Spam detection using a multi-layer perceptron

Projekat u okviru kursa Računarska inteligencija
Matematički fakultet
Univerzitet u Beogradu

Đorđe Milošević
mi19221@alas.matf.bg.ac.rs

April 2021

Sadržaj

1 Opis problema **3**

2 Implementacija **3**
2.1 Obrada ulaznih podataka 3
2.2 Kreiranje TF-IDF matrice 3
2.3 Train-test split 4
2.4 Viseslojna neuronska mreza 4
2.5 Ocena modela 4

3 Rezultati **4**

4 Zaključak **4**

Literatura **5**

1 Opis problema

Data je baza email poruka koja sadrzi 2 kolone. Prva kolona pod nazivom *text* sadrzi datu poruku koju je potrebno obraditi, dok druga kolona pod nazivom *spam* sadrzi vrednosti 1, ukoliko je email poruka spam, ili 0, ukoliko poruka nije spam. Cilj je napraviti model neuronske mreze koji ce pomocu vise slojeva efektivno da odredi da li je email poruka spam ili ne.

2 Implementacija

Podaci ce prvo biti pretprocesirani kako bi bili u pogodnom obliku za dalji rad. Nakon toga je potrebno obraditi dati sadrzaj poruka kako bi njihov oblik bio pogodan za rad sa neuronskim mrezama. Za to cemo koristiti TF-IDF matrice. Potrebno je izdvojiti sve reci koje se javljaju u email porukama i one ce ciniti kolone nase TF-IDF matrice. Zatim ce podaci biti podeljeni na 2 skupa, *train* i *test*, kako bismo mogli da istreniramo nas model na jednom skupu i pravilno da testiramo na drugom. Nakon toga cemo napraviti nasu viseslojnu neuronsku mrezu. Za kraj cemo videti ocenu kvaliteta naseg modela pomocu matrice konfuzije, ocene preciznosti itd.

2.1 Obrada ulaznih podataka

Ulazni podaci se nakon ucitavanja pretprocesiraju. Prvo se uklanjaju duplikati zarad efikasnosti u daljem toku rada, a zatim se proverava da li postoje neke null vrednosti podataka i uklanjaju se ako takve postoje.
Podatke je zatim potrebno podeliti u 2 grupe. Prva grupa ce sadrzati podatke koji se nalaze u koloni *text* i oznacavacemo ih sa X, a druga grupa ce sadrzati podatke koji se nalaze u koloni *spam* i oznacavacemo ih sa y. Ova podela nam dalje sluzi za treniranje podataka, kao i za ocenu samog modela.

2.2 Kreiranje TF-IDF matrice

Za kreiranje tf-idf matrice koristicemo *TfidfVectorizer* iz biblioteke *sklearn*

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

TF oznacava *term frequency*, odnosno koliko puta se data rec javlja u jednom dokumentu u odnosu na ukupan broj reci datog dokumenta. Formula za izracunavanje TF vrednosti za rec t u dokumentu d je:

$$TF(t,d) = \frac{t}{d}$$

gde je t broj ponavljanja reci t u datom mail-u, a d je ukupan broj reci u mail-u.
IDF oznacava *Inverse document frequency* odnosno *inverznu frekvenciju dokumenta*. IDF vrednost se odnosi na skup dokumenata (u ovom slucaju skup mail-ova). IDF nam sluzi da smanji znacaj reci koje se cesto javljaju, a poveca znacaj recima koje su retke. Formula za izracunavanje IDF vrednosti za rec t, u dokumentu d, koji pripada skupu dokumenata D je:

$$IDF(t, d, D) = \log \frac{|D|}{|\{d \in D, t \in d\}|}$$

Na kraju za dobijanje TF-IDF vrednosti potrebno je da pomnozimo dve prethodno dobijene vrednosti:

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

2.3 Train-test split

Da bismo mogli da upotrebimo podatke tako da nas model moze na najbolji nacin da uci nad njima, moramo ih podeliti na *train* i na *test* skupove. Nad *train* skupom cemo da istreniramo nas model, nakon cega ce model to steceno znanje da iskoristi u evaluaciji test podataka.

2.4 Viseslojna neuronska mreza

Viseslojna neuronska mreza

2.5 Ocena modela

Ocena modela

3 Rezultati

4 Zaključak

Literatura

[1]

[2]

[3]

[4]

[5]

[6]