Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

Enero, 2023

Similitud de Dice (1)

- Esta medida de similitud también se le conoce como el Coeficiente de Sørensen. Es una métrica estadística que se utiliza para comparar la similitud de conjuntos de datos.
- La fórmula original de Sørensen está destinada a ser aplicada a datos con presencia/ausencia de valores y se define como sigue:

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|} = D_S = \frac{2C}{A + B}$$

- Donde: A y B son el número de elementos muestra de un conjunto de datos respectivamente y C es el número de especies compartidas por los dos elementos muestra.
- D_S es el coeficiente de similitud y varía entre 0 y 1. Esta expresión se extiende fácilmente a la abundancia en lugar de la presencia/ausencia de especies.
- La versión cuantitativa del índice de Sørensen también se conoce como el índice binario de Czekanowski.

Similitud de Dice (2)

- El conjunto de operaciones se pueden expresar en términos de operaciones vectoriales sobre vectores binarios *A* y *B,* lo cual proporciona el mismo resultado en vectores binarios y también da una similitud más general sobre los vectores en términos generales.

$$D_S = \frac{2|A \cdot B|}{|A|^2 + |B|^2} = s(x, y) = \frac{2\sum_{i=1}^{p} x_i y_i}{\sum_{i=1}^{p} (x_i)^2 + (y_i)^2}$$

- Para el caso de conjuntos por ejemplo, *X* e *Y* de palabras clave utilizadas en la recuperación de la información, el coeficiente puede ser definido como dos veces la información compartida (intersección) sobre la suma de las cardinalidades.
- Entonces el coeficiente se puede utilizar también como medida de similitud entre cadenas. Dadas dos secuencias *x* e *y*, se puede calcular el coeficiente como sigue:

$$D_S = \frac{2n_t}{n_x + n_y}$$

Similitud de Dice (3)

$$D_S = \frac{2n_t}{n_x + n_y}$$

- Donde n_t es el número de dígrafos (n-gramas) (formado por dos caracteres consecutivos) en común a las dos cadenas, n_x es el número de dígrafos en la cadena *x*, y n_y es el número de dígrafos en la cadena *y*. Por ejemplo, para calcular la similitud entre:

night **nacht**

- Se procede primeramente con el cálculo de los dígrafos de cada palabra:

ni , **ig** , **gh** , **ht**

na , **ac** , **ch** , **ht**

- Cada conjunto tiene 4 elementos y su intersección se reduce a un elemento n_t. Con la fórmula dada, se obtiene la similitud.

$$D_S = \frac{2 \cdot 1}{(4+4)} = 0.25$$

Similitud de Dice (4)

- De manera análoga al índice de Jaccard, la medida de distancia se obtiene al restarle a 1 el valor de la similitud.
- Dado un valor del coeficiente de Dice (D), es posible calcular el índice de Jaccard (J) y viceversa, mediante las siguientes ecuaciones.

$$D = \frac{2J}{(1+J)}; \qquad J = \frac{D}{(2-D)}$$

- En cierta forma, se puede observar que el coeficiente de Dice le da mayor peso a los elementos comunes entre ambos conjuntos, lo que se puede apreciar al comparar los resultados de calcular la similitud por ambos métodos.

Similitud de Dice (5)

- La versión cuantitativa de la similitud de Dice o del coeficiente de Sørensen, se conoce como el porcentaje de similitud; en donde se considera como un índice que está basado en datos de abundancia (y no de presencia/ausencia).

$$I_{S\_cuant} = \frac{2p_n}{a_n + b_n}$$

T.Bil	G.Spi	C.Tri	C.Sca	Medidas	Tipo
1	21	11	16	49	A
1	8	3	0	12	B
1	8	3	0	12	p_n

- En este caso p_n representa la sumatoria de la abundancia más baja de cada una de las especies compartidas entre ambos sitios. a_n y b_n es el número total de individuos en el sitio *A* y *B* respectivamente.

$$\boldsymbol{I_{S\_cuant} = \frac{2 \cdot 12}{49 + 12} = 0.3934426}$$

Similitud de Dice (6)

- Conclusiones
  - El coeficiente de Dice o Índice de Sørensen se utiliza como medida de similitud para analizar datos del dominio de la ecología, geografía, biología, medicina entre otros. Este índice expresa el grado en que dos muestras son semejantes por las especies presentes en ellas.
  - En el área de recuperación de información encuentra uso en la lexicografía infográfica, donde interviene directamente en la medición de la puntuación de asociación léxica de dos palabras, así como el análisis de términos, considerando a éstas en un espacio vectorial.
  - La razón de este uso es más empírica que teórica, se puede justificar teóricamente como la intersección de dos conjuntos difusos.
  - En comparación con la distancia euclidiana, esta métrica se ajusta bien para conjuntos de datos heterogéneos y da menos peso a los casos desviados.

