CE888: Data Science and Decision Making

Lab 4: Recommender systems

Ana Matran-Fernandez

4 February 2019

Institute for Analytics and Data Science
University of Essex

Table of contents

1. Setting up
2. Lab exercises

Setting up

CE888 repository

□ If you have changed anything in your local repository since the last time you were in this computer, make sure you do: `git pull` from the repository folder.

□ This will download all the changes you did into your local folder.

Downloading the lab 4 materials

- □ Go to the Moodle page for this week:
- □ `https://moodle.essex.ac.uk/course/view.php?id=6683&section=10`
- □ Download the slides and code for today's practice into your local Github directory (e.g., `/labs/lab4`).
- □ Unzip the code, commit and push it before you make any changes.

Lab exercises

Lab materials

- □ Inside `lab4` you will see 2 ipython notebooks
- □ Open them and see what is inside:
  - □ Rec_correct.ipynb
  - □ Rec_features.ipynb
- □ Have a look. They're basically the implementation of what we saw in today's lecture.
- □ After this, you will create your own notebook and work on a new dataset (see next slide).

Lab exercises

- □ Create a new ipython notebook
- □ Load the data from the file `jester-data-1.csv`
    - □ The data is from http://eigentaste.berkeley.edu/dataset/ and it contains the ratings of 101 jokes from 24,983 users
    - □ The jokes are here
- □ Label approx 10% of the dataset cells as 99, to denote they are part of the validation set. Keep the the actual values of the cells so you can use them later.
- □ Use latent factor modelling to infer the hidden ratings of the users (they are labeled as ”99” in the dataset) on the training set
- □ Calculate the performance of the algorithm on the validation dataset
- □ Change hyper-parameters (i.e. learning rates, number of iterations, number of latent factors etc) as needed so you can get good results
- □ Report the MSE on the test dataset
- □ (Bonus) Use pandas to find the best and the worst rated jokes