CSE 158/258

Web Mining and Recommender Systems

Assignment 1

Assignment 1

- Two recommendation tasks
- Due **Nov 18** (four weeks from today)
- Submissions should be made on Kaggle, plus a short report to be submitted to gradescope

Assignment 1

Data

Assignment data is available on:
http://cseweb.ucsd.edu/classes/fa19/cse258-a/files/assignment1.tar.gz

Detailed specifications of the tasks are available on:
http://cseweb.ucsd.edu/classes/fa19/cse258-a/files/assignment1.pdf
(or in this slide deck)

Data

1. Training data: 200k book reviews from Goodreads

```
userID,bookID,rating
u79354815,b14275065,4
u56917948,b82152306,5
u97915914,b44882292,5
u49688858,b79927466,5
u08384938,b05683889,2
u13530776,b86375465,4
u46307273,b92838791,5
u18524450,b35165110,2
u69700998,b17128180,5
u43359569,b34596567,5
```

Assignment 1

Tasks

1. Estimate **whether** a particular book would be read

u65407115-b69897799 -> 0/1?

f(user,item) -> true/false

Assignment 1

Tasks – CSE158 only

2. Estimate the **category** of a book based on its review

{'n_votes': 0, 'review_id': 'r24440074', 'user_id': 'u08070901', 'review_text': 'Pretty decent. The ending seemed a little rush but a good ending to the first trilogy in this series. The fact that most of the time it is a military fantasy makes it interesting. Also all of the descriptions of food just make me hungry.', 'rating': 5, 'genreID': 2, 'genre': 'fantasy_paranormal'}

f(user,item) -> category

Assignment 1

Tasks – CSE258 only

2. Estimate the **rating** given a user/book pair

u12927896-b38220226 -> 0..5

f(user,item) -> star rating

Assignment 1

Evaluation

1. Estimate whether a book will be read or not

Categorization Accuracy (fraction of correct classifications):

$$\mathrm{CategorizationAccuracy}(\hat{r}, r) = \textstyle\sum_{u,i} \delta(\hat{r}_{u,i} = r_{u,i})$$

predictions (0/1)

Read (1) and
Non-read (0) books)

test set of read /
non-read books

Assignment 1

Evaluation (158 task 2)

2. Estimate the category of a review

Categorization Accuracy (fraction of correct classifications):
5 categories have been selected and are mapped to numbers from 0-4 (see baselines.py)

$$\mathrm{CategorizationAccuracy}(\hat{r}, r) = \sum_{u,i} \delta(\hat{r}_{u,i} = r_{u,i})$$

predictions (0-4)

groundtruth
category

test set of reviews

Assignment 1

Evaluation (258 task 2)

2. Estimate what rating a user would give to a book

$$\mathrm{RMSE}(f) = \sqrt{\tfrac{1}{N}\textstyle\sum_{u,i,t \in \text{test set}} (\underset{\text{model's prediction}}{f(u,i,t)} - \underset{\text{ground-truth}}{r_{u,i,t}})^2}$$

(just like the Netflix prize)

Assignment 1

Test data

It’s a secret! I’ve provided files that include lists of tuples that need to be predicted:

pairs_Read.txt
pairs_Category.txt
pairs_Rating.txt

Test data

Files look like this

(note: not the actual test data):

```
userID-bookID,prediction
u10867277-b35018725,4
u58578865-b45488412,3
u53582462-b60611623,2
u58775274-b02793341,4
u52022406-b80770760,1
u77792103-b62925951,1
u86157817-b67402445,2
u60596724-b61972458,2
u30345190-b26955550,5
u27548114-b46455538,5
u51025274-b82629707,1
```

Assignment 1

Test data

But I've only given you this:

(you need to estimate the final column)

```
userID-bookID,prediction
u10867277-b35018725
u58578865-b45488412
u53582462-b60611623
u58775274-b02793341
u52022406-b80770760
u77792103-b62925951
u86157817-b67402445
u60596724-b61972458
u30345190-b26955550
u27548114-b46455538
u51025274-b82629707
```

last column missing

Assignment 1

Baselines

I've provided some simple baselines that generate valid prediction files

(see baselines.py)

Assignment 1

Baselines

1. Estimate whether a book will be read by a user

- Rank books by popularity in the training data
- Return 1 if a test item is among the top 50% of most popular books, or 0 otherwise

Assignment 1

Baselines

2. Estimate the category of a book

Look for certain words in the review (e.g. if the word “fantasy” appears, classify as “Fantasy”)

Assignment 1

Baselines

2. Estimate what rating a user would give to a book

Use the global average, or the user’s personal average if we have seen that user before

Assignment 1

Kaggle

I've set up a competition webpage to evaluate your solutions and compare your results to others in the class:

https://inclass.kaggle.com/c/cse158258-fa19-read-prediction
https://inclass.kaggle.com/c/cse158-fa19-category-prediction
https://inclass.kaggle.com/c/cse258-fa19-rating-prediction

The leaderboard only uses 50% of the data – your final score will be (partly) based on the other 50%

Assignment 1

Marking

Each of the two tasks is worth **10%** of your grade. This is divided into:

- 5/10: Your performance compared to the simple baselines I have provided. It should be **easy** to beat them by a bit, but **hard** to beat them by a lot
- 3/10: Your performance compared to others in the class on the held-out data
- 2/10: Your performance on the *seen* portion of the data. This is just a consolation prize in case you badly overfit to the leaderboard, but should be easy marks.

- 5 marks: A **brief** written report about your solution. The goal here is not (necessarily) to invent new methods, just to apply the right methods for each task. Your report should just describe which method/s you used to build your solution

Assignment 1

Fabulous prizes!

Much like the Netflix prize, there will be an award for the student with the lowest MSE/highest accuracy on Monday Nov. 18th

(estimated value US$1.29)

Assignment 1

Homework

Homework 3 is intended to get you set up for this assignment

Assignment 1

What worked last year, and what did I change?

Assignment 1

Questions?