CSE 255 – Lecture 8
Data Mining and Predictive Analytics

Assignment 1
Assignment 1

- Two recommendation tasks
- Due **Nov 17** (four weeks -1 day from today)
- Submissions should be made electronically to Sheeraz Ahmad (sahmad@cs.ucsd.edu)
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/fa15/cse255-a/files/assignment1.pdf
(or in this slide deck)
Data

1. Training data: 1M book reviews from Amazon

   {'itemID': 'I572782694', 'rating': 5.0, 'helpful': {'nHelpful': 0, 'outOf': 0}, 'reviewText': 'favorite of the series...May not have been as steamy as some of the others...but the characters, their depth, and believability were amazing. wanted to curl up with Devlin and make it all better(wink wink). an amazing series...found Laura Kate when I stumbled onto Hearts in Darkness(one of my all time faves)...this series ranks up there with my Kresley Cole and Gena Showalter favorites.', 'reviewerID': 'U243261361', 'summary': 'Loved it', 'unixReviewTime': 1399075200, 'category': [['Books']], 'reviewTime': '05 3, 2014'}
1. Estimate how **helpful** people will find a user’s review of a product.

The review text is:

```
favorite of the series...May not have been as steamy as some of the others...but the characters, their depth, and believability were amazing. wanted to curl up with Devlin and make it all better(wink wink). an amazing series...found Laura Kate when I stumbled onto Hearts in Darkness(one of my all time faves)...this series ranks up there with my Kresley Cole and Gena Showalter favorites.'
```

Using the function `f(user, item, outOf) → nHelpful`

```
f(user, item, outOf) → nHelpful
```
Assignment 1

Tasks

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

{itemID: 'T572782694', rating: 5.0, helpful: {'nHelpful': 0, 'outOf': 0}, reviewText: 'favorite of the series...May not have been as steamy as some of the others...but the characters, their depth, and believability were amazing. wanted to curl up with Devlin and make it all better(wink wink). an amazing series...found Laura Kate when I stumbled onto Hearts in Darkness(one of my all time faves)...this series ranks up there with my Kresley Cole and Gena Showalter favorites.', reviewerID: 'U243261361', unixReviewTime: 1399075200, category: [['Books']]

f(user,item) → star rating
1. Estimate how helpful people will find a user’s review of a product

Absolute error:

$$AE(\hat{r}, r) = \frac{1}{N} \sum_{u,i} |\hat{r}_{u,i} - r_{u,i}|$$

predictions (# helpfulness votes)

actual # helpfulness votes
Evaluation

1. Estimate how helpful people will find a user’s review of a product
   • You are given the total number of votes, from which you must estimate the number that were helpful
   • I chose this value (rather than, say, estimating the fraction of helpfulness votes for each review) so that each vote is treated as being equally important
   • The Absolute error is then simply a count of how many votes were predicted incorrectly
Evaluation

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

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

(model's prediction)  (ground-truth)

(just like the Netflix prize)
Test data

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

pairs_Helpful.txt
pairs_Rating.txt
### Test data

Files look like this (note: not the actual test data):

<table>
<thead>
<tr>
<th>userID</th>
<th>itemID</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>U310867277</td>
<td>I435018725</td>
<td>4</td>
</tr>
<tr>
<td>U258578865</td>
<td>I545488412</td>
<td>3</td>
</tr>
<tr>
<td>U853582462</td>
<td>I760611623</td>
<td>3</td>
</tr>
<tr>
<td>U158775274</td>
<td>I102793341</td>
<td>2</td>
</tr>
<tr>
<td>U152022406</td>
<td>I380770760</td>
<td>4</td>
</tr>
<tr>
<td>U977792103</td>
<td>I662925951</td>
<td>4</td>
</tr>
<tr>
<td>U686157817</td>
<td>I467402445</td>
<td>5</td>
</tr>
<tr>
<td>U160596724</td>
<td>I061972458</td>
<td>1</td>
</tr>
<tr>
<td>U830345255</td>
<td>I826955550</td>
<td>1</td>
</tr>
<tr>
<td>U027548114</td>
<td>I046455538</td>
<td>1</td>
</tr>
<tr>
<td>U251025274</td>
<td>I482629707</td>
<td>1</td>
</tr>
</tbody>
</table>
Assignment 1

Test data

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

<table>
<thead>
<tr>
<th>userID</th>
<th>itemID</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>U310867277</td>
<td>I435018725</td>
<td></td>
</tr>
<tr>
<td>U258578865</td>
<td>I545488412</td>
<td></td>
</tr>
<tr>
<td>U853582462</td>
<td>I760611623</td>
<td></td>
</tr>
<tr>
<td>U158775274</td>
<td>I102793341</td>
<td></td>
</tr>
<tr>
<td>U152022406</td>
<td>I380770760</td>
<td></td>
</tr>
<tr>
<td>U977792103</td>
<td>I662925951</td>
<td></td>
</tr>
<tr>
<td>U686157817</td>
<td>I467402445</td>
<td></td>
</tr>
<tr>
<td>U160596724</td>
<td>I061972458</td>
<td></td>
</tr>
<tr>
<td>U830345255</td>
<td>I826955550</td>
<td></td>
</tr>
<tr>
<td>U027548114</td>
<td>I046455538</td>
<td></td>
</tr>
<tr>
<td>U251025274</td>
<td>I482629707</td>
<td></td>
</tr>
</tbody>
</table>

(last column missing)
Baselines

I’ve provided some simple baselines that generate valid prediction files
(see baselines.py)
1. Estimate how helpful people will find a user’s review of a product
   • Predict the global average helpfulness rate, or the user’s average helpfulness rate if we’ve observed this user before
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
Kaggle

I’ve set up a competition webpage 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%
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
Fabulous prizes!

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

(estimated value US$1.29)
Homework

Homework 3 is intended to get you set up for this assignment
(Homework will be released next week)
What worked last year, and what did I change?
What worked last year, and what did I change?
Questions?