Big Data Analytics Assignment -2

Part 2 :- Clustering of Trajectories

Task-7

Consider the attached dataset for each user, which contains day trajectories (a list of (timestamp, longitude, latitude)) for each user. Create a dissimilarity metric that can quantify how dissimilar two day trajectories are from one another. Your measure should be symmetric (i.e., d(x,y) = d(y,x) for all trajectories x,y), and its value should be 0 if and only if two trajectories are the same (i.e., d(x,y) = 0 if and only if x=y). Argue (or prove, if you like) why this measure has these features, is appropriate for this job, and is resistant to data mistakes (e.g., duplicated timepoints, adding timepoints in the middle that do not really change the trajectory).

Solution :-

The dissimilarity measure is a numerical measure of how distinct two data items are, and it normally decreases as objects or data samples get more similar. The lowest dissimilarity is often 0, while the maximum limit varies depending on the amount of variation that may be tolerated.

Dissimilarity metrics are mostly used to find outliers, cluster borders, or exceptions, such as traffic network.

The Dissimilarity metric can be calculated in a variety of methods. Depending on the type of dataset, these many solutions are more appropriate.

The following formula can be used to do the calculations:

$$d(P, Q) = \|P - Q\|_0 = J\overline{\sum_{i=1}^{2} (p_i - q_i)^2}$$

$$= f\overline{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

Where: $P = (p_1, p_2)$ **and** $Q = (q_1, q_2)$

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)

Therefore, based on the values the answer is 0.073 which states that the two randomly chosen data points are more alike.

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

Non-negativity: d(p, q) ≥ 0, for any two distinct observations p and q.
Symmetry: d(p, q) = d(q, p) for all p and q.
d(p, q) = 0 only if p = q.

Task -8

Is your measure a distance metric, i.e., does it satisfy the triangle equality (d(x,y)= d(x,z) + d(z,y) for all trajectories x,y,z). f yes, explain why (or prove it if you choose), and if not, provide a counterexample.

Solution :-

Because it fulfils the triangle inequality d(p, q) d(p, r) + d(r, q) for all p, q, r, we can certainly assert that the dissimilarity measure we picked is a distance metric.

To prove the triangle inequality, let C, D, E be concept descriptions and let $d_{\mathrm{p}}(C, D) = d_1$, $dp(D, E) = d_2$. Then in particular, $d^{\mathrm{d}}(C, {}_1D) \leq d$ and thus $C \sqsubseteq \rho^{d_1}(D)$ by extensivity. Similarly, we obtain $D \sqsubseteq \rho^{d_2}(E)$Since relaxations are non-decreasing we obtain from the

$$C \sqsubseteq \rho^{d_1+d_2}(E) \qquad d^d_\rho(C, E) \leq d_1 + d_2$$

$$d_\rho(C, E) \leq d_1 + d_2 = d_\rho(C, D) + d_\rho(D, E)$$

latter and therefore, i.e. Analogously, it can be shown that and thus .

As a result of the above facts, it is evident that the Euclidean Dissimilarity measure is based on the triangle inequality metric.

Task -9

We'd want to put days with similar trajectories together for each user. In this instance, which of the clustering techniques discussed in the lectures would you use and why?

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 want 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 employ 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 centroid equals the sum of pairwise squared Euclidean distances divided by the number of points, K-Means is implicitly dependent on pairwise Euclidean distances between data points. Euclidean geometry is the source of the name "centroid." In Euclidean space, it is a multivariate mean.

How can k-means be considered a "excellent" solution?
Within cluster variation (WCV) for a specific cluster is how we comprehend this.
The goal of K Means algorithm is to minimize the Within Cluster Variation and maximize the Between Cluster Variation.

Step 1: Out of the points to be clustered, randomly assign n points as the center of the cluster where n is the number of clusters required.

Step 2: Find the distance between all the points with the points chosen as random centers.

Step 3: Assign the points to the cluster which it is closest to.This completes the 1st iteration, now we have n clusters with randomly assigned centers.

The WCV decreases with each iteration of the k-means algorithm.
In principle, we may aim for the lowest overall WCV possible (also called inertia).
However, k-means typically only finds local minima.

Limitation of K Means:-
Only one measure of dissimilarity is supported by means: Distance in Euclidean space
As a result, issues requiring non-Euclidian dissimilarity measurements necessitate the use of a different clustering technique.
Although the Euclidian distance is determined for all real vectors, it is not a useful measure of dissimilarity for all real vectors. E.g.:\sTimeseries: A considerably better metric of dissimilarity is correlation.

Rerfrences:-

https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

https://liverpool.instructure.com/courses/48071/files/6339215?module_item_id=1300069

https://link.springer.com/article/10.1007/s00357-020-09373-2