

Machine Learning

Assignment 1

CLO2

Ida Bagus Dwi Satria Kusuma
1301140297

October 25, 2017

1. Given three vectors $a = (1001101)$, $b = (1101010)$ and $c = (1000011)$
 - (a) **(25 points)** Using Jaccard coefficient, what pair of vectors that have high similarity?

Answer From [10](Introduction to Data Mining), Similarity measures between objects that contain only binary attributes are called similarity coefficient, and typically have values between 0 and 1, which is same similar like vectors given in this case.

Then, let x and y be two objects that consist of n binary attributes. The comparison of two such objects, i.e., two binary vectors, leads to the following four quantities (frequencies):

f_{00} = the number of attribute where x is 0 and y is 0

f_{01} = the number of attribute where x is 0 and y is 1

f_{10} = the number of attribute where x is 1 and y is 0

f_{11} = the number of attribute where x is 1 and y is 1

According to [10], the **Jaccard coefficient**, which is often symbolized by J , is given by the following equation:

$$\begin{aligned}
J &= \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 match}} \\
&= \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \quad (1)
\end{aligned}$$

Then for each pair of vectors, the similarity are :

i. **Pair a and b**

$$a = (1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1)$$

$$b = (1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0)$$

$$f_{00} = 1$$

$$f_{01} = 2$$

$$f_{10} = 2$$

$$f_{11} = 2$$

$$\begin{aligned}
J(a, b) &= \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 match}} \\
&= \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \\
&= \frac{2}{2 + 2 + 2} \\
&= \frac{2}{6} \\
&= \frac{1}{3}
\end{aligned}$$

ii. **Pair a and c**

$$a = (1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1)$$

$$c = (1 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1)$$

$$f_{00} = 2$$

$$f_{01} = 1$$

$$f_{10} = 2$$

$$f_{11} = 2$$

$$\begin{aligned}
J(a, b) &= \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 match}} \\
&= \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \\
&= \frac{2}{1 + 2 + 2} \\
&= \frac{2}{5}
\end{aligned}$$

iii. **Pair b and c**

$$\begin{aligned}
b &= (1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0) \\
c &= (1 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1)
\end{aligned}$$

$$\begin{aligned}
f_{00} &= 2 \\
f_{01} &= 1 \\
f_{10} &= 2 \\
f_{11} &= 2
\end{aligned}$$

$$\begin{aligned}
J(b, c) &= \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 match}} \\
&= \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \\
&= \frac{2}{2 + 1 + 2} \\
&= \frac{2}{5}
\end{aligned}$$

The similarity of pair a and b is $\frac{1}{3}$, pair a and c is $\frac{2}{5}$, and pair b and c is $\frac{2}{5}$. So, pair of vectors that have high similarity are pair a and c and pair b and c .

(b) **(25 points)** Using Simple Matching Coefficient, what pair of vectors that have high similarity?

Answer From [10](Introduction to Data Mining), Similarity measures between objects that contain only binary attributes are called

similarity coefficient, and typically have values between 0 and 1, which is same similar like vectors given in this case.

Then, let x and y be two objects that consist of n binary attributes. The comparison of two such objects, i.e., two binar vectors, leads to the following four quantities (frequencies):

f_{00} = the number of attribute where x is 0 and y is 0
 f_{01} = the number of attribute where x is 0 and y is 1
 f_{10} = the number of attribute where x is 1 and y is 0
 f_{11} = the number of attribute where x is 1 and y is 1

According to [10], the **Simple Matching Coefficient**, which is often symbolized by SMC , is given by the following equation:

$$\begin{aligned}
 SMC &= \frac{\text{number of matching attribute values}}{\text{number of attributes}} \\
 &= \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}
 \end{aligned} \tag{2}$$

