CSE 158

Web Mining and Recommender Systems

Assignment 1

Assignment 1

- Two recommendation tasks
- Due **Nov 20** (four weeks -2 days 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/fa17/cse158-a/files/assignment1.pdf
(or in this slide deck)

Assignment 1

Data

1. Training data: 200k product reviews from Google Local

{'rating': 5.0, 'businessID': 'B408037852', 'reviewText': u"This is where i go to shop for gifts from my mom. She loves this stuff. Cna't get enough. I like that you can customize the items. Store is well alid out and shoppable.", 'userID': 'U093387342', 'reviewTime': u'Mar 24, 2013', 'categories': [u"Women's Clothing Store", u'Fashion Accessories Store', u'Shoe Store'], 'reviewHash': 'R471510664', 'unixReviewTime': 1364143460}

Assignment 1

Tasks

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

{'rating': 5.0, 'businessID': 'B408037852', 'reviewText': u"This is where i go to shop for gifts from my mom. She loves this stuff. Cna't get enough. I like that you can customize the items. Store is well alid out and shoppable.", 'userID': 'U093387342', 'reviewTime': u'Mar 24, 2013', 'categories': [u"Women's Clothing Store", u'Fashion Accessories Store', u'Shoe Store'], 'reviewHash': 'R471510664', 'unixReviewTime':

f(user,business) → true/false

Assignment 1

Tasks – CSE158 only

2. Estimate the **category** of a store based on its review

{'rating': 5.0, 'businessID': 'B408037852', 'reviewText': u"This is where i go to shop for gifts from my mom. She loves this stuff. Cna't get enough. I like that you can customize the items. Store is well alid out and shoppable.", 'userID': 'U093387342', 'reviewTime': u'Mar 24, 2013', 'categories': [u"Women's Clothing Store", u'Fashion Accessories Store', u'Shoe Store'], 'reviewHash': 'R471510664', 'unixReviewTime':

f(user,item) → category

Assignment 1

Tasks – CSE258 only

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

{'rating': 5.0, 'businessID': 'B408037852', 'reviewText': u"This is where i go to shop for gifts from my mom. She loves this stuff. Cna't get enough. I like that you can customize the items. Store is well alid out and shoppable.", 'userID': 'U093387342', 'reviewTime': u'Mar 24, 2013', 'categories': [u"Women's Clothing Store", u'Fashion Accessories Store', u'Shoe Store'], 'reviewHash': 'R471510664', 'unixReviewTime': 1364143460}

f(user,business) → star rating

Assignment 1

Evaluation

1. Estimate whether a business would be visited 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)

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

test set of visited/ non-visited businesses

Assignment 1

Evaluation

2. Estimate the category of a business

Categorization Accuracy (fraction of correct classifications):
10 categories have been selected and are mapped to numbers from 0-9 (see baselines.py)

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

predictions (0-9)

groundtruth category

test set of businesses

Assignment 1

Test data

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

pairs_Visit.txt
pairs_Category.txt
~~pairs_Rating.txt~~

Assignment 1

Test data

Files look like this

(note: not the actual test data):

```
userID-businessID,prediction
U310867277-B435018725,4
U258578865-B545488412,3
U853582462-B760611623,2
U158775274-B102793341,4
U152022406-B380770760,1
U977792103-B662925951,1
U686157817-B467402445,2
U160596724-B061972458,2
U830345190-B826955550,5
U027548114-B046455538,5
U251025274-B482629707,1
```

Assignment 1

Test data

But I've only given you this:

(you need to estimate the final column)

```
userID-businessID,prediction
U310867277-B435018725
U258578865-B545488412
U853582462-B760611623
U158775274-B102793341
U152022406-B380770760
U977792103-B662925951
U686157817-B467402445
U160596724-B061972458
U830345190-B826955550
U027548114-B046455538
U251025274-B482629707
```

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 business would be visited

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

Assignment 1

Baselines

2. Estimate the category of a business

Look for certain words in the review (e.g. if the word “bar” appears, classify as “Bar”)

Baselines

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

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/cse158-258-fa17-visit-prediction
https://inclass.kaggle.com/c/cse158-fa17-category-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. 20th

(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. 13)

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?