AI 705: Midterm Project

I. DESCRIPTION

The following are the steps required for the midterm project. Additonally, you are encouraged to build upon these points by incorporating ideas you learnt in the class. The data set consists of ratings for core courses and electives for each student. The steps are as follows:

1) First, construct a user vector. It can be as simple as just the ratings of each user for all the "core" courses (because they will be present).
2) Based on the user vectors obtained for all the users, perform the "K-means clustering".
3) Now, for each cluster, you can infer the ratings of all electives (by taking average of ratings available for elective).
4) For example, say we have 162×25 incomplete matrix (i.e., many missing entries) initially, where 162 are number of users, 5 are core courses, and 20 are electives. Each user vector will be a 5×1 vector. After performing K-means clustering, with, say, 10 clusters, your final matrix is 10×25, which is complete (i.e., no missing entries).
   Note: You are working on real data. So, you might encounter a situation where, after K-means, your matrix might not be complete. You must handle such situations with a combination of common sense and right mathematical technique.
5) Now perform SVD on this 10×25 matrix as a noise reduction paradigm.
6) The final matrix after SVD is your prediction matrix that is used to make inferences. For example, for a given user, we first find the cluster she/he belongs to. Next, we order the elective courses based on the values in the cluster row she/he belongs to.

II. EVALUATION METRICS

Following are the evaluation metrics used to grade your midterm presentations:

1) **Analysis of algorithms:** Why were a certain number of clusters selected? How many singular values were removed and why?
2) **Analysis of Results:** Through plots and metrics (not just descriptive reasoning), what are the interesting patterns observed in the results? How did SVD help? Reiterating that, your arguments **must** be supported by numbers, metrics, plots.

3) **Novelty:** What novel ideas are incorporated in this project and why? What were your motivations for using novel ideas.

4) **Real Time inference:** On a real test points I provide during inference, explain your results. Answering the conceptual questions asked during the presentation.

5) **Presentation:** The quality of slides and oral presentation, **as a group**.

It is important to note that you will be graded as a group and not at an individual level.

III. SANITY CHECKS

The following are the sanity checks every project **must** pass. Otherwise, your submission will be invalidated:

1) The codes will go through rigorous Plagiarism checks.

2) The plots/numbers/metrics in the presentation slides should match the output when your submitted codes are run. Fudging of results will not be tolerated at any degree.

3) Basic checks will be done to validate that **all** team members are involved in the project during the presentation.

IV. FORMAT OF PRESENTATION

The duration and slots of each presentation will be decided based on number of teams and feasiblity. It will be communicated very soon once all logistics are finalised. Take notice of the following points:

1) Teams must respect their presentation duration and slot with highest precision. Penalty will be levied if a group exceeds their alloted duration (proportional to number of additional seconds consumed).

2) I will decide the order and duration of each speaker in a group. For example, say there are four members in the group A, B, C, D. I can first ask B to present. After a random amount of time, I will ask D to take over, and this process is repeated.

3) The teams must upload their presentation and implementation codes (with instructions on how to run your code) immediately after the presentation. The submission folder will be shared with you in due course.