Web Mining and Recommender Systems

Assignment 1

Assignment 1

- Three recommendation tasks (two per class)
- Due **Nov 15** (four weeks 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://cseweb.ucsd.edu/classes/fa21/cse258-b/files/assignment1.tar.gz

Detailed specifications of the tasks are available on:

http://cseweb.ucsd.edu/classes/fa21/cse258-b/files/assignment1.pdf

(or in this slide deck)

Assignment 1

Data

1. Training data: 500k recipe interactions food.com

```
user_id,recipe_id,date,rating
88348277,03969194,2004-12-23,5
86699739,27096427,2002-01-12,4
03425965,44197323,2012-10-03,5
73973193,24971400,2008-04-09,5
15215209,60170202,2010-10-07,5
75799794,39662395,2012-05-02,5
77745222,88709727,2003-12-07,2
80598779,09359141,2007-10-10,2
35769308,83909791,2012-02-13,4
31763244,20530585,2010-04-30,5
```

Assignment 1

Data

1. Training data (and metadata about the recipes):

{'name': 'sexy fried eggs for sunday brunch', 'minutes': 10, 'contributor_id': '14298494', 'submitte
d': '2004-05-21', 'steps': 'heat a ridged griddle pan\tlightly brush the tomato slices and bread wit
h some olive oil\tcook the tomato slices first , for at least 5 minutes\twhen they are almost ready
, toast the bread in the same pan until well bar-marked\tin the meantime , pour a little olive oil i
nto a small frying pan and crack in the egg\tallow it to set for a minute or so and add the garlic a
nd chilli\tcook for a couple of minutes , spooning the hot oil over the egg until cooked to your lik
ing\tplace the griddled bread on a plate and quickly spoon the tomatoes on top\tthrow the chives
into the egg pan and splash in the balsamic vinegar\tseason well , then slide the egg on to the
tomatoes and drizzle the pan juices on top\tserve immediately , with a good cup of tea !',
'description': 'this is from silvana franco\'s book "family" which i love. i made these for brunch
yesterday and we loved them so much that we had them again today!', 'ingredients': ['plum

Assignment 1

Tasks

1. Estimate **whether** a particular recipe would be made (cooked)

42434461-91625775 -> 0/1?

f(user,item) -> true/false

Assignment 1

Tasks – CSE158 only

2. Estimate the **cook-time** of a recipe based on its description

{'name': 'oatmeal buttermilk pancakes', 'minutes': 21, 'contributor_id': '07092896',
'submitted': '2002-02-20', 'steps': 'mix oats and soda to buttermilk\tlet stand 5
minutes\tsift together flour , baking powder , salt and sugar\tadd sifted dry ingredients
, shortening and eggs to oats mixture\tstir until combined\tfor each pancake , pour
about 1 / 4 cup batte a golden brown ,
turning only once\tse ': '', 'ingredients': ['quick
oats', 'baking soda', 'b der', 'salt', 'sugar',
'shortening', 'eggs'], 'r

f(recipe) -> cook time (minutes)

Assignment 1

Tasks – CSE258 only

2. Estimate the **rating** a user will give to a recipe

44767493-89977206 -> 0..5

f(user,item) -> real value

Assignment 1

Evaluation

1. Estimate whether a recipe will be cooked 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)

Cooked (1) and
Non-cooked (0) recipes

test set of cooked /
non-cooked recipes

Assignment 1

Evaluation (158 task 2)

2. Estimate the cook time of a recipe

RMSE (actually MSE) between predicted and correct cook time:

$$\mathrm{RMSE}(f) = \sqrt{\tfrac{1}{N} \textstyle\sum_{u,i,t \in \text{test set}} (f(u,i,t) - r_{u,i,t})^2}$$

model's prediction

ground-truth

Assignment 1

Evaluation (258 task 2)

2. Estimate the rating a user will give to a recipe

$$\mathrm{RMSE}(f) = \sqrt{\frac{1}{N}\sum_{u,i,t\in\text{test set}}(f(u,i,t) - r_{u,i,t})^2}$$

model's prediction

ground-truth

(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:

stub_Made.txt
stub_Minutes.txt
stub_Rated.txt

Assignment 1

Test data

Files look like this

(note: not the actual test data):

```
user_id-recipe-id,prediction
40867277-35018725,4
58578865-45488412,3
53582462-60611623,2
58775274-02793341,4
52022406-80770760,1
77792103-62925951,1
86157817-67402445,2
60596724-61972458,2
30345190-26955550,5
27548114-46455538,5
51025274-82629707,1
```

Assignment 1

Test data

But I've only given you this:

(you need to estimate the final column)

```
user_id-recipe-id,prediction
40867277-35018725
58578865-45488412
53582462-60611623
58775274-02793341
52022406-80770760
77792103-62925951
86157817-67402445
60596724-61972458
30345190-26955550
27548114-46455538
51025274-82629707
```

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 recipe will be made by a user

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

Assignment 1

Baselines

2. Estimate the cook time of a recipe

Simple linear regression on the recipe length ('steps' field)

Assignment 1

Baselines

2. Estimate a user's rating of a recipe

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://www.kaggle.com/c/cse158258-cooking-prediction/
https://www.kaggle.com/c/cse158-cook-time-prediction/
https://www.kaggle.com/c/cse258-recipe-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:

- 6/10: 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.

- 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!

Usually I give lovely prizes, but hard to do in a (mostly) online class. Can venmo you the price of a coffee if you like?

Assignment 1

Homework

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

Assignment 1

What worked last year, and what did I change?

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?