Paralelno kolaborativno filtriranje velikih razmera za Netfliks nagradu

Stefanović Vuk
Marković Dimitrije

Avgust 2024
Matematicki fakultet

Contents

1 Data Manipulation **3**

2 KNN **3**

3 SVD **4**

4 RBM **4**

5 Feature Engineering **4**

6 Ridge **5**

7 XGB **6**

8 Zaključak **6**

1 Data Manipulation

U ovoj fazi projekta, fokusirali smo se na pripremu i obradu podataka kako bismo omogućili efikasno treniranje modela za preporučivanje filmova.

S obzirom na veličinu originalnog Netflix dataset-a, odlučili smo da radimo sa samo četvrtinom podataka. Ova odluka je doneta kako bismo smanjili vreme obrade i resurse potrebne za treniranje modela, dok istovremeno zadržavamo dovoljno podataka za kreiranje pouzdanih modela.

Pre nego što smo započeli sa daljim koracima obrade, izvršili smo ispitivanje podataka. Ova analiza obuhvatila je proveru distribucije ocena, identifikaciju filmova sa najviše i najmanje ocena, kao i korisnika koji su dali najviše i najmanje recenzija. Ovi uvidi su nam pomogli da bolje razumemo strukturu podataka i da donesemo informisane odluke o daljoj obradi.

Jedan od ključnih koraka u obradi podataka bio je uklanjanje filmova koji su imali najmanji broj ocena. Ovi filmovi često sadrže nedovoljno informacija za treniranje modela, što može dovesti do slabijih performansi. Uklanjanjem ovih filmova, fokusirali smo se na one koji imaju veću relevantnost u korisničkim preporukama.

Slično tome, uklonili smo korisnike koji su dali najmanje recenzija. Korisnici sa malim brojem ocena često pružaju ograničene informacije koje mogu negativno uticati na kvalitet modela. Uklanjanjem ovih korisnika, povećali smo pouzdanost podataka korišćenih za treniranje.

Nakon filtriranja podataka, podelili smo preostale podatke na trening i test skup. Trening skup smo koristili za treniranje modela, dok smo test skup sačuvali za evaluaciju performansi modela na nevidjenim podacima. Ova podela omogućava da se na objektivan način meri preciznost predikcija modela. Svi podaci su zatim sačuvani u odgovarajućim CSV fajlovima za dalju upotrebu.

2 KNN

U ovoj sekciji smo se fokusirali na treniranje i optimizaciju item-based KNN (K-Nearest Neighbors) modela kako bismo postigli što bolje performanse u predikciji ocena filmova.

Kao početnu tačku, odlučili smo se za item-based KNN pristup, gde se sličnost izmedju filmova računa na osnovu ocena koje su im dodelili korisnici. Prednost ovog pristupa je u tome što omogućava preporuke na osnovu sličnosti izmeu filmova, što je često intuitivnije i korisnicima bliže od user-based pristupa.

Proces treniranja uključivao je eksperimentisanje sa različitim vrednostima hiperparametra k, koji odreuje broj najsličnijih filmova koji se uzimaju u obzir prilikom predikcije. Takoe smo isprobavali različite mere sličnosti, uključujući kosinusnu sličnost i Pearson korelaciju, kako bismo utvrdili koji pristup daje najbolje rezultate.

Nakon opsežnog testiranja, zaključili smo da je Pearson-ova korelacija kao mera sličnosti dala najbolje rezultate. Optimalnu vrednost za k smo dobili kada je $k = 18$, pri čemu je model postigao RMSE (Root Mean Square Error) od 0.9258. Ovaj rezultat je pokazao da je model sposoban da precizno predvia ocene na osnovu sličnosti izmeu filmova, što ga čini pogodnim za korišćenje u hibridnom modelu.

3 SVD

Nakon KNN modela, fokusirali smo se na treniranje i optimizaciju Singular Value Decomposition (SVD) modela, koji je jedan od najpopularnijih pristupa u kolaborativnom filtriranju.

SVD model koristi matricu korisnika i filmova da bi dekomponovao podatke u latentne faktore, što omogućava preciznije predikcije ocena. Da bismo postigli najbolje rezultate, eksperimentisali smo sa različitim vrednostima hiperparametara, uključujući broj latentnih faktora ($n\_factors$), broj epoha (n_epochs), stopu učenja (lr_all), i regularizaciju (reg_all).

Nakon opsežnog testiranja, identifikovali smo sledeće optimalne vrednosti za hiperparametre:

- **n_factors**: 10
- **n_epochs**: 30
- **lr_all**: 0.005
- **reg_all**: 0.02

Sa ovim parametrima, SVD model je postigao RMSE od 0.8676, što predstavlja napredak u tačnosti predikcija u poredjenju sa KNN modelom. Ovaj rezultat ukazuje na to da SVD model može efikasno da uhvati latentne odnose izmeu korisnika i filmova, što ga čini izuzetno korisnim za preporučivanje.

4 RBM

U okviru eksperimentisanja sa različitim modelima, isprobali smo i Restricted Boltzmann Machine (RBM) kao potencijalni model za preporuke.

RBM model je specifičan tip neuronske mreže koji se koristi za kolaborativno filtriranje. Model pokušava da nauči latentne faktore iz podataka o korisnicima i filmovima kako bi predviao ocene. Trenirali smo model na našem skupu podataka i evaluirali njegovu performansu na trening i test skupu.

Na trening skupu, RBM je postigao RMSE nešto iznad 0.7, što je u početku izgledalo obećavajuće. Meutim, prilikom testiranja modela na nevienim podacima, RMSE je značajno porastao iznad 3.0, što ukazuje na ozbiljno preprilagoavanje (overfitting) modela.

