Recommender Systems

Aanbevelingen zonder platform of eigen producten

Nicolas De Jaeghere

Proloog

AI is eating the world —Onbekend

- Toepassingen veelal onzichtbaar of *gimmicky*
- Opkomst *meta-producten* met hoofdingrediënt aanbevelingen (o.a. e-commerce, entertainment en relaties)
- Intrigerende technologie

Doel

- Hoe aanbevelingen te genereren en evalueren
- Welke aanbevelingen mogelijk zijn zonder over een platform of eigen producten te beschikken

Terminologie

- Aanbeveling versus reclame
- Gepersonaliseerde versus niet gepersonaliseerde aanbeveling
- Collaborative filtering versus content-based filtering versus hybrid recommender
- Expliciete versus impliciete beoordeling
- Cold start

Dataset

- MovieLens of de MNIST voor recommender systems
- Netflix prize
- Gecureerd

MovieTweetings

I rated Brazil 9/10 https://www.imdb.com/title/tt0088846/ #IMDb —Een Twitter gebruiker

- Gestructureerde tweets getweet door de IMDb app
- Actueel en vrij beschikbaar
- Jong, weinig beoordelingen, zeker voor oudere films
- Weinig meta-data

MovieTweetings: data

Weten met welke data je werkt is weten welke resultaten je mag verwachten

- ratings.dat: gebruiker ID, IMDb ID, beoordeling van 0 t.e.m. 10 en datum/tijd vermoedelijk in lokale tijd
- users.dat: gebruiker ID en Twitter ID
- movies.dat: IMDb ID, titel met jaar van uitgave en genres

MovieTweetings: eigenschappen

Kwaliteit van dataset is hoog

MovieTweetings: eigenschappen vervolg

- Beoordelingen:
  - 867 696 beoordelingen
  - Enkele beoordelingen voor onuitgegeven films (niet opgelost)
- Gebruikers:
  - 67 630 gebruikers
  - 4 832 dubbele gebruikers
- Films:
  - 35 613 films
  - Van 251 films ontbreekt genres (niet opgelost)
  - Twee dubbele films

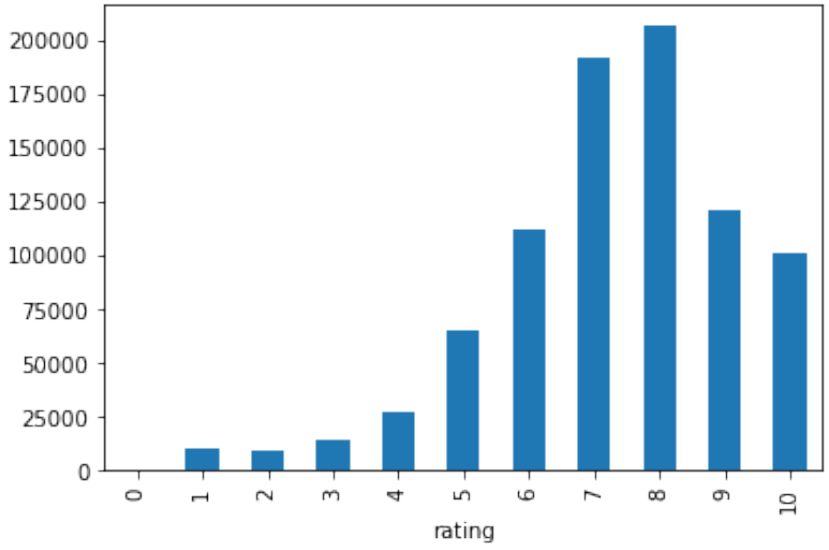
MovieTweetings: beoordelingen

- Meeste gebruikers beoordelen weinig films
- Enkele gebruikers beoordeelden meer dan tweeduizend films
- Meeste films worden weinig beoordeeld
- Enkele films werden rond de drieduizend maal beoordeeld
- Meeste beoordelingen zijn positief
- Uitschieters zijn geen anomalieën

