Decipher 2.0

20.01.2022

Lokesh Dhaundiyal
Team - ld
IIT ISM Dhanbad

Problem statement

To design an algorithm to recommend courses and tests to users based on their experience and performance.

Goals

We want to design an algorithm that can recommend courses/tests to users.

The recommendation made by the algorithm should have the following properties- :

1. The recommendation should be related to the past experience of that user. Like the courses or tests that the users enrolled in the past.
2. The algorithm should also learn from the experience of other users while predicting for a specific user.
3. The recommendation should be dynamic so that they can give the best results with the updated data each time.
4. Must address the problem of cold start and popularity based biased
5. If required then it should be able to predict courses/tests which are more profitable than others much more often or in case of ties in prediction.

Solution Approach

Link to notebook - shorturl.at/cyI12

For designing the algorithm we use the hybrid approach of content-based recommendation plus Collaborative filtering using WALS algorithm with gradient descent as optimization instead of Normal equation and PCA for dimensionality reduction and Matrix Factorization.

Our solution can predict the list of courses and list of tests which are not already taken by the students, in decreasing order of their popularity and based on the past courses and tests taken by that user and other users too.

We calculate the prediction for users using both content-based approach and collaborative filtering and then combine them for final prediction.

By doing so our model doesn't suffer the problem of cold start and also helps users discover new items that are outside the subspace defined by their historical profile.

For Recommending Courses

I. Collaborative filtering through low-rank matrix factorization approach

Collaborative filtering through low-rank matrix factorization is a way of taking a sparse matrix of users and ratings, assuming a certain number of latent factors (k), and factoring out a lower-rank representation of all the users and items.

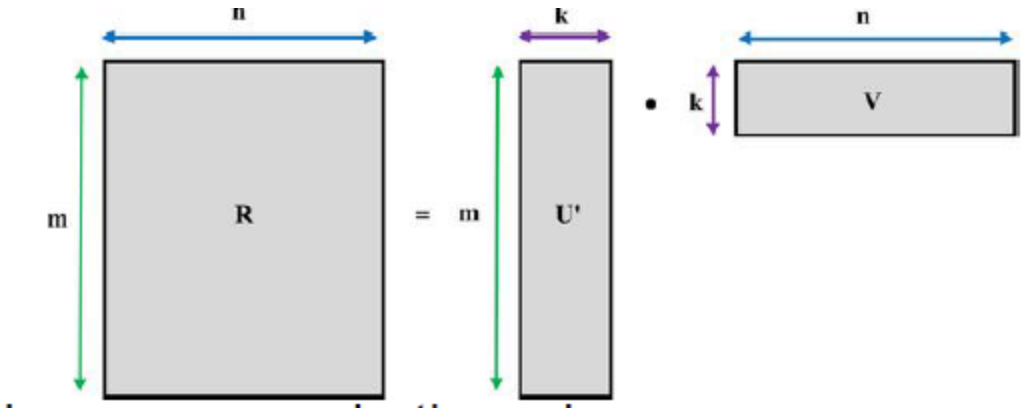
We use the target groups and super target groups as latent factors to decide the recommendation.

We output a matrix of shape (No_of_users, No_of courses) , in which each entry corresponds to a score that decides the rank of that course in the list of courses recommended to that user.

The more the score, the better will be that course for that user.

To find this matrix:

We factorize our matrix into two submatrices -:



User matrix of dim (no_of student * latent_fatcor)

Course_matrix of dim (no_of courses * latent_factor)

Here latent factor is a hyperparameter.

Our final matrix would be the dot product of the student matrix and the transpose of course_matrix.

However, we experiment with two choices-

1. We initialise our user matrix of shape (no_users,no_of target group + supergroup) . Each entry of this matrix contains 0 if that student is not a part of that group/supergroup else it will contain 1. We got a matrix of shape(10000,229).
2. Using much less value for latent_factor .i.e <50 and using PCA to initialize user and course matrix.

The rating matrix that we can obtain from them is:

rating_matrix=(user_matrix)*transpose((course_matrix))

Now we will implement the weighted alternating least square algorithm but with a twist. Instead of using normal equations for its optimization, we will be using gradient descent.

- WALS Algorithm

$$R_{mn} = U_m^T \cdot V_n$$

$$L^w = W \circ \sum_{m,n} (R_{mn} - U_m^T \cdot V_n)^2$$

In gradient descent, We will use a weight of 0.01 for the unobserved data.

After around 120 iterations we will obtain our user and corresponding feature matrix. Multiplying those we will get rating of every course for every students.

