Collaborative Filtering Exercise

CSCI 374 Oberlin College Fall 2017
September 17, 2017

Introduction

For this exercise, we will be using real-world data from Last.fm (available from: http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-360K.html) to see how Collaborative Filtering can be used to recommend artists to music listeners.

Data Set

The following is the number of plays for each of the 10 users who listened the most times to the 5 most popular bands in the data set. This data will serve as our information about other users that will be used to make recommendations. We will consider the number of times a user has played a song by an artist to be an implicit rating, rather than asking them to rate each artist on some scale.

User	Artist	Plays
User 1	the beatles	39655
User 2	muse	44076
User 2	coldplay	962
User 2	radiohead	903
User 3	muse	47468
User 3	coldplay	6051
User 3	radiohead	489
User 4	pink floyd	31957
User 4	the beatles	14975
User 5	the beatles	31526
User 5	pink floyd	5882
User 6	muse	42970
User 7	the beatles	33685
User 7	pink floyd	2351
User 7	radiohead	2304
User 8	coldplay	31121
User 8	radiohead	18652
User 8	muse	690
User 9	coldplay	118857
User 9	the beatles	4
User 10	muse	44036
User 10	coldplay	168

Convert to User Rating Vectors

The first thing we need to do is convert the data set above into a set of ratings vectors, one vector for each user. This is done by copying the data above into the table below. Please fill in this table.

User	The Beatles	Radiohead	Coldplay	Pink Floyd	Muse
User 1					
User 2					
User 3					
User 4					
User 5					
User 6					
User 7					
User 8					
User 9					
User 10					

Users for Recommendation

Here are the ratings histories for two new users for whom we want to make recommendations:

User	The Beatles	Radiohead	Coldplay	Pink Floyd	Muse
User 21	3344	*null*	*null*	22458	*null*
User 101	*null*	6293	2286	*null*	5156

Calculating Similarity

Now that we have users for whom we want to calculate recommendations, as well as the ratings histories of other users, we need to start calculating the similarities between all of the users.

For simplicity, we will use the Manhattan distance as our distance function for this exercise (since it doesn't require any squares or square roots). Recall that Manhattan distance is calculated as:

$$d(\boldsymbol{p^i}, \boldsymbol{p^j}) = \frac{1}{|A|}\sum_{a \in A}\left|p_a^i - p_a^j\right|$$

where A is the set of artists to which both $\boldsymbol{p^i}$ and $\boldsymbol{p^j}$ have listened**.** From distance, we can calculate similarity as:

$$sim(\boldsymbol{p^i}, \boldsymbol{p^j}) = \frac{1}{d(\boldsymbol{p^i}, \boldsymbol{p^j})}$$

Fill in the table below with the similarities between users:

	User 21	User 101
User 1		
User 2		
User 3		
User 4		
User 5		
User 6		
User 7		
User 8		
User 9		
User 10		

Estimating Ratings

Next, we need to estimate the ratings for User 21 and User 101 for the artists for which they have not yet listened. We can use three different formulas for these ratings:

Option 1: Average Rating

$$\hat{r}(\boldsymbol{p}, t) = \frac{1}{k} \sum_{\boldsymbol{p^i} \in \boldsymbol{P_k}} r(\boldsymbol{p^i}, t)$$

where t is an artist that $\boldsymbol{p}$ hasn't rated, k is a number, and $\boldsymbol{P_k}$ are the k users most similar to $\boldsymbol{p}$

Option 2: Weighted Average Rating

$$\hat{r}(\boldsymbol{p}, t) = \frac{1}{z} \sum_{\boldsymbol{p^i} \in \boldsymbol{P}} sim(\boldsymbol{p}, \boldsymbol{p^i}) * r(\boldsymbol{p^i}, t)$$

where P is the set of all previous users (Users 1 – 10) and

$$z = \sum_{p^i \in \boldsymbol{P}} sim(\boldsymbol{p}, \boldsymbol{p^i})$$

Option 3: Adjusted Weighted Average Rating

$$\hat{r}(\boldsymbol{p},t) = \bar{r}_{\boldsymbol{p}} + \frac{1}{z}\sum_{\boldsymbol{p^i} \in \boldsymbol{P}} sim(\boldsymbol{p},\boldsymbol{p^i}) * \left[r(\boldsymbol{p^i},t) - \bar{r}_{\boldsymbol{p_i}}\right]$$

where $\bar{r}_p$ is the average rating by user $\boldsymbol{p}$:

$$\bar{r}_p = \frac{1}{|A|}\sum_{a \in A} r(\boldsymbol{p},a)$$

Note: if another user p^i has not rated artist t, then we leave them out of the above calculations (even if they are one of the k most similar users to p in Option 1 – in that case we substitute for k the number of close neighbors in $\boldsymbol{P_k}$ who have rated t).

Please fill in the below tables:

$\hat{r}(\boldsymbol{p},t)$ for User 21:

	Radiohead	**Coldplay**	**Muse**
Option 1			
Option 2			
Option 3			

$\hat{r}(\boldsymbol{p},t)$ for User 101:

	The Beatles	**Pink Floyd**
Option 1		
Option 2		
Option 3		

Making Recommendations

Choose the song with the highest estimated rating for user (for each option). Circle your choices above (you should circle one song per row). These are the songs you would suggest the user listen to, based on what you know of their preferences and what other users similar to them have enjoyed.