Gegroepeerd op	$\overline{x}$	s	Mediaan
Gebruiker	14	49	2
Film	24	112	2
Beoordeling	7	2	8

MovieTweetings: beoordelingen vervolg

- Onbalans is niet abnormaal (sparsity en de long tail)
- Weerslag op voorspellingen



MovieTweetings: films

Mix van oude en nieuwe films maar voornamelijk films uitgegeven in de laatste twintig jaar

MovieTweetings: augmenteren

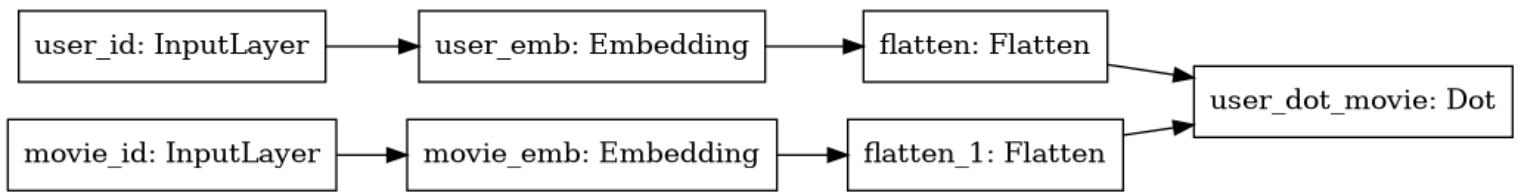
- Beoordelingen reduceren tot vijf klassen van 1 t.e.m. 5
- Koppeling van beoordeling naar film m.b.v. nieuwe movie ID
- Doorlopende user en movie ID
- Test gebruiker aanmaken
- Genres integreren
- Datum/tijd integreren als proxy voor populariteit

MovieTweetings: bruikbaarheid

- Niet alle data is informatie
- Uit een gebruiker met één beoordeling en een film met één beoordeling kan niets worden afgeleid
- Gestart met alle data maar uiteindelijk gereduceerd tot gebruikers met minimaal vijfentwintig beoordelingen (11% gebruikers)
- Weerslag op aantal beoordelingen en films
- Films op andere manier opgevangen

Matrix factorization

- Algoritme voor collaborative filtering
- Oud maar blijft populair
- Onderdelen op zichzelf bruikbaar
- Vinden van *verborgen genres*
- Geïmplementeerd in Keras



Trainen model

- Benaderd als regressie en classificatie probleem
- MSE of categorical crossentropy
- Gelijkaardig aan andere neurale netwerken

Evalueren model

- Error function train versus validation
- MSE, MAE, zowel regressie als classificatie
- Scatterplot werkelijk versus voorspeld
- Accuracy
- Precision, recall en F1
- ROCC en AUC
- Confusion matrix

Optimaliseren hyperparameters

- Manueel, geen grid search
- Kort geëxperimenteerd met cloud computing

Optimaliseren hyperparameters vervolg

- Groote latente representatie
- Breedte en diepte netwerk
- Spatial dropout, dropout en batch normalization
- Sample of class weights
- Meta-data

Optimaliseren hyperparameters: regressie

Model	MSE	MAE
Collaborative baseline	0,88	0,63
Collaborative 1	0,90	0,64
Collaborative 2	0,95	0,65
Collaborative deep	0,45	0,52
Collaborative deep 1	0,83	0,71
Collaborative deep 2	0,92	0,76
Collaborative deep 3	0,49	0,54
Hybrid baseline	0,44	0,51
Hybrid 1	0,43	0,50

Optimaliseren hyperparameters: classificatie

Model	Acc	F1 1	F1 2	F1 3	F1 4	F1 5
Collaborative baseline	0,60	0,06	0,04	0,51	0,67	0,55
Collaborative 1	0,60	0,11	0,08	0,51	0,68	0,54
Collaborative 2	0,60	0,03	0,03	0,52	0,67	0,55
Collaborative deep	0,60	0,00	0,00	0,52	0,67	0,55
Collaborative deep 1	0,50	0,12	0,19	0,50	0,53	0,58
Collaborative deep 2	0,49	0,12	0,20	0,46	0,53	0,58
Collaborative deep 3	0,59	0,07	0,06	0,52	0,67	0,55
Hybrid baseline	0,60	0,00	0,00	0,52	0,68	0,55
Hybrid 1	0,60	0,00	0,00	0,53	0,68	0,53

