Term Project Report

Arnold Jiadong Yu

May 26, 2019

1 Section I: Goal

The goal is to build a recommendation system to the users. With a input such as city and the type of business, give the user recommendations of the top 10 businesses.

2 Section II: Datasets and Preprocessing

The data downloaded from yelp website (https://www.yelp.com/dataset) contains business.json, review,json, user.json, checkin.json, tip.json and photo.json. I only need to use busness.json, review.json, user.json to build our recommendation system. Therefore, I need to preprocess the data into csv form to work on.

While converting json file to csv file, I can check the column names. Most columns in business.csv, review.csv, user.csv will not be needed. Therefore, we need to preprocess them. We also need to handle missing value and string conversion. String in json file was presented in binary, after conversion it will contain b" which represent converted from binary.

1. Preprocess business.csv, size from (192609, 60) $\Rightarrow$ (192127, 11)
2. Preprocess review.csv, size from (6685900, 9) $\Rightarrow$ (6685900, 4)
3. Preprocess user.csv, size from (1637138, 22) $\Rightarrow$ (1637138, 4)

3 Section III: Main Strategies

3.1 Popularity Based

The popularity based recommendation system require 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

This works very similar as a regular 5 stars ranking system. I count negative and positive score based on reviews of that business. I rank the business by taking the difference between positive and negative score. The higher the score, the higher the ranking.

3.2 Content Based

In Popularity Based Recommendation System, when I search for category Chinese and city Phoenix, the result it's not specific enough. Here I introduce Content Based Recommendation System. In regular setting, content based compare similarity by building similarity matrix. Since the input doesn't contain description, I add it manually. (e.g. "I like fresh food, quick order. I also like chinese chicken soup") The ranking is

based on the description I entered.

I used a technique called TFIDF. I built a metric space where center is the description. Therefore, the higher the Euclidean distance to the origin, the higher ranking. This is a modified content based system compare to regular one.

3.3 Collaborative Filtering

3.3.1 Similarity Based

In Content Based Recommendation System, I didn't consider user personal behavior or experience. Here I introduce Collaborative Filtering Recommendation System with both user based and item based. In regular setting, Collaborative Filtering computes Pearson correlation coefficient. I used the following formula to derive the results.

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

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.

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

For computation efficiency, we only keep K nearest neighbors and a $LIMIT$ number for two users to be considered as neighbors.

To avoid cold star and performance efficiency, I consider user had rated one item, two items and three items in the review. I also set $K = 25$ and $LIMIT = 2$ to test accuracy and performance.

Item based Collaborative Filtering is arithmetical equivalent to user based collaborative filtering. I just made a slightly modification to the above two equations to compute the results. I set $K = 20$ and $LIMIT = 5$.

3.3.2 Model Based

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

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

   $$M \approx WH$$

   where W is $m \times k$ and H is $k \times n$, k is latent dimensions.
   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.
   In the code, $\alpha = 0.01$ and $\beta = 0.05$.

2. Deep Neural Network - Neural Network is applied. It is very similar to Matrix Factorization which we work with W and H, here we have two embedding layers. Then we turn them into a sequences by flattening and concatenate them. The activation function is ReLu with mse loss and mse metrics. Also regularization is added.

4 Section IV: Evaluation and Results

4.1 RMSE

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

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

4.2 MAP@K

1. Precision

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

2. Precision @ N : proportion of top-N recommendations that are relevant.

3. Average Precision @ N

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

where m is relevant items in top N items, $P(k)$ is precision $@kth$ in top-N and $rel(k)$ is an indicator function. MAP@N is mean of AP@N of all users.

4.3 MAR@K

1. Recall

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

2. Recall @ N : proportion of relevant recommendations that are in top-N.

3. 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, $P(k)$ is recall $@kth$ in top-N and $rel(k)$ is an indicator function. MAR@N is mean of AR@N of all users.

4.4 Results

1. Popularity Based

```
In [1]: runfile('C:/Users/Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term project2/2.Popluarity_based/
recommendation_popularity_based.py', wdir='C:/Users/Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term
project2/2.Popluarity_based')
MAP@10: 1.000000
MAR@10: 0.550000
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'
```

2. Content Based

```
In [2]: runfile('C:/Users/Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term project2/3.Content_based/
recommendation_content_based.py', wdir='C:/Users/Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term project2/3.Content_based
MAP@10: 1.000000
MAR@10: 0.550000
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'
```

3. Collaborative Filter - Similarity Based, User-User

```
In [4]: runfile('C:/Users/Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term project2/4.Collaborative Filtering/CF_user_user
wdir='C:/Users/Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term project2/4.Collaborative Filtering')
RMSE of Training Set : 0.636866
RMSE of Test Set : 0.786857
MAP@10: 0.042729
MAR@10: 0.033349
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'
```

