Recommender Systems on Yelp Dataset

Presenter: Arnold Jiadong Yu

San Francisco State University

May 21, 2019

Introduction

1. Goal - With input data such as City and type of business, give the user recommedations of the top 10 businesses.
2. Dataset - The dataset can be obtained from Yelp Website. It contains only information of business, reviews, user information, checkin, tip, and photos in moderate amount of cities in America. The business file contains around 200 thousands businesses. The reviews file contains around 7 millions reviews.
3. Challenges - The dataset is large and it takes time to access the data and run the algorithms. It is not easy to figure out the algorithms of Collaborative Filtering as well as how to calculate MAP@N and MAR@N. I have to code most algorithms manually since it varies based on different dataset.

Popularity-based

The popularity based recommendation system requires a input of a type of business and city. (e.g. Chinese, Phoenix).

The ranking follows the rule listed below,

Stars	Negative	Positive	Sum
0	1	0	1
0.5	0.9	0.1	1
1	0.8	0.2	1
1.5	0.7	0.3	1
2	0.6	0.4	1
2.5	0.5	0.5	1
3	0.4	0.6	1
3.5	0.3	0.7	1
4	0.2	0.8	1
4.5	0.1	0.9	1
5	0	1	1

Content-based

The Content-based recommendation system requires a input of a type of business, city and a short description. (e.g. Chinese, Phoenix, I like chicken soup.)

The ranking follows Tf-idf scores.

Collaborative Filtering - Similarity Based

The Collaborative Filtering - Similarity Based recommendation system require a input of user-business matrix.

$$\text{rate}(i,j) = \frac{\sum_{i' \in \Phi_j} w_{ii'} r_{i'j}}{\sum_{i' \in \Phi_j} w_{ii'}}$$

where Φ_j is the set of all users who rated item j, $r_{i'j}$ is the rating that user i' given item j.
$w_{ii'}$ are the weight computed by the neighbor of the user i'.

$$w_{ii'} = \frac{\sum_{j \in \Psi_{ii'}} (r_{ij} - \bar{r}_i)(r_{i'j} - \bar{r}_{i'})}{\sqrt{\sum_{j \in \Psi_i} (r_{ij} - \bar{r}_i)^2} \cdot \sqrt{\sum_{j \in \Psi_{i'}} (r_{i'j} - \bar{r}_{i'})^2}}$$

where Ψ_i is the set of items that user i rated, $\Psi_{ii'}$ is set of items both user i and i' rated and $\bar{r}_i$ is the average of all r_{ij}'s.
Deviation has also been considered since user i rated business j a star 5 may not mean the same as user i' rated business j a star 5.

$$\text{dev}(i,j) = r_{ij} - \bar{i}_i$$

Collaborative Filtering - Model Based

Let **M** be $m \times n$ matrix which represents user-business matrix, m user and n business.

1. Matrix Factorization - Nonnegative Matrix Factorization(NMF) with regularization

   $$M \approx WH$$

   The objective function

   $$d_{Fro}(M, WH) + \alpha\beta||W||_1 + \alpha\beta||H||_1 + \frac{\alpha(1-\beta)}{2}||W||^2_{Fro} + \frac{\alpha(1-\beta)}{2}||H||^2_{Fro}$$

   where α is the intensity of the regularization and β is L1 and L2 ratio.
2. Deep Neural Network

1. Mean Square Error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

where $\hat{Y}_i$ is the predicted value and Y_i is the actual value.

2. Root Mean Square Error

$$\text{RMSE} = \sqrt{\text{MSE}}$$

Mean Average Precision @ N and Mean Average Recall @ N

1. Precision

$$P = \frac{\#\text{our recommendations that are relevant}}{\#\text{items we recommended}}$$

2. Recall

$$P = \frac{\#\text{our recommendations that are relevant}}{\#\text{all the possible relevant items}}$$

3. Average Precision @ N

$$AP@N = \begin{cases} \frac{1}{\min\{m,N\}} \sum_{k=1}^{N} P(k) \cdot rel(k) & m \neq 0 \\ 0 & m = 0 \end{cases}$$

4. Average Recall @ N

$$AR@N = \begin{cases} \frac{1}{\min\{m,N\}} \sum_{k=1}^{N} P(k) \cdot rel(k) & m \neq 0 \\ 0 & m = 0 \end{cases}$$

where m is relevant items in top N items, $rel(k)$ is an indicator function. MAP@N and MAR@N is mean of AP@N and AR@N.