Optimaliseren hyperparameters: classificatie vervolg

Model	MSE	MAE
Collaborative baseline	0,55	0,45
Collaborative 1	0,55	0,45
Collaborative 2	0,55	0,45
Collaborative deep	0,55	0,44
Collaborative deep 1	1,23	0,70
Collaborative deep 2	1,20	0,69
Collaborative deep 3	0,56	0,45
Hybrid baseline	0,54	0,44
Hybrid 1	0,53	0,44

Evalueren voorspellingen

- Doel, populariteit versus nieuwigheid
- Subjectief, doelstellingen en experten
- Minimum aantal beoordelingen vooraleer aanbevelingen kunnen worden gedaan
- Snel verouderd

Evalueren voorspellingen: regressie

Titel	Beoordeling
The Godfather: Part II	4,166014
The Godfather	4,160283
Cinema Paradiso	4,119936
The Green Mile	4,119329
Amadeus	4,114968
The Shawshank Redemption	4,111031
12 Angry Men	4,098608
Avengers: Endgame	4,089588
Saving Private Ryan	4,085806
Il buono, il brutto, il cattivo	4,082156

Evalueren voorspellingen: regressie vervolg

Titel	Beoordeling
The Godfather: Part II	4,166014
The Godfather	4,160283
The Green Mile	4,119329
Saving Private Ryan	4,085805
Il buono, il brutto, il cattivo	4,082156
The Lord of the Rings: The R...	4,066126
Schindler's List	4,055110
La vita è bella	3,997949
A Beautiful Mind	3,979335
Incendies	3,966434

Evalueren voorspellingen: classificatie

Titel	Beoordeling	Overtuiging
The Lord of the Rings: The R...	5	0,476686
Catch Me If You Can	4	0,647723
Reservoir Dogs	4	0,643887
The King's Speech	4	0,635510
Philomena	4	0,633914
Batman Begins	4	0,624551
Moneyball	4	0,620673
The Gentlemen	4	0,610244
Fargo	4	0,608949
Full Metal Jacket	4	0,603370

Toepassingen

- Niet gepersonaliseerde aanbevelingen, seeded recommendations
- Gebruikers of films *in omgeving*, cosine similarity en L_p-norms

Toepassingen: films in omgeving

Titel	Cosine
Alien	1,000000
Aliens	0,791019
Alien: Covenant	0,776472
Alien: Resurrection	0,454473
Alien^3	0,120532
Prometheus	0,076686

Toepassingen: films in omgeving vervolg

Title	L_2
Alien	0,000000
The Dark Knight Rises	0,164957
Star Trek Into Darkness	0,177475
The Bourne Ultimatum	0,178117
Star Wars: Episode VII - The F...	0,186614
Contratiempo	0,190096
The Incredibles	0,202713
The Matrix	0,203655
The Exorcist	0,206783
Terminator 2: Judgment Day	0,209838

Uitbreidingen

- Domein bijzonder rijk
- Afhankelijk van doel
- Uitgaande van MovieTweetings

Uitbreidingen: data

- Herinterpreteren beoordelingen
- Verval beoordelingen

Uitbreidingen: algoritmes

- Collaborative Denoising Auto-encoder (CDAE)
- Ensembles
- Voorspellen van relevantie, Learning to Rank (LTR), Bayesian Personalized Ranking (BPR)

Uitbreidingen: trainen

- Waarde weinig beoordeelde films
- Loss

Uitbreidingen: evalueren

- Coverage
- Personalization
- Intra-list similarity
- t-distributed Stochastic Neighbor Embedding (t-SNE)

Uitbreidingen: overige

- Boosting
- Kant en klare oplossingen

Conclusie

- Inzicht verkregen
- Enigszins bruikbaar