What is the Jaccard Similarity Index and can it be easily explained?

Authored by

stats writer

April 23, 2024

The Jaccard Similarity Index is a statistical measure used to determine the degree of similarity between two sets of data. It is commonly used in the field of data analysis and machine learning. The index calculates the similarity between two sets by dividing the size of the intersection of the sets by the size of the union of the sets. This results in a value between 0 and 1, where 0 indicates no similarity and 1 indicates complete similarity. The Jaccard Similarity Index can be easily explained as a measure of overlap between two sets, making it a useful tool for comparing data sets and identifying patterns.

A Simple Explanation of the Jaccard Similarity Index

The Jaccard Similarity Index is a measure of the similarity between two sets of data.

Developed by , the index ranges from 0 to 1. The closer to 1, the more similar the two sets of data.

The Jaccard similarity index is calculated as:

Jaccard Similarity = (number of observations in both sets) / (number in either set)

Or, written in notation form:

$J(A, B) = |A \cap B| / |A \cup B|$

If two datasets share the exact same members, their Jaccard Similarity Index will be 1. Conversely, if they have no members in common then their similarity will be 0.

The following examples show how to calculate the Jaccard Similarity Index for a few different datasets.

Example 1: Jaccard Similarity

Suppose we have the following two sets of data:

A =
B =

To calculate the Jaccard Similarity between them, we first find the total number of observations in both sets, then divide by the total number of observations in either set:

Number of observations in both: {0, 2, 5, 9} = 4Number of observations in either: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} = 10Jaccard Similarity: 4 / 10 = 0.4

The Jaccard Similarity Index turns out to be 0.4.

Example 2: Jaccard Similarity Continued

Suppose we have the following two sets of data:

C =
D =

Number of observations in both: {} = 0Number of observations in either: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10} = 11Jaccard Similarity: 0 / 11 = 0

The Jaccard Similarity Index turns out to be 0. This indicates that the two datasets share no common members.

Example 3: Jaccard Similarity for Characters

Note that we can also use the Jaccard Similarity index for datasets that contain characters as opposed to numbers.

For example, suppose we have the following two sets of data:

E =
F =

To calculate the Jaccard Similarity between them, we first find the total number of observations in both sets, then divide by the total number of observations in either set:

Number of observations in both: {'monkey'} = 1Number

of observations in either: {'cat', 'dog', hippo', 'monkey', 'rhino', 'ostrich', 'salmon'} = 7Jaccard Similarity: 1 / 7= 0.142857

The Jaccard Similarity Index turns out to be 0.142857. Since this number is fairly low, it indicates that the two sets are quite dissimilar.

The Jaccard Distance

The Jaccard distance measures the dissimilarity between two datasets and is calculated as:

Jaccard distance = 1 - Jaccard Similarity

This measure gives us an idea of the difference between two datasets or the *difference* between them.

For example, if two datasets have a Jaccard Similarity of 80% then they would have a Jaccard distance of 1 - 0.8 = 0.2 or 20%.

The following tutorials explain how to calculate Jaccard Similarity using different statistical software: