Homework 2: Location Popularity Ranking: cases study on Movie Theater and Hotel

Hsun-Ping Hsieh

EE, CCE

UCLAB

Urban Science & Computing Lab

Location Popularity Ranking: cases study on Movie Theater and Hotel

- Goal: given a set of candidate areas in the city to open a store, our aim is to identify the most promising ones in terms of their prospect to attract a large number of check-ins (i.e, popular).

- Motivation: land economy, and its importance in the success of a business

Task

- Assume you are a boss who wants to build movie theater and hotel in New Yorl
- Given:
  - Users' all kinds of check-in records in New York
  - 20 candidate locations with their position(lat, lng)
- You need to
  - Identify the ones can attract more customers
  - Rank them by your estimated popularity

Check-in format (Tab division)

- 1. User ID
- 2. Venue ID
- 3. Venue category ID
- 4. Venue category name
- 5. Latitude
- 6. Longitude
- 7. Timezone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC) **(-300 or -240 in NYC)**
- 8. UTC time

730	439c437bf964a520f02b1fe3	4bf58dd8d48988d1fa931735	Hotel	40.758328	-73.985457	-300	Sat Jan 12 19:00:22 +0000 2013
973	4e713390fa766da6339dc53f	4bf58dd8d48988d1df941735	Bridge	40.70629485823015	-73.99711489677429	-300	Sat Jan 12 19:00:43 +0000 2013
295	4b1edd38f964a520b52024e3	4bf58dd8d48988d179941735	Bagel Shop	40.773601	-73.959708	-300	Sat Jan 12 19:03:00 +0000 2013
674	4ec7349e2c5b532d065dfa93	4bf58dd8d48988d1c0941735	Mediterranean Restaurant	40.738526793717064	-74.00365413986914	-300	Sat Jan 12 19:04:50 +0000 2013
675	4a785af0f964a52071e51fe3	4bf58dd8d48988d1e0931735	Coffee Shop	40.776241837275684	-73.94997418614841	-300	Sat Jan 12 19:05:42 +0000 2013
553	4aea39e4f964a52058ba21e3	4bf58dd8d48988d16c941735	Burger Joint	40.70979201243495	-73.8594102859497	-300	Sat Jan 12 19:06:16 +0000 2013
1081	4c34858f66e40f47009fc98b	4bf58dd8d48988d1d0941735	Dessert Shop	40.739486020568506	-73.78538131713867	-300	Sat Jan 12 19:06:32 +0000 2013
437	4f8cb6bfe4b03758900ba5ab	4bf58dd8d48988d1be941735	Latin American Restaurant	40.72773552228127	-74.00261546349739	-300	Sat Jan 12 19:07:37 +0000 2013
349	4ab94773f964a520c07e20e3	4bf58dd8d48988d1f8941735	Furniture / Home Store	40.77283953431822	-73.98224823531172	-300	Sat Jan 12 19:08:47 +0000 2013
349	4bbe24468ca376b0f45ec77a	4bf58dd8d48988d164941735	Plaza	40.77348003217447	-73.982059	-300	Sat Jan 12 19:08:57 +0000 2013
349	4ae8ccbff964a52068b221e3	4bf58dd8d48988d13b941735	School	40.77365559748017	-73.98337310904346	-300	Sat Jan 12 19:09:29 +0000 2013
372	4f5f54fae4b01e3f069e79c5	4bf58dd8d48988d124941735	Office	40.749125957451646	-73.96870906027378	-300	Sat Jan 12 19:09:39 +0000 2013
416	4a9b03dcf964a520013420e3	4bf58dd8d48988d1e0931735	Coffee Shop	40.72851994502955	-73.98732733228006	-300	Sat Jan 12 19:10:52 +0000 2013
720	4c82cc18dc018cfab730d36c	4bf58dd8d48988d1c6941735	Scandinavian Restaurant	40.923597873945845	-74.07411575317383	-300	Sat Jan 12 19:12:32 +0000 2013
720	4a8ff4acf964a520b71520e3	4bf58dd8d48988d1f8941735	Furniture / Home Store	40.924059944774754	-74.07398700714111	-300	Sat Jan 12 19:12:45 +0000 2013

