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)

Assignment 1

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'}

Assignment 1

Tasks

1. Estimate how **helpful** people will find a user's review of a product

{'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': [['Bo

f(user,item,outOf) → nHelpful

Assignment 1

Tasks

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

{'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', 1399075200, 'category': [['Bo

f(user,item) → star rating

Assignment 1

Evaluation

1. Estimate how helpful people will find a user's review of a product
Absolute error:

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

predictions (# helpfulness votes)

actual # helpfulness votes

Assignment 1

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

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_Helpful.txt
pairs_Rating.txt

Assignment 1

Test data

Files look like this

(note: not the actual test data):

```
userID-itemID,prediction
U310867277-I435018725,4
U258578865-I545488412,3
U853582462-I760611623,3
U158775274-I102793341,2
U152022406-I380770760,4
U977792103-I662925951,4
U686157817-I467402445,5
U160596724-I061972458,1
U830345255-I826955550,1
U027548114-I046455538,1
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
U830345255-I826955550
U027548114-I046455538
U251025274-I482629707
```

last column missing

Baselines

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

(see baselines.py)

Assignment 1

Baselines

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

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/cse-255-255-fa15-assignment-1-task-1-helpfulness-prediction/
https://inclass.kaggle.com/c/cse-255-fa15-assignment-1-task-2-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 on Wednesday Nov. 18th

(estimated value US$1.29)

Assignment 1

Homework

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

(Homework will be released next week)

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?