Clustering Methods

Exercises 4/7, 15.2.2022

Five tasks required. If done more, each earn one bonus point.

1. Give equations for calculating (a) sum of squared error (SSE) for a given cluster, (b) sum of all pairwise distances in the cluster. How these two relate to each other?

2. Show counter examples of data for which the following cost functions would fail: (a) single link, (b) complete link, (c) average link, (d) SSE.

3. Show an example of a cluster in 2-d Euclidean space where Medoid would be better cluster representative than Centroid. Show also an example when Medoid would not be a good choice?

4. Implement possibility to use text data in your clustering algorithm. Implement edit distance function and calculate all pairwise distances of the items within these two clusters:

   A={ireadL, relanE, rlanZd, irelLITnd};
   B={fiInVlLand, filanNM, finPAlaQd, finlCnUd}.

5. Find Medoids for the above two clusters. What is the cost of these two clusters? Would it be possible to find better cluster representative with lower cost? Use the edit distance function of previous task.

Bonus tasks:

6. Implement function CentroidDistance that takes two clusters centroids as input and calculates their squared distance multiplied by their sizes. Make another function that calculates pairwise distances of clusters. Show the results for the k-means result for S2 dataset.

Bonus task: