CSE 255 – Lecture 5

Data Mining and Predictive Analytics

Assignment 2

Assignment 2

- Three recommendation tasks
- Due **March 9** (four weeks from today, the last week of class)
- Submissions should be made electronically to Dongcai (doshen@cs.ucsd.edu)

Assignment 2

Data

Assignment data is available on:
http://jmcauley.ucsd.edu/cse255/data/assignment2.tar.gz

Detailed specifications of the tasks are available on:
http://cseweb.ucsd.edu/~jmcauley/cse255/slides/lecture3_assignment2.pdf

Assignment 2

Data

1. Training data: 1,000,000 electronics reviews from Amazon

{'itemID': 'I502326793', 'rating': 3.0, 'helpful': {'nHelpful': 1, 'outOf': 1}, 'reviewText': "I like the look of the case and how it holds everything together but I dislike that it allows the tablet to slide down a bit. I'm a bit of a perfectionist when it comes to this and my tablet seems to fall where say 1% of the screen is unseen but it'll be a bit crooked and that drives me completely crazy lol. I also wish it could be stood up and not always propped up as sometimes the screen direction is necessary for certain apps and this renders the stand useless.", 'reviewerID': 'U081291677', 'summary': 'Practical but not Perfect', 'unixReviewTime': 1377388800, 'category': [['Electronics', 'Computers & Accessories', 'Touch Screen Tablet Accessories', 'Cases & Sleeves', 'Cases']], 'reviewTime': '08 25, 2013'}

Assignment 2

Data

1. Training data: 1,000,000 electronics reviews from Amazon

{'itemID': 'I502326793', 'rating': 3.0, 'helpful': {'nHelpful': 1, 'outOf': 1}, 'revie e and how it holds everything togethe ablet to slide down a bit. I'm a to this and my tablet seems to fa een but it'll be a bit crooked and th also wish it could be stood up es the screen direction is neces rs the stand useless.", 'reviewe ctical but not Perfect', 'unixReviewTime': 1377388800, 'category': [['Electronics', 'Computers & Accessories', 'Touch Screen Tablet Accessories', 'Cases & Sleeves', 'Cases']], 'reviewTime': '08 25, 2013'}

Data is from a 2014 scrape of Amazon and **should not** overlap with the Amazon data I've previously provided

Assignment 2

Tasks

1. Estimate what **rating** a user will give to a product

{'itemID': 'I502326793', 'rating': 3.0, 'helpful': {'nHelpful': 1, 'outOf': 1}, 'reviewText': "I like the look of the case and how it holds everything togeth[illegible]ablet to slide down a bit. I'm a[illegible] to this and my tablet seems to f[illegible]een but it'll be a bit crooked and t[illegible] also wish it could be stood up and not always propped up as sometimes the screen direction is necessary for certain apps and this renders the stand useless.", 'reviewerID': 'U081291677', 'summary': 'Practical but not Perfect', 'unixReviewTime': 1377388800, 'category': [['Electronics', 'Computers & Accessories', 'Touch Screen Tablet Accessories', 'Cases & Sleeves', 'Cases']], 'reviewTime': '08 25, 2013'}

f(user,item) → rating

Assignment 2

Tasks

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

f(user,item) → purchased/not purchasd

{'itemID': 'I502326793', 'rating': 3.0, 'helpful': {'nHelpful': 1, 'outOf': 1}, 'reviewText': "I like the look of the case and how it holds everything tog t to slide down a bit. I' this and my tablet seems t ut it'll be a bit crooked an o wish it could be stood he screen direction is necessary for certain apps and this renders the stand useless.", 'reviewerID': 'U081291677', 'summary': 'Practical but not Perfect', 'unixReviewTime': 1377388800, 'category': [['Electronics', 'Computers & Accessories', 'Touch Screen Tablet Accessories', 'Cases & Sleeves', 'Cases']], 'reviewTime': '08 25, 2013'}

Assignment 2

Tasks

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

{'itemID': 'T502326793', 'rating': 3.0, 'helpful': {'nHelpful': 1, 'outOf': 1}, 'reviewText': "I like the look of the case and how it holds everything tog t to slide down a bit. I' this and my tablet seems t ut it'll be a bit crooked an o wish it could be stood he screen direction is necessary for certain apps and this renders the stand useless.", 'reviewerID': 'U081291677', 'summary': 'Practical but not Perfect', 'unixReviewTime': 1377388800, 'category': [['Electronics', 'Computers & Accessories', 'Touch Screen Tablet Accessories', 'Cases & Sleeves', 'Cases']], 'reviewTime': '08 25, 2013'}

f(user,item,outOf) → nHelpful

Assignment 2

Evaluation

1. Estimate what rating a user will give to a product

RMSE:

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

predictions (star ratings)

(test) ratings

user/item pairs in test set

Assignment 2

Evaluation

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

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 2

Evaluation

2. 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 2

Evaluation

3. 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 2

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 2

Test data

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

labeled_Rating.txt
labeled_Purchase.txt
labeled_Helpful.txt

Test data

Files look like this

(note: not the actual test data):

```
userID-itemID,prediction
U310867277-I435018725,2.0
U258578865-I545488412,3.0
U853582462-I760611623,5.0
U158775274-I102793341,5.0
U152022406-I380770760,5.0
U977792103-I662925951,4.0
U686157817-I467402445,5.0
U160596724-I061972458,5.0
U830345190-I826955550,2.0
U027548114-I046455538,4.0
U251025274-I482629707,1.0
```

Assignment 2

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 2

Baselines

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

(see baselines.py)

Assignment 2

Baselines

1. Estimate what rating a user will give to a product

- Predict the average, or the *user* average if we've seen this user before, basically

$$f(u, i) = \alpha + \beta_u$$

Assignment 2

Baselines

2. 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 2

Baselines

3. 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 2

Kaggle

We'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-assignment-2-task-1-rating-prediction/
https://inclass.kaggle.com/c/cse-255-assignment-2-task-2-purchase-prediction/
https://inclass.kaggle.com/c/cse-255-assignment-2-task-3-helpfulness-prediction/

The leaderboard only uses 50% of the data – your final score will be (partly) based on the other 50%

Assignment 2

Marking

Each of the three 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 2

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 2

Homework

Homework 6 and 7 are intended to get you set up for this assignment. They require you to implement some simple approaches of your own, using a **different** dataset (Amazon Video Games) whose format is exactly the same

(Homework is due on the morning of Wed. Feb 25, or in class on the 23rd)

Assignment 2

Questions:

For problems accessing the Kaggle page please contact Dongcai. Other question to Piazza.