SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 000

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`.

Zahvaljujem se Donaldu E. Knuthu, Bogu računarske znanosti.

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 probleme.

Osim navedenih modela, još jedan model koji je sposoban efikasno obraditi nelinearne podatke je **stroj s potpornim vektorima** (engl. *Support Vector Machine*). 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 drugom poglavlju bit će predstavljen pregled područja, povijest modela stroja s potpornim vektorima te problem analize sentimenta. U trećem poglavlju detaljnije će se obraditi model SVM. Bit će opisana motivacija i interpretacija modela. Nadalje, detaljnije će se pojasniti algoritmi optimizacije modela. U četvrtom poglavlju 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. Šesto poglavlje 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. U potpoglavlju 3.1 pojasnit će se ideja razdvajajuće hiperravnine. Uz pomoć hiperravnine margine, u potpoglavlju 3.2 iznijet će se geometrijska interpretacija modela.

3.1. 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 može razdvojiti podatke.

Neka je zadan prostor *X* dimenzije n. Tada je **hiperravnina** definirana kao potprostor dimenzije $n - 1$ unutar prostora *X*. Primjerice, u dvodimenzijskom prostoru hiperravnina predstavlja bilo koji pravac koji leži u ravnini. Analogno, pojam hiperravnine vrijedi i za prostore većih dimenzija.

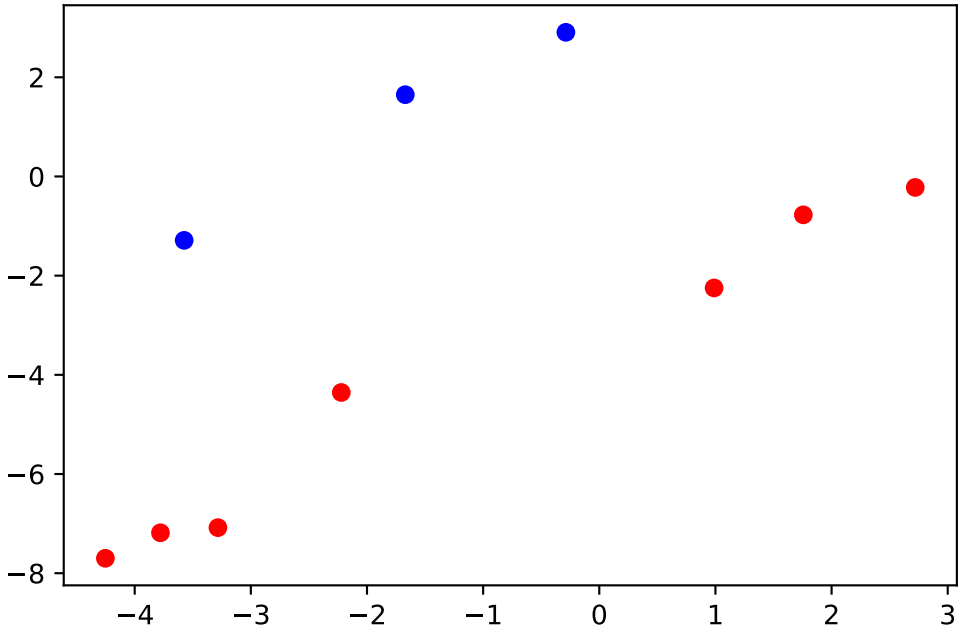
Za hiperravninu zadanom jednadžbom $f(x) = \beta_0 + \beta^T x = 0$ vrijede sljedeća svojstva:

1. Za svaku točku T na hiperravnini vrijedi: $\beta_0 = -\beta^T x$
2. Normala zadana izrazom: $\mathbf{n} = \frac{\beta}{\|\beta\|}$
3. Udaljenost točke P od hiperravnine iznosi: $d = \frac{f(x)}{\|\beta\|}$

Osim pojma hiperravnine, potrebno je definirati i pojam binarnog klasifikatora. Neka je upravo prostor *X* ulazni prostor koji definira potpun skup primjera. Svakom primjeru $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ pridružena je oznaka razreda y. Neka y poprima vrijednosti iz skupa *C*. Ako se klasificiraju podaci koji su podijeljeni u dva razreda, tj. $|C| = 2$, tada se govori o binarnoj klasifikaciji. **Binarni klasifikator** je klasifikator koji može provesti binarnu klasifikaciju.

Hiperravnine same po sebi nisu pretjerano interesantne. No, za klasifikaciju interesantan je određen podskup margina. Hiperravnina koja razdvaja dva razreda podataka naziva se **razdvajajuća hiperravnina.**

Na slici 3.1 prikazani su linearno razdvojivi podaci. Također, prikazane su i dvije razdvajajuće hiperravnine. Valja primijetiti kako je moguće konstruirati beskonačno mnogo razdvajajućih hiperravnina.



Slika 3.1: Razdvajajuće hiperravnine

3.2. Geometrijska interpretacija modela

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.