Similitud de Jaccard (1)

- A diferencia de otras métricas, el índice o coeficiente de similitud de Jaccard opera sobre conjuntos, por lo que comúnmente se utiliza para comparar sentencias o párrafos completos como un conjunto de palabras.
- Sin embargo también puede ser utilizado para comparar palabras considerándolas como conjuntos de letras o caracteres. Cualquiera que sea el nivel de *tokenización* sobre el que se utilice, es interesante notar que la posición que ocupa el elemento no tiene relevancia y que elementos repetidos son considerados como uno solo dentro del conjunto.
- El coeficiente de similitud de Jaccard mide la similitud entre dos conjuntos de muestras. Se define como la relación entre el tamaño de la intersección de ambos conjuntos y el tamaño de la unión y es una medida de la similitud entre ambos; es decir, es la división entre el número de elementos en común que tienen los dos conjuntos sobre el número de elementos únicos que tiene la unión de ambos conjuntos.

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

Similitud de Jaccard (2)

- Por su parte, la *distancia de Jaccard* es el resultado de restarle a 1 el valor de la similitud.

$$d_J(A,B) = 1 - sim_j(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

- El índice de similitud de Jaccard, compara individuos (instancias) de dos conjuntos para ver individuos compartidos y distintos.
- Es una medida de similitud para los dos conjuntos de datos, con un rango de [0 a 1]. Cuanto mayor sea el valor que se acerca a 1, más similares serán las dos poblaciones.
- El índice Jaccard es una estadística que sirve para comparar y medir qué tan similares son dos conjuntos diferentes entre sí. Aunque es fácil de interpretar, es susceptible de tamaños de muestra pequeños.

Similitud de Jaccard (3)

- Puede dar resultados erróneos, especialmente con muestras más pequeñas o conjuntos de datos con observaciones faltantes.
- Para medir el grado de similitud entre un documento y/o una consulta debemos llevar o extender esta ecuación, que está expresada en función de conjuntos de términos, a una expresión en función de vectores de términos.
- Esta forma extendida del coeficiente de Jaccard también se conoce con el nombre de coeficiente de Tanimoto.

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

Similitud de Jaccard (4)

- Expresada en términos vectoriales: $sim_J(d_j, q) = \dfrac{\overrightarrow{d_j} \cdot \vec{q}}{\left|\overrightarrow{d_j}\right| + |\vec{q}| - \overrightarrow{d_j} \cdot \vec{q}}$

- Sustituyendo los términos en los conjuntos *X* e *Y*, se tiene que:

$$sim_J(x, y) = \frac{\sum_{i=1}^{p} x_i y_i}{\sum_{i=1}^{p} (x_i)^2 + \sum_{i=1}^{p} (y_i)^2 - \sum_{i=1}^{p} x_i y_i}$$

- En resumen, en ciencias de la computación se utiliza para medir distancia entre vectores definidos sobre un espacio vectorial booleano (componentes del vector 0 o 1).

$$sim_j(A, B) = \frac{|A \wedge B|}{|A \vee B|}$$

- Donde:
  - ∧ y ∨ son respectivamente las operaciones × (AND) y + (OR) de la lógica booleana y $|A| = \sum a_i$

Similitud de Jaccard (5)

- Hay dos alternativas principales para encontrar la métrica de similitud en textos:
  - El enfoque *basado en caracteres* se ocupa de los caracteres individuales presentes en el documento con la secuencia adecuada.
  - El enfoque *basado en términos* se ocupa de la palabra completa.
    - Las palabras a menudo se simplifican o *lematizan* antes de realizar la prueba según el proceso de limpieza de datos inicial utilizado para el propósito específico.
    - Por tanto, métrica de similitud de Jaccard es adecuada para este enfoque.
    - La similitud de Jaccard puede aplicarse a vectores de documentos que ya están en formato de bolsa de palabras.
    - La definición de la distancia es uno menos el tamaño de la intersección sobre el tamaño de la unión de los vectores (en este caso la distancia puede ser alta).
    - La similitud de Jaccard permite aceptar listas pares (es decir, documentos) como entradas. Cuando son los mismos vectores, el valor devuelto es 0; lo que significa que la distancia es 0 y los dos documentos son idénticos.

