Assignment #3: Dimensionality Reduction methods
Submission Date: During Week 14 Lab (Tuesday, December 31, 2024)
10% Assignment

1. Instructions

- The total number of pages if this file is 5 pages.
- Students complete the assignment in order.
- Each student must use his/her GitHub account that was created in ASSIGNMENT 1.
- Each student **MUST** prepare the homework solution and write the report **BEFORE ARRIVING WEEK 14 LAB**. Your write-up includes the requirements stated in Section 3.5. HANDWRITTEN scans for your write-up ARE NOT ACCEPTED.
- **Each student must finish/compose the assignment during week 14 lab and post the solution and the pre-prepared report to GitHub.**
- Failure to attend week 14 lab and complete the assignment will result in a ZERO in this assignment.

2. Introduction

This assignment consists of the following THREE parts.
Part 1: PCA Method with Mean-Filling.
Part 2: PCA Method with Maximum Likelihood Estimation.
Part 2: Singular Value Decomposition (SVD) method.

3. Assignment Description and requirements

3.1. General requirements:

1. Use the dataset you generated in ASSIGNMENT 1.
2. Adjust the rating on a 1-to-5 scale.
3. Count the total number of users in your dataset and save it a variable called "*Tnu*".
4. Count the total number of items in your dataset and save it a variable called "*Tni*".
5. Count the number of ratings for every product in this dataset.
6. Draw the distribution of ratings (count of ratings against rating values [0, 1, ...,5]). Show if the matrix is sparse or not. Show if there is bias or not, and the level of bias if it exists.
7. Choose the two lowest rated items (name them I1 and I2) and consider them target items.
8. Save all the above results for later use.

3.2. Part 1: PCA Method with Mean-Filling:

Use the PCA method with mean-filling technique and compute the covariance matrix then compute the rating prediction for the target items I1 and I2.

Part 1 Requirements and questions

3.2.1. Calculate the average rating for each of the target items (I1 and I2).
3.2.2. Use the mean-filling method to replace the unspecified ratings of each of the target items (I1 and I2) with its corresponding mean value.
3.2.3. Calculate the average rating for each item.
3.2.4. For each item, calculate the difference between ratings and the mean rating of the item.
3.2.5. Compute the covariance for each two items.

3.2.6. Generate the covariance matrix.
3.2.7. Determine the top 5-peers and top 10-peers for each of the target items (I1 and I2) using the transformed representation (covariance matrix).
3.2.8. Determine reduced dimensional space for each user in case of using the top 5-peers.
3.2.9. Use the results from point 3.2.8 compute the rating predictions of the original missing rating for each of the target items (I1 and I2) using the top 5-peers.
3.2.10. Determine reduced dimensional space for each user in case of using the top 10-peers.
3.2.11. Use the results from point 3.2.10 to compute the rating predictions of the original missing rating for each of the target items (I1 and I2) using the top 10-peers.
3.2.12. Compare the results of point 3.2.9 with results of point 3.2.11. Comment on your answer.

3.3. Part 2: PCA Method with Maximum Likelihood Estimation

Use the PCA method with MLE technique and compute the covariance matrix then compute the rating prediction for the target items I1 and I2.

For simplicity, assume the Maximum Likelihood Estimate of the covariance between each pair of items is estimated as the covariance between only the specified entries. i.e, only the users that have specified ratings for a particular pair of items are used to estimate the covariance. If there are no users in common between a pair of items, the covariance is estimated to be 0.

Part 2 Requirements and questions

3.3.1. Generate the covariance matrix.
3.3.2. Determine the top 5-peers and top 10-peers for each of the target items (I1 and I2) using the transformed representation (covariance matrix).
3.3.3. Determine reduced dimensional space for each user in case of using the top 5-peers.
3.3.4. Use the results from point 3.3.3 compute the rating predictions of the original missing rating for each of the target items (I1 and I2) using the top 5-peers.
3.3.5. Determine reduced dimensional space for each user in case of using the top 10-peers.
3.3.6. Use the results from point 3.3.5 to compute the rating predictions of the original missing rating for each of the target items (I1 and I2) using the top 10-peers.
3.3.7. Compare the results of point 3.3.3 with results of point 3.3.6. Comment on your answer.
3.3.8. Compare the results of point 3.2.9 with results of point 3.3.4. Comment on your answer.
3.3.9. Compare the results of point 3.2.11 with results of point 3.3.6. Comment on your answer.

