Clustering Methods

Exercises 1/7, 18.1.2022

1. Select your favorite programming language and implement clustering tool that is capable of reading text data form this page: http://cs.uef.fi/sipu/datasets/ (e.g. s1.txt), and output clustering result in two output files: centroid.txt and partition.txt. Implement dummy clustering algorithm that selects *k* (user given parameter) random data points and (a) outputs them as the centroids. (b) Select the partition of each data point randomly (random number from 1 to *k*) and output the result into the partition file accordingly.

   Output format for partition should be a list of integers each on its own line. For centroids, each centroid should be on its own line where the attribute values (integer or float) are separated by spaces.

2. Create your own data set using this tool: http://cs.uef.fi/paikka/Radu/tools/demonstrator/. Save the data and put it into the *Clusterator*: http://cs.uef.fi/paikka/Radu/clusterator/. Using Clusterator perform clustering of your data (by pressing the play button) and make screenshot of the result. Then compare the result to your dummy algorithm by drag-and-dropping (a) the centroid.txt into the centroid box; (a) the partition.txt to the data labels box. Make similar screenshots and calculate how much worse is the *mse* (in percentage) of the dummy result compared to the real clustering result.

3. Implement two functions to your software: (a) distance function, (b) sum-of-squared errors. The first one takes any two data objects (or centroids) as input and output their Euclidean distance. The second takes both data *and* the set of centroids as input. It then calculates the sum of *squared* distances between each data object and one *randomly* chosen centroid. Using your data, calculate pairwise distances between all data points, and report the average distance.

4. Use the tool https://cs.uef.fi/~ashfaq/ and select k-means, dataset S2, and 15 clusters. (a) What is the smallest SSE- and CI-values you can reach? (b) What are the highest SSE- and CI-values you can reach? Only values when the status=*converged* can be used.

5. Four different data is shown below. Analyze each of them according to: (a) what is the dimensionality of the data, (b) what are attributes you can find, (c) how many clusters?