ENGG4030/ESTR4300 Spring 2018 Homework 4

Due date: Apr 22 (Sunday), 2018 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 [50 marks + 20 bonus]: K-means with PCA

(a) [50 marks] Refer to pg. 70 of the lecture notes on "PCA with EigenFaces". By applying similar PCA techniques on the training dataset of the handwritten digits in Q1 of Homework#3, one can (approximately) represent each 28x28-pixel image of handwritten digit as the linear combination of M (e.g. = 25) principal "eigen-digits".
Re-do the K-means cluster as well as the handwritten digit classification in Q1 of Homework#3 under the reduced dimensional space (25 dimension). Compare your results with those from Homework#3 and explain your observations. Submit both your code and results. (You may extend either your own codes from Homework#3 or the codes provided in our suggested solution for Homework#3).

(b) [20 Bonus marks for ENGG4030 and Mandatory for ESTR4300] Implement GMM with PCA. Re-do the handwritten digit classification in Q1 of Homework#3 under the reduced dimensional space (25 dimension), and replace the K-means algorithm with the Gaussian Mixture Model (refer to pg. 38 of the lecture nodes on "Clustering"). Compare the classification results of K-means with PCA and GMM with PCA. Submit both your code and result.

Note:

1. You don't have to use MapReduce to implement PCA (of course you are welcome to have a try). As for the K-means and GMM algorithm, you still need to use MapReduce to implement them.
2. You should redo all three questions of Q1 of Homework#3, for both K-means with PCA and GMM with PCA. Follow the same instructions and requirements of Q1 of Homework#3.

Q2 [50 marks]: Recommender Systems

Consider the following incomplete movie rating matrix:

	Movie A	Movie B	Movie C	Movie D	Movie E	Movie F
User I	1	1	6	4	4	
User II		3	?	4	5	4
User III	6			2	4	4
User IV	2	1	4	5		5

User V	4	4	2		3	1

(a) [10 marks] Perform Item-to-Item Collaborative Filtering to predict the rating of User II on Movie C based on the cosine similarity metric. State any additional parameters you choose for your operation. Show your steps and explain how you handle the negative effect of missing ratings in the matrix.

(b) [10 marks] Perform User-to-User Collaborative Filtering to redo the prediction of the rating of User II on Movie C based on the cosine similarity metric. State any parameters you choose for your operation. Show your steps. Explain the difference of the prediction result (if any).

(c) [20 marks] Compared to Collaborative Filtering, Matrix Factorization techniques are usually more effective because they allow us to discover the latent features underlying the interactions between users and items. A matrix factorization example and python code are provided in Ref [1]. Please read the blog in [1] to understand the code and then use the programme given in the blog to predict the rating of User II on Movie C.

(d) [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 II on Movie C and compare the result with that in part (c) 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/