Similarity Computations

The goal is to compare the movie ratings produced by two people, let's call them X and Y, to find out how similar the two people are. We will compare two columns of the movie ratings matrix.

Note: the same methods work for comparing any vectors of paired values. You can compare the daily high temperatures between New York and Chicago, for example. Or you can compare two movies (not two raters), in other words two *rows* of the movie rating matrix, to see whether the pattern of how-people-rated-them is similar.

In this section we will look at two different methods that have two different uses: Euclidean distance and Pearson correlation. The temperature in New York is generally warmer than Chicago. But they both go up in the summer and down in the winter. If you want to find how different they are absolutely, then something like the Euclidean distance is useful. If you want to find out whether they tend to go up and down together, Pearson is the right computation.

Let's imagine that X and Y have each rated six movies like this, (we also added person Z):

Movie	X	Y	Z
M1	5	3	2
M2	3	2	3
M3	1	1	4
M4	5	3	2
M5	1	1	4
M6	3	2	3

Euclidean Distance

In this case we treat a person's ratings is a vector of numbers. First compute the distance between each person's ratings vectors of the same movies:

$$\text{dist} \quad = \quad \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

In the example case, that is:

$$\text{dist} \quad = \quad \sqrt{(5-3)^2 + (3-2)^2 + (1-1)^2 + (5-3)^2 + (1-1)^2 + (3-2)^2}$$

$$
\begin{aligned}
&= \sqrt{10} \\
&\approx 3.16
\end{aligned}
$$

Clearly if the two sets of ratings are the same then dist is zero. Every time the two people disagree on a movie a positive number is added into the sum and the distance is larger. The distance is then converted to a *similarity* number. Zero distance means X and Y were in perfect agreement, we want to call this similarity = 1.0. Bigger distances should produce smaller similarity numbers.

$$
\begin{aligned}
d &= \frac{1}{1+\text{dist}} \\
&= \frac{1}{1+\sqrt{10}} \\
&\approx 0.24
\end{aligned}
$$

Pearson Correlation r

On a straight numerical comparison X and Y are quite different. They agreed on only two movies (the worst ones) in the set. X gave five stars to 2 movies, while Y never gave higher than 3 stars to any movie.

However the only difference is the grading scale. They both agree on which are the best movies, the middle movies, and the worst movies. The way to capture this similarity is to get away from the raw numbers and convert the ratings to *standard score*, which means to rewrite each rating as standard deviations above or below that person's mean. So for person X we compute the mean $\bar{x}$ (in the usual way) and standard deviation: $s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$.

Movie	X	Y	$(x_i - \bar{x})/s_x$	$(y_i - \bar{y})/s_y$
M1	5	3	1.12	1.12
M2	3	2	0.0	0.0
M3	1	1	−1.12	−1.12
M4	5	3	1.12	1.12
M5	1	1	−1.12	−1.12
M6	3	2	0.0	0.0
mean	$\bar{x} = 3$	$\bar{y} = 2$		
stdev	$s_x = \sqrt{3.2}$	$s_y = \sqrt{0.8}$		

Now the two sets of ratings have been put on an equal scale, and we can see that in this case they are identical.

This is a version of the formula for r. It converts the individual ratings to standard scores then combines the two sets of standard scores into one overall similarity measure:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

The sum captures how much X and Y *agree*. Each term in the sum has an X rating multiplied by the corresponding Y rating (where both have been converted to standard scores).

- The sum increases when X and Y agree. If they both rate the movie higher than average (two positive scores), or both lower than average (two negative scores), the product is positive.
- The sum decreases when X and Y disagree, when one rating is positive and the other negative.

Ultimately $-1.0 \leq r \leq 1.0$, where 1 means perfect agreement, -1 means perfect disagreement, and 0 means there is no relationship (they sometimes agree and sometimes disagree, in equal amounts). In the above example:

$$\begin{aligned} r &\approx \frac{1}{5}(1.12^2 + 0^2 + -1.12^2 + 1.12^2 + -1.12^2 + 0^2) \\ &\approx \frac{1}{5}5 \\ &\approx 1.0 \end{aligned}$$

Now compare person X to person Z. It is clear that Z gives lower marks to the movies that X likes, and vice versa. What does the Pearson r show?