Zbog ovako velike razlike u grešci izmedju trening i test skupa, odlučili smo da RBM model ne uključimo u finalni hibridni model. Iako je RBM pokazao dobre performanse na treningu, njegova sposobnost generalizacije na nevidjene podatke bila je vrlo loša, što ga čini nepouzdanim za preporučivanje u realnim uslovima.

5 Feature Engineering

Da bismo unapredili tačnost našeg hibridnog modela, posvetili smo se kreiranju skupa relevantnih atributa koji će se koristiti kao ulazi za različite modele unutar hibridnog pristupa.

U ovoj fazi, fokusirali smo se na identifikaciju i generisanje skupa atributa koji će omogućiti hibridnom modelu da bolje razume karakteristike filmova i korisnika. Cilj je bio da se integrišu podaci o filmovima, korisnicima i predikcijama različitih modela kako bi se postigla što veća tačnost u preporukama.

Jedan od ključnih koraka bio je generisanje predikcija za ocene filmova korišćenjem dva modela: KNN i SVD. Ove predikcije su zatim uključene kao atributi u hibridni model. Ideja je bila da se kombinuju snage oba modela kako bi se poboljšala preciznost finalnih predikcija.

Pored predikcija KNN-a i SVD-a, kreirali smo i nekoliko dodatnih atributa koji pružaju važne informacije o filmovima i korisnicima. To uključuje:

- **Prosečna ocena filma**: Srednja vrednost svih ocena koje je film dobio.
- **Prosečna ocena koju korisnik daje**: Prosečna ocena koju odredjeni korisnik dodeljuje filmovima.
- **Broj ocena korisnika**: Ukupan broj ocena koje je korisnik dao.
- **Broj ocena za film**: Ukupan broj ocena koje je film dobio.
- **Varijansa**: Varijansa ocena filma, koja pruža uvid u disperziju ocena.

Ovi atributi su korišćeni kako bi se dodatno obogatio skup podataka i omogućila bolja diferencijacija izmeu različitih filmova i korisnika.

Nakon što smo generisali sve relevantne atribute, kreirali smo trening, validacioni i test skup podataka. Trening skup je korišćen za treniranje modela, validacioni skup za izbor najboljih hiperparametara, dok je test skup korišćen za konačnu evaluaciju performansi modela. Svi ovi skupovi su sačuvani u CSV formatu, čime je obezbedjena njihova dostupnost za kasniju upotrebu i analizu.

6 Ridge

Kao deo našeg hibridnog pristupa, koristili smo Ridge regresiju kako bismo dodatno poboljšali preciznost predikcija. Ridge regresija je linearni model koji uključuje regularizaciju kako bi se smanjila mogućnost preprilagodjavanja, što je posebno važno kada se koristi veliki broj atributa.

Za treniranje Ridge modela koristili smo skup atributa koje smo kreirali u fazi *Feature Engineering*. Integracijom ovih različitih izvora informacija, model je mogao bolje da uhvati odnose izmeu korisnika i filmova.

Prilikom treniranja Ridge modela, eksperimentisali smo sa različitim vrednostima hiperparametra α, koji kontroliše stepen regularizacije. Optimizacijom ovog parametra, postigli smo balans izmedju složenosti modela i njegove sposobnosti da generalizuje na nevidjene podatke.

Najbolji rezultat smo postigli sa RMSE vrednošću od 0.73 na test skupu. Ovaj rezultat pokazuje da je Ridge regresija bila vrlo efikasna u smanjenju greške u predikcijama, čineći je pouzdanim delom našeg hibridnog modela. Regularizacija je omogućila da model zadrži preciznost, a da pri tome izbegne preprilagoavanje.

7 XGB

Kao deo naših napora da poboljšamo tačnost predikcija, koristili smo XGBoost, jedan od najefikasnijih algoritama za gradijentno pojačavanje. XGBoost je poznat po svojoj brzini i performansama, što ga čini idealnim za rad sa velikim skupovima podataka.

Za treniranje XGBoost modela, koristili smo skup atributa koje smo generisali u fazi *Feature Engineering*. Primenili smo Grid Search kako bismo pronašli najbolje hiperparametre za model, testirajući različite kombinacije broja stabala (*n_estimators*), stope učenja (*learning_rate*), maksimalne dubine stabala (*max_depth*), i poduzorka podataka (*subsample*).

Eksperimentisali smo sa sledećim vrednostima hiperparametara:

- **n_estimators**: [50, 100, 200]
- **learning_rate**: [0.01, 0.1, 0.2]
- **max_depth**: [3, 5, 7]
- **subsample**: [0.8, 0.9, 1.0]

Nakon opsežnog testiranja, optimalna kombinacija hiperparametara za naš model bila je:

- **learning_rate**: 0.2
- **max_depth**: 7
- **n_estimators**: 200
- **subsample**: 0.8

Sa ovim optimalnim hiperparametrima, XGBoost model je postigao RMSE od 0.6973 na trening skupu i 0.7002 na test skupu. Ovi rezultati pokazuju da je model bio u stanju da generalizuje vrlo dobro, sa minimalnom razlikom izmeu grešaka na trening i test podacima. To ukazuje na stabilnost modela i njegovu sposobnost da precizno predvidja ocene na osnovu danih atributa.

8 Zaključak

Naš rad je pokazao da kombinacija različitih tehnika može unaprediti tačnost preporuka u sistemima poput Netflix-a. Važno je napomenuti da smo radili samo na četvrtini dostupnih podataka, što može uticati na konačne rezultate i ostavlja prostor za dalja istraživanja na većem skupu podataka.