3.4. Part 3: Singular Value Decomposition (SVD) method

In this part you MUST use the SVD technique to generalize eigenvalue decomposition to non-square matrices. i.e., decompose the incomplete ratings matrix R into a product of three matrices, i.e. $\boldsymbol{R} = \boldsymbol{U\Sigma V^T}$, where $\boldsymbol{U}$ that is $\boldsymbol{m \; x \; m}$ **orthogonal** (***its transpose is its inverse***), $\boldsymbol{V}$ that is $\boldsymbol{n \; x \; n}$ **orthogonal**, and $\sum$ that is $\boldsymbol{m \; x \; n}$ **diagonal** (***non-zero values only at its diagonal***).

Apply the concept of **truncated SVD** on the ratings matrix you generated in ASSIGNMENT 1 by adopting the **low-rank** assumption, to approximate your full dimensional matrix using a **small number** of **factors (k-features)** that capture the data's essential structure and main patterns.

Part 3 Requirements and questions

3.4.1. Calculate the average rating for each item.

3.4.2. Use the mean-filling method to replace the unspecified ratings for each item.

3.4.3. Compute the eigenvalues $(\lambda_1, \lambda_2, \lambda_3, \cdots \lambda_n)$ and their corresponding eigenvectors $(\vec{v}_1, \vec{v}_2, \vec{v}_3, \cdots \vec{v}_n)$ of the ratings matrix you generated in ASSIGNMENT 1.

3.4.4. Make sure that the set of eigenvectors that you calculated in step 3.4.3 are mutually orthogonality (i.e. the dot product of every pair of vectors is equal to zero). If they are orthogonal go to step 3.4.6, if not go to step 3.4.5.

3.4.5. Perform Vector Normalization to convert them to orthogonal, and linearly independent vectors.

3.4.6. Check if the set of eigenvectors that you calculated in step 3.4.3 are orthonormal (i.e. every vector MUST have magnitude 1). If they are orthonormal, assume the new eigenvalues equals the original eigenvalues $(\sigma_1 = \lambda_1, \sigma_2 = \lambda_2, \cdots \sigma_n = \lambda_n)$ and go to step 3.4.7.2, if not go to step 3.4.7.

3.4.7. Apply Gram-Schmidt method to convert them into an orthonormal set of vectors. This done by:

3.4.7.1. Assuming the first orthonormal vector equals one of the orthogonal vectors (you may assume it is the one that corresponds to the highest eigenvalue) $\vec{u}_1 = \vec{v}_1$.

3.4.7.2. Normalizing $\vec{u}_1$ $(\vec{e}_1 = \frac{\vec{u}_1}{|\vec{u}_1|})$.

3.4.7.3. Assigning the highest eigenvalue to the first new eigenvalue $(\sigma_1 = \lambda_x)$.

3.4.7.4. Calculating the new predicted vector $\hat{u}_1 = \frac{Ae_1}{\sigma_1}$.

3.4.7.5. Calculating the projection of the vector $\vec{u}_1$ on $\vec{v}_2$ $(Proj_{\vec{u}_1}\vec{v}_2)$.

3.4.7.6. Constructing the new orthonormal vector as $(\vec{u}_2 = \vec{v}_2 - Proj_{\vec{u}_1}\vec{v}_2)$.

3.4.7.7. Normalizing $\vec{u}_2$ $(\vec{e}_2 = \frac{\vec{u}_2}{|\vec{u}_2|})$.

3.4.7.8. Calculating the new eigenvalue $\sigma_2 = \sqrt{\frac{A^T A \cdot \vec{e}_2}{\vec{e}_2}}$.

3.4.7.9. Calculating the new predicted vector $\hat{u}_2 = \frac{Ae_2}{\sigma_2}$.

3.4.7.10. Repeat steps 3.4.7.4 to 3.4.7.7 to calculate the rest of normalized orthonormal vectors $(\vec{e}_3, \vec{e}_4, \cdots \vec{e}_n)$ and their corresponding new predicted vector $(\hat{u}_1, \hat{u}_2, \cdots \hat{u}_n)$ and new eigenvalues $(\sigma_3, \sigma_4, \cdots \sigma_n)$.

3.4.8. Construct the predicted waiting matrix $\widehat{\boldsymbol{\Sigma}}$ from the eigenvalues $(\sigma_1, \sigma_2, \cdots \sigma_n)$ on the main diagonal.

