Assignment part – 2

Clustering of Trajectories

7. Solution:

The dissimilarity measure is a numerical representation of how different two data items are, and it typically decreases as objects or data samples become more similar. The minimum dissimilarity is frequently 0, whereas the maximum limit varies depending on how much variance may be accepted. Outliers, cluster boundaries, and exceptions, such as traffic networks, are typically found using dissimilarity measurements.

We can deduce from the provided information that we will be performing our computations in 2D, therefore it makes more sense to use the Euclidean Distance dissimilarity measure to determine the dissimilarity between two points, P1(x, y) and P2(x,y) (x,y).

The following formula can be used to do the calculations:

𝒅 (𝑷, 𝑸) = ‖𝑷 − 𝑸‖𝟎 **=** J∑𝟐 (𝒑𝒊 − 𝒒𝒊)𝟐

𝒊=𝟏

**= ƒ**(𝒑𝟏 − 𝒒𝟏)𝟐 + (𝒑𝟐 − 𝒒𝟐)𝟐

**Where:** 𝑷 = (𝒑𝟏, 𝒑𝟐) **and** 𝑸 = (𝒒𝟏, 𝒒𝟐)

Let the two be P(x, y) and Q(x,y) where x is longitude and y is latitude.

P (39.90729,116.370055) & Q(39.97503333,116.3421)

As a result of the values, the response is 0.073, indicating that the two randomly selected data points are more similar.

Because the data points satisfy the three needed characteristics, we can call given dissimilarity a measure.

• Non-negativity: for every two different observations p and q, d(p, q) 0.

• Symmetry: for any p and q, d (p, q) = d(q, p).

• If p = q, d (p, q) = 0.

8. Solution:

We can be certain that the dissimilarity measure we chose is a distance metric because it fulfils the triangle inequality d (p, q) d(p, r) + d(r, q) for all p, q, r.

Let C, D, and E be concept descriptions, and dp (C, D) = d1, dp (D, E) = d2 to show the triangle inequality. Then, in particular, dd (C, D) d and, as a result, by extensibility. In the same way, we get. We obtain from the relaxations since they are non-decreasing.

In a similar vein, it may be demonstrated that and thus.

The triangle inequality metric is clearly used in the Euclidean Dissimilarity measure as a result of the following findings.

9. Solution:

The K means clustering approach will be the most acceptable and suitable clustering algorithm to apply based on our main aim, which is "for any given user, we would like to group days that have similar trajectories."

The main reason for this is that the K-Means approach - a vector quantization method commonly used as a clustering method - does not explicitly use pairwise distances between data points (in contrast to hierarchical and other clustering methods, which allow for any proximity metric). It entails assigning points to the nearest centroid over and over again, based on the Euclidean distance between data points and a centroid. Because the total of squared deviations from the centroid equals 1, K-Means is implicitly based on pairwise Euclidean distances between data points.

What makes k-means such a "great" solution?

We understand this as within cluster variation (WCV) for a single cluster.

The K Means algorithm aims to reduce within-cluster variation while increasing between-cluster variation.

Step 1: Choose n points at random from the points to be clustered as the cluster's centre, where n is the number of clusters necessary.

Step 2: Calculate the distance between all of the spots using the random centres as a guide.

Step 3: Assign the points to the cluster with which it is most closely associated.

The first iteration is now complete, and we have n clusters with randomly allocated centres.

References

1. <https://liverpool.instructure.com/courses/48071/files/6339215?module_item_id=1300069>
2. <https://link.springer.com/article/10.1007/s00357-020-09373-2>
3. <https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>
4. 17 types of similarity and dissimilarity measures used in data science. By Mahmoud Harmouch