

Web Mining and Recommender Systems

Assignment 1

Assignment 1

- Three recommendation tasks (two per class)
- Due **Nov 14** (four weeks from today)
- Submissions should be made on ~~Kaggle~~ gradescopes, along with a short report

Assignment 1

Data

Assignment data is available on:

<http://cseweb.ucsd.edu/classes/fa22/cse258-a/files/assignment1.tar.gz>

Detailed specifications of the tasks are available on:

<http://cseweb.ucsd.edu/classes/fa22/cse258-a/files/assignment1.pdf>

(or in this slide deck)

Assignment 1

Data

1. Training data: book interactions from Goodreads

```
userID,bookID,rating  
u67805239,b61372131,4  
u54531895,b75189008,4  
u76549666,b75446982,4  
u03186275,b23482469,2  
u21322233,b09979253,3  
u00402241,b68456479,1  
u88999268,b49553867,0  
u39455611,b40151793,5  
u90502882,b01672704,4  
u92679832,b26246971,4
```

Assignment 1

Data

1. Training data (and data about reviews/categories):

```
{'user_id': 'u75242413', 'review_id': 'r45843137', 'rating': 4, 'review_text': "a clever book with a deeply troubling premise and an intriguing protagonist. Thompson's clean, sparse prose style kept each page feeling light even as some rather heavy existential questions dropped upon them. I enjoyed it. \n and that cover design is boom-pow gorgeous.", 'n_votes': 1, 'genre': 'mystery_thriller_crime', 'genreID': 3}
{'user_id': 'u72358746', 'review_id': 'r38427923', 'rating': 2, 'review_text': "A little too much retconning for me, to be honest. Wolverine's past has mostly been a mystery and for the most part, I am content with that. Saying he formed a proto-X-Men group doesn't feel right, and neither does the part Xavier plays so far (I didn't think he really established a school before he was crippled) .", 'n_votes': 0, 'genre': 'comics_graphic', 'genreID': 1}
{'user id': 'u55827211', 'review id': 'r97393610', 'rating': 5, 'review text': "So glad I finally got around
```

Assignment 1

Tasks

1. Estimate **whether** a particular book would be read
u37758667,b99713185 -> 0/1?

f(user,item) ->
true/false

Assignment 1

Tasks – CSE158 only

2. Estimate the **category** of a book based on text in its review (or other metadata)

```
{'user_id': 'u75242413', 'review_id': 'r45843137', 'rating': 4, 'review_text': "a clever book with a deeply troubling premise and an intriguing protagonist. Thompson's clean, sparse prose style kept each page feeling light even as some rather heavy existential questions dropped upon them. I enjoyed it. \n and that cover design is boom-pow gorgeous.", 'n_votes': 1, 'genre': 'mystery_thriller_crime', 'genreID': 3}
```

f(review) ->
category (0..4)

Assignment 1

Tasks – CSE258 only

2. Estimate the **rating** a user will give to a book

u17832892,b51986055 -> 0..5

$f(\text{user,item}) \rightarrow \text{real value}$

Assignment 1

Evaluation

1. Estimate whether a book will be read or not

Categorization Accuracy (fraction of correct classifications):

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

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

test set of read /
non-read books

Assignment 1

Evaluation

2. Estimate the category of a book (CSE158 only)

Categorization Accuracy (fraction of correct classifications):

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

predictions (0..4)
category (0..4)

test set of reviews


Assignment 1

Evaluation

2. Estimate the rating a user will give to a book (258 only)

$$\text{RMSE}(f) = \sqrt{\frac{1}{N} \sum_{u,i,t \in \text{test set}} (f(u, i, t) - r_{u,i,t})^2}$$

model's prediction ground-truth



(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.csv
pairs_Rating.csv
pairs_Category.csv

Assignment 1

Test data

Files look like this

(note: not the actual test data):

```
userID,bookID,prediction
u17832892,b51986055,4
u94058414,b95439113,3
u54876772,b61970919,5
u27182378,b29199360,4
u00343094,b17138341,4
u55453694,b70912031,4
u53021409,b04222499,3
u26001504,b15025576,4
u48139087,b56425922,2
u70455688,b09902724,5
u28953105,b07142228,5
```

Assignment 1

Test data

But I've only given you this:
(you need to estimate the final column)

```
userID,bookID,prediction
```

```
u17832892,b51986055
```

```
u94058414,b95439113
```

```
u54876772,b61970919
```

```
u27182378,b29199360
```

```
u00343094,b17138341
```

```
u55453694,b70912031
```

```
u53021409,b04222499
```

```
u26001504,b15025576
```

```
u48139087,b56425922
```

```
u70455688,b09902724
```

```
u28953105,b07142228
```

last column missing



Baselines

I've provided some simple baselines that
generate valid prediction files
(see `baselines.py`)

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

Simple solution that looks for a few "category-specific" words

Assignment 1

Baselines

2. Estimate a user's rating of 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://www.kaggle.com/c/cse158258-cooking-prediction/>
<https://www.kaggle.com/c/cse158-cook-time-prediction/>
<https://www.kaggle.com/c/cse258-recipe-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:

- 6/10: 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
 - 2/10: 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!

Usually I give lovely prizes, but hard to do in a (mostly) online class. Can venmo you the price of a coffee if you like?

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

What worked last year, and what did I change?

Assignment 1

What worked last year, and what did I change?

Assignment 1

Questions?