavom bojom. Ovime se kažnjavaju modeli koji pretjerano slijede sve primjerke, ne vodeći računa o generalnoj strukturi podataka.

Valja navesti dvije najosnovnije funkcije kazne koje se koriste za regularizaciju SVM modela. Funckija kazne kod L1-SVM modela zadana je izrazom $max(1 - y^{(i)}\mathbf{w}^T\mathbf{x}^{(\mathbf{i})}, 0)$, a kod L2-SVM funkcija kazne je zadana izrazom $max(1 - y^{(i)}\mathbf{w}^T\mathbf{x}^{(\mathbf{i})}, 0)^2$. U nastavku će se koristiti L1-SVM.

Nakon uvođenja regularizacije, optimizacijski problem može se redefinirati na sljedeći

način:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s obzirom na} \quad & y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi_i,\ i = 1, \ldots, N \\ & \xi_i \geq 0,\ i = 1, \ldots, N. \end{aligned} \tag{3.5}$$

Valja obratiti pažnju na regularizacijski parametar C čiji iznos izražava spremnost na progrešku klasifikaciju primjerka iz skupa za učenje. Za velike iznose parametra C, optimizacija će inzistirati na ispravnoj klasifikaciji primjeraka iz skupa za učenje, makar pod cijenu pretreniranosti. Za manje iznose, optimizacija će pogrešno klasificirati neke od primjeraka iz skupa za učenje kako bi se pronašla hiperravnina koja dobro opisuje generalnu strukturu podataka.

3.6. Primarni i dualni optimizacijski problem

Nakon što je optimizacijski problem postavljen, vrijedi ga pokušati i riješiti. Postupak koji će se koristiti u rješavanju ovog problema je metoda Lagrangeovih multiplikatora. Neka je zadan **primarni** optimizacijski problem oblika:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s obzirom na} \quad & g_i(x) \leq 0,\ i = 1, \ldots, m \\ & h_i(x) = 0,\ i = 1, \ldots, n. \end{aligned} \tag{3.6}$$

Za dani optimizacijski problem, može se postaviti Lagrangeova funkcija oblika:

$$\mathcal{L}(x, \alpha, \beta) = f(x) + \sum_{i=1}^{m}\alpha_i g_i(x) + \sum_{i=1}^{n}\beta_i h_i(x) \tag{3.7}$$

gdje su α i β Lagrangeovi multiplikatori. Budući da se traži ekstrem funkcije, Lagrangeova funkcija se parcijalno derivira po varijablama x, α i β te se parcijalne derivacije izjednače s nulom. Time se dobivaju tri jednadžbe s tri različite nepoznanice.

Nakon definiranja Lagrangeove funkcije, valja definirati i još jednu vrijednost:

$$\theta_{\mathcal{P}}(x) = \max_{\alpha,\beta,\alpha_i \geq 0} \quad \mathcal{L}(x, \alpha, \beta). \tag{3.8}$$

Može se pokazati kako je iznos $\theta_{\mathcal{P}}(x)$ jednak vrijednosti funkcije $f(x)$ u slučaju kada su svi uvjeti zadovoljeni. U slučaju da neki od uvjeta nisu zadovoljeni, $\theta_{\mathcal{P}}(x)$ iznosi nula. Primjerice neka je $g_i(x) > 0$. Tada se može izabrati α_i za koji desna strana jednadžbe 3.6 iznosi ∞. Analogno vrijedi i za $h_i(x) \neq 0$.

Minimizacijom vrijednosti $\theta_{\mathcal{P}}(x)$ dobije se problem jednak primarnom. Vrijednost primarnog problema dana je izrazom:

$$p = \min_{x} \quad \theta_{\mathcal{P}}(x). \tag{3.9}$$

Za pronalazak $\theta_{\mathcal{P}}(x)$ interesantan je bio maksimum Lagrangeove funkcije u odnosu na parametre α i β. Umjesto toga, problem se može modificrati da umjesto traženja maksimuma u odnosu na parametre α i β, traži se minimum u odnosu na x. **Dualni** problem definira se na sljedeći način:

$$\max_{\alpha,\beta,\alpha_i \geq 0} \theta_{\mathcal{D}}(x) = \max_{\alpha,\beta,\alpha_i \geq 0} \min_{x} \mathcal{L}(x, \alpha, \beta). \tag{3.10}$$

Valja primijetiti kako je dualni problem jednak primarnom, uz zamjenu poretka funkcije min i funkcije max. Vrijednost $\theta_{\mathcal{D}}(x)$ je dana izrazom:

$$d = \max_{\alpha,\beta,\alpha_i \geq 0} \quad \theta_{\mathcal{D}}(x). \tag{3.11}$$

Uzimajući u obzir odnos između funkcija max i min jasno je da vrijedi relacija $d \leq p$. No, posebno su interesantni slučajevi gdje su vrijednosti primarnog i dualnog problema jednake.

