Contents

1 Euclidean Distance

The Euclidean distance between two points, x and y, with k dimensions is calculated as:

$$\sqrt{\sum_{j=1}^{k}(x_j - y_j)^2}$$

The Euclidean distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity.

1.1 Example

Compute the Euclidean Distance between the following points: $X = \{1, 5, 4, 3\}$ and $Y = \{2, 1, 8, 7\}$

x_j	y_j	$x_j - y_j$	$(x_j - y_j)^2$
1	2	-1	1
5	1	4	16
4	8	-4	16
3	7	-4	16
			49

The Euclidean Distance between the two points is $\sqrt{49}$ i.e. 7.

1.2 Squared Euclidean Distance

The Squared Euclidean distance between two points, x and y, with k dimensions is calculated as:

$$\sum_{j=1}^{k}(x_j - y_j)^2$$

The Squared Euclidean distance may be preferred to the Euclidean distance as it is slightly less computational complex, without loss of any information.

2 Standardized Euclidean distance

Let us consider measuring the distances between between two points using the three continuous variables pollution, depth and temperature. Let us suppose that a difference of 4.1 in terms of pollution is considered quite large and unusual, while a difference of 48 in terms of depth is large, but not particularly unusual. What would happen if we applied the Euclidean distance formula to measure distance between two cases.

Variables	case 1	case 2
Pollution	6.0	1.9
Depth	51	99
Temp	3.0	2.9

Here is the calculation for Euclidean Distance:

$$d = \sqrt{(6.0 - 1.9)^2 + (51 - 99)^2 + (3.0 - 2.9)^2}$$

$$d = \sqrt{16.81 + 2304 + 0.01} = \sqrt{2320.82} = 48.17$$

The contribution of the second variable depth to this calculation is huge one could say that the distance is practically just the absolute difference in the depth values (equal to —51-99— = 48) with only tiny additional contributions from pollution and temperature. These three variables are on completely different scales of measurement and the larger depth values have larger differences, so they will dominate in the calculation of Euclidean distances.

The approach to take here is **standardization**, which is is necessary to balance out the contributions, and the conventional way to do this is to transform the variables so they all have the same variance of 1. At the same time we ***center*** the variables at their means this centering is not necessary for calculating distance, but it makes the variables all have mean zero and thus easier to compare.

The transformation commonly called standardization is thus as follows:

$$\text{standardized value} = \frac{\text{observed value} \;\; \text{mean}}{\text{standard deviation}}$$

Variables	Case 1	Case 2	Mean	Std. Dev	Case 1 (std)	Case 2 (std)
Pollution	6.0	1.9	4.517	2.141	0.693	-1.222
Depth	51	99	74.433	15.615	-1.501	1.573
Temp	3.0	2.9	3.057	0.281	-0.201	-0.557

$$d_{std} = \sqrt{(0.693 - (-1.222))^2 + (-1.501 - 1.573)^2 + (-0.201 - (-0.557))^2}$$

$$d_{std} = \sqrt{3.667 + 9.449 + 0.127} = \sqrt{13.243} = 3.639$$

Pollution and temperature have higher contributions than before but depth still plays the largest role in this particular example, even after standardization. But this contribution is justified now, since it does show the biggest standardized difference between the samples.

3 Manhattan (City Block) Distance

The City block distance between two points, x and y, with k dimensions is calculated as:

$$\sum_{j=1}^{k} |x_j - y_j|$$

The City block distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity.

3.1 Example

Compute the Manhattan Distance between the following points: $X = \{1, 3, 4, 2\}$ and $Y = \{5, 2, 5, 2\}$

x_j	y_j	$x_j - y_j$	$\lvert x_j - y_j \rvert$
1	5	-4	4
3	2	1	1
4	5	-1	1
2	2	0	0
			6

The Manhattan Distance between the two points is 6.

4 Cluster Analysis : Proximity Matrices

Using ***nearest neighbour*** linkage, describe how the agglomeration schedule based on the following proximity matrix. With nearest neighbour, a case is assigned to the cluster of the case with which it has the shortest distance. Cluster are also joined on this basis.

