Deterministic Models and Optimization
Clustering methods

Programming assignment. Fall 2020

1 Basics

The task is to implement two popular clustering algorithms, to test them on benchmark data, and to visualize the results. We assume that we have numerical input data: N points in dimension d, that is, each data is given by a vector in $\mathbb{R}^d$. The distance between two point x and y is the usual Euclidean distance

$$d(x,y) = \left(\sum_{i=1}^{d} (x_i - y_i)^2 \right)^{1/2}.$$

Informally, the goal is to group the N points into K clusters (groups), so that points in each group are closed to each other, and points in different groups are far apart. We discuss two classical clustering methods. In both cases the number K of clusters is predetermined (but in practice it has to be adjusted to the given data).

As a general reference for clustering (and much more), see The elements of statistical learning, by Trevor Hastie, Robert Tibshirani and Jerome Friedman, available on-line at `https://web.stanford.edu/ hastie/ElemStatLearn/`

1.1 MST clustering

As discussed in class, this is based on Kruskal's algorithm for finding a Minimum Spanning Tree. The associated graph has one vertex for each point and all edges among them. The weight of an edge xy between points x and y is the distance $d(x,y)$. Points are merged according to increasing distance with the condition that no cycle is created. If we stop the algorithm at the i-th step, then there are $N - i$ components, which are taken as the clusters.

1.2 K-means clustering

Start with K initial points as the seeds of the K clusters. Iteratively assign each point to the center of its closest cluster. The algorithm terminates when the clusters do not change; see the reference above for details.

1.3 Measures for assessing the quality of a clustering

There are many measures that produce a concrete measure for how good is a clustering based on the principle "point in the same cluster are close, and those in different clusters are far". In this project we will use two measures.

The Davies-Bouldin index. It is defined as follows. For a cluster C_i its centroid is $A_i = \frac{1}{|C_i|}\sum_{x\in C_i} x$, the average of the points in the cluster. The dispersion of C_i is a measure of the scatter within the cluster and is defined as

$$S_i = \left(\frac{1}{|C_i|}\sum_{x\in C_i} d(x, A_i)\right)^{1/2},$$

where d is the Euclidean distance. The separation between two clusters C_i and C_j is defined as the distance between their centroids:

$$M_{i,j} = d(A_i, A_j).$$

Now let $R_{i,j} = \frac{S_i+S_j}{M_{i,j}}$ and $D_i = \max_{j\neq i} R_{i,j}$. Finally define the Davies-Bouldin index as

$$DB = \frac{1}{K}\sum_{i=1}^{K} D_i.$$

A low DB index is an indication of a good clustering.

The Dunn index. We let $\Delta_i = \max_{x,y\in C_i} d(x, y)$ be the maximum distance within a cluster C_i, and $\delta(C_i, C_j) = \min_{x\in C_i, y\in C_j} d(x, y)$ the distance between clusters C_i and C_J. Then the Dunn index is

$$DU = \frac{\min_{i\neq j}\delta(C_i, Cj)}{\max_i \Delta_i}.$$

A high DB index is an indication of a good clustering.

2 The assignment

- Write code for the MST clustering and for K-means clustering. The input are the N points given by their coordinates and the number K of clusters to be produced. The output is the collection of K clusters, given by the index of the points. For instance, one cluster could be $\{2, 3, 5, 8\}$, and another one $\{1, 4, 712, 13\}$.
- You don't to need to code from scratch standard procedures such as Sorting or Union-find, you can import the code from a free library. In this case you must make sure to use the right format of the parameters when calling one of these procedures.
- For each data set, compare the "quality" of the clustering by computing the DB and the DU indices.
- When the dimension is 2, draw the data sate in the plane using different colors for each cluster.

Further considerations.

- The preferred language for the code is Python, although R, C++ and Java are also allowed.
- The code must be understandable to an external reader to understand, with enough comments on what is done at each step and how is the flow of the algorithm.
- Good writing is important, both English usage and readability of the text, and will be taken into consideration in the marks.

3 Benchmark data

There two data sets, each of them given by integer valued vectors in dimensions 2 and 5, respectively. They are in the text files `Synthetic` and `Thyroid`, each row corresponding to the coordinates of one point. The number k of clusters in `Synthetic` is 15. It is unspecified in `Thyroid`; one must try different values of k and decide the value depending on the outcome of the indexes of quality. In both cases the quality indexes must be computed. For the data in `Synthetic`, the cluster must be represented graphically with different colors for different clusters.