Kako bi vrijednost primarnog i dualnog problema bila jednaka, potrebno je postaviti neka ograničenja na funkcije f, g_i i h_i. Neka je f konveksna funkcija. Nadalje, neka sje g_i konveksne funkcija i neka za svaku od njih vrijedi $g_i(x) < 0$. Također, neka je h_i linearna funkcija. Ako su ti uvjeti zadovoljeni, tada postoje α, β i x za koje vrijedi sljedeća jednakost:

$$p = d = \mathcal{L}(x, \alpha, \beta). \tag{3.12}$$

Osim gornje jednakosti, za parametre vrijede i Karush-Kuhn-Tucker (KKT) uvjeti:

$$\frac{\partial \mathcal{L}(x, \alpha, \beta)}{\partial x_i} = 0, \ i = 1, \ldots, p \tag{3.13}$$

$$\frac{\partial \mathcal{L}(x, \alpha, \beta)}{\partial \beta_i} = 0, \ i = 1, \ldots, n \tag{3.14}$$

$$\alpha_i g_i(x) = 0, \ i = 1, \ldots, m \tag{3.15}$$

$$g_i(x) \leq 0, \ i = 1, \ldots, m \tag{3.16}$$

$$\alpha_i \geq 0, \ i = 1, \ldots, m \tag{3.17}$$

Valja napomenuti kako se kod optimizacije stroja s potpornim vektorima u pravilu koristi dualni problem. Također, za SVM interesantan je uvjet zadan jednakošću 3.15. U pravilu $\alpha_i \neq 0$ iz čega slijedi $g_i(x) = 0$. Taj uvjet je ključan za pronalazak potpornih vektora.

3.7. Optimizacija stroja s potpornim vektorima

Primjenjujući prethodno poglavlje na optimizacijski problem 3.5 primarna Lagrangeova funkcija može se zapisati na sljedeći način:

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (y^{(i)}(\mathbf{w}^T \mathbf{x^{(i)}} + b) - (1 - \xi_i)) - \sum_{i=1}^{N} \mu_i \xi_i. \tag{3.18}$$

Vidljivo je kako je ograničenje $y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi_i$ zapisano kao:

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

Gore navedeno ograničenje ima zanimljivo svojstvo. U slučaju da vrijedi $g_i(\mathbf{w}, b) = 0$ za neki primjerak $(x^{(i)}, y^{(i)})$ tada taj primjerak se naziva **potpornim vektorom**. Razumna je pretpostavka kako je potpornih vektora relativno malo u odnosu na cijeli skup primjeraka za učenje. Vodeći tom pretpostavkom te uvjetom 3.15 može se zaključiti kako samo za potporne vektore $\alpha_i \neq 0$ dok za ostale primjerke vrijedi $\alpha_i = 0$.

Primarni problem dan jednadžbom 3.18 može se derivirati po $\mathbf{w}, b$ i ξ_i i izjednačiti s nulom:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)} \tag{3.19}$$

$$0 = \sum_{i=1}^{N} \alpha_i y^{(i)} \tag{3.20}$$

$$\alpha_i = C - \mu_i, \forall i. \tag{3.21}$$

Ubacivanjem gornjih jednadžbi u 3.18, dolazi se do dualne Lagrangeove funkcije:

$$\mathcal{L}(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)}. \tag{3.22}$$

Uz ograničenje 3.20 te ograničenje $0 \leq \alpha_i \leq C$ dualni problem može se zapisati na sljedeći način:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)} \\ \text{s obzirom na} \quad & 0 \leq \alpha_i \leq C, \ i = 1, \ldots, N \\ & \sum_{i=1}^{N} \alpha_i y^{(i)} = 0. \end{aligned} \tag{3.23}$$

Ako se pogleda uvjet 3.20 može se zaključiti kako samo potporni vektori utječu na odabir hiperravnine. U slučaju linearno nerazdvojivih podataka za potporne vektore koji se nalaze na margini, Lagrangeov multiplikator iznosi $0 < \alpha_i < C$. U slučaju da potporni vektor se ne nalazi na margini, vrijednost pripadajućeg multiplikatora iznosi C.