Case	1	2	3	4	5	6	7	8	9	10
1	0.00	**4.82**	89.39	85.97	46.26	71.87	56.42	23.75	31.57	11.70
2	**4.82**	0.00	94.24	38.96	**5.55**	35.07	74.52	71.27	61.84	**4.84**
3	89.39	94.24	0.00	57.65	27.27	25.31	20.89	**2.84**	63.50	89.39
4	85.97	38.96	57.65	0.00	**22.94**	**7.13**	70.49	23.09	**12.75**	85.97
5	46.26	**5.55**	27.27	**22.94**	0.00	39.44	17.43	79.22	14.47	46.26
6	71.87	35.07	25.31	**7.13**	39.44	0.00	27.50	30.65	13.34	71.87
7	56.42	74.52	20.89	70.49	17.43	27.50	0.00	91.16	44.92	**6.42**
8	23.75	71.27	**2.84**	23.09	79.22	30.65	91.16	0.00	**3.18**	23.75
9	31.57	61.84	63.50	**12.75**	14.47	13.34	44.92	**3.18**	0.00	31.57
10	11.70	**4.84**	89.39	85.97	46.26	71.87	**6.42**	23.75	31.57	0.00

- The closest pair in terms of distance (2.84) are cases 3 and 8. So this is the first linkage.
- The next closest pair (3.18) are 8 and 9. The next linkage joins case 9 to 3 and 8.
- The next closest pair (4.82) are 1 and 2. So this is the next linkage. [So far (3,8,9) and (2,10)]
- The next closest pair (4.84) are 2 and 10. The next linkage joins case 1 to 2 and 10.
- The next closest pair (5.55) are 2 and 5. The next linkage joins case 5 to 1, 2 and 10. [So far (3,8,9) and (1,2,5,10)]
- The next closest pair (6.42) are 7 and 10. The next linkage joins case 7 to 1, 2, 5 and 10.
- The next closest pair (7.13) are 4 and 6. The next linkage joins case 4 to 6. [So far (3,8,9), (4,6) and (1,2,5,10) All cases are in clusters. This is a 3 cluster solution.]

- The next closest pair (11.70) are 1 and 10. Disregard, because they are already clustered together.
- The next closest pair (19.44) are 4 and 9. This joins cluster (4,6) to cluster (3,8,9) [So far (3,4,6,8,9) and (1,2,5,10). This is a 2 cluster solution.]
- The next closest pairing is 4 and 5. This linkage joins all cases together in one cluster.

5 Logistic Regression: Odds Ratios and Log-Odds

Suppose that in a sample of 100 men, 90 drank wine in the previous week, while in a sample of 100 women only 20 drank wine in the same period. The odds of a man drinking wine are 90 to 10, or 9:1, while the odds of a woman drinking wine are only 20 to 80, or 1:4 = 0.25:1. The odds ratio is thus 9/0.25, or 36, showing that men are much more likely to drink wine than women. The detailed calculation is:

$$\frac{0.9/0.1}{0.2/0.8} = \frac{0.9 \times 0.8}{0.1 \times 0.2} = \frac{0.72}{0.02} = 36$$

This example also shows how odds ratios are sometimes sensitive in stating relative positions: in this sample men are 90/20 = 4.5 times more likely to have drunk wine than women, but have 36 times the odds.

The logarithm of the odds ratio, the difference of the logits of the probabilities, tempers this effect, and also makes the measure symmetric with respect to the ordering of groups. For example, using natural logarithms, an odds ratio of 36/1 maps to 3.584, and an odds ratio of 1/36 maps to -3.584.

6 Logistic Regression: Logits

The logit transformation is given by the following formula:

$$\eta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

To inverse of the logit transformation is given by the following formula:

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

6.1 Example 1

Given that $\pi_i = 0.2$, compute η_i.

$$\eta_i = \log\left(\frac{0.2}{1-0.2}\right) = \log\left(\frac{0.2}{0.8}\right)$$

$$\eta_i = \log(0.25) = -1.386$$

6.2 Example 2

Given that $\eta_i = 2.3$, compute π_i.

$$\pi_i = \frac{e^{2.3}}{1 + e^{2.3}} = \frac{9.974}{1 + 9.974} = 0.908$$

7 Logistic Regression

In logistic regression, the logit may be computed in a manner similar to linear regression:

$$\eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

7.1 Example 2

Let us suppose that the probability of survival of a marine species of fauna is dependent on pollution, depth and water temperature. Suppose the logit for the logistic regression was computed as follows:

$$\eta_i = 0.14 + 0.76x_1 - 0.093x_2 + 1.2x_3$$

Variables	case 1	case 2
Pollution(x_1)	6.0	1.9
Depth (x_2)	51	99
Temp (x_3)	3.0	2.9

Compute the probability of success for both case 1 and case 2.

- case $1 \eta_1 = 0.14 + (0.76 \times 6) - (0.093 \times 51) + (1.2 \times 3) = 3.557$
- case $2 \eta_2 = 0.14 + (0.76 \times 1.9) - (0.093 \times 99) + (1.2 \times 2.9) = -4.143$

The probabilities for success are therefore:

$$\pi_1 = \frac{e^{3.557}}{1 + e^{3.557}} = \frac{35.057}{1 + 35.057} = 0.972$$

$$\pi_2 = \frac{e^{-4.143}}{1 + e^{-4.143}} = \frac{0.0158}{1 + 0.0158} = 0.0156$$