Yelp Rating Prediction

Yingzi Xu

What data we have ?

- Business (id, name, location, category, open …)
- User (id, name, votes …)
- Check In (business id, # check in during each time period)
- Review (business id, user id, stars, dates, votes…)
  - Split into 2 subsets
  - 80% training , 20% testing

Model 1 - 4

- Model 1
- Overall Mean
- RMSE
- 1.217

- Model 2
- User Mean
- RMSE
- 1.266

- Model 3
- Business Mean
- RMSE
- 1.153

- Model 4
- User-
- Business
- Mean
- RMSE
- 1.219

Model 5

- Business-based collaborative filtering recommendation

$$\hat{r}_{ui} = \bar{r} + \frac{\sum_{j \in N_u(i)} w_{ij}(r_{uj} - \bar{r}_j)}{\sum_{j \in N_u(i)} w_{ij}}$$

- $\bar{r}$ is the average rating of all users have given to business i
- for user u, the rating for business i is the weighted average of the same user's ratings on similar business
- RMSE = 1.275
- How do we define similar ??

Model 5

- Similarity measure : Jaccard Index $J_{ik} = \frac{|I \cap K|}{|I \cup K|}$

	User 1	User 2	User3	User 4	User 5	User 6
Bus A	4			5	1	
Bus B	5	5	4			
Bus C				2	4	5
Bus D		3				

J_{AB} = 1/5
J_{AC} = 2/4

Business A appears closer to C .
WRONG !

- Rounding the ratings

- Consider ratings of 3, 4, 5 as a "1" ; Consider ratings 1 ,2 as unrated.

	User 1	User 2	User3	User 4	User 5	User 6
Bus A	1			1		
Bus B	1	1	1			
Bus C					1	1
Bus D		1				

J_{AB} = 1/4
J_{AC} = 0

Business A appears closer to B.

1.137

Model 6

- Blend model 1- 5 together

$$\hat{r}_{ui} = \frac{\sum_k (w_k \hat{r}_{uik})}{\sum_k w_k}, k = 1,2,3,4,5$$

$$w_k = \frac{1}{RMSE_k^3}$$

- RMSE = 1.137
- The best !

Thank you.