RecSys Challenge 2016

Sumit Sidana

April 15, 2016

1 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

2 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

3 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"active_during_test" It is waste to recommend those items which have "active_during_test = 0" because the items are no longer active during the test week.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.

4 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
- Still running on Tiger

5 Next Immidiate Step

- Spark: again with the interesting observation

6 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