Big Data Analytics Assignment 4

Due Date: 23:59 Jul 21, 2020

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

I declare that the assignment submitted 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.

Signed(Student__________________________) Date:_________________________________

Name____________________________________SID__________________________________

Submission notice:

- Submit your homework to Email xuhl@dgut.edu.cn

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.

Matrix Factorization

In this question, you will implement a Probabilistic Matrix Factorization algorithm to predict the potential ratings that users would assign to items. Distributed with this file are “data.tar.bz2” which contains two files. The file named “training.dat” contains all the training data. Each line of this file is a tab separated triplet, of which the first column is the user id, the second column is the movie id and the last column is the rating the user assigned to the movie. The other file named “testing.dat” contains all the challenges. Each line of this file is a tab separated numbers, of which the first is the user id and the second is the movie id.

Based on the training data given in “training.dat”, you are to implement a Probabilistic Matrix Factorization method. Train your model on the training data, and output your predicted ratings for all the challenges, in the exactly same order as the “testing.dat”. Only output the predicted number (Don’t print the user id and movie id, you can output integers or floats), one number each line.

Your score for this question will be based on your RMSE value you can achieve. The definition of RMSE you can find from here: http://en.wikipedia.org/wiki/Root-mean-square_deviation

Your prediction will be compared with the ground truth. So try to be as accurate as possible. You can modify the PMF algorithm where you see as appropriate. You will submit two files for this assignment: one file containing the source code of your program and another txt file containing your response to all the challenges. Be sure to comment your source code so that the tutor can follow it.

References:

[1] DataSet

https://xuhappy.github.io/courses/BigData/homework/data/MF.zip