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 mod 6	3x + 2 mod 6	5x+2 mod 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.

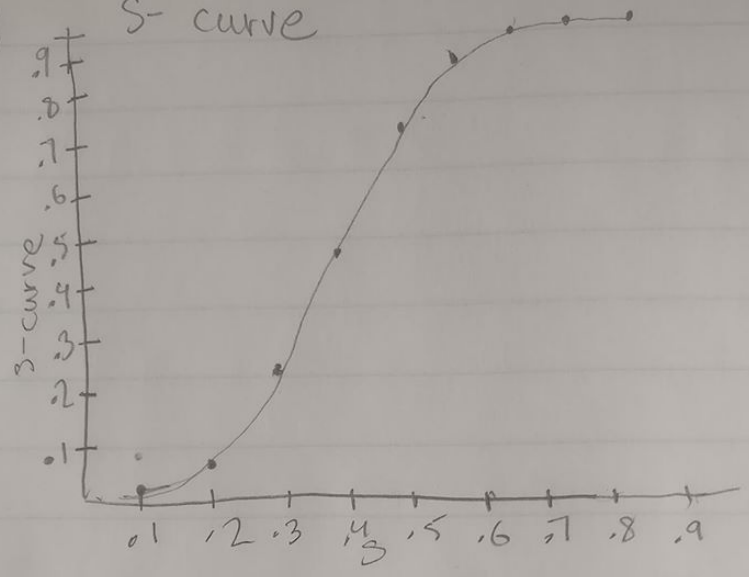
3.4.1

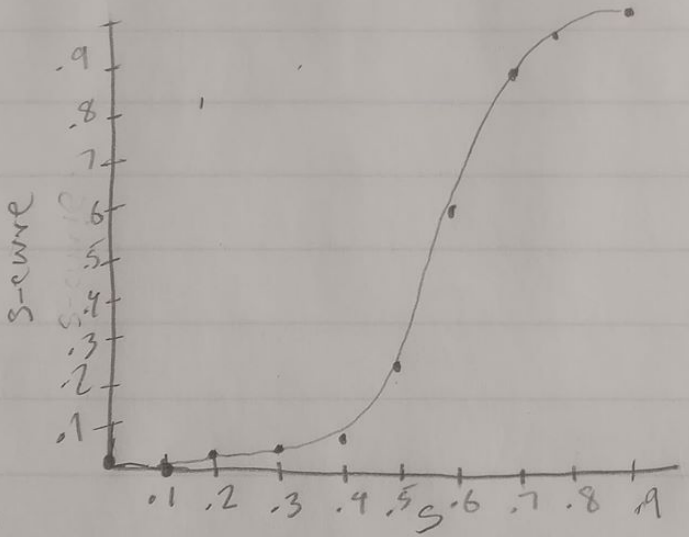
- $1 - (1 - \mathbf{.1}^3)^{10} = 0.00996$
  $1 - (1 - \mathbf{.2}^3)^{10} = 0.07718$
  $1 - (1 - \mathbf{.3}^3)^{10} = 0.23945$
  $1 - (1 - \mathbf{.4}^3)^{10} = 0.48387$
  $1 - (1 - \mathbf{.5}^3)^{10} = 0.73692$
  $1 - (1 - \mathbf{.6}^3)^{10} = 0.91227$
  $1 - (1 - \mathbf{.7}^3)^{10} = 0.98502$
  $1 - (1 - \mathbf{.8}^3)^{10} = 0.99923$
  $1 - (1 - \mathbf{.9}^3)^{10} = 1.0$

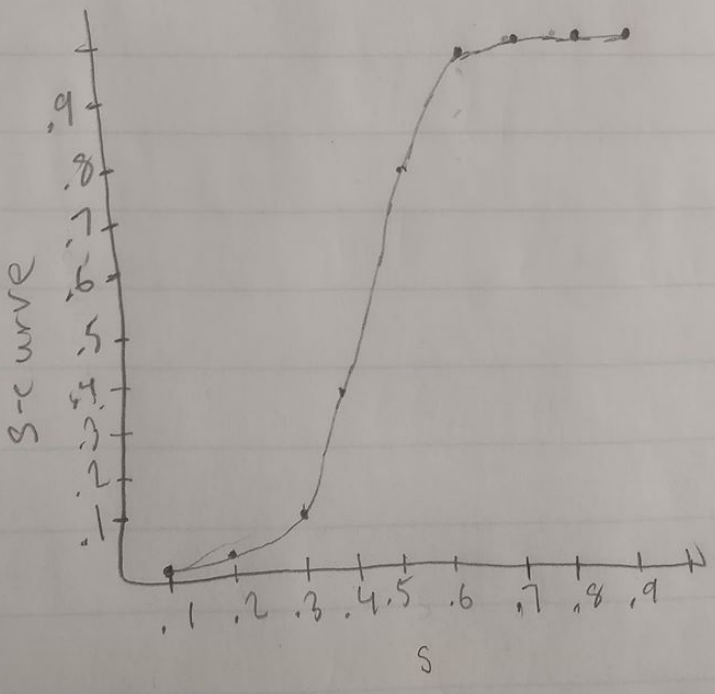
- $1 - (1 - .1^6)^{20} = 0.00000$
  $1 - (1 - .2^6)^{20} = 0.00128$
  $1 - (1 - .3^6)^{20} = 0.01448$
  $1 - (1 - .4^6)^{20} = 0.07881$
  $1 - (1 - .5^6)^{20} = 0.27019$
  $1 - (1 - .6^6)^{20} = 0.61541$
  $1 - (1 - .7^6)^{20} = 0.91819$
  $1 - (1 - .8^6)^{20} = 0.99771$
  $1 - (1 - .9^6)^{20} = 1.0$