3.4.9. Constructing the items matrix $\hat{V}$, its columns are the set of orthonormal vectors $(\vec{e}_3, \vec{e}_4, \cdots \vec{e}_n)$.

3.4.10. Construct the predicted user matrix $(\widehat{U})$, its columns are the predicted vectors $(\hat{u}_1, \hat{u}_2, \cdots \hat{u}_n)$.

3.4.11. Use the results from points 3.4.8 -to- 3.4.10 to construct the newly reduced rating matrix $\hat{R} = \widehat{U}\hat{\Sigma}\hat{V}^T$.

3.4.12. Use the results from point 3.4.11 to find missing ratings in the original rating matrix for each of the target items (I1 and I2).

3.5. Discussion and Conclusion:

During the homework and **BEFORE YOU COME TO LAB 13**, prepare a report that include the following

1- A section called "Summary and Comparison" summarizes and compares the results of part 1, part 2 and part 3, emphasizing the accuracy of predicting the missing rating, and the pros and cons of each method.

2- A "conclusion" section, which summarizes your own comments and conclusion, shows the impact of matrix factorization method.

4. Method and significance of the assessment

- This coursework accounts for **10%** of the course mark and is done **INDIVIDUALLY**.
- This assessment will demonstrate progression of the student's academic ability and assess students' achievement intended learning outcomes (ILOs) of the course.
- The marks will be awarded pro-rata, depending on the details, evidence of academic and technical talents, professionalism, information organization, and discussion skills.
- A **40%** reduction will be applied if evidence of thoroughness is not adequate.
- If the student appears to be performing it right but makes a mistake, they will receive a maximum of 70% for any of the requirements.

5. Marking schema:

- 10% of the mark for the report as described in Section 3.5, including evidence of the presented knowledge, topic understanding, completeness of the information, analysis, and the contribution of the student.
- 80% of the mark for the implementation, experiments, and results according to the requirements.
- 10% of the mark for posting the report, code, dataset, and the PLAGIARISM REPOR on GitHub.

6. Report (paper) Format:

Report must follow the following.

- be word-processed,
- a report format standard,
- use correct paragraphing, formal grammar, tenses, and spelling,

- be submitted on A4,
- The title page includes the Assignment title <AIE425 Intelligent Recommender Systems, Fall Semester 24/25> on the first line, <Assignment #2: Significance Weighting-based Neighborhood CF Filters> on the second line, and <Student ID, Full Name> on the third line. Use double line space, centered contents and without page numbering.
- All other pages are 1.5 line spaced,
- use 12-point Arial font size for normal text, 12-point Arial font size, bolded for headings,
- have page numbers centered on the bottom of each page in the format <Page X of Y>,
- section headings are with the same font, size and alignment.

7. Plagiarism and Academic Honesty

- INTEGRITY and COLLABORATION: Student are encouraged to discuss issues related to the assignment with other students, but genuine collaboration on all or part of the assignment must be explicitly acknowledged, or he/she will be penalized.
- PLAGIARISM is strictly prohibited and may result in failure in this course.
- This is an exercise, so submit your own work. If you submit material that is not entirely your own, you must state this clearly in your submission.
- A PLAGIARISM REPORT is required for the submitted report. A similarity ratio of greater than 30% is not acceptable.
- Code and/or written material should NOT be shared. Even a single line of code without reference will be considered plagiarism.
- Please do not use external code unless authorized. This is easier to detect than you might think. If you must do so for any standard initialization and embedding parts, this must be referenced and properly marked/highlighted.
- ASK Teaching Staff If you have any queries or are confused whether specific activities constitute dishonesty. It's better to be safe than sorry.

8. Feedback given to students in response to assessed work.

- Specific oral feedback on the assessed components.
- The written component will be assessed directly through annotations on the page.
- Feedback for programming will also be placed on the coursework assessment sheet returned with the coursework mark.
- During contact hours, students will receive oral generic comments on every part of their assignment..
- If students need more input, they are encouraged to speak with the teaching staff.

9. Deliverables

- The whole assignment (report, dataset, and code) will be submitted on GitHub during week 14 lab.
  - All files MUST be named as follow: <StudentID_StudentFirstName_FileName>.
  - Put your name, and ID at the top of each file even in programs.
- Late submission, if there is a solid reason, is permissible during the next day subject to the arrangement with the teaching staff, however you will lose 50% of the coursework mark.
- more than 48 hours delay is not accepted, and you will get **ZERO** in this assignment.

----------- END -----------