CSE 190 – Lecture 8

Data Mining and Predictive Analytics

Assignment 1

Assignment 1

- Two recommendation tasks
- Due **May 20** (four weeks -1 day from today)
- Submissions should be made electronically to Long Jin (longjin@cs.ucsd.edu)

Assignment 1

Data

Assignment data is available on:
http://jmcauley.ucsd.edu/cse190/data/assignment1.tar.gz

Detailed specifications of the tasks are available on:
http://cseweb.ucsd.edu/~jmcauley/cse190/files/assignment1.pdf
(or in this slide deck)

Assignment 1

Data

1. Training data: 500,000 clothing reviews from Amazon

{'itemID': 'I557295048', 'rating': 5.0, 'helpful': {'nHelpful': 1, 'outOf': 1}, 'reviewText': 'I never have to bother with retying or tripping on my shoelaces -- and, the rainbow colors suit my whimsy in me!', 'reviewerID': 'U893540719', 'summary': 'Where have they been all my life!', 'unixReviewTime': 1395273600, 'category': [['Health & Personal Care', 'Medical Supplies & Equipment', 'Daily Living Aids', 'Dressing Aids', 'Shoe Fasteners & Laces'], ['Clothing, Shoes & Jewelry', 'Novelty, Costumes & More', 'Shoe Care & Accessories', 'Shoelaces']], 'reviewTime': '03 20, 2014'}

Assignment 1

Tasks

1. Estimate whether a user would **purchase** (really review) a product or not

f(user,item) → purchased/not purchasd

{'itemID': 'T5 ful': 1, 'outOf': 1}, ' tying or tripping on my whimsy in me!', 'reviewe hey been all my life!', 'un alth & Personal Care', 'Medical Supplies & Equipment', 'Daily Living Aids', 'Dressing Aids', 'Shoe Fasteners & Laces'], ['Clothing, Shoes & Jewelry', 'Novelty, Costumes & More', 'Shoe Care & Accessories', 'Shoelaces']], 'reviewTime': '03 20, 2014'}

Assignment 1

Tasks

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

{'itemID': 'I557295048', 'rating': 5.0, 'helpful': {'nHelpful': 1,
'outOf': 1}, 'reviewText': 'I never have to bother with retying or
tripping on my
me!', 'reviewe
my life!', 'un
Personal Care'
'Dressing Aids
Jewelry', 'Novelty, Costumes & More', 'Shoe Care & Accessories',
'Shoelaces']], 'reviewTime': '03 20, 2014'}

f(user,item,outOf) → nHelpful

Assignment 1

Evaluation

1. Estimate whether a user would purchase (really review) a product or not

1 - Hamming loss (fraction of misclassifications):

$$\mathrm{HammingLoss}(\hat{r}, r) = \frac{1}{N} \sum_{u,i} \frac{\delta(\hat{r}_{u,i} \neq r_{u,i})}{2}$$

predictions (0/1)

purchased (1) and non-purchased (0) items)

test set of purchased/ non-purchased items

Assignment 1

Evaluation

1. Estimate whether a user would purchase (really review) a product or not

For this task, the test set has been constructed such that exactly 50% of pairs (u,i) correspond to purchased items and 50% to non-purchased items

Assignment 1

Evaluation

2. 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

3. 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

Test data

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

pairs_Purchase.txt
pairs_Helpful.txt

Assignment 1

Test data

Files look like this

(note: not the actual test data):

```
userID-itemID,prediction
U310867277-I435018725,1
U258578865-I545488412,0
U853582462-I760611623,0
U158775274-I102793341,0
U152022406-I380770760,1
U977792103-I662925951,1
U686157817-I467402445,0
U160596724-I061972458,0
U830345190-I826955550,0
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
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 a user would purchase (really review) a product or not
   - Predict 1 if the item is among the top 50% of most popular items, or 0 otherwise

Assignment 1

Baselines

2. 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

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-190-assignment-1-task-1-purchase-prediction/
https://inclass.kaggle.com/c/cse-190-assignment-1-task-2-helpfulness-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:

- 3/10: 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
- 3/10: 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
- 2/10: 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.

Assignment 1

Fabulous prizes!

Much like the Netflix prize, there will be an award for the student with the lowest MSE on the day of the last lecture

(estimated value US$1.29)

Assignment 1

Homework

Homework 3 is intended to get you set up for this assignment. It requires you to implement some simple approaches of your own, using a **different** dataset (Amazon Video Games) whose format is exactly the same

(Homework will be released on Tuesday; if you're keen you can probably guess the URL)

Assignment 1

Questions?