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	**Task 1**	**Task 2**
User 33	I23	I21, I22	I17	I16, I18	I24	**I16, I18, I24**	**I16, I18**
User 44	I17	I35, I63	I77, I88	I131	I135, I222	**I131, I222**	**I131**
User 55	I18, I19	I14	I15	I17, I20	I14	**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
==========
-
bodybuilding_supplements
energy_food_products
supplements
==========
breakfast_cereal_subcat
==========
-
breakfast_cereal
==========
side_dish_subcat
==========
-
antipasto
fries
legumes
maize
pasta
potatoes
salad
soup
==========
condiment_subcat
==========
-
condiment
==========
beverage_subcat
==========
-
alcohol
chocolate
coffee
energy_drink
fruit_juice
hot_beverage
lemon_beverage
milk_substitutes
rice_beverage
soft_drink

snack_subcat
==========
-
snack
==========
dessert_subcat
==========
-
cake
chinese_desserts
confectionery
cookies
custards
doughnuts
frozen_desserts
ice_cream
pastries
pies
puddings
==========
fast_food_brand_subcat
==========
-
fast_food
==========
snack_brand_subcat
==========
-
snack_brand
==========
herb_spice_subcat
==========
-
herbs
mixtures
peppers
spices

staple_subcat
==========
-
banana
maize
meat
other_cereal
rice
root_tuber
wheat
==========
meat_subcat
==========
-
beef
fish
game
lamb
meat
meatball
pork
poultry
sausage
seafood
shellfish
==========
egg_dairy_subcat
==========
-
dairy_product
egg
==========
vegetable_subcat
==========
-
bulb_stem_vegetables
flowers_flower_buds
fruits
leafy_salad
podded_vegetables
root_tuberous_vegetables
sea_vegetables
vegetables

mushroom_subcat
==========
-
mushroom
==========
fruit_subcat
==========
-
berry
mediterranean
temperate
tropical
==========
nut_seed_subcat
==========
-
gymnosperm_seed
nut
other
pseudocereal
==========
bean_legume_subcat
==========
-
legume
soy_product
==========
preparation_subcat
==========
-
device
dry
fat
mechanical
mixed
non_heat
wet

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

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

  - 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

$$MAP@\mathrm{K} = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{m}\sum_{\mathrm{i}=1}^{\mathrm{K}} P_{\mathrm{u}}(\mathrm{i})rel_{\mathrm{u}}(\mathrm{i})$$

Evaluation : MAP@5 Example

Rank	Hit?
1	
2	X
3	X
4	X
5	

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

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

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