3.8. Jezgreni trikovi

Dosad je optimizacija stroja s potpornim vektorima mogla pronaći razdvajajuću hiperravninu jedino u slučaju linearno razdvojivih podataka. Želja je izgraditi klasifikator koji će moći ispravno klasificirati podatke koji su nisu linearno razdvojivi.

Kod metoda poput logističke regresije svaki vektor značajki može se, koristeći mapiranje značajki, preslikati u neki drugi vektor značajki. Neka je $\mathbf{x} = (x_1, \ldots, x_p)$ vektor značajki. Radi jednostavnosti neka postoji samo jedna značajka tj. $\mathbf{x} = (x_1)$. Neka funkcija ϕ vrši polinomijalno mapiranje značajki do trećeg stupnja tj. $\phi(x) = (x, x^2, x^3)$. Tada vektor značajki $\mathbf{x}$ može se preslikati u $\mathbf{x}' = \phi(x_1) = (x_1, x_1^2, x_1^3)$. Ova ideja može se iskoristiti i kod stroja s potpornim vektorima. Svaku pojavu vektora značajki $\mathbf{x}$ moguće je zamijeniti funkcijom $\phi(\mathbf{x})$.

Valja se na trenutak vratiti na dualnu Lagrengeovu funkciju zadanu jednakošću 3.22. Može se primijetiti kako dio funkcije zadan izrazom $x^{(i)T}x^{(j)}$ je zapravo skalarni umnožak vektora značajki dvaju primjera. Nadalje će se pisati $\langle x^{(i)}, x^{(j)} \rangle$. Vektore značajki moguće je preslikati u neke nove vektore značajki koristeći funkciju ϕ. Vrijedi:

$$\mathcal{L}(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle.$$

Primjerak $(x^{(i)}, y^{(i)})$ tada se klasificira na sljedeći način:

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

Vidljivo je kako i dualni problem i funkcija klasifikacije koriste funkciju oblika $\langle \phi(x), \phi(x') \rangle$, a ne $\phi(x)$. Može se definirati **jezgrena funkcija** oblika:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle. \tag{3.24}$$

Zamjenu skalarnog umnoška vektora x i x' s jezgrenom funkcijom $K(x, x')$ naziva se **jezgreni trik**. Najčešće korištene jezgrene funkcije su radijalna bazna funkcija (RBF), polinomijalna i sigmoidalna jezgrena funkcija. Radijalna bazna funkcija ima oblik:

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

gdje je σ parametar funkcije. Sigmoidalna jezgrena funkcija dana je oblikom:

$$K(\mathbf{x}, \mathbf{x}') = \tanh(\gamma \mathbf{x}^T \mathbf{x}' + r)$$

gdje su γ i r parametri funkcije. Polinomijalna jezgrena funkcija zadana je jednadžbom:

$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + k)^d,$$

gdje je k proizvoljna konstanta, a d stupanj polinoma. Na primjer, zadan je vektor značajki $\mathbf{x} = (x_1, x_2)$. Neka je $k = 1$ i $d = 2$. Jezgrena funkcija tada iznosi:

$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + 1)^2 = 1 + 2x_1 x_1' + 2x_2 x_2' + (x_1 x_1')^2 + (x_2 x_2')^2 + 2x_1 x_1' x_2 x_2'.$$

Vidljivo je kako smo vektor značajki s dva svojstva mapirali na vektor značajki sa šest svojstava. Općenito, polinomijalna jezgrena funkcija d-tog reda, preslika vektor značajki duljine p u vektor značajki duljine $\binom{n+d}{d}$. Iz povećanja dimenzionalnosti proizlazi i jedan od problema jezgrenih funkcija. Naime, neka su podaci linearno razdvojivi u slučaju prostora koji sadrži interakcijsku značajku $x_1 x_1'$. Tada je moguće, koristeći gore navedenu polinomijalnu funkciju drugog stupnja, pronaći razdvajajuću hiperravninu. No, uz ponalazak težine za tu značajku optimizacija traži i težine za ostalih pet značajki. Ovaj problem s povećanjem reda jezgrene funkcije postaje sve izraženiji, pogotovo ako je inicijalni prostor svih primjeraka velike dimenzije.