- $1 - (1 - .1^5)^{50} = 0.0005$
  $1 - (1 - .2^5)^{50} = 0.01588$
  $1 - (1 - .3^5)^{50} = 0.11454$
  $1 - (1 - .4^5)^{50} = 0.40228$
  $1 - (1 - .5^5)^{50} = 0.79555$
  $1 - (1 - .6^5)^{50} = 0.98253$
  $1 - (1 - .7^5)^{50} = 0.99990$
  $1 - (1 - .8^5)^{50} = 1.0$
  $1 - (1 - .9^5)^{50} = 1.0$

341 S- curve


S-curve
.9
.8
.7
.6
.5
.4
.3
.2
.1
.1 .2 .3 .4 .5 .6 .7 .8 .9
S


S-curve
.9
.8
.7
.6
.5
.4
.3
.2
.1
.1 .2 .3 .4 .5 .6 .7 .8 .9
S


S-curve
.9
.8
.7
.6
.5
.4
.3
.2
.1
.1 .2 .3 .4 .5 .6 .7 .8 .9
S

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) / (sqrt(3^2 + -1^2 + 2^2) * sqrt(-2^2 + 3^2 + 1^2))

   -6 – 3 +2/(sqrt(14) * sqrt(14)) = -7/14 = **-0.5**

   Cos-1(-0.5) = **120 degrees**

b. 1(2) + 2(4) + 3(6)/ (SQRT(1^2 + 2^2 + 3^2) * SQRT(2^2 + 4^2 + 6^2))

   2 + 8 + 18/(SQRT(14) * SQRT(56)) = 28/SQRT(784) = 28/28 = **1**

   Cos-1(1) = **0 degrees**

c. 5(-1) + 0(-6) – 4(2)/(sqrt(5^2 + 0^2 + -4^2) * sqrt(-1^2 + -6^2 + 2^2))

   -5 – 8/(sqrt(25 + 16)*sqrt(41)) = -13/41 = **-0.317**

   Cos-1(-13/41) = **108.486 degrees**

d. 0(0) + 1(0) + 1(1) + 0(0) + 1(0) + 1(0)/(sqrt(1+1+1+1)*sqrt(1)) = ½ = **0.5**

   Cos-1(.5) = **60 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

Jaccard similarity = x and y / x or y = 2/5

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

Cosine Similarity = (x*y)/(sqrt(x*x)*sqrt(x*x)) = (x and y)*z/(x or y)*z

SMC closely matches humming distance (as the numerator) and Cosine similarly 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**

corr(x,y) = corr((1,1,1,1),(2,2,2,2)) = cos(a,b)

a = (1,1,1,1) - mean(1,1,1,1) = (1,1,1,1)- 1 = (0,0,0,0)

b = (2,2,2,2) - 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 = **undef**

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**

corr(x,y) = corr((0,1,0,1),(1,0,1,0)) = cos(a,b)

a = (0,1,0,1) - mean(0,1,0,1) = (0,1,0,1) - .5 = (-.5,.5,-.5,.5)

b = (1,0,1,0) - mean(1,0,1,0) = (1,0,1,0) - .5 = (.5,-.5,.5,-.5)

cos(a,b) = cos((-.5,.5,-.5,.5), (.5,-.5,.5,-.5)) = -1/sqrt(1)8sqrt(1) = -1/1 = **-1**

eucid(x,y) = sqrt((0-1)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2) = sqrt(4) = **2**

jaccard(x,y) = jacc((0,1,0,1),(1,0,1,0)) = 0/4 = **0**

c. cos(x,y) = cos((0,-1,0,1),(1,0,-1,0)) = 0/sqrt(2)*sqrt(2) = 0/2 = **0**

corr(x,y) = corr((0,-1,0,1),(1,0,-1,0)) = cos(a,b)

a = (0,-1,0,1) - mean(0,-1,0,1) = (0,-1,0,1) - 0 = (0,-1,0,1)

b = (1,0,-1,0) - 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**

eucid(x,y) = sqrt((0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2) = sqrt(4) = **2**

