

# CSE 158

Web Mining and Recommender Systems

## Assignment 1

# Assignment 1

- Two recommendation tasks
- Due **Feb 27** (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/wi17/cse158-a/files/assignment1.pdf>

(or in this slide deck)

# Assignment 1

## Data

### 1. Training data: 200k clothing reviews from Amazon

```
{'categoryID': 0, 'categories': [['Clothing, Shoes & Jewelry', 'Women', 'Clothing', 'Lingerie, Sleep & Lounge', 'Intimates', 'Bras', 'Everyday Bras'], ['Clothing, Shoes & Jewelry', 'Women', 'Petite', 'Intimates', 'Bras', 'Everyday Bras']], 'itemID': 'I241092314', 'reviewerID': 'U023577405', 'rating': 4.0, 'reviewText': 'I love the look of this bra, it is what I wanted, however, it is about a cup size AND band size too small. The cups are sheer, which is what I wanted and the look is very sexy and it arrived much quicker than promised. I plan to order another one, but in a larger size.', 'reviewHash': 'R800651687', 'reviewTime': '02 7, 2013', 'summary': 'Beautiful but size runs small', 'unixReviewTime': 1360195200, 'helpful': {'outOf': 0, 'nHelpful': 0}}
```

# Assignment 1

## Tasks

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

```
{'categoryID': 0, 'categories': [['Clothing, Shoes & Jewelry', 'Women', 'Clothing', 'Lingerie, Sleep & Lounge', 'Intimates', 'Bras', 'Everyday Bras'], ['Clothing, Shoes & Jewelry', 'Women', 'Petite', 'Intimates', 'Bras', 'Everyday Bras']], 'itemID': 'I241092314', 'reviewerID': 'U023577405', 'rating': 4.0, 'reviewText': 'I love the look of this bra, it is what I wanted, however, it is about a cup size AND band size too small. The cups are sheer, which is what I wanted and the look is very different than promised. I plan to order another one. Helpful but size runs small', 'reviewHash': 'R800651687', 'reviewTime': '2012-07-11T14:14:14Z', 'helpful': {'outOf': 0, 'nHelpful': 0}}
```

f(user,item,outOf)  
→ nHelpful

# Assignment 1

## Tasks

### 2. Estimate the **category** of a product given its review/metadata

```
{'categoryID': 0, 'categories': [['Clothing, Shoes & Jewelry', 'Women', 'Clothing', 'Lingerie, Sleep & Lounge', 'Intimates', 'Bras', 'Everyday Bras'], ['Clothing, Shoes & Jewelry', 'Women', 'Petite', 'Intimates', 'Bras', 'Everyday Bras']], 'itemID': 'I241092314', 'reviewerID': 'U023577405', 'rating': 4.0, 'reviewText': 'I love the look of this bra, it is what I wanted, however, it is about a cup size AND band size too small. The cups are sheer, which is what I wanted and the look is very different than promised. I plan to order another one. Helpful but size runs small', 'reviewHash': 'R800651687', 'reviewTime': 1234567890, 'helpful': {'outOf': 0, 'nHelpful': 0}}
```

f(user,item,outOf)  
→ nHelpful

# Assignment 1

## Tasks – CSE258 only

### 2. Estimate the **rating** given a user/item pair

```
{'categoryID': 0, 'categories': [['Clothing, Shoes & Jewelry', 'Women', 'Clothing', 'Lingerie, Sleep & Lounge', 'Intimates', 'Bras', 'Everyday Bras'], ['Clothing, Shoes & Jewelry', 'Women', 'Petite', 'Intimates', 'Bras', 'Everyday Bras']], 'itemID': 'I241092314', 'reviewerID': 'U023577405', 'rating': 4.0, 'reviewText': 'I love the look of this bra, it is what I wanted, however, it is about a cup size AND band size too small. The cups are sheer, which is what I wanted and the look is very sexy and it arrived much quicker than I expected. I bought the one, but in a larger size.', 'reviewTime': '2013-02-07T13:00:00', 'summary': 'Beautiful bra', 'unixReviewTime': 1360195200,
```

$f(\text{user}, \text{item}) \rightarrow \text{star rating}$

# Assignment 1

## Evaluation

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



# 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 the category of a product

#### 1 - Hamming loss (fraction of correct classifications):

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

predictions (0/1) →  $\hat{r}$

purchased (1) and non-purchased (0) items →  $r$

test set of purchased/non-purchased items →  $u, i$


# Assignment 1

## 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)

# 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\_Category.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-I760611623,2
U158775274-I102793341,4
U152022406-I380770760,1
U977792103-I662925951,1
U686157817-I467402445,2
U160596724-I061972458,2
U830345190-I826955550,5
U027548114-I046455538,5
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 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

## **Baselines**

### 2. Estimate the category of a product

Look for certain words in the review (e.g. if the word "baby" appears, classify as "Baby Clothes")

# 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-helpfulness-prediction>  
<https://inclass.kaggle.com/c/cse158-categorization>

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 Monday Feb. 27th

(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 Feb. 20)

# 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?**