Target instances(Number of instances: 20)

- Venue ID latitude longitude

4da5dd2fcda1c55f755f88c5	40.7592973036631	-73.9953273548198
4a7b8beaf964a52058eb1fe3	40.7555958969709	-73.9738011360168
44d7b1f3f964a52070361fe3	40.7365071557598	-73.9888040721416
4b8d436bf964a520a2f032e3	40.7041314104303	-74.1860818862915
4f22ca77e4b0ed3396a83a05	40.7150236306864	-74.0158423847851
41575800f964a520311d1fe3	40.759829140949	-73.9859557023833
4a967ab2f964a520362620e3	40.7653700556837	-73.9760568737984
49d18dfdf964a5208f5b1fe3	40.7641599692142	-73.9737885148353
491301a9f964a52066521fe3	40.7600712229529	-73.9863689140948
4d0304fc54d0236ac1a2e6d5	40.96550875	-74.0628835833333
4ae6f117f964a520a6a721e3	40.7450768413166	-73.9886759964726
4ec6d0b4be7ba4fc6da4febd	40.7278344793782	-73.9908969674878
4cec13060f196dcb7b8e5bae	40.7198270119281	-74.0000009536743
4ab79e30f964a520397a20e3	40.7605669795532	-73.9847734528531
4bbf8a28f8219c74a127b010	40.7438084396906	-73.9829885866209
4ac2d629f964a520d79a20e3	40.7238332903496	-74.0052609209414
4e0cfe6ae4cd27fc7d21976c	40.7440300852635	-73.9840388475374
3fd66200f964a520bee71ee3	40.7680829580543	-73.9849857991787
49d3d4a7f964a5201a5c1fe3	40.7408839808236	-74.0076595904337
4d9c92f4baae54815f2cde64	40.7419897287103	-74.0035843849182

Effectiveness

- NDCG@20
  - relevance judgments are in a scale of [0,r], r>2
  - Evaluate the quality of ranking

Summarize a Ranking: DCG

- Cumulative Gain (CG) at rank n
  - Let the ratings of the n locations be r_1, r_2, …r_n (in ranked order)
  - CG = r_1+r_2+…r_n
- Discounted Cumulative Gain (DCG) at rank n
  - DCG = $r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots r_n/\log_2 n$
    - We may use any base for the logarithm, e.g., base=2

Discounted Cumulative Gain

- DCG is the total gain accumulated at a particular rank p:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- Alternative formulation:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

  - used by information retrieval
  - emphasis on retrieving highly relevant documents

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:
  - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- discounted gain:
  - 3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
  - = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- DCG:
  - DCG_1=3, DCG_2=3+2=5, DCG_3=3+2+1.89=6.89…

Summarize a Ranking: NDCG

- Normalized Cumulative Gain (NDCG) at rank n
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
  - Must be normalized by the DCG of idea ranking

    3, 3, 3, 2, 2, 2, 1, 0, 0, 0

NDCG - Example

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

NDCG – in HW2

- Assume I want evaluate k=20 candidate locations

- $DCG_{GT20}=19 + 18/\log_2 2 + 17/\log_2 3 + 16/\log_2 4 \ldots +0/\log_2 20$
  [idea]:19,18,17,16,15….0
- $DCG_{\text{your method20}} = 18+19/\log_2 2+0/\log_2 3+\ldots.17/\log_2 20$
  [Your method]:18,19,0,…17
- $NDCG@20=DCG_{\text{your method20}}/DCG_{GT20}$
- $NDCG@10=DCG_{\text{your method10}}/DCG_{GT10}$

Submitted file formation

- Two files, movie theater and hotel
- Each row has only one location_id
- E.g.
  locationID1
  locationID2
  locationID3
  .
  .
  .
  .

