

2018 SDML HW2

Prof. Shou-De Lin

TA: Nancy Cheng

Introduction

- Given food consumption data of each user, can we predict:
 - Task 1: what kind of new items will be consumed in the future?
 - Task 2: what kind of items will be consumed the next day?
- There are three types of information we provide
 - user/food consumption history (i.e. implicit ratings)
 - user profiles containing user features
 - food profiles containing food features

Task 1 - Implicit recommender

- Given: the food consumption history of each user
- Predict: what kind of 'new' food each user will consume in the future
- Historical consumption data for each user can be found in rating_train.csv
 - Since it is a consumption diary, users may eat repeated food items
- What to predict:
 - csv file, separating each column by comma, **must have header (userid, foodid)**
 - eg. userid,foodid
 0,99 3 1 75 20
 1,2 50 30 23 9
 - **For each user, need to do return k=20 most probable items (i.e. a total of 2608 * 20 entries)**
- Evaluation:
 - MAP@20

Data Example

Training data

Testing data

Ideal Answers

Userid/Itemid	9/14	9/15	9/16	9/17	9/18
User 33	I23	I21, I22	I17	I16, I18	I24
User 44	I17	I35, I63	I77, I88	I131	I135, I222
User 55	I18, I19	I14	I15	I17, I20	I14

Task 1	Task 2
I16, I18, I24	I16, I18
I131, I222	I131
I17, I20	I15

Task 2 - Predict consuming food item set for next day

- For each user, we will try to predict a set of food items which may be consumed by the user for the **next day**
- More detail spec will be announced on next week
- Output file format:
 - csv file, seperating each column by comma, **must have header (userid, foodid)**
 - eg.
userid,foodid
0,99 3 1 75 20
1,2 50 30 23 9
 - **For each user, need to do return k=20 most probable items (i.e. a total of 2608 * 20 entries)**
- Evaluation:
 - MAP@20

Data description

- The dataset is from a 6-month recording period between October 2014 to March 2015
 - Remove items taken by fewer than 30 different users and keep top 3000 users who consumed the most unique food items
 - Remove users who don't have user profile records
- It contains:
 - User-item implicit rating data (rating_train.csv)
 - User profile data (user.csv)
 - Item profile data (food.csv)

Implicit ratings(rating_train.csv): total 2681494 entries

date	userid	foodid
Diary date	User id	Food id as it appears in the diary

- On the average,
 - Training data will cover 86% of unique food items consumed by each user
 - A user has 1028 entries
 - A user consumes 202 unique items
 - A food item is consumed by 95 unique users

Item features (food.csv)

- Annotated food name is in JSON's key-value pairs format
 - Key name indicates main category, sub-category, and constituent name, separated by double underscores ("__"); value indicates occurrences
 - Example:

```
{""egg_dairy"": 1, ""egg_dairy__dairy_product__yogurt"": 1, ""herb_spice"": 1,
""herb_spice__spices"": 1, ""egg_dairy__dairy_product"": 1,
""herb_spice__spices__vanilla"": 1}
```
- There are total 5532 unique food items, each has 56 features (mostly categorical), some values are missing

dietary_supplement_subcat	snack_subcat	staple_subcat	mushroom_subcat
=====	=====	=====	=====
-	-	-	-
bodybuilding_supplements	snack	banana	mushroom
energy_food_products	=====	maize	=====
supplements	dessert_subcat	meat	fruit_subcat
=====	=====	other_cereal	=====
breakfast_cereal_subcat	-	rice	-
=====	cake	root_tuber	berry
-	chinese_desserts	wheat	mediterranean
breakfast_cereal	confectionery	=====	temperate
=====	cookies	meat_subcat	tropical
side_dish_subcat	custards	=====	=====
=====	doughnuts	-	nut_seed_subcat
-	frozen_desserts	beef	=====
antipasto	ice_cream	fish	-
fries	pastries	game	gymnosperm_seed
legumes	pies	lamb	nut
maize	puddings	meat	other
pasta	=====	meatball	pseudocereal
potatoes	fast_food_brand_subcat	pork	=====
salad	=====	poultry	bean_legume_subcat
soup	-	sausage	=====
=====	fast_food	seafood	-
condiment_subcat	=====	shellfish	legume
=====	snack_brand_subcat	=====	soy_product
-	=====	egg_dairy_subcat	=====
condiment	-	=====	preparation_subcat
=====	snack_brand	-	=====
beverage_subcat	=====	dairy_product	-
=====	herb_spice_subcat	egg	device
-	=====	=====	dry
alcohol	-	vegetable_subcat	fat
chocolate	herbs	=====	mechanical
coffee	mixtures	-	mixed
energy_drink	peppers	bulb_stem_vegetables	non_heat
fruit_juice	spices	flowers_flower_buds	wet
hot_beverage		fruits	
lemon_beverage		leafy_salad	
milk_substitutes		podded_vegetables	
rice_beverage		root_tuberous_vegetables	
soft_drink		sea_vegetables	
		vegetables	

User profiles(user.csv) : 2608 unique users

Values can be missing !!

