IEMS5709 Fall2024
Homework 4

Release date: Nov 25, 2024
Due date: Dec 9, 2024 (Monday) 11:59 PM
No late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student__________________________) Date:_______________________________

Name___________________________________ SID________________________________

Submission notice:

- Submit your report in a single PDF document on Elearning

General homework policies:

A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value, and justify any assumptions you make. You will be graded not only on whether your answer is correct, but also on whether you have done an intelligent analysis.

Q1 [20 marks]: Singular Value Decomposition (SVD) for Dimensionality Reduction

The table below shows 5 individuals' support level for distinct football teams. Each row of the table represents a person's support level over 6 teams (from 5 to 1 star), and the columns of the table correspond to distinct teams.

	Paris Saint-Germain	**Manchester City**	**Arsenal**	**Bayern Munich**	**Real Madrid**	**Barcelona**
A	4	3	5	4	3	5
B	5	4	3	1	4	2
C	3	1	5	5	1	5
D	3	4	1	3	4	3
E	4	3	4	5	2	5

(a) **[10 marks]** Originally, each person's support level is presented in a 6-D space. Using SVD, this set can be approximately embedded into a 2-D space instead. Show the resultant U, Σ, V^T for achieving this goal. (Feel free to perform the SVD by hand or using any other package, e.g. Matlab, Mathematica, and Python.)

(b) **[10 marks]** Based on the SVD representation, one can observe that there are 2 dominant "concepts". For the rest of this question, we will use the corresponding two dimensional concept space to approximate the original dataset.

1. **[5 marks]** For a new person with support levels [5 3 4 3 5 1], what is the representation of his/her performance in the "concept" space?
2. **[5 marks]** Compute the cosine similarities between A and C based on their vectors in the original space and the new "concept" space.

Q2 [50 marks]: K-means with PCA and Eigendigits

(a) [35 marks] Refer to page 85 of lecture notes Dimension Reduction [1] on "PCA with EigenFaces". By applying similar PCA techniques on the training dataset of the handwritten digits in Q2 of Homework#3, one can (approximately) represent each 28x28-pixel image of a handwritten digit as the linear combination of M (e.g. = 20) principal "eigendigits" (i.e., eigenvectors).

Re-do the K-means cluster (K=10) as well as the handwritten digit classification in Q2(b) of Homework#3 under the reduced M-dimensional space, with the following requirements:

1. Plot out the values of eigenvalues in decreasing order. (x-axis: index; y-axis: eigenvalue) Hint: The resultant figure will be similar to page 84 of lecture notes, with (1, λ1), (2, λ2) … corresponding to the data points of the largest eigenvalue, second largest eigenvalue, etc.
2. Try multiple choices of M (i.e. the number of eigendigits to be kept) before applying K-means clustering. Then:
   (i) Compare the corresponding performance (in terms of *Classification Accuracy*) when using the top 4, 8, 16, 32, 64 eigendigits, as the principal vectors (and **(ii)** state the corresponding "energy kept" in each case, see page 52 of lecture notes), and **(iii)** compare with the *Classification Accuracy* result from Homework#3-Q2(b).
3. Explain your observations and the trade-offs involved in picking different M.

Note:

1. Set K=10 for K-means in all cases.
2. Use the centroids (seeds) `random_seed_1` provided in HW#3's `random_seeds.rar` [2] for centroid initialization.
3. You can implement the PCA (dimension reduction) part on your local machine.
4. You can call existing PCA libraries for this task.
5. If you had trouble implementing K-Means in HW#3 Q2(b), you may consult TAs for the suggested solution of that previous question, and based on which implement your own code for this task.
6. Submit both your code and results.

(b) [15 marks] Visualization. Having implemented PCA, we obtain **M** principal vectors of the training set. Similar to the discussion on "eigenfaces" in the lecture, we can also visualize the "eigendigits" in this question.

Please display "the images" corresponding to the **top 64** principal vectors ("eigendigits"). (You may implement the visualization program by yourself, or use the Python code provided in [2] for visualization.)

Submit the results (visualized images) of your visualization as well as your observations (and your self-implemented visualization programs if any).

Q3 [30 marks]: Recommender Systems

Consider the following incomplete book rating matrix:

	Book A	Book B	Book C	Book D	Book E	Book F
User 1	2	1	5	4	3	
User 2		2		3	5	4
User 3	5		4	1	4	2
User 4	2	3	4	5	?	
User 5		4	1		3	2

(a) [10 marks] Memory-based Collaborative Filtering. Calculate the predicted rating of **User 4** on **Book E** using:

(i) **[5 marks]** Item-Item Collaborative Filtering

(ii) **[5 marks]** User-User Collaborative Filtering

Note: Please

1. Select the **top 2** nearest neighbors when computing the predicted rating;
2. Use **Pearson correlation** as the similarity metric.

(b) [10 marks] Model-based Collaborative Filtering. Matrix Factorization techniques are effective to discover the latent features underlying the interactions between users and items. A matrix factorization example and its Python code are provided in the blog of Ref [4]. Please read the blog in [4] to understand the python code and then use it to predict the rating of **User 4** on **Book E**. Compare the result with the ones you obtained in part (a).

Note: Specify the number of features (K in the Python code) as 3.

(c) [10 marks] Bug Correction. Actually, there is a “bug” in the source code provided in [4]. The bug is related to a common mistake during the implementation of Gradient Descent. Identify the mistake and correct it. Use the corrected code to predict the rating of **User 4** on **Book E** again and compare the result with that in part (b) in terms of the final objective value.

References

[1] Lecture Slides on Dimension Reduction:
https://mobitec.ie.cuhk.edu.hk/ierg4300Fall2024/static_files/slides/DimReductionIERG4300IEMS5709Fall2024.pdf

[2] Two seeds for K-means initialization:
https://mobitec.ie.cuhk.edu.hk/ierg4300Fall2024/static_files/homework/random_seeds.zip

[3] Principal Vectors Visualization:
https://mobitec.ie.cuhk.edu.hk/ierg4300Fall2024/static_files/homework/viz_principal_vectors.py

[4] Matrix Factorization: A Simple Tutorial and Implementation in Python:
http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/