Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Eirini Ntoutsi
PD Dr. Matthias Schubert

Knowledge Discovery in Databases II

WS 2015/2016

Übungsblatt 4: Cluster Analysis in High-Dimensional Data – CASH

Aufgabe 4-1 4C: Computing Clusters of Correlation Connected Objects

In this exercise, you will implement the algorithm 4C based on the code template *Py_4C_template.py*. The main algorithm is already coded, but there are four methods which need to be completed before the algorithm will work.

(a) Download the template and the data set. Study the code to see what it does and to understand the interfaces of the missing methods.

(b) Write a method ϵ-range query to determine the local environment of vector q.
Input: Dataset D, query vector q and range *epsilon*.
Output: A numpy matrix where each row is a close by feature vector.

(c) Implement a method to compute the local correlations for all data objects and store them in a list.
Input: Dataset D, range *epsilon*, weight of low variant dimensions *kappa*, and decision threshold *delta*.
Output: A list containing all local correlation distance matrices.

(d) Write a function for computing the correlation distance between x having local distance matrix $S1$ and y having local distance matrix$S2$.

(e) Implement a 2nd ϵ-range query on D using the local correlation distances.
Note: This time the query is given as the row index in D to allow easy localization of its local correlation matrix.

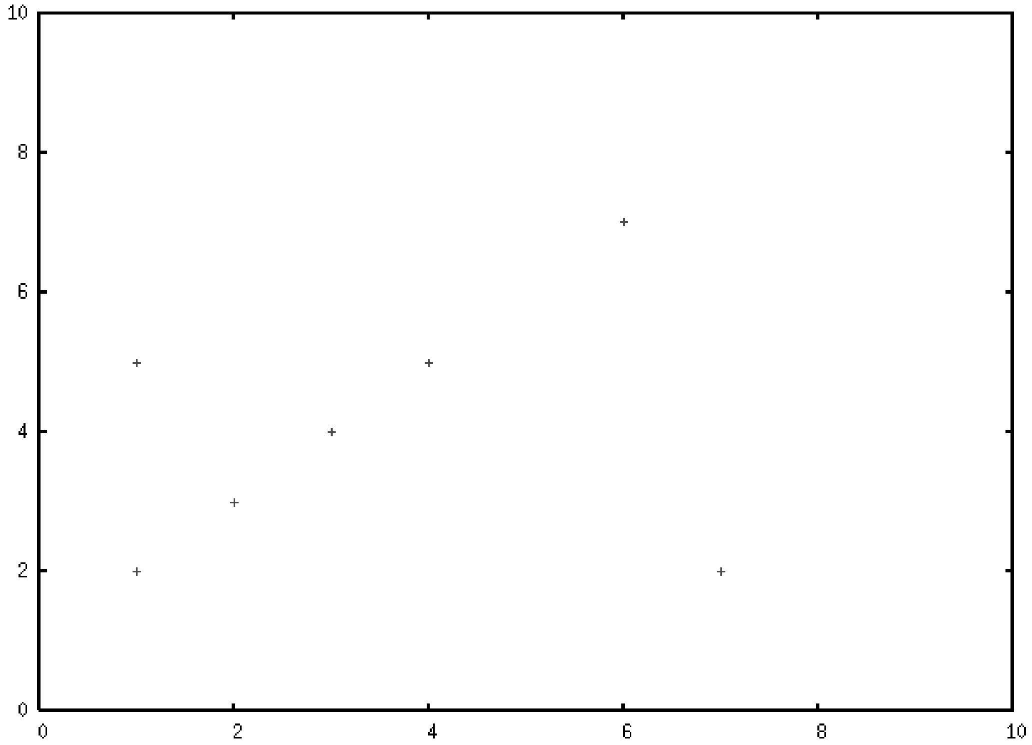
(f) Try out several parameters for delta to find the two linear correlation clusters using the 4C algorithm.

Aufgabe 4-2 CASH: Hough-Transform

Consider the data set “`cashDaten.txt`”, from the lecture website.

(To visualize the data space, use the following gnuplot command:

```
 plot [0:10][0:10] ``cashDaten.txt'' title ''  )
```



Determine the parameter space associated with this data space, i.e. for each point a parameter function of the following form:

$$f_p(\alpha_1, \ldots, \alpha_{d-1}) \quad = \quad \sum_{i=1}^{d} p_i \cdot \left(\prod_{j=1}^{i-1} \sin(\alpha_j) \right) \cdot \cos(\alpha_i)$$

(Note: $\alpha_d = 0$).

Visualize the parameter functions. Where are dense regions located?