**ujMetrics**

* **Cosine, Correlation, Distance, Similarity, Normalization, Norm, etc**

What does each of condition tell you?

**(Q01)** (Cosine) Two attributes have a cosine of 1.

The angle difference between the attributes is nonexistent, they are the same angle

**(Q02)** (Correlation) Two attributes have a correlation of 1.

A correlation of one means that the two attributes have perfect positive correlation and show 1:1 similarity based on the average direction of the data set. It also means that neither attribute has a standard deviation of zero

**(Q03)** (Jaccard) Two attributes have a jaccard of 1.

There was at least one matching presence

**(Q04)** (Distance) Two attributes have a (normalized) distance of 1.

The square root of the sum of the 2 attributes squared is 1. The vectors are not on the same space

**(Q05)** (Similarity) Two attributes have a (normalized) similarity of 0.

The two attributes are completely dissimilar

Let’s further explore cosine, correlation, and z-score with the following questions.

**(Q06)** What is the range of values that are possible for the cosine measure?

Between 0 and 1 inclusive

**(Q07)** What is the range of values that are possible for the Pearson’s correlation measure?

-1 to 1 with values ranging from a totally opposite correlation to an identical one.

**(Q08)** If two objects have a cosine measure of 1, are they identical? Explain.

(Hint) What is the cosine of x=(2, 2, 2, 2) and y=(4,4,4,4)?

What do you infer from this result?

Two objects having a cosine of 1 means that they face the same way, that is all we can infer from this value. As seen in the hint we can have different values for the vectors and still get the same angle. No, two objects with a cosine measure of 1 are not necessarily identical.

**(Q09)** What is the relationship among cosine, correlation, & z-score, if any? (Hint: Refer to the definition of each measure and try to explain how you can make a connection among the three measures.)

Cosine measures direction similarity between vectors by dividing their dot product by the product of their magnitude

Correlation tells you the covariance between the vectors divided by the product of the standard deviation of both

z-score tells you how far a data point is from the mean, divided by the standard deviation

Cosine and correlation are ways of measuring the similarities between vectors, z-score tells you the similarities between the values in a vector in relation to the mean in terms of the standard deviation

Answer the questions for the following vectors x and y.

**(Q10)** x =(2, 2, 4, 4)

* 1-norm normalization of vector x
* How do you interpret 1-norm normalization?
* 2-norm normalization of vector x
* How do you interpret 2-norm normalization?

1. (2+2+4+4)^1/1 = 12, 1-norm-normalized(x) = (2/12, 2/12, 4/12, 4/12)
2. Projection of x onto hyperspace
3. (2^2 + 2^2 + 4^2 + 4^2)^(1/2) = (4 + 4 + 16 + 16)^(1/2) = sqrt(40),2-norm-normalization(x) = (2/sqrt(40), 2/sqrt(40), 4/sqrt(40), 4/sqrt(40))
4. Projection of x onto 2d hyperspace (surface of a hypersphere)

**(Q11)** y=(1, 2, 3 ,4, 5)

* Mean-centering of vector x
* What is the characteristic of mean-centering?
* z-score transformation of vector x
* What is/are characteristic(s) of z-score transformation?

1. Mean-centering of x = each attr in x gets subtracted by the mean
   1. Mean of x = (1+2+3+4+5)/5 = 3
   2. Mean-centered(y) = (-2,-1,0,1,2)
2. Mean centering pulls every attribute in the vector towards the mean, you can use it to very easily identify outliers in the vector
3. Zscore = (value in y – mean) / stdev
   1. Stdev = sqrt((4 + 1 + 0 + 1 + 4)/5) = sqrt(2)
   2. Transform every vector value into z-score:
   3. y = (-2/sqrt(2), -1/sqrt(2), 0, 1/sqrt(2), 2/sqrt(2) )
4. the z-score transformation tells us the relationship of each value to the mean in terms of the standard deviation

**(Q12)** x =(0, 1, 0, 1, 0, 1), y=(1, 0, 1, 0, 1, 0)

* Cosine
* Correlation
* Euclidian
* Jaccard

1. Cosine similarity = x\*y/ (norm(x)\*norm(y))
   1. x\*y = 0\*1+1\*0+0\*1+1\*0+0\*1+1\*0 = 0
   2. norm(x) = sqrt(0\*0+1\*1+0\*0+1\*1+0\*0+1\*1) = sqrt(3)
   3. norm(y) = sqrt(1\*1+0\*0+1\*1+0\*0+1\*1+0\*0) = sqrt(3)
   4. cosine similarity = 0/3 = 0
2. correlation = cov(x,y)/(stdev(x)\*stdev(y)
   1. mean(x) = (0+1+0+1+0+1)/6 = 0.5
   2. mean(y) = (1+0+1+0+1+0)/6 = 0.5
   3. cov(x,y) = 1/(6-1)\*((0-0.5)(1-0.5)+(1-0.5)(0-0.5)+(0-0.5)(1-0.5)+(1-0.5)(0-0.5) +(0-0.5)(1-0.5)+(1-0.5)(0-0.5)) = 1/5 \* (-1.5) = -0.3
   4. stdev(x) = sqrt(((-0.5^2 + 0.5^2 + -0.5^2 + 0.5^2 + -0.5^2 + 0.5^2)/6) = 0
   5. stdev(y) = 0
   6. correlation is undefined as stdev for both vectors is zero
3. Euclidian Distance = sqrt((1-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2)
   1. = sqrt(1-1+1-1+1-1) = sqrt(0) = 0
4. Jaccard = (number of matching presences) / (number of attributes not involved in 00 matches)
   1. = (f11)/(f01 + f10 + f11)
   2. = 0/(3 + 3 + 0) = 0