# Clustering Methods

Exercises 3/7, 20.3.2017

1. Implement Centroid Index (CI) with one of the following tools: **C**, **Java**, **Matlab**, **R, Excel**.

2. Study the behaviour of k-means by measuring TSE and CI values using selected datasets among A1, A2, A3, S1, S2, S3, S4, Birch1, Birch2, Dim32 and Unbalance. Report the results. If you did not have own implementation, use any existing one. The datasets and ground truth centroids can be found here: http://cs.uef.fi/sipu/datasets/

3. K-means can be initialized in the following ways:
   - Select *k* random points as centroids
   - Assign points randomly into k clusters. Calculate the centroids of the clusters.
   - Select center point of entire dataset as first cluster. Then select furthest point to this as the 2nd centroid. Then at every step, calculate the distance of every point to its nearest centroid. Select the one with biggest distance as the centroid.
   - For every point *x*, finding its 5-nearest neighbors. Calculate the average distance between x and its neighbors. This value represents the density of the point. Then sort the points in descending ordeer according to its density. Select the k biggest densities as the centroids.

   Select one method that you think is bad, and another one that you think is the best among these. Show by counter example why the bad method does not work. Also explain why you expect the other method works best.

4. Implement your favourite initialization method (other than the *k* random points) in your k-means solution and show its behaviour with selected datasets from those above.

5. K-means can be improved simply by repeating it 100 times. For which datasets you can reach CI=0 values by this simple trick, and which not?

6. Give proof that the optimal cluster centroid for minimizing total square error (TSE) in euclidean space is the average vector. Give also a counter example that it is not optimal if when minimizing total **absolute** error.

7. Consider the following two clustering tasks: (a) list of F1 drivers, (b) programming languages. The data to be used is from Wikipedia:

   https://en.wikipedia.org/wiki/List_of_Formula_One_drivers
   https://en.wikipedia.org/wiki/Comparison_of_programming_languages

   What is the size of data (N) and dimensionality (D)? Give the type of each attribute and tell whether **average**, **mode**, and **medoid** would be sensible to use for it. Any attribute for which distribution would be useful?

8. Using the following tool, create dataset containing five of each letter: {A, A, A, A, A, B, B, …}. Divide the letters into four clusters so that the entropy value in the middle is (a) maximized, (b) minimized.

   http://cs.uef.fi/paikka/Radu/clusterator/cat.php

Send your answers as PPT file to Sami Sieranoja (samisi@cs.uef.fi) by Monday 10.00 latest. In your email, have title "*Clustering exercises 3*"