Similitud de Jaccard (6)

- En otras áreas como las ciencias médico-biológicas, la similitud o índice de Jaccard se utiliza para medir la similitud entre muestras o poblaciones, utilizando la siguiente fórmula:

$$I_j = \frac{c}{a + b - c}$$

- Donde:
  - a = Número de especies o individuos presentes en la muestra A.
  - b = Número de especies o individuos presentes en la muestra B.
  - c = Número de especies presentes en ambas muestras.
- En este sentido, el valor de 0, indica que las muestras no presentan especies en común y tiende a 1 a medida que aumenta el número de especies compartidas.

Similitud de Jaccard (7)

- Ejemplos:
  - Con base en la fórmula: $sim_j(A,B) = \frac{|A \cap B|}{|A \cup B|}$
  - Si dos conjuntos de datos comparten exactamente los mismos miembros, su índice de similitud Jaccard será 1. Por el contrario, si no tienen miembros en común, su similitud será 0.
  - Ejemplos de cómo calcular el índice de similitud de Jaccard para datasets diferentes.
    - Supongamos que se tienen los siguientes dos conjuntos de datos:
    - **A = [0, 1, 2, 5, 6, 8, 9]**; **B = [0, 2, 3, 4, 5, 7, 9]**
    - Para calcular la similitud de Jaccard entre ellos, primero encontramos el número total de observaciones en ambos conjuntos, luego dividimos por el número total de observaciones en cualquiera de los conjuntos:
    - **Número de observaciones en ambos:** {0, 2, 5, 9} = 4
    - **Número total de observaciones:** {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} = 10
    - $sim_j(A,B) = \frac{4}{10} = 0.4$

Similitud de Jaccard (8)

- Ejemplos:
  - Supongamos que se tienen los siguientes dos conjuntos de datos:
  - **C = [0, 1, 2, 3, 4, 5]**; **D = [6, 7, 8, 9, 10]**
  - Para calcular la similitud de Jaccard entre ellos, primero encontramos el número total de observaciones en ambos conjuntos, luego dividimos por el número total de observaciones en cualquiera de los conjuntos:
  - **Número de observaciones en ambos:** {} = 0
  - **Número total de observaciones:** {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10} = 11
  - $sim_j(A, B) = \frac{0}{11} = 0$
  - El índice de similitud de Jaccard resulta ser 0. Esto significa que los dos conjuntos de datos no comparten miembros comunes.

Similitud de Jaccard (9)

- Ejemplos:
  - También podemos usar el índice de similitud de Jaccard para conjuntos de datos que contienen caracteres en lugar de números. Por ejemplo, supongamos que tenemos los siguientes conjuntos:
  - **E = ['gato', 'perro', 'hipopótamo', 'mono']**
  - **F = ['mono', 'rinoceronte', 'avestruz', 'salmón']**
  - Para calcular la similitud de Jaccard entre ellos, primero encontramos el número total de observaciones en ambos conjuntos, luego dividimos por el número total de observaciones en cualquiera de los conjuntos:
    - **Número de observaciones en ambos:** {'mono'} = 1
    - **Número total de observaciones:** {'gato', 'perro', hipopótamo', 'mono', 'rinoceronte', 'avestruz', 'salmón'} = 7
    - $sim_j(A, B) = \frac{1}{7} = 0.142857$
  - El índice de similitud de Jaccard resulta ser 0.142857 . Dado que este número es bastante bajo, indica que los dos conjuntos son bastante diferentes.

Similitud de Jaccard (10)

- Ejemplos:
  - La distancia de Jaccard mide la diferencia entre dos conjuntos de datos y se calcula como: $d_J(A,B) = 1 - sim_j(A,B)$
  - Esta medida nos proporciona una idea de la diferencia entre dos conjuntos de datos o la diferencia entre ellos.
  - Por ejemplo, si dos conjuntos de datos tienen una similitud de Jaccard del 80%, entonces tendrían una distancia de Jaccard de 1 – 0.8 = 0.2 o equivalente al 20%.