Machine Learning
Assignment 1
CLO2

Ida Bagus Dwi Satria Kusuma
1301140297

September 11, 2017

1. Given three vectors $a = (1001101)$, $b = (1101010)$ and $c = (1000011)$

   (a) **(25 points)** Using Jaccard coeffcient, 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 attributs. The comparison of two such objects, i.e., two binar vectors, leads to the following four quantities (frequencies):

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

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 match}} \tag{1}$$
$$= \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Then for each pair of vectors, the similarity are :

i. **Pair a and b**
$a = (1\ 0\ 0\ 1\ 1\ 0\ 1)$
$b = (1\ 1\ 0\ 1\ 0\ 1\ 0)$

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

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

ii. **Pair a and c**
$a = (1\ 0\ 0\ 1\ 1\ 0\ 1)$
$c = (1\ 0\ 0\ 0\ 0\ 1\ 1)$

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

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

iii. **Pair b and c**
  $b = (1\ 1\ 0\ 1\ 0\ 1\ 0)$
  $c = (1\ 0\ 0\ 0\ 0\ 1\ 1)$

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

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

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 attributs. The comparison of two such objects, i.e., two binar vectors, leads to the following four quantities (frequencies):

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

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

Then for each pair of vectors, the similarity are :

i. **Pair a and b**
$a = (1\ 0\ 0\ 1\ 1\ 0\ 1)$
$b = (1\ 1\ 0\ 1\ 0\ 1\ 0)$

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

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

4

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:

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

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

   Then for each pair of vectors, the similarity are :

i. **Pair p and q**

$$p = (0.1\;0.8\;-0.2)$$
$$q = (0.1\;-0.3\;0.6)$$

$$p \cdot q = 0.1 * 0.1 + 0.8 * -0.3 + -0.2 * 0.6$$
$$= 0.01 + (-0.24) + (-0.12)$$
$$= (-0.35)$$

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

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

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

ii. **Pair p and r**

$$p = (0.1 \, 0.8 - 0.2)$$
$$r = (-0.1 \, 0.5 \, 0.3)$$

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

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

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

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

iii. **Pair q and r**

$$q = (0.1 - 0.30.6$$
$$r = (-0.10.50.3)$$

$$q \cdot r = 0.1 * (-0.1) + (-0.3) * 0.5 + 0.6 * 0.3$$
$$= (-0.01) + -0.15 + 0.18$$
$$= 0.02$$

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

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

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

(b) **25 points** Using Euclidean distance, what pair of vectors that have high similarity?