Metrics and examples

**For binary data**

- **L1 distance** corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors.

- **Jaccard similarity** is a measure of the similarity between two binary vectors.

For given two binary vectors,

x = 010101001

y = 010011000

**Hamming distance** = 3;

there are 3 binary numbers different between the x and y.

**Jaccard coefficient**:

J= (number of matching presences) / (number of attributes not involved in 00 matches)

J=(f11)/(f01 + f10 + f11)

f01 = 1 the number of attributes where x was 0 and y was 1

f10 = 2 the number of attributes where x was 1 and y was 0

f00 = 5 the number of attributes where x was 0 and y was 0

f11 = 2 the number of attributes where x was 1 and y was 1

J= (2) / (1 + 2 + 2)

J = 2/5 = 0.4

**Simple Matching Coefficient**:

SMC = (f11 + f00) / (f01 + f10 + f11 + f00)

= (2 + 5) / (1 + 2 +2 +5)

= 7/10 = 0.7

For the following vectors, x and y, calculate the indicated similarity or distance measures.

**x^T = (1,1,1,1)**

**y^T = (2,2,2,2)**

**Cosine**

x^T y = 1\*2 + 1\*2 + 1\*2 + 1\*2 = 8

||x|| = sqrt(1\*1 + 1\*1 + 1\*1 + 1\*1) = sqrt (4) = 2

||y|| = sqrt(2\*2 + 2\*2 + 2\*2 + 2\*2) = sqrt (16) = 4

cos(x,y) = (x^T y) / (||x||\*||y||) = (8)/ (2\*4)

cos(x,y) = 1

**Correlation**

corr(x, y) = [covariance(x,y)] / [standard deviation(x) \* standard deviation(y)]

Mean of x =(1+1+1+1) / 4 = 1

Mean of y = (2+2+2+2) / 4 = 2

covariance(x,y) = 1/(4 -1) [(1-1)(2-2) + (1-1)(2-2) + (1-1)(2-2) + (1-1)(2-2)] = 0

Standard deviation (x) = sqrt[((1/(4-1))) \* {(1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2}]

= sqrt[(1/3) \* 0] = 0

Standard deviation (y) = sqrt[((1/(4-1))) \* {(2-2)^2 + (2-2)^2 + (2-2)^2 + (2-2)^2}]

= sqrt[(1/3) \* 0] = 0

corr(x,y) = 0/0 = undefined

**Euclidean**

d(x, y) = sqrt((1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2)

Euclidean distance = 2

**x^T = (0,1,0,1)**

**y^T = (1,0,1,0)**

**Cosine**

x^T y = 0\*1 + 1\*0 + 0\*1 + 1\*0 = 0

||x|| = sqrt(0\*0 + 1\*1 + 0\*0 + 1\*1) = sqrt (2)

||y|| = sqrt(1\*1 + 0\*0 + 1\*1 + 0\*0) = sqrt (2)

cos(x,y) = (x^T y) / (||x||\*||y||) = (0)/ (sqrt (2) \*sqrt (2))

cos(x,y) = 0

**Correlation**

corr(x, y) = [covariance(x,y)] / [standard deviation(x) \* standard deviation(y)]

Mean of x = (0+1+0+1) / 4 = ½ = 0.5

Mean of y = (1+0+1+0) / 4 = ½ = 0.5

Covariance(x,y) = 1/(4-1) \* [(0-½)(1-½) + (1-½)(0-½) +(0-½)(1-½) +(1-½)(0-½) ]

Covariance(x,y) = (1/3) \* [(-1/4) + (-1/4) + (-1/4) + (-1/4)]

Covariance(x,y) = -1/3

Standard\_deviation (x) = sqrt[((1/(4-1))) \* {(1-1/2)^2 + (0-1/2)^2 + (1-1/2)^2 + (0-1/2)^2}]

= sqrt[(1/3) \* 1] = 0.57735

Standard\_deviation (y) = sqrt[((1/(4-1))) \* {(0-1/2)^2 + (1-1/2)^2 + (0-1/2)^2 + (1-1/2)^2}]

= sqrt[(1/3) \* 1] = 0.57735

Corr(x,y) = (-1/3) / (0.57735 \* 0.57735)

Corr(x,y) = -1

**Euclidean**

d(x, y) = sqrt((0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2)

Euclidean distance = 2

**Jaccard**

J= (numbar of matching presences) / (number of attributes not involved in 00 matches)

J=(f11)/(f01 + f10 + f11)

f01 = 2 the number of attributes where x was 0 and y was 1

f10 = 2 the number of attributes where x was 1 and y was 0

f00 = 0 the number of attributes where x ws 0 and y was 0

f11 = 0 the number of attributes where x was 1 and y was 1

J= (0) / (2 + 2 + 0)

J = 0

**x^T = (0,-1,0,1)**

**y^T = (1,0,-1,0)**

**Cosine**

x^T y = 0\*1 + (-1)\*0 + 0\*(-1) + 1\*0 = 0

||x|| = sqrt(0\*0 + (-1)\*(-1) + 0\*0 + 1\*1) = sqrt (2)

||y|| = sqrt(1\*1 + 0\*0 + (-1)\*(-1) + 0\*0) = sqrt (2)

cos(x,y) = (x^T y) / (||x||\*||y||) = (0)/ (sqrt (2) \*sqrt (2)) = 0

**Correlation**

corr(x, y) = [covariance(x,y)] / [standard deviation(x) \* standard deviation(y)]

Mean of x = (0+(-1)+0+1) / 4 = 0

Mean of y = (1+0+(-1)+0) / 4 = 0

Covariance(x,y) = 1/(4-1) \* [(0-0)(1-0) + (-1-0)(0-0) +(0-0)(-1-0) +(1-0)(0-0) ]

