Similarity Notes

Aaron Tuor

September, 2015

1 Similarity Metrics

Many recommendation algorithms employ some form of similarity metric in the generation of ratings predictions. Similarity metrics are often associated with some form of distance measure.

Definition 1.0.1. Let δ be a function $\delta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$. Then δ is a ***distance measure*** if it satisfies the following four properties.

(d1) $\delta(\mathbf{x}, \mathbf{y}) \geq 0$ (no negative distances).

(d2) $\delta(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$ (different vectors can't be in the same position).

(d3) $\delta(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x}, \mathbf{y})$ (distance is symmetric).

(d4) $\delta(\mathbf{x}, \mathbf{y}) \leq \delta(\mathbf{z}, \mathbf{x}) + \delta(\mathbf{z}, \mathbf{y})$ (triangle inequality).

1.1 Manhattan Distance

The Manhattan Distance, δ_M, between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is the sum of the magnitudes of the differences in each dimension, i.e.,

$$\delta_M(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|.$$

1.2 Euclidean Distance

The most commonly found distance measure for real valued vectors is the Euclidean distance. For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the Euclidean distance, δ_E, between $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$\delta_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

1.3 L_r norm (Minkowski distance)

Euclidean distance, and Manhattan distance are examples of L_r norms. For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the L_r norm, δ_r, between $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$\delta_r(\mathbf{x}, \mathbf{y}) = \Big(\sum_{i=1}^{n} |x_i - y_i|^r\Big)^{\frac{1}{r}}.$$

By the above definition, Manhattan distance is the L_1 norm and Euclidean distance is the L_2 norm. Another related distance measure is the L_∞ norm, δ_∞, defined as:

$$\delta_\infty(\mathbf{x}, \mathbf{y}) = \lim_{r \to \infty} \Big(\sum_{i=1}^{n} |x_i - y_i|^r\Big)^{\frac{1}{r}}.$$

As r gets larger, only the dimension with the largest difference matters, so the equation above turns out to be equivalent to:

$$\max(|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|).$$

1.4 Mahalanobis distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})\sigma^{-1}(\mathbf{x} - \mathbf{y})^T}$$

1.5 Jaccard Coefficient

The Jaccard Coefficient, δ_J, between two sets, U and V, is defined as:

$$\frac{\Big|U \cap V\Big|}{\Big|U \cup V\Big|}$$

.

In the context of recommendation systems this type of distance measure makes sense when the information about user preferences is binary in nature, such as a rating system of like or dislike, or implicit preference data such as user purchases and items viewed for a significant amount of time.

1.6 Tanimoto Coefficient (Extended Jaccard Coefficient)

Measures the similarity of two sets by comparing the size of the overlap against the size of the two sets. In the case of binary attributes reduces to the Jaccard Coefficient.

$$T(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

1.7 Log-likelihood

1.8 Cosine Distance

The Cosine distance between two vectors, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, is defined as the size of the angle between them, i.e.,

$$\arccos \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} = \arccos \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

1.9 Pearson-correlation coefficient (PCC)

Unlike the preceding measures the Pearson correlation coefficient is not a distance measure. It is a measure of the linear correlation (dependence) between two random variables X and Y, giving a value between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. The Pearson correlation coefficient of variables X and Y is defined as:

$$\frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Let x, and y be two users and vectors, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, consist of ratings of the n items which both x and y have rated. Consider $\mathbf{x}$ and $\mathbf{y}$ as samples from the distributions of ratings of x and y. Now we can substitute the sample covariance and standard deviations based on $\mathbf{x}$ and $\mathbf{y}$ into the equation above. We end up with the sample Pearson correlation coefficient, $r_{\mathbf{x},\mathbf{y}}$, below:

$$r_{\mathbf{x},\mathbf{y}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_Y)^2}} = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_X)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_Y)^2}}$$

$$\text{Let } \mathbf{x}' = \mathbf{x} - \begin{bmatrix} \mu_X \\ \vdots \\ \mu_X \end{bmatrix} \text{ and } \mathbf{y}' = \mathbf{y} - \begin{bmatrix} \mu_Y \\ \vdots \\ \mu_Y \end{bmatrix}.$$

$$\text{Then, } r_{\mathbf{x},\mathbf{y}} = \frac{\mathbf{x}' \cdot \mathbf{y}'}{\|\mathbf{x}'\| \|\mathbf{y}'\|}.$$

This is just the cosign of the angle between our normalized ratings vectors. Users may have different conventions for employing a given ratings scale. For instance for a 1-5 discrete scale one user may only rate items they like, rating these 5, whereas another may only employ values 1,3, and 5 to signify items they like, are ambivalent about, and dislike respectively, while a third user may employ the full range of the rating scale. The centered ratings obtained by subtracting the mean rating of the user from each rating in the vector counteracts the effects heterogeneous usage conventions.

The correlation coefficient gives a similarity score between -1 and 1, where ratings close to zero signify little correlation, ratings closer to 1 signify that the two users rate items similarly, and items closer to -1 signify that the two users rate items dissimilarly.

The PCC as well as cosine similarity don't take into account the length of the vectors being compared and so by themselves may provide counterintuitive results as regards similarity of users. For instance consider two users who have rated a lot of items but only a few in common and the ratings of these common items for the two users happen to be very close. This would give a PCC close to 1 when these users have obviously different rating behaviors.

To counteract these types of effects a weight may employed when constructing a metric to discount vectors with few entries. For users with n items commonly rated, a general formula for the weight may take the form $\frac{n}{n+\lambda}$ where λ is a parameter which determines how much effect the shortness of ratings vectors detracts from the overall similarity.

If the ratings are normalized by the mean rating of the commonly reviewed items of a user instead of the mean rating of a user in general then the PCC is undefined when users overlap with a single item (as the mean rating equals the value of the rating for the single item) or when either user has the same rating for all overlapping items (as the mean rating equals the value of every item rated).

1.10 Spearman rank correlation

The Spearman correlation performs the same calculation as PCC, except that ratings are first mapped to ranks in the following manner. The highest rating gets rank 1, the next highest rank 2 and so on. If items have the same rating then each receives an average rank. So, for instance, if 3 is the second highest rating and there are 4 items with a rating of 3 their ranking is computed as follows:

$$\frac{2+3+4+5}{4}$$