Non-Euclidian Distance

B ANIRUDH SRINIVSAN
COE17B019

1. Introduction

Euclidean distances are special because they conform to our physical concept of distance. But there are many other distance measures which can be defined between multivariate samples. These non-Euclidean distances are of different types

Metric

Non-Metric

2. Metric and Non-Metric distances

2.1 Metric

Distances that follow the following axioms are called metric distances,

1. $d_{ab} = d_{ba}$
2. $d_{ab} > 0 \; and = 0 \; if \; and \; only \; if \; a \; = \; b$
3. $dabd_{ac} + d_{ca}$

where, d_{ab} is the distance measured from point a to point b.

2.1.1 Some Example of Metric Distances

Mahalanobis distance

The Mahalanobis distance is a measure of the distance between a point P and a distribution D. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D. This distance is zero if P is at the mean of D, and grows as P moves away from the mean along each principal component axis. If each of these axes is re-scaled to have unit variance, then the Mahalanobis distance corresponds to standard Euclidean distance in the transformed space. The Mahalanobis distance is thus unitless and scale-invariant, and takes into account the correlations of the data set.

The Mahalanobis distance of an observation $\vec{x} = (x_1, x_2, x_3, \ldots, x_N)^T$ *from a set of observations with mean* $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$ *and covariance matrix* S *is defined as* :

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})S^{-1}(\vec{x} - \vec{\mu})} \tag{1}$$

Mahalanobis distance (or "generalized squared interpoint distance" for its squared value) can also be defined as a dissimilarity measure between two random vectors and of the same distribution with the covariance matrix S:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}. \tag{2}$$

Example,

$$\begin{bmatrix} 64.0 & 580.0 & 29.0 \\ 66.0 & 570.0 & 33.0 \\ 68.0 & 590.0 & 37.0 \\ 69.0 & 660.0 & 46.0 \\ 73.0 & 600.0 & 55.0 \end{bmatrix}$$

we need to find the Mahalanobis Distance of point (66, 640, 44) is from this data mean of the data = (68.0, 600.0, 40.0) clearly, number of observation, N = 5

$$Z = \begin{bmatrix} x1 - \mu_x & y1 - \mu_y & z1 - \mu_z \\ x2 - \mu_x & y2 - \mu_y & z2 - \mu_z \\ x3 - \mu_x & y3 - \mu_y & z3 - \mu_z \\ x4 - \mu_x & y4 - \mu_y & z4 - \mu_z \\ x5 - \mu_x & y5 - \mu_y & z5 - \mu_z \end{bmatrix} \tag{3}$$

Therefore,

$$Z = \begin{bmatrix} -4 & -20 & -11 \\ -2 & -30 & -7 \\ 0 & -10 & -3 \\ 1 & 60 & 6 \\ 5 & 0 & 15 \end{bmatrix}$$

$$S = \frac{1}{N-1} Z^T Z \tag{4}$$

In our case,

$$S = \begin{bmatrix} 11.5 & 50 & 34.75 \\ 50 & 1250 & 205 \\ 34.25 & 205 & 110 \end{bmatrix}$$

Similarly,

$$S^{-1} = \begin{bmatrix} 3.6885 & 0.0627 & -1.2821 \\ 0.0627 & 0.0022 & 0.0240 \\ -1.2821 & -0.0240 & 0.4588 \end{bmatrix}$$

By eq(2),

$$D = \sqrt{[66 - 68 \;\; 640 - 600 \;\; 44 - 40]^T S^{-1} [66 - 68 \;\; 640 - 600 \;\; 44 - 40]}$$

$$D = 5.33$$

Jaccard Similarity Measure

The Jaccard similarity index (sometimes called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It's a measure of similarity for the two sets of data, with a range from 0% to 100%. The higher the percentage, the more similar the two populations. Although it's easy to interpret, it is extremely sensitive to small samples sizes and may give erroneous results, especially with very small samples or data sets with missing observations.

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{5}$$

A similar statistic, the Jaccard distance, is a measure of how dissimilar two sets are. It is the complement of the Jaccard index and can be found by subtracting the Jaccard Index from 100% or 1.

$$D(X,Y) = 1 - J(X,Y) \tag{6}$$

Example,
A = {0,1,2,5,6}, B = {0,2,3,4,5,7,9}
According to eq(5) and eq(6),

$$\begin{aligned} J(A,B) &= \frac{|\{0,2,5\}|}{|\{0,1,2,3,4,5,6,7,9\}|} \\ J(A,B) &= 0.33 \\ D(A,B) &= 1 - 0.33 \\ D(A,B) &= 0.67 \end{aligned}$$

2.2 Non-Metric

Distances that fail to follow any one of the axioms are deemed as non-metric.