Hint1: Regression method

- For each candidate location, predicting the total number of checkins by extracting their neighborhood variables
  - For example, number of surrounding stadium -> Competitiveness for Hotel
  - For example, big transportation stations -> Helpful for Hotel
  - For example, many number of checkins at night -> Helpful for movie theater
  - Neighborhood definition: radius = k meters

Geographic Features & Mobility Features

- Static
  - Number of location for certain type
  - Density
- Dynamic
  - Area Popularity
  - Incoming flow
  - Number of checkins for certain kind of location

Category Hierarchy

- https://developer.foursquare.com/categorytree
  - Arts & Entertainment->**Movie Theater** ->Drive-in Theater
  - Travel & Transport->**Hotel** -> Motel

Submitted report formats(1)

Movie Theater	Physical meaning
Variable 1	
Variable 2	
Variable 3	
Variable 4	
Variable 5	
Variable 6	
Variable 7	
Variable 8	
Variable 9	
Variable 10	

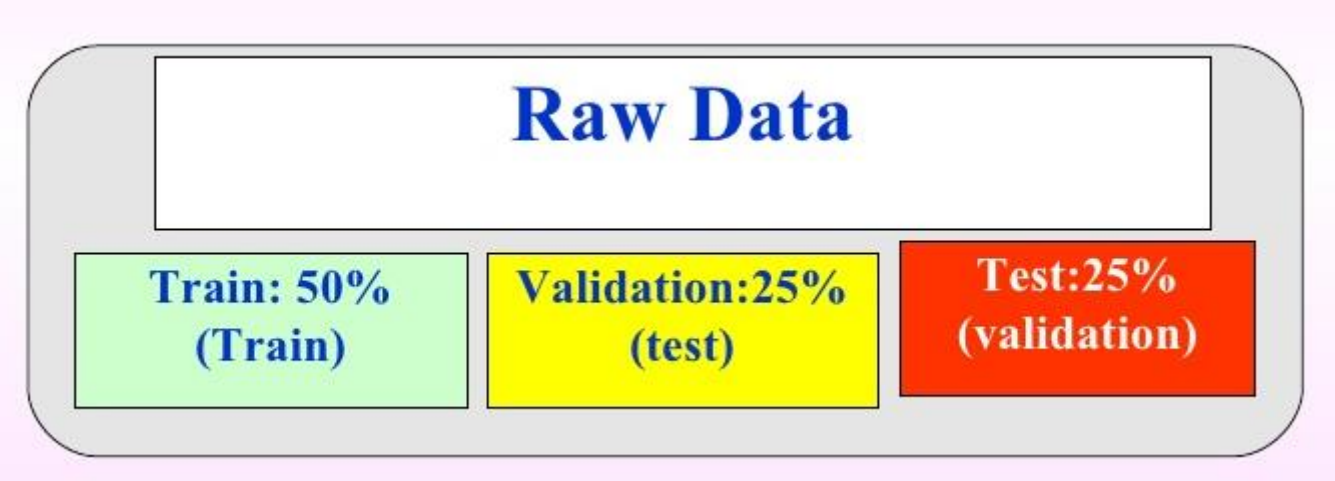
Hotel	Physical meaning
Variable 1	
Variable 2	
Variable 3	
Variable 4	
Variable 5	
Variable 6	
Variable 7	
Variable 8	
Variable 9	
Variable 10	

Submitted report formats(2)

- Questions
  - How do you combine proposed variables, anything worth to mention for your method?

Hint2: validation & testing

- If you want to tune parameters…(e.g. radius or SVR's parameters)
  - You need to divide validation set from the training set



The training data I gave you
(Train_MoiveTheater.txt & Train_Hotel.txt)

This week's suggestions

- Please be quickly to extract **all locations (and corresponding checkins)** using **big** range around target locations, both for training and testing locations

Policy

- Deadline: 12/13 23:59pm
  - Penalty: each day late -5
- Submit your report
- Submit your answer(two files)
  - Ranking of these 20 candidate locations for Movie Theater and Hotel
- 15% of your final grade
- Grade: NDCG@n: 50%, report: 50%
  - Normalized score
- Encourage you guys to propose your own variables(features)
- Discuss with your teammate(1-3)

Please discuss