

## CS 422: Data Mining

Department of Computer Science  
Illinois Institute of Technology  
Jithin Joyson

### Spring 2019: Homework 4

## 1 Exercises (3.1.1, 3.2.1, 3.3.3, 3.4.1, 3.5.4, 3.5.5)

### 1.1 3.1.1

$a = \{1, 2, 3, 4\}$

$b = \{2, 3, 5, 7\}$

$c = \{2, 4, 6\}$

Jaccard Similarity(a,b) = a and b / a or b  
=  $2/6$   
=  $1/3$   
Jaccard Similarity(a,c) = a and c / a or c  
=  $2/5$   
Jaccard Similarity(b,c) = b and c / b or c  
=  $1/6$

### 3.2.1

	Words		Characters
[1]	The most effective	[1]	'The'
[2]	most effective way	[2]	'he '
[3]	effective way to	[3]	'e m'
[4]	way to represent	[4]	' mo'
[5]	to represent documents	[5]	'mos'
[6]	represent documents as	[6]	'ost '
[7]	documents as sets	[7]	'st '
[8]	as sets for	[8]	't e'
[9]	sets for the	[9]	' ef'
[10]	for the purpose	[10]	'eff'

### 3.3.3

(a)

Row	$S_1$	$S_2$	$S_3$	$S_4$	$2x + 1 \bmod 6$	$3x + 2 \bmod 6$	$5x+2 \bmod 6$
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

0	$S_1$	$S_2$	$S_3$	$S_4$
$h_1$	Infinity	1	Infinity	1
$h_2$	Infinity	2	Infinity	2
$h_3$	Infinity	2	Infinity	2

1	$S_1$	$S_2$	$S_3$	$S_4$
$h_1$	Infinity	1	Infinity	1
$h_2$	Infinity	2	Infinity	2
$h_3$	Infinity	1	Infinity	2

2	$S_1$	$S_2$	$S_3$	$S_4$
$h_1$	5	1	Infinity	1
$h_2$	2	2	Infinity	2
$h_3$	0	1	Infinity	0

3	$S_1$	$S_2$	$S_3$	$S_4$
$h_1$	5	1	1	1
$h_2$	2	2	5	2
$h_3$	0	1	5	0

4	$S_1$	$S_2$	$S_3$	$S_4$
$h_1$	5	1	1	1
$h_2$	2	2	2	2
$h_3$	0	1	4	0

5	$S_1$	$S_2$	$S_3$	$S_4$
$h_1$	5	1	1	1
$h_2$	2	2	2	2
$h_3$	0	1	4	0

Row	$S_1$	$S_2$	$S_3$	$S_4$
$h_1$	5	1	1	1
$h_2$	2	2	2	2
$h$	0	1	4	0

(b)  $h(3)$  is only true permutation since it contains over all the rows  $\{0,1,2,3,4,5\}$

(c) Column combinations: (1,2), (1,3), (1,4), (2,3), (2,4), (3,4)  
 True Similarities: 0/4, 0/4, 1/4, 0/4, 1/4, 1/4  
 Estimated Similarities: 1/3, 1/3, 2/3, 2/3, 2/3, 2/3

None of the similarities are alike. Thus, the transposition doesn't save Jaccard similarities.

