Assignment 4 part 1

PART I -

UNDERSTANDING RECOMMENDER SYSTEMS

(5 points)

Read the paper “Matrix Factorization Techniques for Recommender Systems” by Koren et al

available from any of the sources below

<http://ieeexplore.ieee.org/document/5197422/>

<http://dl.acm.org/citation.cfm>?id=1608614

https://datajobs.com/data-science-repo/Recommender-Systems-%5BNetflix%5D.pdf

Answer the following questions:

1. Recommender Systems (RS) are based on one of two strategies – content filtering and collaborative

filtering. Write down key features and properties of each of them.

The content filtering approach creates a profile for each user or product to characterize its nature

• Profile based

• require gathering external information that might not be available or easy to collect.

Collaborative filtering analyzes relationships between users and interdependencies among products to identify new user-item associations.

• Domain free

• More accurate than content based

• It’s a cold start problem due to its inability to address the system’s new products and users

2. Two primary areas of collaborative filtering are neighborhood methods and latent factor models. Write down key features of each of them.

Neighbor- hood methods are

• centered on computing the relationships between items or, alternatively, between users.

• item- oriented approach evaluates a user’s preference for an item based on ratings of “neighboring” items by the same user.

• the user-oriented approach identifies like-minded users who can complement each other’s

ratings.

latent factor models:

• alternative approach that tries to explain the ratings by characterizing both items and users

• less well-defined dimensions such as depth of character development or quirkiness;

• completely uninterpretable dimensions.

3. What do the “factors” in latent factor represent? How are these factors discovered?

Factors represent the measure of dimensions. Factors are inferred from the rating patterns.

4. What is the difference between explicit and implicit feedback for RS? What do you think

are advantages and disadvantages of each?

Recommender systems rely on different types of input data, which are often placed in a matrix with one

dimension representing users and the other dimension representing items of interest.

The most convenient data is high-quality explicit feedback, which includes explicit input by users regarding their interest in product. Usually, explicit feedback comprises a sparse matrix, since any single user is likely to have rated only a small percentage of possible items.

When explicit feedback is not available, recommender systems can infer user preferences using implicit feedback, which indirectly reflects opinion by observing user behavior including purchase history, browsing history, search patterns, or even mouse movements. Implicit feedback usually denotes the presence or absence of an event, so it is typically represented by a densely-filled matrix.

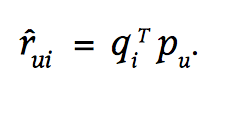
5. In the latent factor model for dimension f, the item i and user u are represented as vectors

. How is the interaction between item i and user u calculated?

The elements of **qi** measure the extent to which the item possesses those factors, positive or negative.

For a given user u, the elements of **pu** measure the extent of interest the user has in items that are high on the corresponding factors, again, positive or negative.

interaction between item i and user u:



6. What does the learning system try to minimize? Understand the meaning of each term in

the equation.

The learning system tries to minimize the regularized squared error on the set of known ratings.

7. There are 2 learning algorithms for latent factorization – stochastic gradient descent (SGD)

and alternating least squares (ALS). What are the advantages of ALS over SGD?

ALS is favorable in at least two cases. The first is when the system can use parallelization. In ALS, the

system computes each qi independently of the other

item factors and computes each p independently of the u other user factors. This gives rise to potentially massive parallelization of the algorithm.9 The second case is for systems centered on implicit data. Because the training set cannot be considered sparse, looping over each single training case—as gradient descent does—would not be practical. ALS can efficiently handle such cases.

8. Read about the Netflix competition and the authors’ entry. What were the most descriptive dimensions (features) that their models discovered? Summarize briefly. Also, mention what metric do they use to check the performance of their models.

Factorizing the Netflix user-movie matrix allows us to discover the most descriptive dimensions for predicting movie preferences. We can identify the first few most important dimensions from a matrix decomposition and explore the movies’ location in this new space.

Movies are placed per their factor vectors. Someone familiar with the movies shown can see clear meaning in the latent factors. The first factor vector (x-axis) has on one side lowbrow comedies and horror movies, aimed at a male or adolescent audience (Half Baked, Freddy vs. Jason), while the other side contains drama or comedy with serious undertones and strong female leads (Sophie’s Choice, Moonstruck). The second factorization axis (y-axis) has independent, critically acclaimed, quirky films (Punch-Drunk Love, I Heart Huckabees) on the top, and on the bottom, mainstream formulaic films (Armageddon, Runaway Bride).