Ciencia de Datos e Inteligencia de Negocios

Índices de Similitud o Similaridad

Los índices de similitud o similaridad, son medidas que se calculan para determinar qué tan similar es un vector o conjunto de datos con otro. Los índices de similaridad o similitud dependen mucho del tipo de datos que se están modelando.

1. Datos binarios

   Para datos binarios, se debe de considerar que las medidas de similitud propuestas hasta ahora, parte de los conteos que se muestran en la siguiente tabla:

		Individuo i: Presente (1)	Individuo i: Ausente (0)	
Individuo i	Presente (1)	a	b	a+b
	Ausente (0)	c	d	c+d
		a+c	b+d	a+b+c+d

   Esta matriz de generalmente es llamada matriz de confusión y esta implementada en la paquetería de python sklearn.

   from sklearn.metrics import confusion_matrix

   y_true = [0, 1, 0, 1, 0, 0]

   y_pred = [1, 1, 1, 0, 0, 0]

   confusion_matrix(y_true, y_pred)

   a. Emparejamiento simple
      El emparejamiento esta implementado en Python en la paquetería sklearn.
      Ejemplo:

```
>>> import numpy as np
>>> from sklearn.metrics import accuracy_score
>>> y_pred = [0, 1, 0, 1, 0, 0]
>>> y_true = [1, 1, 1, 0, 0, 0]
```

```
>>> accuracy_score(y_true, y_pred)
>>> accuracy_score(y_true, y_pred, normalize=False)
```

b. Jaccard

El índice de Jaccard se encuentra en la paquetería sklearn en Python.

Ejemplo 1: Sklearn

```
>>> import numpy as np
>>> from sklearn.metrics import jaccard_similarity_score
>>> y_pred = [0, 1, 0, 1, 0, 0]
>>> y_true = [1, 1, 1, 0, 0, 0]
>>> jaccard_similarity_score(y_true, y_pred)
0.5
>>> jaccard_similarity_score(y_true, y_pred, normalize=False)
2
```

Ejemplo 2: Scipy

from scipy.spatial.distance import jaccard

jaccard(u, v)

c. Dice

Ejemplo 1: Scipy

from scipy.spatial.distance import dice

dice(u, v)

d. Kulczynski

Ejemplo 1: Scipy

from scipy.spatial.distance import kulsinski

kulsinski (u, v)

e.

2. Datos Multiestado

from sklearn.metrics import confusion_matrix
y_true = [2, 0, 2, 2, 0, 1]
y_pred = [0, 0, 2, 2, 0, 2]
confusion_matrix(y_true, y_pred)
array([[2, 0, 0],
[0, 0, 1],
[1, 0, 2]])

a. Emparejamiento simple

```
>>> import numpy as np
>>> from sklearn.metrics import accuracy_score
>>> y_pred = [0, 2, 1, 3]
>>> y_true = [0, 1, 2, 3]
>>> accuracy_score(y_true, y_pred)
>>> accuracy_score(y_true, y_pred, normalize=False)
```

b. Jaccard

```
>>> import numpy as np
>>> from sklearn.metrics import jaccard_similarity_score
>>> y_pred = [0, 2, 1, 3]
>>> y_true = [0, 1, 2, 3]
>>> jaccard_similarity_score(y_true, y_pred)
>>> jaccard_similarity_score(y_true, y_pred, normalize=False)
```

c.

3. Datos Cuantitativos
   a. Distancia Euclideana
      Ejemplo 1: Numpy

      import numpy as np
      a = np.array([1,2,3])
      b = np.array([2,3,4])
      dist = numpy.linalg.norm(a-b)

      Ejemplo 2: Sklearn

>>> from sklearn.metrics.pairwise import euclidean_distances
>>> X = [[0, 1], [1, 1]]
>>> # distance between rows of X
>>> euclidean_distances(X, X)
>>> # get distance to origin
>>> euclidean_distances(X, [[0, 0]])

Ejemplo 3: Scipy
from scipy.spatial.distance import euclidean
euclidean(u,v)

b. Coeficiente de Pearson
Ejemplo 1: Scipy
#Find a correlation
>>> from scipy.stats.stats import pearsonr
# The dependent variable
>>> x = [1, -2, 2, 3, 1]
# The independent variable
>>> y = [7.5, -3.5, 14.5, 19, 6.6]
#First value is the r-value, 2nd is the p-value
>>> pearsonr(x,y)
(0.98139984935586166, 0.0030366388199721478)

c. Bray-Curtis

Ejemplo 1: Scipy
from scipy.spatial.distance import braycurtis
braycurtis(u, v)

d. Minkowski

Ejemplo 1: Scipy
from scipy.spatial.distance import minkowski
minkowski(u, v)

e.

f.

4.