3.9. Višerazredna klasifikacija

Stroj s potpornim vektorima je binarni klasifikator te sam model nije u mogućnosti klasificirati primjerke u više od dva razreda. To ograničenje je intuitivno budući da hipperravnina može razdijeliti prostor u dva potprostora. Za rješenje ovog problema nude se dvije često korištene metode višerazredne klasifikacije.

Prva metoda višerazredne klasifikacije je metoda "jedan-naspram-ostalih" (engl. *one-vs-all*). Koristeći ovu metodu gradi se sustav klasifikatora. Ako je zadano N različitih razreda, sustav će biti izgrađen od $N - 1$ klasifikatora. Ovom metodom problem se razbija u više binarnih klasifikacija. Svaki stroj vrši klasifikaciju za pojedini razred. Vrijednost 1 označava pripadnost tom razredu dok vrijednost -1 kaže kako primjerak ne pripada razredu. U slučaju da svih $N - 1$ klasifikatora daju negativan izlaz, tada primjerak pripada zadnjem, N-tom razredu. Problem kod ove metode je mogućnost pozitivnog izlaza za više klasifikatora. Tada sustav nije u mogućnosti odrediti razred kojemu primjerak pripada.

Druga metoda višerazredne klasifikacije je metoda "jedan-naspram-jedan" (engl. *one-vs-one*). Kod ove metode također se gradi sustav klasifikatora, ali kod ove metoda jedan klasifikator vrši usporedbu između dvaju razreda. Ukupan broj izgađenih klasifikatora iznosi $\frac{N(N-1)}{2}$. Očito je kako je ova metoda vremenski složenija od "jedan-naspram-ostalih" metode. No, ova metoda je robusnija u slučaju linearno nerazdvojivih podataka te je robusnija na gore navedeni problem jedan-naspram-ostalih" metode. Sustav radi na principu glasanja. Prilikom klasifikacije, svaki klasifikator daje glas nekom od relevantnih razreda. Primjerak pripada nekom razredu u slučaju da je taj razred dobio najviše glasova.

3.10. Primjena SVM-a kod regresijskih problema

Stroj s potpornim vektorima je primarno binarni klasifikator. No, Drucker je 1997. godine predložio proširenje stroja s potpornim vektorima na probleme regresije[1].

Neka je zadan linearan model oblika:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b. \tag{3.25}$$

Moguće je zapisati funkciju optimizacije na sljedeći način:

$$f(\mathbf{w}, b) = \sum_{i=1}^{N} L(y^{(i)} - f(\mathbf{x}^{(i)})) + \frac{C}{2}\|\mathbf{w}\|^2 \tag{3.26}$$

gdje je L funkcija gubitka, a C regularizacijski parametar. Funkciju gubitka L moguće je izabrati na nekoliko načina. Neka je d razlika između dane vrijednosti $y^{(i)}$ i vrijednosti izračunate modelom $f(\mathbf{x}^{(i)})$ za neki primjerak $(\mathbf{x}^{(i)}, y^{(i)})$. U slučaju da je greška manja od neke proizvoljne vrijednosti ϵ tada taj primjerak ne pridonosi ukupnoj grešci. U slučaju pogreške veće od ϵ, primjerak pridonosi ukupnoj grešci s $|d| - \epsilon$. Uspoređujući ovu funkciju gubitka s načinom klasifikacije može se ustvrditi sličnost. Za primjerke koji su jako udaljeni od hiperravnine klasifikator je veoma siguran da su oni ispravno klasificirani te ne utječu na optimizaciju, za razliku od potpornih vektora. Kod regresije, primjerci koji su relativno blizu svojoj očekivanoj vrijednosti ne sudjeljuju u funkciji gubitka. Može se pisati:

$$L(d) = \begin{cases} 0 & |d| < \epsilon \\ |d| - \epsilon & \text{inače.} \end{cases} \tag{3.27}$$