Column	Description
userid	User ID
username	Username
age	Age
gender	Gender
location	Location
city	City
state	State
title	Title
about_me	“About Me”
reasons	“Why I want to get in shape”
inspirations	“My inspiration”
friends_count	Number of friends

Kaggle Competition and submission

HTC Computation resource

- Please fill your email in the following google sheet **by 2018/10/19 pm 12:00**
- <https://docs.google.com/spreadsheets/d/13INUxRuvUeLQabBcJfJE7am-ydiL-7clJwOVqIu5VNw/edit?usp=sharing>
- We will have 150 hours per account on HTC AI Education platform for computation purpose

Evaluation : MAP@K

- AP@K

$$AP@K = \frac{1}{m} \sum_{i=1}^K (P(i) \text{ if } i^{th} \text{ item was relevant}) = \frac{1}{m} \sum_{i=1}^N P(i) rel(i)$$

- m is the number of relevant items in the full space
- rel(k) is just an indicator that says whether that kth item was relevant (rel(k)=1) or not (rel(k)=0)

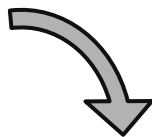
- MAP@K: average the AP@K metric over all your $|U|$ users

$$MAP@K = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{1}{m} \sum_{i=1}^K P_u(i) rel_u(i)$$

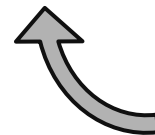
Evaluation : MAP@5 Example

Rank	Hit?
1	
2	X
3	X
4	X
5	

$$AP = \frac{1}{3} \left(\frac{1}{1} + \frac{2}{4} + \frac{3}{5} \right) = \frac{21}{30} = 0.7$$



$$AP = \frac{1}{3} \left(\frac{1}{2} + \frac{2}{3} + \frac{3}{4} \right) = \frac{23}{36} \approx 0.639$$



Rank	Hit?
1	X
2	
3	
4	X
5	X

- MAP@5 for 2 users = $(0.639 + 0.7) / 2$

Kaggle Competitions (same as HW1)

- **Both 2.1 and 2.2 can be conducted as a team of at most 3 persons**
- Testing data will be divided into private and public testing.
- Maximum 5 submissions a day per task are permitted.
- Choose one final submission among all your valid submissions before the deadline.
- Remember to declare your team on the Kaggle platform
- **Using extra data from the Internet is prohibited.**
- Please use the <TeamLeader-ID>_<Chinese Name> as the Kaggle nickname to show on the leaderboards.
 - For example, r05922000_王小明 as the nickname.
- Competition pages:
 - Task 1: <https://www.kaggle.com/c/ntucsie-sdml2018-2-1> (open on 10/18 pm 10)
 - Task 2: <https://www.kaggle.com/c/ntucsie-sdml2018-2-2> (open on 10/24 pm 10)

Grading Policy (same as HW1)

Task 1: 50%, Task 2: 50%. For each task,

- Performance (50%)
 - Performance Ranking: all the participant will be ranked according to the Kaggle testing scores.
 - Baselines: you need to beat our baseline for a basic score.
- Report (50%)
 - Coverage (25%): #methods you tried; please describe and analyze the approaches with experiment results.
 - Novelty (25%): how novel is your model designed.
(Ensemble techniques are valid, however we encourage novel single models.)

CEIBA Submissions (same as HW1)

You should submit your source code along with reports to the corresponding CEIBA entries.

- Including any third-party source codes you used.
- A .zip file should be uploaded for each task. (i.e., you should upload two .zip files in total for HW2)
- Your CEIBA submissions should match your final output on the Kaggle platform. That is, your source codes should be able to reproduce your final performance scores on Kaggle.

Reports (same as HW1)

Two report files should be submitted for the two tasks, respectively.

- Your reports should be formatted in PDF files.
- Only digital submissions on CEIBA are acceptable.
- The reports should include:
 - Official name and the student ID of each member.
 - Attempted approaches to solve specific problems.
 - Analyses and observations based on experiment results.
 - Difficulties encountered, unsolved issues, etc.
 - No page limit.
 - Feel free to include all the experiment results, reference theorems or other appendices.

Format of CEIBA Submissions (same as HW1)

[student-id].zip (team leader's student id, e.g., r05922000.zip)

| -- src/ (the source codes written by you)

| -- lib/ (all the libraries, third-party source codes you used)

| -- report.pdf

| -- README (a 'plaintext' file to explain how to reproduce your results)

(You must submit this .zip to get the 50% performance points.)

<Important> You should upload in .zip format.

.rar, .tar, .gz, .7z, or any other formats will receive 0 points without grading.

Submission Deadlines (no extension will be granted)

- Due time (task 1): 2018/11/06 23:59:59 (Taiwan time)
 - According to the system times of Kaggle and CEIBA
- Due time (task 2): 2018/11/13 23:59:59 (Taiwan time)
- Report due on: 2018/11/18 23:59:59
 - Only team leader is responsible for submitting the .zip & report.
- HW2 presentation: 11/15
- For the delayed submissions:
 - Within 24 hours: original task score * 0.5
 - More than 24 hours: zero point for that task.

Contact TA

Nancy Cheng 程子玲

email: nancy.cheng.tl@gmail.com

TA hour: (Fri) pm 2 - 3