PARTI- 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.

Content filtering: Creates a profile for each user or product to characterize its nature. Require gathering external information that might not be available or easy to collect.

Collaborative filtering: Relies only on past user behavior, without requiring the creation of explicit profiles. Analyzes relationships between users and interdependencies among products to identify new user-item associations.

A major appeal of collaborative filtering is that it is domain free, yet it can address data aspects that are often elusive and difficult to profile using content filtering. While generally more accurate than content-based techniques, collaborative filtering suffers from what is called the cold  
start problem, due to its inability to address the system’s new products and users. In this aspect, content filtering is superior.

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

Neighborhood methods are centered on computing the relationships between items or, alternatively, between users. The item - oriented approach evaluates a user’s preference for an item based on ratings of “neighboring” items by the same user. A product’s neighbors are other products that tend to get similar ratings when rated by the same user.

Latent factor models are an alternative approach that tries to explain the ratings by characterizing both items and users on, say, 20 to 100 factors inferred from the ratings patterns.

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

In a sense, such factors comprise a computerized alternative to the aforementioned human - created song genes.

For movies, the discovered factors might measure obvious dimensions such as comedy versus drama, amount of action, or orientation to children; less well-defined dimensions such as depth of character development or quirkiness; or completely uninterpretable dimensions.

For users, each factor measures how much the user likes movies that score high on the corresponding movie factor.

4. What is the difference between explicit and implicit feedback for RS? What do you think are advantages and disadvantages of each?

The most convenient data is high-quality explicit feedback, which includes explicit input by users regarding their interest in products. 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.

The advantage of explicit feedback is that it is more accurate and convenient to make conclusion. The disadvantage is that it is hard to get the feedback and data from users, and only a small percentage of possible items can be rated by users.

The advantage of implicit feedback is that it is easy to collect from users, and it can be represented by a densely filled matrix. The disadvantage is that implicit feedback may not be very accurate, since it is derived from implicit feedback.

5. In the latent factor model for dimension f, the item i and user u are represented as vectors 𝑞𝑖 and 𝑝𝑢 ∈ 𝑅𝑓 . How is the interaction between item i and user u calculated?

The resulting dot product, qi^T\*pu, captures the interaction between user u and item i — the user’s overall interest in the item’s characteristics. This approximates user u’s rating of item i, which is denoted by rui, leading to the estimate ^rui = qi^T\*pu

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

The system minimizes 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?

While in general stochastic gradient descent is easier and faster than ALS, 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 pu independently of the other user factors. This gives rise to potentially massive parallelization of the algorithm. 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 people to discover the most descriptive dimensions for predicting movie preferences.

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). There are interesting intersections between these boundaries: On the top left corner, where indie meets lowbrow, are Kill Bill and Natural Born Killers, both arty movies that play off violent themes. On the bottom right, where the serious female-driven movies meet the mainstream crowd-pleasers, is The Sound of Music. And smack in the middle, appealing to all types, is The Wizard of Oz.

Plain factorization, adding biases, enhancing user profile with implicit feedback, and two variants adding temporal components. The accuracy of each of the factor models improves by increasing the number of involved parameters, which is equivalent to increasing the factor model’s dimensionality, denoted by numbers on the charts.