IERG4300 /ESTR4300/ IEMS5709 Fall 2021
Homework 4

Release date: Nov 23, 2021

Due date: Dec 7, 2021 (Tuesday) 11:59pm

The solution will be posted right after the deadline, so 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+10 (bonus) marks]: Numerical Example of PCA

A numerical example may clarify the mechanics of principal component analysis. Let us analyze the following 4-variate dataset with 8 observations. Each observation consists of 4 measurements.

$$X = \begin{bmatrix} 7 & 4 & 3 & 4 \\ 4 & 1 & 8 & 3 \\ 6 & 3 & 5 & 2 \\ 8 & 3 & 2 & 10 \\ 4 & 5 & 0 & 9 \\ 1 & 3 & 2 & 5 \\ 6 & 6 & 3 & 2 \\ 8 & 3 & 3 & 6 \end{bmatrix}$$

(a) [20 marks] In this part of the question, you need to represent (with approximation) the 4-variate dataset (containing 8 data points) shown above using only **two** variables instead. You should perform PCA with the help of standard numerical analysis routines/ packages like Matlab or Mathematica or Numpy. Show the (approximate) position of the original 8 observations in a 2-D plot after performing dimension reduction. Include all the key steps and intermediate results in your submission.

(b) [10 marks - Bonus for IERG4300 and Mandatory for ESTR4300 and IEMS5709] Re-do part (a) based on SVD instead of PCA. Compare your answer with that of part (a).

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

(a) [35 marks] Refer to page. 78 of the lecture notes 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 "eigendigit".

Re-do the K-means cluster as well as the handwritten digit classification in Q2 of Homework#3 under the reduced dimensional space (let **M=20** dimensions). Compare your results with those from Homework#3-Q2(b) and explain your observations. You can implement the PCA (dimension reduction) part in your local machine. Submit both your code and results. (You may extend either your own codes from Homework#3 or refer to the codes provided in our suggested solutions for Homework#3 from Tutorial 11).

(b) [15 marks] 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 **first 10** 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 movie rating matrix:

	Movie A	Movie B	Movie C	Movie D	Movie E	Movie F
User I	2	1	5	4	3	
User II		2		3	5	4
User III	5		4	1	4	2
User IV	2	3	4	5	**?**	
User V		4	1		3	2

(a) Calculate the predicted rating of User **IV** on Movie **E** using:

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

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

Please select the **top 2** nearest neighbors when computing the predicted rating.

(b) [10 marks] 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 [1]. Please read the blog in [1] to understand the python code and then use it to predict the rating of User IV on Movie E. Compare the result with the ones you obtained in part (a).

(c) [10 marks] Actually, there is a “bug” in the source code provided in [1]. 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 **IV** on Movie **E** again and compare the result with that in part (b) in terms of the final objective value and the number of iterations required for convergence.

Reference

[1]http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/
[2] http://mobitec.ie.cuhk.edu.hk/ierg4300Fall2021/homework/viz_principal_vectors.py