II. Content-based approach

Content-based recommender relies on items features for predicting items. They tend to be more robust against popularity bias and the cold start problem.

Here for predicting items, we use the cosine similarity between the user embedding represented by the one-hot encoding of the target group/super target group in which it is present and item embeddings which are represented similarly as above.

User embedding shape=[no_of_user , no_of_tgt grp + no_of_super tgt grp]

Item embeddings shape=[no_of_items , no_of_tgt grp + no_of_super tgt grp]

we return the items which are most similar to that user i.e have high cosine similarity.

III. Ensembling both collaborative and content-based results

By ensembling collaborative and content-based results we are able to make recommendations that can hopefully draw from the strengths of both methods.

We output the top 3 predictions from both the recommendations.

IV. Analysis of the results

Let us analyse our prediction for some randomly picked student ids

1. Student id = 807abc753fd9d5fd20282aad3389a45f

   Target groups or super target groups with which this student is connected-:

   [{'language': 'English', 'value': 'SSC MTS'}, {'language': 'Hindi', 'value': ''}]

   [{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

Top 5 predicted course by collaborative filtering-:

top 5 courses for studetnt having student id 807abc753fd9d5fd20282aad3389a45f

8dbdf3dc81630ef31b50e7846d1bdb9e score 1.517145130248537

details

[{'language': 'English', 'value': 'SSC MTS'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Selection Post'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRC Group D'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CHSL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CGL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB NTPC'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

8c87c97dfc97b6a69995a93ae5763413 score 1.4865124477539622

details

[{'language': 'English', 'value': 'SSC MTS'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'Railways Exams'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRC Group D'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CHSL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CGL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB NTPC'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

577e5a0e808d8cba932172b7796e3d08 score 1.455592331352638

details

[{'language': 'English', 'value': 'RRB ALP'}]

[{'language': 'English', 'value': 'SSC MTS'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'Railways Exams'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CHSL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CGL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB NTPC'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CPO'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB JE'}]

[{'language': 'English', 'value': 'SSC JHT'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

efb1474cc7b29b9c06669a01331015e0 score 1.4208442977591653

details

[{'language': 'English', 'value': 'Delhi Police Constable'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC MTS'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'Railways Exams'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRC Group D'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CHSL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CGL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB NTPC'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'Bihar Police SI'}, {'language': 'Hindi', 'value': ''}]

5425e4a2b9250f9098b947796a494c5c score 0.8205995303509955

details

[{'language': 'English', 'value': 'SSC Selection Post'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

Top 5 predicted courses by content filtering-:

efb1474cc7b29b9c06669a01331015e0 score 2

details

[{'language': 'English', 'value': 'Delhi Police Constable'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC MTS'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'Railways Exams'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRC Group D'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CHSL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CGL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB NTPC'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'Bihar Police SI'}, {'language': 'Hindi', 'value': ''}]

8dbdf3dc81630ef31b50e7846d1bdb9e score 2

details

[{'language': 'English', 'value': 'SSC MTS'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Selection Post'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRC Group D'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CHSL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CGL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB NTPC'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

8c87c97dfc97b6a69995a93ae5763413 score 2

details

[{'language': 'English', 'value': 'SSC MTS'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'Railways Exams'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRC Group D'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CHSL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CGL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB NTPC'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

577e5a0e808d8cba932172b7796e3d08 score 2

details

[{'language': 'English', 'value': 'RRB ALP'}]

[{'language': 'English', 'value': 'SSC MTS'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'Railways Exams'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CHSL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CGL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB NTPC'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CPO'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB JE'}]

[{'language': 'English', 'value': 'SSC JHT'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

ff9317caee0d6c22fc730a881f0fe1b2 score 1

details

[{'language': 'English', 'value': 'RRC Group D'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CHSL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CGL'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'RRB NTPC'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC CPO'}, {'language': 'Hindi', 'value': ''}]

[{'language': 'English', 'value': 'SSC Exams'}, {'language': 'Hindi', 'value': ''}]

For Recommending Tests

We can use the same approach as above for recommending tests .but here we can also use the marks obtained by users in their user-item rating matrix instead of 0 and 1.

Other Variations that we can incorporate in our algorithm

1. We can add contextual embeddings to evaluate the similarity between different courses by their titles .for generating embeddings we can either fine-tune present state of the art NLP models on our datasets or train them from scratch.

2. We can also incorporate some other knowledge-based features regarding our courses and users age, gender, location and pass them with other parameters through a neural network to calculate content-based prediction.