### 3.4.1

- $1 - (1 - .1^3)^{10} = 0.00996$   
 $1 - (1 - .2^3)^{10} = 0.07718$   
 $1 - (1 - .3^3)^{10} = 0.23945$   
 $1 - (1 - .4^3)^{10} = 0.48387$   
 $1 - (1 - .5^3)^{10} = 0.73692$   
 $1 - (1 - .6^3)^{10} = 0.91227$   
 $1 - (1 - .7^3)^{10} = 0.98502$   
 $1 - (1 - .8^3)^{10} = 0.99923$   
 $1 - (1 - .9^3)^{10} = 1.0$
- $1 - (1 - .1^6)^{20} = 0.00000$   
 $1 - (1 - .2^6)^{20} = 0.00128$   
 $1 - (1 - .3^6)^{20} = 0.01448$   
 $1 - (1 - .4^6)^{20} = 0.07881$   
 $1 - (1 - .5^6)^{20} = 0.27019$   
 $1 - (1 - .6^6)^{20} = 0.61541$   
 $1 - (1 - .7^6)^{20} = 0.91819$   
 $1 - (1 - .8^6)^{20} = 0.99771$   
 $1 - (1 - .9^6)^{20} = 1.0$
- $1 - (1 - .1^5)^{50} = 0.0005$   
 $1 - (1 - .2^5)^{50} = 0.01588$   
 $1 - (1 - .3^5)^{50} = 0.11454$   
 $1 - (1 - .4^5)^{50} = 0.40228$   
 $1 - (1 - .5^5)^{50} = 0.79555$   
 $1 - (1 - .6^5)^{50} = 0.98253$   
 $1 - (1 - .7^5)^{50} = 0.99990$   
 $1 - (1 - .8^5)^{50} = 1.0$   
 $1 - (1 - .9^5)^{50} = 1.0$

#### 3.5.4

a.  $\{2,3,4\}/\{1,2,3,4,5\} = 3/5$

$$1 - 3/5 = 2/5$$

b.  $\{\}/\{1,2,3,4,5,6\} = 0/6$

$$1 - 0 = 1$$

#### 3.5.5

a.  $3(-2) + -1(3) + 2(1) / (\text{sqrt}(3^2 + -1^2 + 2^2) * \text{sqrt}(-2^2 + 3^2 + 1^2))$

$$-6 - 3 + 2 / (\text{sqrt}(14) * \text{sqrt}(14)) = -7/14 = \mathbf{-0.5}$$

$$\text{Cos-1}(-0.5) = \mathbf{120 \text{ degrees}}$$

b.  $1(2) + 2(4) + 3(6) / (\text{SQRT}(1^2 + 2^2 + 3^2) * \text{SQRT}(2^2 + 4^2 + 6^2))$

$$2 + 8 + 18 / (\text{SQRT}(14) * \text{SQRT}(56)) = 28/\text{SQRT}(784) = 28/28 = \mathbf{1}$$

$$\text{Cos-1}(1) = \mathbf{0 \text{ degrees}}$$

c.  $5(-1) + 0(-6) - 4(2) / (\text{sqrt}(5^2 + 0^2 + -4^2) * \text{sqrt}(-1^2 + -6^2 + 2^2))$

$$-5 - 8 / (\text{sqrt}(25 + 16) * \text{sqrt}(41)) = -13/41 = \mathbf{-0.317}$$

$$\text{Cos-1}(-13/41) = \mathbf{108.486 \text{ degrees}}$$

d.  $0(0) + 1(0) + 1(1) + 0(0) + 1(0) + 1(0) / (\text{sqrt}(1+1+1+1) * \text{sqrt}(1)) = 1/2 = \mathbf{0.5}$

$$\text{Cos-1}(.5) = \mathbf{60 \text{ degrees}}$$

## 1.2 Exercises {14,18,19,20}

14. Euclidian distance after the standardizing the attributes since the ranges of weight and height versus any other attribute will be off scale. Since they are no asymmetric, no cosine similarity and all the attributes are measurements, so the magnitude matters which means it cannot be correlation similarity.

18.

a. Humming distance = # different bits = 3

$$\text{Jaccard similarity} = x \text{ and } y / x \text{ or } y = 2/5$$

b. SMC = # different bit/ # bits = humming distance/ # bits

$$\text{Cosine Similarity} = (x*y)/(\sqrt{x*x}*\sqrt{y*y}) = (x \text{ and } y)/\sqrt{(x \text{ or } y)*z}$$

SMC closely matches humming distance (as the numerator) and Cosine similarity matches Jaccard similarity (elements that are looked at).

c. Jaccard since this metric measures the similarity of the two genes without spitting out how different the genes are without explaining how different.

d. Humming distance since there is a reference of how similar the genes and this metric is quite high (99.9%). Thus, it is better to focus on the differences in the genes.

19. a.  $\cos(x,y) = \cos((1,1,1,1),(2,2,2,2)) = 8/2*4 = 8/8 = 1$

$$\text{corr}(x,y) = \text{corr}((1,1,1,1),(2,2,2,2)) = \cos(a,b)$$

$$a = (1,1,1,1) - \text{mean}(1,1,1,1) = (1,1,1,1) - 1 = (0,0,0,0)$$

$$b = (2,2,2,2) - \text{mean}(2,2,2,2) = (2,2,2,2) - 2 = (0,0,0,0)$$

$$\cos(a,b) = \cos((0,0,0,0),(0,0,0,0)) = 0/0 = \text{undef}$$

$$\text{eucid}(x,y) = \sqrt{(1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2} = \sqrt{4} = 2$$

b.  $\cos(x,y) = \cos((0,1,0,1),(1,0,1,0)) = 0/\sqrt{2}*\sqrt{2} = 0/2 = 0$

$$\text{corr}(x,y) = \text{corr}((0,1,0,1),(1,0,1,0)) = \cos(a,b)$$

$$a = (0,1,0,1) - \text{mean}(0,1,0,1) = (0,1,0,1) - .5 = (-.5,.5,-.5,.5)$$

$$b = (1,0,1,0) - \text{mean}(1,0,1,0) = (1,0,1,0) - .5 = (.5,-.5,.5,-.5)$$

$$\cos(a,b) = \cos((-0.5, 0.5, -0.5, 0.5), (0.5, -0.5, 0.5, -0.5)) = -1/\sqrt{1}8\sqrt{1} = -1/1 = -1$$

$$\text{euclid}(x,y) = \sqrt{((0-1)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2)} = \sqrt{4} = 2$$

$$\text{jaccard}(x,y) = \text{jacc}((0,1,0,1),(1,0,1,0)) = 0/4 = 0$$

$$\text{c. } \cos(x,y) = \cos((0,-1,0,1),(1,0,-1,0)) = 0/\sqrt{2}*\sqrt{2} = 0/2 = 0$$

$$\text{corr}(x,y) = \text{corr}((0,-1,0,1),(1,0,-1,0)) = \cos(a,b)$$

$$a = (0,-1,0,1) - \text{mean}(0,-1,0,1) = (0,-1,0,1) - 0 = (0,-1,0,1)$$

$$b = (1,0,-1,0) - \text{mean}(1,0,-1,0) = (1,0,-1,0) - 0 = (1,0,-1,0)$$

$$\cos(a,b) = \cos((1,0,-1,0), (0,-1,0,1)) = 0/\sqrt{2}*\sqrt{2} = 0$$

$$\text{euclid}(x,y) = \sqrt{((0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2)} = \sqrt{4} = 2$$

$$\text{d. } \cos(x,y) = \cos((1,1,0,1,0,1),(1,1,1,0,0,1)) = 3/\sqrt{4}*\sqrt{4} = 3/4 = 0.75$$

$$\text{corr}(x,y) = \text{corr}((1,1,0,1,0,1),(1,1,1,0,0,1)) = \cos(a,b)$$

$$a = (1,1,0,1,0,1) - \text{mean}(1,1,0,1,0,1) = (1,1,0,1,0,1) - 2/3 = (.3, .3, -.6, .3, -.6, .3)$$

$$b = (1,1,1,0,0,1) - \text{mean}(1,1,1,0,0,1) = (1,1,1,0,0,1) - 2/3 = (.3, .3, .3, -.6, -.6, .3)$$

$$\cos(a,b) = \cos((.3, .3, -.6, .3, -.6, .3), (.3, .3, .3, -.6, -.6, .3)) = (1/3)/(12/9) = 1/4$$

