CSE 158

Web Mining and Recommender Systems

Assignment 1

Assignment 1

- Two recommendation tasks
- Due **Nov 19** (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://jmcauley.ucsd.edu/data/assignment1.tar.gz

Detailed specifications of the tasks are available on:
http://cseweb.ucsd.edu/classes/fa18/cse158-a/files/assignment1.pdf
(or in this slide deck)

Assignment 1

Data

1. Training data: 200k product reviews from Amazon

```
{'reviewTime': '09 26, 2013', 'reviewText': "The model in this picture
has them rolled up at the top because they are actually very high
waisted! that's my only complaint though, because they are very good
quality, and fit really well! I am 5'2&#34; 120lbs with thick thighs and
i love them i can't wait to wear them out!", 'helpful': {'nHelpful': 0,
'outOf': 0}, 'reviewerID': 'U490934656', 'reviewHash': 'R798569390',
'categories': [['Clothing, Shoes & Jewelry', 'Women'], ['Clothing, Shoes
& Jewelry', 'Novelty, Costumes & More', 'Novelty', 'Clothing', 'Women',
'Leggings']], 'unixReviewTime': 1380153600, 'itemID': 'I402344648',
'rating': 4.0, 'summary': 'High Waisted', 'categoryID': 0}
```

Assignment 1

Tasks

1. Estimate **whether** a particular item would be reviewed

```
{'reviewTime': '09 26, 2013', 'reviewText': "The model in this picture
has them rolled up at the top because they are actually very high
waisted! that's my only complaint though, because they are very good
quality, and fit really well! I am 5'2&#34; 120lbs with thick thighs and
i love them i can't wait to wear them out!", 'helpful': {'nHelpful': 0,
'outOf': 0}, 'reviewerID': 'U490934656', 'reviewHash': 'R798569390',
                                   es & Jewelry', 'Women'], ['Clothing, Shoes
                                  & More', 'Novelty', 'Clothing', 'Women',
                                   1380153600, 'itemID': 'I402344648',
                                   Waisted', 'categoryID': 0}
```

f(user,item) → true/false

Assignment 1

Tasks – CSE158 only

2. Estimate the **category** of an item based on its review

```
{'reviewTime': '09 26, 2013', 'reviewText': "The model in this picture
has them rolled up at the top because they are actually very high
waisted! that's my only complaint though, because they are very good
quality, and fit really well! I am 5'2&#34; 120lbs with thick thighs and
i love them i can't wait to wear them out!", 'helpful': {'nHelpful': 0,
'outOf': 0}, 'reviewerID': 'U490934656', 'reviewHash': 'R798569390',
'categories': [['Clothing, Shoes & Jewelry
& Jewelry', 'Novelty, Costumes & More', 'N
'Leggings']], 'unixReviewTime': 138015360
'rating': 4.0, 'summary': 'High Waisted',
```

f(user,item) → category

Assignment 1

Tasks – CSE258 only

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

```
{'reviewTime': '09 26, 2013', 'reviewText': "The model in this picture
has them rolled up at the top because they are actually very high
waisted! that's my only complaint though, because they are very good
quality, and fit really well! I am 5'2&#34; 120lbs with thick thighs and
i love them i can't wait to wear them out!", 'helpful': {'nHelpful': 0,
'outOf': 0}, 'reviewerID': 'U490934656', 'reviewHash': 'R798569390',
'categories': [['Clothing, Shoes & Jewelry', 'Women'], ['Clothing, Shoes
& Jewelry', 'Novelty, Costumes & More', 'Novelty', 'Clothing', 'Women',
'Leggings']], 'unixReviewTime': 1380153600, 'itemID': 'I402344648',
'rating': 4.0, 'summary': 'High Waisted', 'categoryID': 0}
```

f(user,item) → star rating

Assignment 1

Evaluation

1. Estimate whether an will be purchased/reviewed or not

Categorization Accuracy (fraction of correct classifications):

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

predictions (0/1)

visited (1) and
non-visited (0) business)

test set of visited/
non-visited businesses

Assignment 1

Evaluation

2. Estimate what rating a user would give to an item

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

model's prediction

ground-truth

(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_Purchase.txt
pairs_Category.txt
~~pairs_Rating.txt~~

Test data

Files look like this

(note: not the actual test data):

```
userID-itemID,prediction
U310867277-I435018725,4
U258578865-I545488412,3
U853582462-I760611623,2
U158775274-I102793341,4
U152022406-I380770760,1
U977792103-I662925951,1
U686157817-I467402445,2
U160596724-I061972458,2
U830345190-I826955550,5
U027548114-I046455538,5
U251025274-I482629707,1
```

Assignment 1

Test data

But I’ve only given you this:

(you need to estimate the final column)

```
userID-itemID,prediction
U310867277-I435018725
U258578865-I545488412
U853582462-I760611623
U158775274-I102793341
U152022406-I380770760
U977792103-I662925951
U686157817-I467402445
U160596724-I061972458
U830345190-I826955550
U027548114-I046455538
U251025274-I482629707
```

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 an will be purchased/reviewed

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

Assignment 1

Baselines

2. Estimate the category of an item

Look for certain words in the review (e.g. if the word “daughter” appears, classify as “Girl's clothing”)

Assignment 1

Baselines

2. Estimate what rating a user would give to an item

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-fa18-purchase-prediction
https://inclass.kaggle.com/c/cse258-fa18-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/accuracy on Monday Nov. 19th

(estimated value US$1.29)

Assignment 1

Homework

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

(Homework is already out, but not due until Nov. 12)

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?