= (1/3) \* 0 = 0

corr(x,y) = 0

**Eulidean**

d(x, y) = sqrt((0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2)

Euclidean distance = 2

**x^T = (1,1,0,1,0,1)**

**y^T = (1,1,1,0,0,1)**

**Cosine**

x^T y = 1\*1 + 1\*1 + 0\*1 + 1\*0 + 0\*0 + 1\*1 = 3

||x|| = sqrt(1\*1 + 1\*1 + 0\*0 + 1\*1 + 0\*0 + 1\*1) = 2

||y|| = sqrt(1\*1 + 1\*1 + 1\*1 + 0\*0 + 0\*0 + 1\*1) = 2

cos(x,y) = (x^T y) / (||x||\*||y||) = (3)/ (2 \* 2) = 0.75

**Correlation**

corr(x, y) = [covariance(x,y)] / [standard deviation(x) \* standard deviation(y)]

Mean of x = (1+1+0+1+0+1) / 6 = 4/6

Mean of y = (1+1+1+0+0+1) / 6 = 4/6

Covariance(x,y) = 1/(6-1) \* [(1-4/6)(1-4/6) + (1-4/6)(1-4/6) +(0-4/6)(1-4/6) +(1-4/6)(0-4/6) + (0-4/6)(0-4/6) + (1-4/6)(1-4/6) ] = (1/5)(1/3)=1/15

Standard\_deviation (x) = sqrt[((1/(6-1))) \* {(1-4/6)^2 + (1-4/6)^2 + (0-4/6)^2 + (1-4/6)^2 + (0-4/6)^2 + (1-4/6)^2}] = sqrt[(1/5) \* (4/3)] = 0.5164

Standard\_deviation (y) = sqrt[((1/(6-1))) \* {(1-4/6)^2 + (1-4/6)^2 + (1-4/6)^2 + (0-4/6)^2 + (0-4/6)^2 + (1-4/6)^2}] = sqrt[(1/5) \* (4/3)] = 0.5164

Corr(x,y) = (1/15) / (0.5164 \* 0.5164)

Corr(x,y) = 0.25

**Jaccard**

J= (numbar of matching presences) / (number of attributes not involved in 00 matches)

J=(f11)/(f01 + f10 + f11)

f01 = 1 the number of attributes where x was 0 and y was 1

f10 = 1 the number of attributes where x was 1 and y was 0

f00 = 1 the number of attributes where x ws 0 and y was 0

f11 = 3 the number of attributes where x was 1 and y was 1

J= (3) / (1 + 1 + 3)

J = 3/5 = 0.6

**(x^T = (2,-1,0,2,0,-3)**

**y^T = (-1,1,-1,0,0,-1)**

**Cosine**

x^T y = 2\*(-1) + (-1)\*1 + 0\*(-1) + 2\*0 + 0\*0 + (-3)\*(-1) = 0

||x|| = sqrt(2\*2 + (-1)\*(-1) + 0\*0 + 2\*2 + 0\*0 + (-3)\*(-3)) = sqrt (18)

||y|| = sqrt((-1)\*(-1) + 1\*1 + (-1)\*(-1) + 0\*0 + 0\*0 + (-1)\*(-1)) = 2

cos(x,y) = (x^T y) / (||x||\*||y||) = (0)/ (sqrt (18) \* 2)

cos(x,y) = 0

**Correlation**

corr(x, y) = [covariance(x,y)] / [standard deviation(x) \* standard deviation(y)]

Mean of x = (2+(-1)+0+2+0+(-3)) / 6 = 0

Mean of y = ((-1)+1+(-1)+0+0+(-1)) / 6 = -1/6

Covariance(x,y) = 1/(6-1) \* [(2-0)(-1+1/6) + (-1-0)(1+1/6) +(0-0)(-1+1/6) + (2-0)(0+1/6) + (0-0)(0+1/6) + (-3-0)(-1+1/6) ] = (1/5) \* 0 = 0

Corr(x,y) = 0

References:

<http://csucidatamining.weebly.com/assign-3.html>

Introduction to Data Mining,By: Pang-Ning Tan, Michael Steinbach, Vipin Kumar - Addison Wesley, 2005.

**Spearman’s rank-order correlation**

<https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide-2.php>

We then complete the following table:

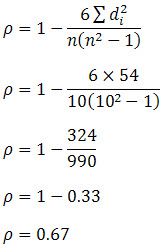
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Grade1 | Grade2 | Rank1 | Rank2 | d | d^2 |
| 66 | 66 | 9 | 4 | 5 | 25 |
| 70 | 70 | 3 | 2 | 1 | 1 |
| 40 | 40 | 10 | 10 | 0 | 0 |
| 60 | 60 | 4 | 7 | 3 | 9 |
| 65 | 65 | 6 | 5 | 1 | 1 |
| 56 | 56 | 5 | 9 | 4 | 16 |
| 59 | 59 | 8 | 8 | 0 | 0 |
| 77 | 77 | 1 | 1 | 0 | 0 |
| 67 | 67 | 2 | 3 | 1 | 1 |
| 63 | 63 | 7 | 6 | 1 | 1 |

Where d = difference between ranks and d2 = difference squared.

We then calculate the following:

[Spearman Formula](https://statistics.laerd.com/statistical-guides/img/spearman-4.jpg)

We then substitute this into the main equation with the other information as follows:

[](https://statistics.laerd.com/statistical-guides/img/spearman-5.jpg)

as *n* = 10. Hence, we have a ρ (or *r*s) of 0.67. This indicates a strong positive relationship between the ranks individuals obtained in the two grades. That is, the higher you ranked in grade1, the higher you ranked in grade2 also, and vice versa.