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 [1]

3. Stroj s potpornim vektorima

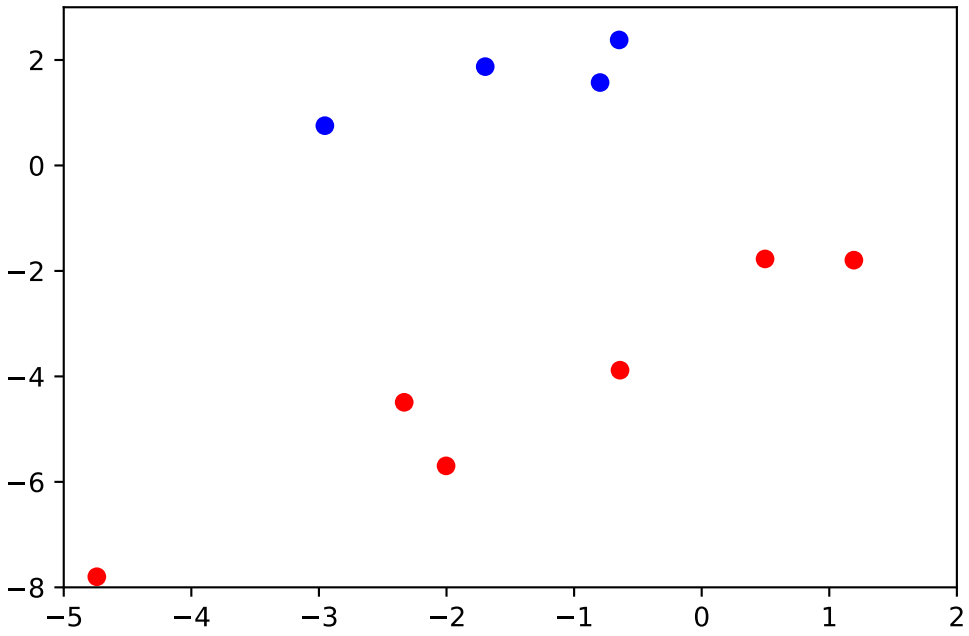
U ovom poglavlju bit će predstavljen model stroja s potpornim vektorima. Potpoglavlje 3.1 definira pojam klasifikacije i pojašnjava razliku između klasifikacije i regresije. U potpoglavlju 3.2 pojasnit će se ideja razdvajajuće hiperravnine. Potpogljavlje 3.3 predstavit će pojam margine razdvajajuće hiperravnine 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 3.5 opisuje transformaciju prostora značajki koristeći jezgrene trikove. Jezgreni trikovi su efikasne metode koje omogućuju razdvajanje originalno linearno nerazdvojivih podataka. U potpoglavlju 3.6 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

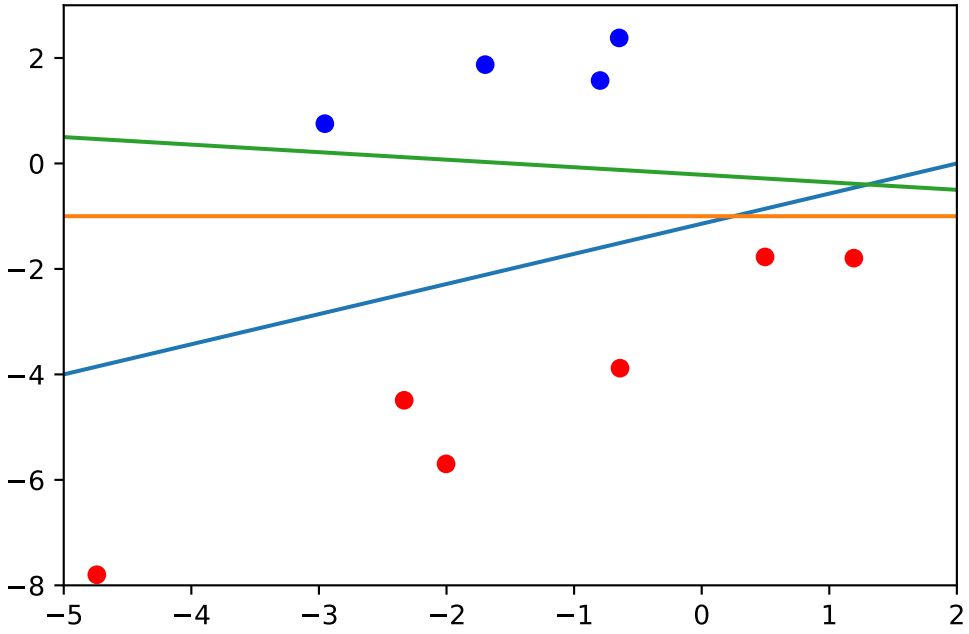
Interpretaciju modela stroja s potpornim vektorima potrebno je započeti s pojmom koji nije strogo vezan uz sam model. Primjerice model logističke regresije, iako temeljen na vjerojatnosti, u konačnici pronalazi hiperravninu kojom razdvaja podatke.

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:

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

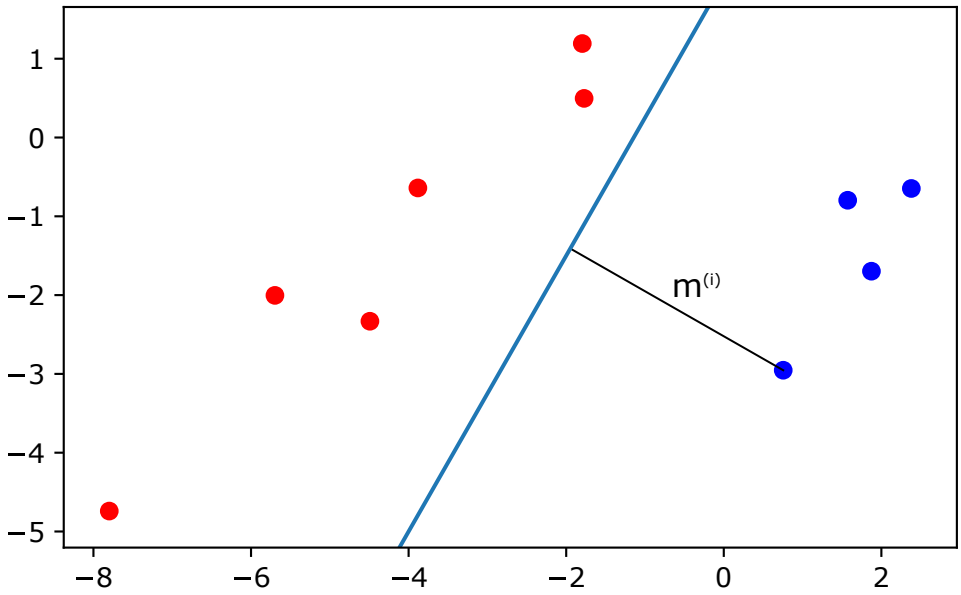
gdje je $\mathbf{x}^{(i)}$ primjer iz skupa podataka, a $y^{(i)}$ je oznaka razreda pridružena primjeru $\mathbf{x}^{(i)}$.

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.

3.5. Jezgreni trikovi

3.6. Regularizacija

4. Analiza sentimenta

5. Eksperiment

6. Zaključak

Zaključak.

LITERATURA

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

Primjena stroja s potpornim vektorima za analizu sentimenta korisničkih recenzija

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Title

Abstract

Abstract.

Keywords: Keywords.