Punto A

Domanda

For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

x = 0101010001

y = 0100011000

Risposta

In [ ]:
```
import numpy as np
```

In [ ]:
```
x = np.array([0, 1, 0, 1, 0, 1, 0, 0, 0, 1])
y = np.array([0, 1, 0, 0, 0, 1, 1, 0, 0, 0])
```

In [ ]:
```
# Calcolo della distanza di Hamming
hamming_distance = np.sum(np.logical_xor(x,y))
print("Distanza di Hamming tra x e y:", hamming_distance)
```

```
Distanza di Hamming tra x e y: 3
```

In [ ]:
```
# Calcolo della similarità di Jaccard
jaccard_similarity = np.sum(np.logical_and(x, y)) / np.sum(np.logical_or(x, y))
print("Similarità di Jaccard tra x e y:", jaccard_similarity)
```

```
Similarità di Jaccard tra x e y: 0.4
```

Punto B

Domanda

Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)

Risposta

La distanza di Hamming è simile al SMC perché entrambi i metodi esaminano l'intero set di dati e cercano quando i dati sono simili o dissimili uno a uno. La distanza di Hamming

fornisce il risultato di quante attributi sono diversi, infatti può esserr equiparato a uno xor, mentre il coefficiente di matching semplice fornisce il risultato del rapporto di quanti attributi sono uguali sull'intero insieme di campioni. Le informazioni risultanti rivelano le stesse informazioni: una rivela quanti sono diversi, l'altra rivela quanti sono uguali. In questo senso, una rivela l'informazione inversa dell'altra.

In [ ]:
```
# Calcolo di SMC
numeratore = np.sum(x == y)
denominatore = numeratore + np.sum(x!=y)
smc_similarity = numeratore / denominatore
print("Simple Matching Coefficient tra x e y:", smc_similarity)
```

```
Simple Matching Coefficient tra x e y: 0.7
```

In [ ]:
```
# Calcolo della similarità del coseno
norm_x = np.linalg.norm(x)
norm_y = np.linalg.norm(y)
cosine_similarity = np.dot(x, y) / (norm_x * norm_y)

print("Similarità del coseno tra x e y:", cosine_similarity)
```

```
Similarità del coseno tra x e y: 0.5773502691896258
```

In un contesto binario, le misure di Jaccard e della similarità del coseno possono essere considerate più simili tra loro rispetto alla misura di Hamming. Entrambe considerano i bit uguali diversi da zero, ma con interpretazioni leggermente diverse: Jaccard conta i bit uguali in relazione al totale dei bit diversi da zero, mentre il coseno considera i bit uguali in relazione alla norma dei vettori. Ad esempio, Jaccard utilizza il coefficiente M_{11} nel numeratore, che rappresenta i bit uguali tra i due vettori su tutti i bit diversi da zero, mentre il coseno calcola il prodotto scalare dei due vettori, fornendo 1 se ci sono bit uguali e 0 se ci sono bit diversi. Hamming, invece, si limita a fare lo xor dei due vettori, senza metterli in relazione con gli altri elementi dei due vettori. Inoltre, la metrica del coseno di similitudine può essere estesa in modo tale da produrre il coefficiente di Jaccard nel caso di attributi binari. Si ottiene così il coefficiente di Tanimoto $T(A, B)$, rappresentato da:

$$T(A, B) = \frac{A \cdot B}{\|(A)^2\| + \|(B)^2\| - A \cdot B}$$

Punto C

Domanda

Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

Risposta

Jaccard è la misura più appropriata, perchè valuta la proporzione di geni che sono comuni a entrambi gli organismi rispetto al totale dei geni distinti presenti nei due organismi. Infatti, calcola il rapporto tra il numero di geni condivisi, M_{11} e il numero totale di geni presenti in almeno uno dei due organismi, $M_{01} + M_{10} + M_{11}$.

Punto D

Domanda

If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.)

Risposta

Tra tutte, la distanza di Hamming è ideale in questo contesto perché tiene conto delle differenze dirette nei geni tra i due individui, senza considerare il conteggio dei geni condivisi. Poiché gli esseri umani condividono più del 99.9% dei geni, la maggior parte delle differenze genetiche tra di loro si concentrerebbe solo su una piccola frazione dei geni totali.