Recommender Systems – Evaluation Measures

Exercise sheet

Keti Korini & Alexander Brinkmann

For this exercise, we use real-world data from Last.fm[1] to evaluate the recommendations of a fictitious recommender system. We evaluate the recommender system's performance using evaluation measures for numerical ratings, categorical ratings and ranked results.

1 Dataset

We start with the data from Table 1. The table shows the play counts for each band in the data set of 2 users – assume empty cells to correspond to no (0) plays for the user-artist pair. We use these implicit ratings as ground truth to evaluate the performance of the fictitious recommender system.

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

	The Rolling Stones	John Lennon	Bob Dylan	Franz Ferdinand	The Doors
User 21	-	-	5482	-	2041
User 101	-	-	-	1817	-

Table 1: Ground Truth

Additionally, we have the predictions of a recommender system as shown in table 2.

	The Beatles	Radiohead	Coldplay	Pink Floyd	Muse
User 21	11659	-	26328	24878	-
User 101	-	387	1067	-	4403

	The Rolling Stones	John Lennon	Bob Dylan	Franz Ferdinand	The Doors
User 21	1422	-	-	-	7580
User 101	-	-	-	504	-

Table 2: Recommender System Predictions

2 Tools

To generate recommendations to the two users, we will use the excel spreadsheet evaluation-measures.xlsx. The spreadsheet contains the data set as well as an implementation of the different evaluation measures.

[1] `http://ocelma.net/MusicRecommendationDataset/lastfm-360K.html`

Task 1: Numerical Ratings

Calculate mean absolute error (MAE) and root mean squared error (RMSE) for the two recommended lists of results. Please use the following formula to calculate the MAE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \tag{1}$$

where p_i and r_i describe the i-th prediction and ground truth of the recommender system. Please use the following formula to calculate the RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2} \tag{2}$$

where p_i and r_i describe the i-th prediction and ground truth of the recommender system.

Task 2: Categorical Ratings

Calculate precision, recall and F_1 for the two recommended lists of results. Assume that play counts > 0 indicate a relevant band. Please use the following formula to calculate precision:

$$precision = \frac{tp}{tp + fp} = \frac{|good\ bands\ recommended|}{|all\ recommendations|} \tag{3}$$

Please use the following formula to calculate recall:

$$recall = \frac{tp}{tp + fn} = \frac{|good\ bands\ recommended|}{|good\ bands|} \tag{4}$$

Please use the following formula to calculate F_1:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

Additionally, calculate average scores for all three categorical evaluation measures.

Task 3: Ranked Results

Calculate Average Precision (AP) and normalized Discounted Cumulative Gain (nDCG) for the two recommended lists of results. Sort all lists of bands descending by the implicit ratings. Good (relevant) bands have an implicit rating > 0. For the normalized Discounted Cumulative Gain the relevance scores are shown in table 3 and table 4. If the band is not mentioned in these tables a relevance score of 0 is assumed.

	Pink Floyd	Bob Dylan	The Beatles	The Doors
User 21	5	2	1	1

Table 3: Relevance of bands for User 21

Please use the following formula to calculate AP:

$$AP = \frac{1}{n}\sum_{k=1}^{n} Precision(good\ band_k) \tag{6}$$

	Radiohead	Muse	Coldplay	Franz Ferdinand
User 101	3	2	1	1

Table 4: Relevance of bands for User 101

where $good\ band_k$ describes a good band at position k.
Please use the following formula to calculate DCG(k) where k=5:

$$DCG(k) = \sum_{i=1}^{k} \frac{2^{rel\ good\ band_i} - 1}{log_2(i+1)} \tag{7}$$

where $rel\ good\ band_i$ describes the relevance score of a good band at position i.
For the idealized DCG use the same formula as for the DCG, but instead of using the ranking of the prediction, use the ideal ranking of bands from the ground truth.
Finally, the nDCG can be calculated using the following formula:

$$nDCG(k) = \frac{DCG(k)}{IDCG(k)} \tag{8}$$