Measures of Similarity

Joshua Bernhard

Why Do We Care?

Why do we care about measures of similarity?

- In order to personalize recommendations, many methods require you to find either similar users or items.

- Not every measure of similarity will return the same users or items as most similar

Measures of Similarity

- Pearson's Correlation Coefficient & Cosine Similarity
- Spearman's Correlation Coefficient
- Jaccard Similarity

Pearson Correlation Coefficient & Cosine Similarity

Pearson Correlation Coefficient & Cosine Similarity

- Pearson Correlation Coefficient measures the strength and direction of the linear relationship between two users or two items

- Does not work well if items or users don't have a lot of ratings/views

- Closer to -1 the stronger a negative relationship

- Closer to 1 the stronger a positive relationship

Pearson Correlation Coefficient & Cosine Similarity

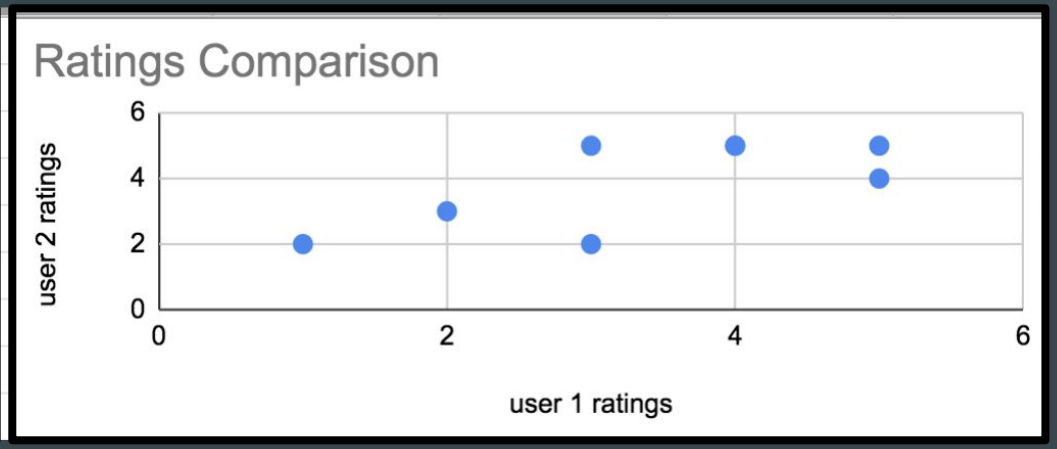
- Cosine Similarity and Pearson Correlation Coefficient are equal when the scores are standardized (subtract mean and divide by standard deviation)

- This works well when you do not have binary events, but rather, user-item ratings (and lots of them).

Pearson Correlation Coefficient & Cosine Similarity

- Given two vectors (could be all ratings for two movies, or all ratings for two individuals)

user1 ratings	user2 ratings
5	5
4	5
3	5
1	2
5	4
2	3
3	2
4	5

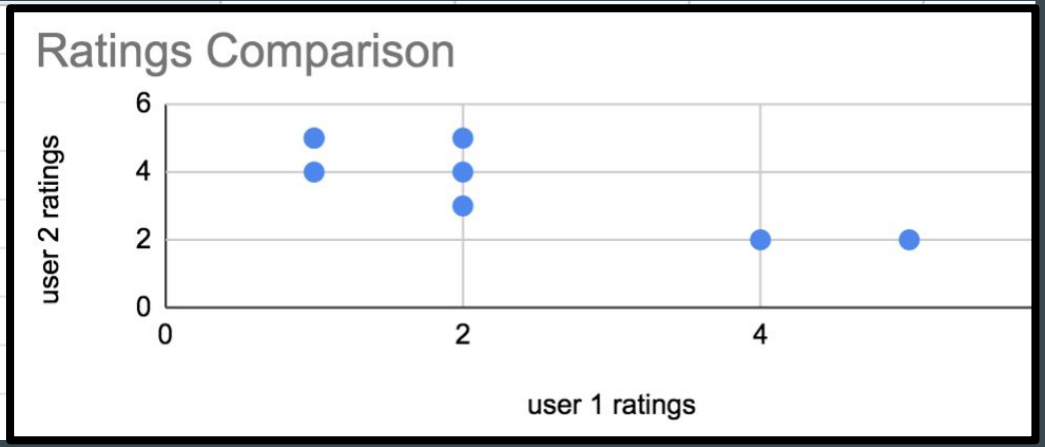


Pearson correlation of 0.7

Pearson Correlation Coefficient & Cosine Similarity

- Given two vectors (could be all ratings for two movies, or all ratings for two individuals)