Then for each pair of vectors, the similarity are :

i. **Pair a and b**

$a = (1\ 0\ 0\ 1\ 1\ 0\ 1)$
 $b = (1\ 1\ 0\ 1\ 0\ 1\ 0)$

$f_{00} = 1$
 $f_{01} = 2$
 $f_{10} = 2$
 $f_{11} = 2$

$$\begin{aligned}
 SMC(a, b) &= \frac{\text{number of matching attribute values}}{\text{number of attributes}} \\
 &= \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} \\
 &= \frac{2 + 1}{7} \\
 &= \frac{3}{7}
 \end{aligned}$$

ii. **Pair a and c**

$$a = (1\ 0\ 0\ 1\ 1\ 0\ 1)$$

$$c = (1\ 0\ 0\ 0\ 0\ 1\ 1)$$

$$f_{00} = 2$$

$$f_{01} = 1$$

$$f_{10} = 2$$

$$f_{11} = 2$$

$$\begin{aligned} SMC(a, c) &= \frac{\text{number of matching attribute values}}{\text{number of attributes}} \\ &= \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} \\ &= \frac{2 + 2}{7} \\ &= \frac{4}{7} \end{aligned}$$

iii. **Pair b and c**

$$b = (1\ 1\ 0\ 1\ 0\ 1\ 0)$$

$$c = (1\ 0\ 0\ 0\ 0\ 1\ 1)$$

$$f_{00} = 2$$

$$f_{01} = 1$$

$$f_{10} = 2$$

$$f_{11} = 2$$

$$\begin{aligned} SMC(b, c) &= \frac{\text{number of matching attribute values}}{\text{number of attributes}} \\ &= \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} \\ &= \frac{2 + 2}{7} \\ &= \frac{4}{7} \end{aligned}$$

The similarity of pair a and b is $\frac{3}{7}$, pair a and c is $\frac{4}{7}$, and pair b and c is $\frac{4}{7}$. So, pair of vectors that have high similarity are pair a and c and pair b and c .

2. Given three vectors $p = (0.1 \ 0.8 \ -0.2)$, $q = (0.1 \ -0.3 \ 0.6)$ and $r = (-0.1 \ 0.5 \ 0.3)$

(a) **(25 points)** Using Cosine similarity, what pair of vectors that have high similarity?

Answer According to [10], if x and y are two document vectors, the **Cosine Similarity** is given by the following equation:

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|} \quad (3)$$

where \cdot indicates the vector dot product, $x \cdot y = \sum_{k=1}^n x_k y_k$, and $\|x\|$ is the length of vector x , $\|x\| = \sqrt{\sum_{k=1}^n x_k^2} = \sqrt{x \cdot x}$.

Then for each pair of vectors, the similarity are :

i. **Pair** p and q

$$p = (0.1 \ 0.8 \ -0.2)$$

$$q = (0.1 \ -0.3 \ 0.6)$$

$$\begin{aligned} p \cdot q &= 0.1 * 0.1 + 0.8 * -0.3 + -0.2 * 0.6 \\ &= 0.01 + (-0.24) + (-0.12) \\ &= (-0.35) \end{aligned}$$

$$\begin{aligned} \|p\| &= \sqrt{0.1 * 0.1 + 0.8 * 0.8 + (-0.2) * (-0.2)} \\ &= \sqrt{0.01 + 0.64 + 0.04} \\ &= \sqrt{0.69} \\ &= 0.83 \end{aligned}$$

$$\begin{aligned} \|q\| &= \sqrt{0.1 * 0.1 + (-0.3) * (-0.3) + 0.6 * 0.6} \\ &= \sqrt{0.01 + 0.09 + 0.36} \\ &= \sqrt{0.46} \\ &= 0.68 \end{aligned}$$

$$\begin{aligned} \cos(p, q) &= \frac{p \cdot q}{\|p\| \|q\|} \\ &= \frac{-0.35}{0.83 * 0.68} \\ &= -0.6201 \end{aligned}$$

ii. **Pair** p and r

$$\begin{aligned}p &= (0.1 \ 0.8 \ -0.2) \\r &= (-0.1 \ 0.5 \ 0.3)\end{aligned}$$

$$\begin{aligned}p \cdot r &= 0.1 * -0.1 + 0.8 * 0.5 + -0.2 * 0.3 \\&= (-0.01) + 0.40 + (-0.06) \\&= 0.33\end{aligned}$$

$$\begin{aligned}\|p\| &= \sqrt{0.1 * 0.1 + 0.8 * 0.8 + (-0.2) * (-0.2)} \\&= \sqrt{0.01 + 0.64 + 0.04} \\&= \sqrt{0.69} \\&= 0.83\end{aligned}$$