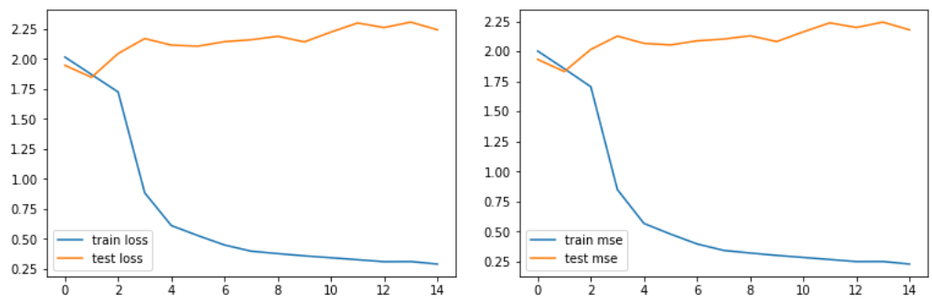
4. Collaborative Filter - Similarity Based, Business-Business

```
In [5]: runfile('C:/Users/Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term project2/4.Collaborative Filtering/
CF_business_business.py', wdir='C:/Users/Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term project2/4.Collaborative Filteri
RMSE of Training Set : 1.330710
RMSE of Test Set : 1.362074
MAP@10: 0.993712
MAR@10: 0.550816
1 'Hong Kong Kitchen''510 E Baseline Rd, Ste D3''Phoenix''AZ''85042'
2 "Wong's Chinese Dining"'1139 E Buckeye Rd''Phoenix''AZ''85034'
3 'Loving Hut''3239 E Indian School Rd''Phoenix''AZ''85018'
4 'Snoh Ice Shavery''914 E Camelback Rd, Unit 4B''Phoenix''AZ''85014'
5 'Harmony and Health Acupuncture''4020 N 20th St, Ste 212''Phoenix''AZ''85016'
6 'Sun Tree Healing Arts''7227 N 16th St, Ste 216''Phoenix''AZ''85020'
7 'The Clinic of Quan Acupuncture & Pain Relief''4550 E Bell Rd, Bldg 8''Phoenix''AZ''85032'
8 'Hive Healing House''1357 W Mulberry Dr''Phoenix''AZ''85013'
9 'King Wong Chinese Food''2545 N 32nd St''Phoenix''AZ''85008'
10 'China Chan Restaurant''10227 N Metro Pkwy E''Phoenix''AZ''85051'
```

5. Collaborative Filter - Model Based, Matrix Factorization

```
In [6]: runfile('C:/Users/Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term project2/5.MF and NN/Matrix_F.py', wdir='C:/User
Home/Desktop/SFSU/Spring 2019/CSC869/Projects/Term project2/5.MF and NN')
RMSE of DataSet : 0.221195
MAP@10: 0.981614
MAR@10: 0.553948
1 'China Chili''302 E Flower St''Phoenix''AZ''85012'
2 'Maxim Restaurant''3424 N 19th Ave''Phoenix''AZ''85015'
3 'Thai Basil''3110 N Central Ave''Phoenix''AZ''85012'
4 'Mu Shu Asian Grill''1502 W Thomas Rd''Phoenix''AZ''85015'
5 'Great Wall Cuisine''3446 W Camelback Rd, Ste 155''Phoenix''AZ''85017'
6 'Siu Wok''2801 N Central Ave''Phoenix''AZ''85012'
7 'International House of Food''1402 W Van Buren St''Phoenix''AZ''85007'
8 'Loving Hut''3239 E Indian School Rd''Phoenix''AZ''85018'
9 'The Prime Chinese Restaurant''24 W Camelback Rd, Ste H''Phoenix''AZ''85013'
10 'Super Dragon''1212 E Northern Ave''Phoenix''AZ''85020'
```

6. Collaborative Filter - Model Based, Deep Neural Network




```
MAP@10: 0.986161
MAR@10: 0.551446
1 'Que Huong Restaurant''3424 N 19th Ave, Ste 8''Phoenix''AZ''85015'
2 'Snoh Ice Shavery''914 E Camelback Rd, Unit 4B''Phoenix''AZ''85014'
3 "Yi's Chinese Restaurant"'1512 W Bell Rd, Ste 7''Phoenix''AZ''85023'
4 "George & Son's Asian Cuisine"'3049 W Agua Fria Fwy''Phoenix''AZ''85027'
5 'The Clinic of Quan Acupuncture & Pain Relief''4550 E Bell Rd, Bldg 8''Phoenix''AZ''85032'
6 'Blazin Mongolian BBQ''9620 N Metro Pkwy W''Phoenix''AZ''85051'
7 'Big Heng''1739 W Glendale Ave''Phoenix''AZ''85021'
8 'Noodle & Rice''2017 E Cactus Rd, Ste G''Phoenix''AZ''85022'
9 'Dragon Bowl''814 E Union Hills Dr, Ste C-14''Phoenix''AZ''85024'
10 'International House of Food''1402 W Van Buren St''Phoenix''AZ''85007'
```

4.5 Results Analysis

All the results from collaborative Filter is to predict the same user. Clearly, it gives different recommendations. Matrix Factorization gives the best and stable results. The business to business similarity based has the highest rmse score. The deep Neural Network cost the longest to run and the results are not stable since the input data is very sparse.

5 Section V: Pros and Cons of the Main Strategies

5.1 Pros and Cons of Popularity Based

1. Pros: Easy to understand and implement. Hyper parameter free. Scalability, less computational power needed.

2. Cons: Same Recommendation for all users. It doesn't consider user's personal hobbies nor behaviors.

5.2 Pros and Cons of Content Based

1. Pros: Unlike Popularity Based, it considers user's personal hobbies and behaviors. Unlike Collaborative Filtering, if the items have enough descriptions, we can avoid cold start. Content based are varied and they open up for different techniques of text analysis.

2. Cons: Content Based tends to over-specialization. They will recommend items similar to those already consumed. They can have some hyper parameter for text analysis. Scalability, the computational time is moderate.

5.3 Pros and Cons of Collaborative Filtering

1. Pros: It gives the more accurate recommendations between these three types.

2. Cons: Cold start, it requires a large amount of data on a user in order to make a accurate recommendations. More hyper parameters. Scalability, a large amount of computation power is often required to calculate recommendations since it involves millions of users and items.

6 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 semantics analysis for content based. Also, relationships between users could be considered 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. The different number of hyper parameters, more layers and varies of activation functions can be tested too.