

JACCARD COEFFICIENT CALCULATIONS

Maria Ingold
12693772
Unit 5
Machine Learning
University of Essex Online
25 November 2024

CONTENTS

JACCARD.....	3
Jaccard Coefficient.....	3
Intersection and Union.....	4
Jaccard Distance.....	4
(Jack, Mary): $J = 0.67$, $d_J = 0.33$	5
Jaccard Coefficient = 0.67	5
Jaccard Distance = 0.33	5
(Jack, Jim): $J = 0.33$, $d_J = 0.67$	6
Jaccard Coefficient = 0.33	6
Jaccard Distance = 0.67	6
(Jim, Mary) = $J = 0.25$, $d_J = 0.75$	6
Jaccard Coefficient = 0.25	6
Jaccard Distance = 0.75	6

JACCARD

The table shows the pathological test results for three individuals.

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	A
Mary	F	Y	N	P	A	P	N
Jim	M	Y	P	N	N	N	A
	Symmetric	Asymmetric	Asymmetric	Asymmetric	Asymmetric	Asymmetric	Asymmetric

Best guess:

- **Gender:** M (Male), F (Female)
- **Fever:** Y (Yes), N (No)
- **Cough:** N (Negative), P (Positive)
- **Test-1,2,3,4:** N (Negative), P (Positive), A (Absent – no result?)

Symmetry:

- Symmetric: No preference as to mapping to 0 or 1, same weight, like M/F.
- Asymmetric: Not equal importance. Preference to map.

Ignore Symmetric.

Jaccard Coefficient

Jaccard Coefficient measures similarities between two sets.

- Asymmetric data. Outcomes not equal importance.
- Binary data
- Range: 0 to 1
- 0: No common elements between two sets (disjoint)
- 1: Identical sets.

The Jaccard Coefficient is defined as

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

$J(A,B)$ = intersection divided by union.

This is also represented as:

$$J = f_{11} / (f_{01} + f_{10} + f_{11})$$

Intersection and Union

- **Intersection** is common
- **Union** is when at least one of the sets has element.
 - 1 means attribute present
 - 0 means attribute absent

Jaccard Distance

The assignment and lecturecast use the Jaccard Distance formula, even though it says Jaccard Coefficient.

Jaccard Distance is represented as:

$$d_J = f_{01} + f_{10} / (f_{01} + f_{10} + f_{11})$$

$$\text{or } d_J(A, B) = 1 - J(A, B) = (|A \cup B| - |A \cap B|) / |A \cup B|$$

Not all binary. Group:

- **Fever:** N (0), Y (1)
- **Cough:** N (0), P (1)
- **Test-1,2,3,4:** N or A (0), P (1)

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	1	0	1	0	0	0
Mary	1	0	1	0	1	0
Jim	1	1	0	0	0	0

Calculate Jaccard Coefficient for the following pairs:

(Jack, Mary): J = 0.67, d_J = 0.33

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	1	0	1	0	0	0
Mary	1	0	1	0	1	0

- Intersection = $|A \cap B| = 2$
- Union = $|A \cup B| = 3$

Jaccard Coefficient = 0.67

$$J(A,B) = |A \cap B| / |A \cup B| = 2/3 = 0.67$$

Or

$$J = f_{11} / (f_{01} + f_{10} + f_{11}) = 2 / (1 + 0 + 2) = 2/3 = 0.67$$

Jaccard Distance = 0.33

$$d_J = f_{01} + f_{10} / (f_{01} + f_{10} + f_{11}) = (1 + 0) / (1 + 0 + 2) = 1/3 = 0.33$$

or

$$d_J(A, B) = 1 - J(A,B) = 1 - 0.67 = 0.33$$

or

$$d_J(A, B) = (|A \cup B| - |A \cap B|) / |A \cup B| = (3 - 2) / 3 = 1/3 = 0.33$$

(Jack, Jim): $J = 0.33$, $d_J = 0.67$

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	1	0	1	0	0	0
Jim	1	1	0	0	0	0

- Intersection = $|A \cap B| = 1$
- Union = $|A \cup B| = 3$

Jaccard Coefficient = 0.33

$$J = |A \cap B| / |A \cup B| = f_{11} / (f_{01} + f_{10} + f_{11}) = 1 / 3 = 0.33$$

Jaccard Distance = 0.67

$$d_J = 1 - J(A,B) = f_{01} + f_{10} / (f_{01} + f_{10} + f_{11}) = 2 / 3 = 0.67$$

(Jim, Mary) = $J = 0.25$, $d_J = 0.75$

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jim	1	1	0	0	0	0
Mary	1	0	1	0	1	0

- Intersection = $|A \cap B| = 1$
- Union = $|A \cup B| = 4$

Jaccard Coefficient = 0.25

$$J = |A \cap B| / |A \cup B| = f_{11} / (f_{01} + f_{10} + f_{11}) = 1 / 4 = 0.25$$

Jaccard Distance = 0.75

$$d_J = 1 - J(A,B) = f_{01} + f_{10} / (f_{01} + f_{10} + f_{11}) = 3 / 4 = 0.75$$