Koristeći gore navedene podatke, može se pisati optimizacijski funkcija:

$$f(\alpha, \alpha') = \epsilon \sum_{i=1}^{N}(\alpha_i' + \alpha_i) - \sum_{i=1}^{N} y^{(i)}(\alpha_i' - \alpha_i) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)\langle x^{(i)}, x^{(j)}\rangle$$

i pripadajući optimizacijski problem:

$$\begin{aligned} \min_{\alpha,\alpha'} \quad & f(\alpha, \alpha') \\ \text{s obzirom na} \quad & \alpha_i \geq 0, \\ & \sum_{i=1}^{N}(\alpha_i' - \alpha_i) = 0, \\ & \alpha_i' \leq \frac{1}{C}, \\ & \alpha_i\alpha_i' = 0. \end{aligned} \tag{3.28}$$

4. Analiza sentimenta

4.1. Definicija

Analiza sentimenta (engl. *Sentiment Analysis*) je područje posvećeno prikupljanju, obradi i analizi subjektivnih informacija, posebice stavova. Subjektivne informacije podrazumijevaju kratkoročna stanja poput emocija i raspoloženja, dugoročna stanja poput stavova pa čak i psihičkih osobina. S druge strane, objektivne rečenice izriču činjenice. Primjerice objektivna rečenica *"Danas je sunčano."* izražava činjenicu te se iz nje ne mogu iščitati emocije, stavovi i slično. Rečenica poput *"Volim sunčan dan."* govori kako izvor mišljenja ima pozitivan stav prema danima koji sunčani te je stoga ova subjektiva. Moguće je da subjektivna rečenica ne izriče stav. Rečenica *"Mislim da će sutra biti sunčan dan."* ne predstavlja nikakvu činjenicu, već subjektivni dojam, no ne izriče nikakav sentiment.

Najčešće promatrana stanja su stavovi pojedinca koji predstavljaju enormni izvor korisnih informacija za zainteresirani subjekt. Koristeći ove informacije, političari mogu osvajati i gubiti izbore, poduzeća plasirati svoje usluge i proizvode, a potrošći informirati se o istima. uporabom metoda obrade prirodnog teksta i metoda strojnog učenja, cilj je automatizirati obradu i analizu subjektivnih informacija. Upravo subjektivnost čini ovo područje teškim i istovremeno zanimljivim.

Problem analize sentimenta može se zadati i na formalniji način. Neka je dana uređena petorka (o, k, i, s, v) kojom se može definirati sentiment. Varijabla o predstavlja promatrani objekt. Primjerice, ako se analizira kvaliteta mobilnih uređaja, promatrani objekt upravo bi bili mobilni uređaji. Varijablom k dana je karakteristika promatranog objekta. U slučaju mobilnih uređaja to bi mogla biti kamera, ekran, zvuk, itd. Ako se analizira uređaj u cjelini, tada ova varijabla može biti zanemarena. Varijabla i predstavlja izvor mišljenja, korisnika. Ključni dio problema, sentiment, dan je varijablom s. Sentiment predstavlja mišljenje korisnika i o nekoj karakteristici k proizvoda o. Mišljenja, stavovi i emocije mijenjaju se s vremenom pa je važno uključiti i vrijeme u sam problem. Vrijeme je predstavljeno varijablom v.

Evo i jednog primjera korisničke recenzije operacijskog sutava [1]:
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.

Nakon čitanja korisničke recenzije, jasno je kako se ovdje radi o negativnoj recenziji proizvoda. Prva rečenica recenzije je najbitnija za određivanje sentimenta prema cijelom proizvodu. Izvor mišljenja, korisnik Marko, ocjenjuje cijeli sustav *užasnim*. Izbor pridjeva, u ovom slučaju *užasan*, sigurno utječe na intenzitet sentimenta. Kad bi se umjesto navedenog pridjeva našao pridjev *loš* i dalje bi izjava bila klasificirana kao negativna, no s puno manjim intenzitetom u odnosu na izjavu koja sadrži pridjev *užasan*. Nakon prve rečenice, slijede dvije rečenice negativnog sentimenta usmjerene na pojedine karakteristike sustava. Četvrta rečenica navodi jednu pozitivnu karakteristiku sustava. U slučaju analize sentimenta na razini cijele recenzije ova rečenica uvodi šum te predstavlja jedan od problema u analizi sentimenta. Iz zadnje rečenice, iako svakako manjeg intenziteta od ostalih, također se može iščitati negativan stav prema proizvodu. Valja napomenuti kako je ova recenzija napisana 2016. godine te upravo godina predstavlja vremensku komponentu problema. Moguće je da se s vremenom stav korisnika promijeni. Tada ova recenzija više neće biti aktualna. Ukupno se može iščitati četiri sentimenta:

(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.)

4.2. Tipovi mišljenja

4.3. Razine analize sentimenta

Sentiment korisnika može se proučavati na više razina. U pravilu sentiment korisnika proučava se na sljedeće tri razine:

1. **Analiza na razini dokumenta**. Cilj ove razine je pronaći generalni sentiment u sklopu cijelog teksta. Provodi se najćešće kada zainteresirane strane ne zanima sentiment neke karakteristike proizvoda ili usluge već generalan stav. Također, pretpostavlja

[1] Napomena: Iako nije dano ime operacijskog sustava, čitatelj može s jednostavnošću zaključiti o kojem se operacijskom sustavu radi

se da korisnik ocjenjuje samo jedan proizvod budući da iz njega nije moguće iščitati mišljenje o više proizvoda. Primjer analize na razini dokumenta je analiza korisničkih recenzija filmova [3].

2. **Analiza na razini rečenica**. Cilj analize na razini rečenica je odrediti pozitivnost, neutralnost ili negativnost svake rečenice. Za ovu razinu analize važno je razlikovati objektivne i subjektivne rečenice [5]. No, postoje iznimke kada rečenice, iako objektivne, nose sentiment. Primjerice, *"Capacitet baterije ovoga mobitela drastično opada već nakon dva mjeseca."* predstavlja objektivnu rečenicu, ali i implicitno izlaže nezadovoljstvo korisnika.

3. **Analiza na razini karakteristika**. Najdetaljnija i najkompleksnija analiza sentimenta vrši se na razini karakteristika. Pretpostavka je da svako mišljenje sadrži sentiment i objekt te da mišljenje nije striktno vezano uz formu tj. uz rečenicu, paragraf, cijeli dokument [2]. Na primjer, rečenica *"Zadovoljan sam s mobitelom marke A, no nije kao onaj marke B."* izražava pozitivan stav prema oba mobitela, no intenzitet sentimenta nije jednak prema oba mobitela. Viljivo je kako postoji dva sentimenta i dva različita objekta što se ne može utvrditi na razini rečenice, a pogotovo na razini dokumenta. Rečenica *"Auto je brz, ali troši previše goriva."* izražava sentiment o jednom objektu, autu, ali za više njegovih karakteristika. Upravo je pronalazak karakteristika i objekata veoma složen problem koji čini ovu razinu još kompleksnijom u odnosu na razine dokumenta i rečenice.

Iako su sve tri razine analize sentimenta zanimljive i složene, u ovom radu fokus je stavljen na razinu dokumenta.

4.4. Sentiment na razini dokumenta

4.5. Problemi kod analize sentimenta

5. Implementacija i rezultati

6. Zaključak

Zaključak.

LITERATURA

[1] Harris Drucker, Chris J. C. Burges, Linda Kaufman, Alex Smola, i Vladimir Vapnik. Support vector regression machines. U *Advances in Neural Information Processing Systems 9*, stranice 155–161. MIT Press, 1997.

[2] Minqing Hu i Bing Liu. Mining and summarizing customer reviews. U *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, stranice 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1. doi: 10.1145/1014052.1014073. URL `http://doi.acm.org/10.1145/1014052.1014073`.

[3] Bo Pang, Lillian Lee, i Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. U *Proceedings of EMNLP*, stranice 79–86, 2002.

[4] V. Vapnik i A. Lerner. Pattern recognition using generalized portrait method. *Avtomatika i Telemekhanika*, 24(6):774–780, 1963.

[5] Janyce M. Wiebe, Rebecca F. Bruce, i Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. U *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, stranice 246–253, College Park, Maryland, USA, June 1999. Association for Computational Linguistics. doi: 10.3115/1034678.1034721. URL `http://www.aclweb.org/anthology/P99-1032`.

Primjena stroja s potpornim vektorima za analizu sentimenta korisničkih recenzija

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Application of Support Vector Machine for Users' Reviews Sentiment Analysis

Abstract

Abstract.

Keywords: Keywords.