Web Mining and Recommender Systems

Assignment 1

Assignment 1

- Three recommendation tasks (two per class)
- Due **Nov 20** (Three weeks minus one day)
- Submissions should be made on gradescope, along with a short report

Assignment 1

Data

Assignment data is available on:
http://cseweb.ucsd.edu/classes/fa23/cse258-a/files/assignment1.tar.gz

Detailed specifications of the tasks are available on:
http://cseweb.ucsd.edu/classes/fa23/cse258-a/files/assignment1.pdf
(or in this slide deck)

Assignment 1

Data

1. Training data: 175k video game reviews from Steam

{'hours': 0.3, 'gameID': 'b96045472', 'hours_transformed': 0.3785116232537298 3, 'early_access': False, 'date': '2015-04-08', 'text': '+1', 'userID': 'u01561183'}
{'userID': 'u88836191', 'early_access': False, 'hours': 63.5, 'hours_transformed': 6.011227255423254, 'found_funny': 1, 'text': 'If you want to sit in queue for 10-20min and have 140 ping then this game is perfect for you :)', 'gameID': 'b19457938', 'user_id': '76561198030408772', 'date': '2017-05-20'}
{'hours': 0.2, 'gameID': 'b09870670', 'hours_transformed': 0.2630344058337938, 'early_access': False, 'date': '2017-01-27', 'text': 'I was really not a fan of the gameplay. Games should at least try a little to be enjoyable and not so tedious and boring from the jump. I was just confused and bored, and that is not what I look for in my hobby.', 'userID': 'u36851597'}

Assignment 1

Tasks

1. Estimate **whether** a particular game would be played (or bought)

u42434461-b91625775 -> 0/1?

f(user,item) -> true/false

Assignment 1

Tasks

2. Estimate the **amount of time** a game will be played (really log_2(time + 1))

u44767493-b89977206 -> 0..\infty

f(user,item) -> real value

Assignment 1

Evaluation

1. Estimate whether a game will be played or note

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)

Played (1) and
Non-played (0) games)

test set of played /
non-played games

Assignment 1

Evaluation

2. Estimate how long a user would play a game (log_2(time + 1))

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

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_Played.txt
pairs_Hours.txt

Test data

Files look like this

(note: not the actual test data):

```
userID,reviewID,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
```

Test data

But I've only given you this:

(you need to estimate the final column)

```
userID,reviewID,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

Baselines

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

(see baselines.py)

Assignment 1

Baselines

1. Estimate whether a game will be played by a user

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

Baselines

2. Estimate how long a user would play a game

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

Assignment 1

Gradescope

The assignment is set up as a competition to evaluate your solutions and compare your results to others in the class

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.

- 2 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 an online class. Maybe I'll gift you a subscription to my channel?!?

Assignment 1

Homework

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

Assignment 1

What worked last year?

Assignment 1

What worked last year?

Assignment 1

Questions?