$$\begin{aligned}\|r\| &= \sqrt{(-0.1) * (-0.1) + 0.5 * 0.5 + 0.3 * 0.3} \\&= \sqrt{0.01 + 0.25 + 0.09} \\&= \sqrt{0.35} \\&= 0.59\end{aligned}$$

$$\begin{aligned}\cos(p, r) &= \frac{p \cdot r}{\|p\| \|r\|} \\&= \frac{0.33}{0.83 * 0.59} \\&= 0.6739\end{aligned}$$

iii. **Pair q and r**

$$q = (0.1 \quad -0.3 \quad 0.6)$$

$$r = (-0.1 \quad 0.5 \quad 0.3)$$

$$\begin{aligned} q \cdot r &= 0.1 * (-0.1) + (-0.3) * 0.5 + 0.6 * 0.3 \\ &= (-0.01) + -0.15 + 0.18 \\ &= 0.02 \end{aligned}$$

$$\begin{aligned} \|q\| &= \sqrt{0.1 * 0.1 + (-0.3) * (-0.3) + 0.6 * 0.6} \\ &= \sqrt{0.01 + 0.09 + 0.36} \\ &= \sqrt{0.46} \\ &= 0.68 \end{aligned}$$

$$\begin{aligned} \|r\| &= \sqrt{(-0.1) * (-0.1) + 0.5 * 0.5 + 0.3 * 0.3} \\ &= \sqrt{0.01 + 0.25 + 0.09} \\ &= \sqrt{0.35} \\ &= 0.59 \end{aligned}$$

$$\begin{aligned} \cos(q, r) &= \frac{p \cdot r}{\|p\| \|r\|} \\ &= \frac{0.02}{0.68 * 0.59} \\ &= 0.0499 \end{aligned}$$

The cosine similarity of pair p and r is -0.6201 (which means the vectors is more opposite), pair p and r is 0.6739 , and pair q and r is 0.0499. The biggest cosine similarity value is from pair p and r , which means the highest similarity pair is p and r .

- (b) **(25 points)** Using Euclidean distance, what pair of vectors that are close to each other?

The Euclidean distance can be obtained using this equation :

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^n |x_k - y_k|^2 \right)^{1/2} \quad (4)$$

For three dimensions :

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2} \quad (5)$$

And for each pair, the euclidean distance are :

i. **Pair p and q**

$$p = (0.1 \ 0.8 \ -0.2)$$

$$q = (0.1 \ -0.3 \ 0.6)$$

$$\begin{aligned} d(p, q) &= \sqrt{(0.1 - 0.1)^2 + (0.8 - (-0.3))^2 + (-0.2 - 0.6)^2} \\ &= \sqrt{0^2 + 1.1^2 + (-0.8)^2} \\ &= \sqrt{1.21 + 0.64} \\ &= \sqrt{1.85} \\ &= 1.3601 \end{aligned}$$

ii. **Pair p and r**

$$p = (0.1 \ 0.8 \ -0.2)$$

$$r = (-0.1 \ 0.5 \ 0.3)$$

$$\begin{aligned} d(p, r) &= \sqrt{(0.1 - (-0.1))^2 + (0.8 - 0.5)^2 + (-0.2 - 0.3)^2} \\ &= \sqrt{0.2^2 + 0.3^2 + (-0.5)^2} \\ &= \sqrt{0.04 + 0.09 + 0.25} \\ &= \sqrt{0.38} \\ &= 0.6164 \end{aligned}$$

iii. **Pair q and r**

$$q = (0.1 \ -0.3 \ 0.6)$$

$$r = (-0.1 \ 0.5 \ 0.3)$$

$$\begin{aligned} d(p, q) &= \sqrt{(0.1 - (-0.1))^2 + ((-0.3) - 0.5)^2 + (0.6 - 0.3)^2} \\ &= \sqrt{0.2^2 + (-0.8)^2 + 0.3^2} \\ &= \sqrt{0.04 + 0.64 + 0.09} \\ &= \sqrt{0.77} \\ &= 0.8775 \end{aligned}$$

From the calculation above, the euclidean distance of pair p and q is 1.3601, pair p and r is 0.6164, and pair q and r is 0.8775. The closest pair of vectors is pair p and r , which is 0.6164.