2.2.1 Some Examples of Non-Metric Distances

Cosine Similarity and Distance

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine similarity $S_C(A,B)$ and $D_C(A,B)$ for two vectors A and B are given by,

$$S_C(A,B) = cos(\Theta) = \frac{A.B}{||A||.||B||} \tag{7}$$

$$D_C(A,B) = 1 - S_C(A,B) \tag{8}$$

The term "cosine similarity" is sometimes used to refer to a different definition of similarity provided below. However the most common use of "cosine similarity" is as defined

above and the similarity and distance metrics defined below are referred to as "angular similarity" and "angular distance" respectively. The angular similarity$S_A(X,Y)$ and angular distance $D_A(X,Y)$ for two vectors X and Y are,

When the vector elements may be positive or negative:

$$D_A(X,Y) = \frac{cos^{-1}(S_C(X,Y))}{\pi} \tag{9}$$

$$S_A(X,Y) = 1 - D_A(X,Y) \tag{10}$$

Or, if the vector elements are always positive:

$$D_A(X,Y) = 2 * \frac{cos^{-1}(S_C(X,Y))}{\pi} \tag{11}$$

$$S_A(X,Y) = 1 - D_A(X,Y) \tag{12}$$

Example,
$\vec{A} = (1,0,1)$, $\vec{B} = (2,1,2)$
By eq(9) and eq(10)

$$cos(\Theta) = \frac{(1,0,1).(2,1,2)}{|(1,0,1)| * |(2,1,2)|}$$

$$cos(\Theta) = \frac{4}{3 * \sqrt{2}}$$

$$cos(\theta) = 0.94$$

$$D_C(A,B) = 1 - 0.94 = 0.06$$

Bhattacharyya distance

In statistics, the Bhattacharyya distance measures the similarity of two probability distributions. It is closely related to the Bhattacharyya coefficient which is a measure of the amount of overlap between two statistical samples or populations.

The coefficient can be used to determine the relative closeness of the two samples being considered. It is used to measure the separability of classes in classification and it is considered to be more reliable than the Mahalanobis distance, as the Mahalanobis distance is a particular case of the Bhattacharyya distance when the standard deviations of the two classes are the same. Consequently, when two classes have similar means but different standard deviations, the Mahalanobis distance would tend to zero, whereas the Bhattacharyya distance grows depending on the difference between the standard deviations.

For probability distributions p and q (discrete distribution) over the same domain X, the Bhattacharyya distance is defined as:

$$D_B(p,q) = -ln(BC(p,q)) \tag{13}$$

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)} \tag{14}$$

For probability distributions p and q (continuous distribution) over the same domain X, the Bhattacharyya distance is defined as:

$$D_B(p,q) = -ln(BC(p,q)) \tag{15}$$

$$BC(p,q) = \int \sqrt{p(x)q(x)}dx \tag{16}$$

Example,
P = [0.25 0.5 0.25], Q = [0.3 0.4 0.3]
From eq(13) and eq(14),

$$BC(P,Q) = \sqrt{0.25 * 0.3} + \sqrt{0.5 * 0.4} + \sqrt{0.25 * 0.3}$$
$$BC(P,Q) = 0.883$$
$$D_C(P,Q) = -ln(0.833) \qquad D_C(P,Q) = 0.124$$

Pearson correlation

In statistics, the Pearson correlation coefficient is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and 1, where 1 is total positive linear correlation, 0 is no linear correlation, and 1 is total negative linear correlation. It is widely used in the sciences.

Pearson's correlation coefficient when applied on a pair of random variables (X,Y), is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{17}$$

where,

- cov is the covariance.
- σ_X is the standard deviation of X.
- $\sigma_Y is the standard deviation of Y.$

The Pearson correlation distance is defined as,

$$D_P(X,Y) = 1 - \rho(X,Y) \tag{18}$$

Example,

Table 1: Example dataset

x	y
43	99
21	65
25	79
42	75
57	87
59	81

Table 2: Calculating xy, $x^2 and y^2$

x	y	xy	x^2	y^2
43	99	4257	1849	9801
21	65	1365	441	4225
25	79	1975	625	6241
42	75	3150	1764	5625
57	87	4959	3249	7569
59	81	4779	3481	6561

Table 3: Summation of all values

	x	y	xy	x^2	y^2
	43	99	4257	1849	9801
	21	65	1365	441	4225
	25	79	1975	625	6241
	42	75	3150	1764	5625
	57	87	4959	3249	7569
	59	81	4779	3481	6561
$\sum$	247	486	20485	11409	40022

From eq(17) and (18),

$$\rho_{x,y} = \frac{2868}{5413.27}$$
$$\rho_{x,y} = 0.53$$
$$D_p(X,Y) = 1 - 0.53$$
$$D_p(X,Y) = 0.47$$