user1 ratings	user2 ratings
2	5
4	2
1	5
1	4
2	4
2	3
5	2
1	5



Pearson correlation of -0.86

Pearson Correlation Coefficient & Cosine Similarity

Pearson's Correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Cosine Similarity

$$cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}}$$

Spearman’s Correlation Coefficient

Spearman's Correlation Coefficient

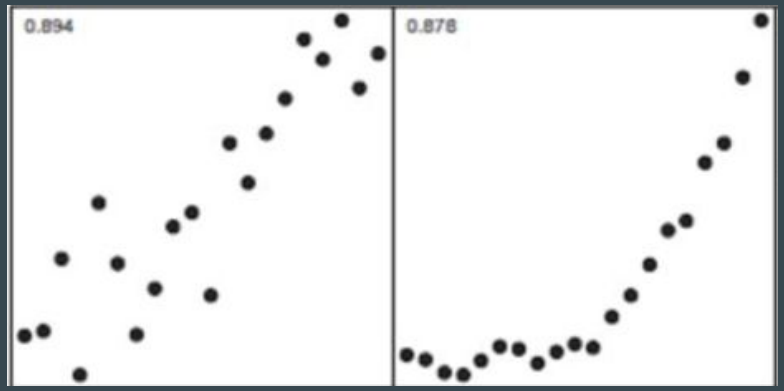
- Spearman's correlation is a non-parametric measure of similarity that does not look for a linear relationship.

- Therefore, if two users (or items) have ratings that are high for the same items and low for the same items, but not with the exact same values, Spearman's coefficient will do a better job of picking up this trend.

- Spearman's coefficient is also between 1 and -1.

Spearman's Correlation Coefficient

- For the first plot, Spearman and Pearson's coefficients will be similar at ~0.9



- In the second plot, the relationship does not fit on a line well, so Spearman will do a better job of capturing these are related with corr ~0.9, while Pearson ~0.8.

Spearman's Correlation Coefficient

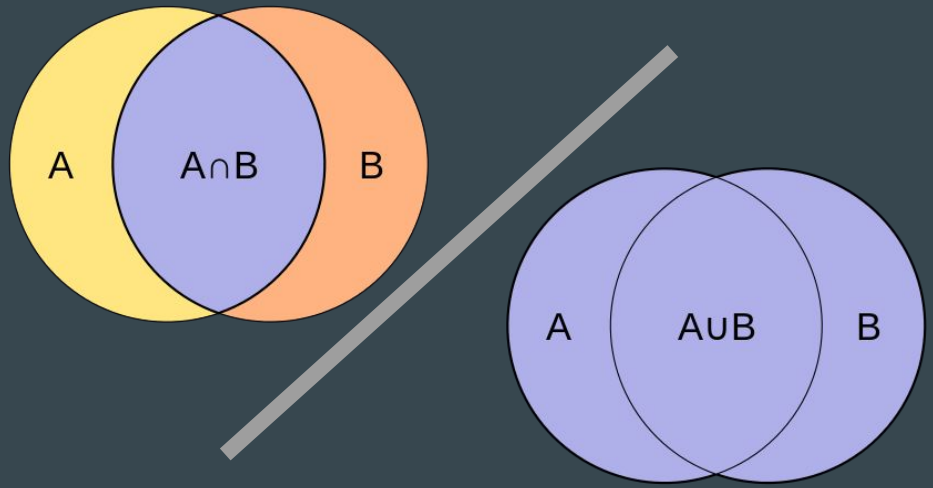
$$\rho = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)\cdot\left(R(y_i) - \overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i) - \overline{R(y)}\right)^2\right)}}$$

Jaccard Similarity

Jaccard Similarity

- Jaccard Similarity is one of the most popular methods to use when comparing two vectors with only binary outcomes.

- Bound between 0 (not similar) and 1 (very similar)



Jaccard Similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Jaccard = 1

	movie 1	movie 2	movie 3	movie 4	movie 5	movie 6	movie 7
user 1	1	0	1	1	1	1	0
user 2	1	0	1	1	1	1	0

Jaccard = 0

	movie 1	movie 2	movie 3	movie 4	movie 5	movie 6	movie 7
user 1	1	0	1	1	0	0	0
user 2	0	1	0	0	0	1	1

Recap

- Measuring item-item or user-user similarity requires knowledge of different similarity metrics, and when to use them.

- There are a number of popular techniques, but the most popular are Pearson/Cosine Similarity (continuous data) and Jaccard Similarity (binary data). However, there are A LOT of techniques we did not look at here.

Additional Note:

Often these methods don't scale very well, so you may have to be creative in implementing for large scale problems.