RecSys Challenge 2016

May 15, 2016

1 Data Stats

- #Users who were displayed **at least** one impression: 2755168
- #Users who did **at least** one click: 522535

2 One interesting observation

Field by the name: "active_during_test" It is waste to recommend those items which have "active_during_test = 0" because the items are no longer active during the test week. But does that mean we don't train on them and drop them? Nope because score further drops down (I checked!). So it is important to train on them because they give us important signals about users but while testing we do not test on them.

3 Most Popular Approach

Algorithm 1 Evaluation Using Mean Squared Error

1: Filter out columns which have "Interaction type = 4"
2: Change all the interaction types to "1" columns which have
3: group by $item_{id}$ and reverse sort

Score: 43278.55

4 Active Popular Items

- Again take 30 most popular items
- But this time take active_during_test = 1 into account
- For each user predict 30 most popular active items
- Score: 73253.63

5 Matrix Factorization Using ALS of SPARK

- Number of Latent features in both items and users: 10
- Number of iterations: 10
- Score is Surpisingly low: 15550.28

6 Matrix Factorization Using ALS of SPARK

- Number of Latent features in both items and users: 100
- Number of iterations: 100
- Number of predicted items for each user: 100
- Score: 25429.54.28
- Score (active): 38646.14
- This can serve as matrix-factorization baseline
- FileName: als_v2
- Pro's:
  - Can now combine this along with content based approach

7 Matrix Factorization Using ALS of SPARK

- Number of Latent features in both items and users: 100
- Number of iterations: 100
- Number of predicted items for each user: 100
- Combined with interesting observation
- Score: 38646.14
- This can serve as matrix-factorization baseline
- FileName: als_v2_1
- Pro's:
  - Can now combine this along with content based approach after predicting 1000 items

8 Matrix Factorization in Python3 using SGD

- Again number of latent features in both items and users: 10
- Learning rate of sgd: 0.0005
- No regularization
- But taking care of our observation
- Number of Latent features in both items and users: 10
- Number of iterations: 10
- Score: 14808.21
- Surprisingly, all predicted items were same, some bug?

9 Personalized Activeness Popularity

- Sort items by taking into account users's past interactions
- Recommend by sorting into descending order
- Take into account active_during_test = 1
- Recommend at least 30 items by completing the list on the basis of popularity
- Score: 287524.88

10 Further Ideas

- Cross Validation
- Increase number of iterations, Increase number of latent factors
- Normalize rating between 1 and 5
- Regularization
- Hybridize between matrix factorization and content fitering and popularity
- Weighting the interactions