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**

corr(x,y) = corr((1,1,0,1,0,1),(1,1,1,0,0,1)) = cos(a,b)

a = (1,1,0,1,0,1) - 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) - 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**

jaccard(x,y) = jacc((1,1,0,1,0,1), (1,1,1,0,0,1)) = **3/5**

e. cos(x,y) = cos((2,-1,0,2,0,-3),(-1,1,-1,0,0,-1)) = 0/sqrt(18)*sqrt(4) = **0**

corr(x,y) = corr((2,-1,0,2,0,-3),(-1,1,-1,0,0,-1)) = cos(a,b)

a = (2,-1,0,2,0,-3) - 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) - 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), 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. Euclid(x,y) = sqrt(sum((x-y)^2) → n)

       Expanding the square:

       = sqrt(sum((x^2 – 2xy + y^2) → 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 - 2cos(x,y) +1)

       Some factoring:

       = sqrt(2(1-cos(x,y))

    g. Euclid(x,y) = sqrt(sum((x-y)^2) → n)

       Expanding the square:

       = sqrt(sum((x^2 – 2xy + y^2) → n)

       Assuming the mean of the vectors to be 0, the corr(x,y) = cos(x – mean, y-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-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

A	5/18
B	5/18
C	8/18

PR2

A	25/108
B	17/54
C	49/108

PR3

A	19/81
B	197/648
C	299/648

PR(10)

A	**0.23**
B	**0.31**
C	**0.46**

5.1.2

TM = Beta * TM + (1-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

A	**0.26**
B	**0.31**
C	**0.43**

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 2nd node will be 1/2 (PR of the second node) * 1 (Contribution of the 2nd 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

Centrality(1) = 1/(5-1) = **1/4**

Centrality(2) = 3/(5-1) = **3/4**

Centrality(3) = 2/(5-1) = **2/4 = 1/2**

Centrality(4) = 4/(5-1) = **4/4 = 1**

Centrality(5) = 2/(5-1) = **2/4 = 1/2**

(b) Closeness(v) = (n-1) *1/sum(dist(v,j)) -> j

Closeness(1) = (5-1)*1/(1+2+2+2) = **4/7**

Closeness(2) = (5-1)*1/(1+1+1+2) = **4/5**

Closeness(3) = (5-1)*1/(1+2+1+2) = 4/6 = **2/3**

Closeness(4) = (5-1)*1/(1+1+1+1) = 4/4 = **1**

Closeness(5) = (5-1)*1/(1+1+2+2) = 4/6 = **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)

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

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

C(2) = 2 * .5 = 1

Normalize = 1/ [2 * (4 Choose 2)] = 2/12 = **1/12**

Betweenness(3): (1,2) (1,4) (1,5) (2,4) (2,5)(4,5) = 0/6 = **0**

Betweenness(4): **(1,2) (1,3) (1,5)** (2,3) (2,5) (3,5) **(3,5)** = 3.5

C(4) = 2 * 3.5 = 7

Normalize = 7/[2 * (4 Choose 2)] = **7/12**

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

b.

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

Centrality(1) = 2/(5-1) = **2/4 = 1/2**

Centrality(2) = 3/(5-1) = **3/4**

Centrality(3) = 2/(5-1) = **2/4 = 1/2**

Centrality(4) = 2/(5-1) = **2/4 = 1/2**

Centrality(5) = 3/(5-1) = **3/4**

(b) Closeness(v) = (n-1) *1/sum(dist(v,j)) -> j

Closeness(1) = (5-1)*1/(1+1+2+2) = 4/6 = **2/3**

Closeness(2) = (5-1)*1/(1+1+1+2) = **4/5**

Closeness(3) = (5-1)*1/(1+2+1+2) = 4/6 = **2/3**

Closeness(4) = (5-1)*1/(1+1+2+2) = 4/6 = **2/3**

Closeness(5) = (5-1)*1/(1+1+1+2) = **4/5**

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

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

Normalize = 1/ [2 * (4 Choose 2)] = **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

Normalize = 3/ [2 * (4 Choose 2)] = 3/12 = **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

Normalize = 1/[2 * (4 Choose 2)] = **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

Normalize = 3/ [2 * (4 Choose 2)] = 3/12 = **1/4**