Experiment Analysis

1. Popularity-based: MAP@10, 1.0, MAR@10, 0.55
2. Content-based: MAP@10, 1.0, MAR@10, 0.55
3. Similarity-based: user-user,train rmse, 0.636866,test rmse, 0.786857 and MAP@10, 0.042729, MAR@10, 0.033349.
   business-business, train rmse, 1.330710, test rmse, 1.362074 and MAP@10, 0.993712, MAR@10, 0.550816
4. Matrix Factorization-based: rmse, 0.221195. MAP@10, 0.981614, MAR@10, 0.553948
5. NN-based: train rmse 0.47424, test rmse 1.467, MAP@10,0.986425, MAR@10, 0.551932

Results

Popularity Based.

```
1 'Chino Bandido''15414 N 19th Ave, Ste K''Phoenix''AZ''85023'
2 'Clever Koi''4236 N Central Ave, Ste 100''Phoenix''AZ''85012'
3 'Snoh Ice Shavery''914 E Camelback Rd, Unit 4B''Phoenix''AZ''85014'
4 "George & Son's Asian Cuisine"'3049 W Agua Fria Fwy''Phoenix''AZ''85027'
5 'China Chili''302 E Flower St''Phoenix''AZ''85012'
6 'Iron Chef''10810 N Tatum Blvd, Ste 106''Phoenix''AZ''85028'
7 "Wong's Chinese Cuisine"'10540 W Indian School Rd, Ste 4''Phoenix''AZ''85037'
8 'International House of Food''1402 W Van Buren St''Phoenix''AZ''85007'
9 'Great Wall Cuisine''3446 W Camelback Rd, Ste 155''Phoenix''AZ''85017'
10 'Loving Hut''3239 E Indian School Rd''Phoenix''AZ''85018'
```

Results

Content Based.

```
1 'Chino Bandido''15414 N 19th Ave, Ste K''Phoenix''AZ''85023'
2 'China Chili''302 E Flower St''Phoenix''AZ''85012'
3 'Iron Chef''10810 N Tatum Blvd, Ste 106''Phoenix''AZ''85028'
4 'Clever Koi''4236 N Central Ave, Ste 100''Phoenix''AZ''85012'
5 "George & Son's Asian Cuisine"'3049 W Agua Fria Fwy''Phoenix''AZ''85027'
6 'Dragon Palace''13825 N 32nd St''Phoenix''AZ''85032'
7 'Thai Basil''3110 N Central Ave''Phoenix''AZ''85012'
8 'Desert Jade''3215 E Indian School Rd''Phoenix''AZ''85018'
9 'Szechwan Palace''668 N 44th St, Ste 108, Cofco Chinese Cultural Center''Phoenix''AZ''85008'
10 'Maxim Restaurant''3424 N 19th Ave''Phoenix''AZ''85015'
```

Results

User-User Based.

```
In [36]: findTop10_and_Print(dict_3_train["'-318sKiQDgbjLzF4FCU1XA'"])
1 'Pink Lotus Express''18635 N 35th Ave''Phoenix''AZ''85027'
2 'Abacus Inn Chinese Restaurant''3509 W Thunderbird Rd''Phoenix''AZ''85053'
3 'Eastern Buffet''1617 E Thomas Rd''Phoenix''AZ''85016'
4 'Noodle & Rice''2017 E Cactus Rd, Ste G''Phoenix''AZ''85022'
5 'Sun Asian Kitchen''2070 E Baseline Rd, Ste D112''Phoenix''AZ''85042'
6 'Wahsun Chinese Restaurant''8056 N 19th Ave''Phoenix''AZ''85021'
7 'Silver Dragon Chinese Restaurant''1739 W Glendale Ave''Phoenix''AZ''85021'
8 'Kwan & Wok Chinese Restaurant''1702 W Camelback Rd''Phoenix''AZ''85015'
9 'International House of Food''1402 W Van Buren St''Phoenix''AZ''85007'
10 'El Mesquite Cocina Mexicana''26 E Baseline Rd, Ste 120''Phoenix''AZ''85042'
```

Limitations

1. Popularity Based - It doesn't consider the user personal hobbies nor behaviors, but it is fast, and can work with any kind of user since it only suggests the most popular businesses.
2. Content Based - It also doesn't consider the user personal hobbies nor behaviors, but it is more accurate than popularity based since it will consider reviews of the businesses. The running time is moderate.
3. Collaborative Filtering - It has cold start. It is the most accurate but it also requires the most running time.

Conclusion and Future Work

The recommender systems covered almost all techniques that are used nowadays. These techniques also give us a understanding of trade-off between complexity and results.

Future work can be considering relationships between users in collaborative filtering. For example, a graph of users can be built. Features of the graph can be calculated and the weights can be more accuracy in similarity based model.