$$\text{jaccard}(x,y) = \text{jacc}((1,1,0,1,0,1), (1,1,1,0,0,1)) = 3/5$$

$$\text{e. } \cos(x,y) = \cos((2,-1,0,2,0,-3),(-1,1,-1,0,0,-1)) = 0/\sqrt{18}*\sqrt{4} = 0$$

$$\text{corr}(x,y) = \text{corr}((2,-1,0,2,0,-3),(-1,1,-1,0,0,-1)) = \cos(a,b)$$

$$a = (2,-1,0,2,0,-3) - \text{mean}(2,-1,0,2,0,-3) = (2,-1,0,2,0,-3) - 0 = (2,-1,0,2,0,-3)$$

$$b = (-1,1,-1,0,0,-1) - \text{mean}(-1,1,-1,0,0,-1) = (-1,1,-1,0,0,-1) + 1/3$$

$$= (-.6, 1.3, -.6, .3, .3, -.6)$$

$$\cos(a,b) = \cos((2,-1,0,2,0,-3), (-.6, 1.3, -.6, .3, .3, -.6)) = 0/\sqrt{18}*\sqrt{30/9}$$

$$= 0$$

20. a. The range is mimicked by the cosine function. Thus, it goes from -1 to +1.
- b. They can be identical but not always since the 1 forces the objects to differ by a constant.
- c. Correlation is cosine similarity of the vectors offset by the mean of those vectors. If the means are 0 (vector is centered around 0),  $\text{corr}(x,y)$  is the same as the  $\cos(x,y)$
- d. There seems to be an inverse relationship between cosine similarity and Euclidian distances. If the cosine similarity increases, the Euclidian distance decreases and vice versa.
- e. There seems to be an inverse relationship between correlation and Euclidean distances. If the correlation increases, the Euclidian distance decreases and vice versa.
- f.  $\text{Euclid}(x,y) = \sqrt{\sum((x-y)^2) \rightarrow n}$

Expanding the square:

$$= \sqrt{\sum((x^2 - 2xy + y^2) \rightarrow n)}$$

Summing n times creates the dot product between x and y. Since x and y are unit vectors and norms are 1:  $\cos(x,y) = x*y/(|x| * |y|) = x*y/1 = x*y$ .

Furthermore, the elements in the unit vector summed over all elements squared is 1 (by definition the magnitude is 1). Thus  $x^2$  and  $y^2$  go to 1. This results in:

$$= \sqrt{1 - 2\cos(x,y) + 1}$$

Some factoring:

$$= \sqrt{2(1-\cos(x,y))}$$

g.  $\text{Euclid}(x,y) = \sqrt{\sum((x-y)^2) \rightarrow n}$

Expanding the square:

$$= \sqrt{\sum((x^2 - 2xy + y^2) \rightarrow n)}$$

Assuming the mean of the vectors to be 0, the  $\text{corr}(x,y) = \cos(x - \text{mean}, y - \text{mean}) = \cos(x - 0, y - 0) = \cos(x,y)$ . Thus, following the previous principle of the vectors being unit vectors, the equation devolves to the following:

$$= \sqrt{2n(1-\text{corr}(x,y))}$$

Since the stdev is division by n, n is preserved



## 1.3 Exercises {5.1.1,5.1.2,5.1.6}

### 5.1.1

$$PR(A) = PR(B)/C(B) + PR(A)/C(A) = (1/3)/2 + (1/3)/3 = 5/18$$

$$PR(B) = PR(A)/C(A) + PR(C)/C(C) = (1/3)/3 + (1/3)/2 = 5/18$$

$$PR(C) = PR(A)/C(A) + PR(B)/C(B) + PR(C)/C(C) = (1/3)/3 + (1/3)/2 + (1/3)/2 = 8/18$$

Transition Matrix:

	A	B	C
A	1/3	1/2	0/2
B	1/3	0/2	1/2
C	1/3	1/2	1/2

Power Iteration:

PR0:

A	1/3
B	1/3
C	1/3

Transition Matrix \* PRN

	PR1	PR2	PR3	PR(10)
A	5/18	25/108	19/81	<b>A 0.23</b>
B	5/18	17/54	197/648	<b>B 0.31</b>
C	8/18	49/108	299/648	<b>C 0.46</b>

### 5.1.2

$$TM = \text{Beta} * TM + (1-\text{Beta})TM$$

$$(0.8*TM + 0.2*TM) * PRN$$

4/15	6/15	0
4/15	0	6/15
4/15	6/15	6/15

+

1/15	1/15	1/15
1/15	1/15	1/15
1/15	1/15	1/15

Transition Matrix:

1/3	7/15	1/15
1/3	1/15	7/15
1/3	7/15	7/15

Power Iteration:

PR0:

A	1/3
B	1/3
C	1/3

Transition Matrix \* PRN

PR1

A	13/45
B	13/45
C	19/45

PR2

A	7/27
B	211/675
C	289/675

PR3

A	2641/10125
B	3109/10125
C	35/81

PR10

<b>A</b>	<b>0.26</b>
<b>B</b>	<b>0.31</b>
<b>C</b>	<b>0.43</b>

#### 5.1.6

Since all the dead ends are eliminated, only the first node will remain which has an out going edge to itself. Thus, the PageRank for this node will be 1. This node is the predecessor of the next node which will have a PageRank of  $(1 * (1/2)) = 1/2$ . The nodes after the 2<sup>nd</sup> node will be  $1/2$  (PR of the second node) \* 1 (Contribution of the 2<sup>nd</sup> node) =  $1/2$ .

Thus, the first node (all the way to the left) PR will be 1. All the nodes after will have a PR of  $1/2$ .

## 1.4 Centrality Measures

a.

(a) Centrality( $v$ ) = degree( $v$ )/ $n-1$  where  $n$  = # nodes

$$\text{Centrality}(1) = 1/(5-1) = \mathbf{1/4}$$

$$\text{Centrality}(2) = 3/(5-1) = \mathbf{3/4}$$

$$\text{Centrality}(3) = 2/(5-1) = \mathbf{2/4 = 1/2}$$

$$\text{Centrality}(4) = 4/(5-1) = \mathbf{4/4 = 1}$$

$$\text{Centrality}(5) = 2/(5-1) = \mathbf{2/4 = 1/2}$$

(b) Closeness( $v$ ) =  $(n-1) * 1/\text{sum}(\text{dist}(v,j)) \rightarrow j$

$$\text{Closeness}(1) = (5-1) * 1/(1+2+2+2) = \mathbf{4/7}$$

$$\text{Closeness}(2) = (5-1) * 1/(1+1+1+2) = \mathbf{4/5}$$

$$\text{Closeness}(3) = (5-1) * 1/(1+2+1+2) = 4/6 = \mathbf{2/3}$$

$$\text{Closeness}(4) = (5-1) * 1/(1+1+1+1) = 4/4 = \mathbf{1}$$

$$\text{Closeness}(5) = (5-1) * 1/(1+1+2+2) = 4/6 = \mathbf{2/3}$$

(c) Betweenness( $v$ ) = # shortest paths that go through  $v$

Shortest Paths: (1,2) (1,3) (1,4) (1,5) (2,3) (2,4) (2,5) (3,4) (3,5) (3,5) (4,5)

$$\text{Betweenness}(1): (2,3) (2,4) (2,5) (3,4) (3,5) (3,5) (4,5) = \mathbf{0}$$

$$\text{Betweenness}(2): (1,3) (1,4) (1,5) (3,4) (\mathbf{3,5}) (3,5) (4,5) = .5$$

$$C(2) = 2 * .5 = 1$$

$$\text{Normalize} = 1 / [2 * (4 \text{ Choose } 2)] = 2/12 = \mathbf{1/12}$$

$$\text{Betweenness}(3): (1,2) (1,4) (1,5) (2,4) (2,5) (4,5) = 0/6 = \mathbf{0}$$

$$\text{Betweenness}(4): (\mathbf{1,2}) (\mathbf{1,3}) (\mathbf{1,5}) (2,3) (2,5) (3,5) (\mathbf{3,5}) = 3.5$$

$$C(4) = 2 * 3.5 = 7$$

$$\text{Normalize} = 7/[2 * (4 \text{ Choose } 2)] = \mathbf{7/12}$$

$$\text{Betweenness}(5): (1,2) (1,3) (1,4) (2,3) (2,4) (3,4) = \mathbf{0}$$

b.

(a)  $\text{Centrality}(v) = \text{degree}(v)/n-1$  where  $n = \# \text{ nodes}$

$$\text{Centrality}(1) = 2/(5-1) = \mathbf{2/4 = 1/2}$$

$$\text{Centrality}(2) = 3/(5-1) = \mathbf{3/4}$$

$$\text{Centrality}(3) = 2/(5-1) = \mathbf{2/4 = 1/2}$$

$$\text{Centrality}(4) = 2/(5-1) = \mathbf{2/4 = 1/2}$$

$$\text{Centrality}(5) = 3/(5-1) = \mathbf{3/4}$$

(b)  $\text{Closeness}(v) = (n-1) * 1/\text{sum}(\text{dist}(v,j)) \rightarrow j$

$$\text{Closeness}(1) = (5-1) * 1/(1+1+2+2) = 4/6 = \mathbf{2/3}$$

$$\text{Closeness}(2) = (5-1) * 1/(1+1+1+2) = \mathbf{4/5}$$

$$\text{Closeness}(3) = (5-1) * 1/(1+2+1+2) = 4/6 = \mathbf{2/3}$$

$$\text{Closeness}(4) = (5-1) * 1/(1+1+2+2) = 4/6 = \mathbf{2/3}$$

$$\text{Closeness}(5) = (5-1) * 1/(1+1+1+2) = \mathbf{4/5}$$

(c)  $\text{Betweenness}(v) = \# \text{ shortest paths that go through } v$

$$\text{Shortest Paths: } (1,2) (1,2) (1,3) (1,4) (1,5) (2,3) (2,4) (2,5) (3,4) (3,5) (4,5)$$

$$(4,5)$$

Betweenness(1): (2,3) (2,4) (2,5) (3,4) (3,5) (4,5) **(4,5)** = .5

$$C(1) = 2 * .5 = 1$$

$$\text{Normalize} = 1 / [ 2 * (4 \text{ Choose } 2)] = \mathbf{1/12}$$

Betweenness(2): (1,3) (1,4) (1,5) **(3,4)** (3,5) (4,5) **(4,5)** = 1.5

$$C(2) = 2 * 1.5 = 3$$

$$\text{Normalize} = 3 / [ 2 * (4 \text{ Choose } 2)] = 3/12 = \mathbf{1/4}$$

Betweenness(3): (1,2) (1,2) (1,4) (1,5) (2,4) (2,5) (4,5) (4,5) = **0**

Betweenness(4): **(1,2)** (1,2) (1,3) (1,5) (2,3) (2,5) (3,5) = .5

$$C(4) = 2 * .5 = 1$$

$$\text{Normalize} = 1/[2 * (4 \text{ Choose } 2)] = \mathbf{1/12}$$

Betweenness(5): (1,2) **(1,2)** **(1,3)** (1,4) (2,3) (2,4) (2,5) (3,4) = 1.5

$$C(5) = 2 * 1.5 = 3$$

$$\text{Normalize} = 3 / [ 2 * (4 \text{ Choose } 2)] = 3/12 = \mathbf{1/4}$$