

Knowledge Discovery in Databases II
 SS 2017

Exercise 4: High Dimensional Data Clustering

Exercise 4-1 Density-based Subspace-Clustering (SubClu)

Show that the following statement (monotonicity of the core point property) holds:

Let D be a set of d -dimensional feature vectors, \mathcal{A} the set of all attributes (dimensions/features). Further let $p \in D$ and $S \subseteq \mathcal{A}$ be a subspace (attribute subset).

Then the following holds for arbitrary $\epsilon \in \mathbb{R}^+$ and $minPts \in \mathbb{N}$:

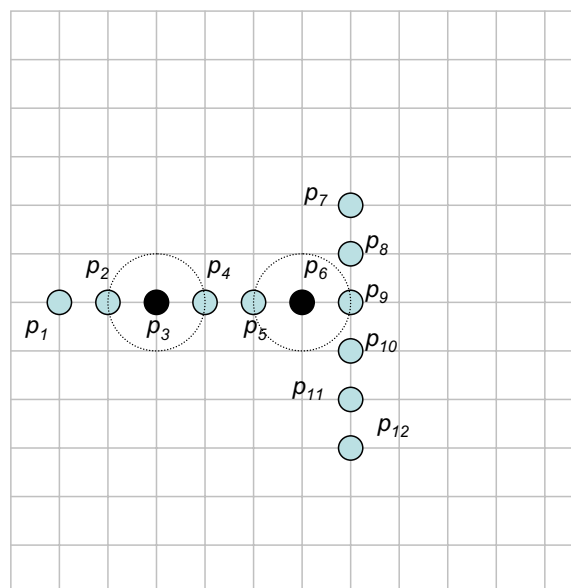
$$\forall T \subseteq S : |\mathcal{N}_\epsilon^S(p)| \geq minPts \Rightarrow |\mathcal{N}_\epsilon^T(p)| \geq minPts$$

with $|\mathcal{N}_\epsilon^S(p)| := \{q \in D \mid L_P(\pi_S(p), \pi_S(q)) \leq \epsilon\}$.

Exercise 4-2 Density-based Projected-Clustering (PreDeCon)

The algorithm PreDeCon is closely related to 4C. Instead of the expensive PCA, it uses variance analysis and a weighted Euclidean distance function: For the points in a candidate's ϵ -neighborhood, each dimension whose variance is below δ is weighted more heavily (κ).

Consider the 2D data set shown below. Assume the width of the grid to be 1 unit, use the Euclidean distance function to determine a point's ϵ -neighborhood.



Calculate, if p_3 and p_6 are core points. Assume the following parameter values: $minPts = 3, \epsilon = 1, \delta = 0.25, \lambda = 1, \kappa = 100$

Exercise 4-3 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 ϵ .

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 ϵ , weight of low variant dimensions κ , and decision threshold δ .

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.

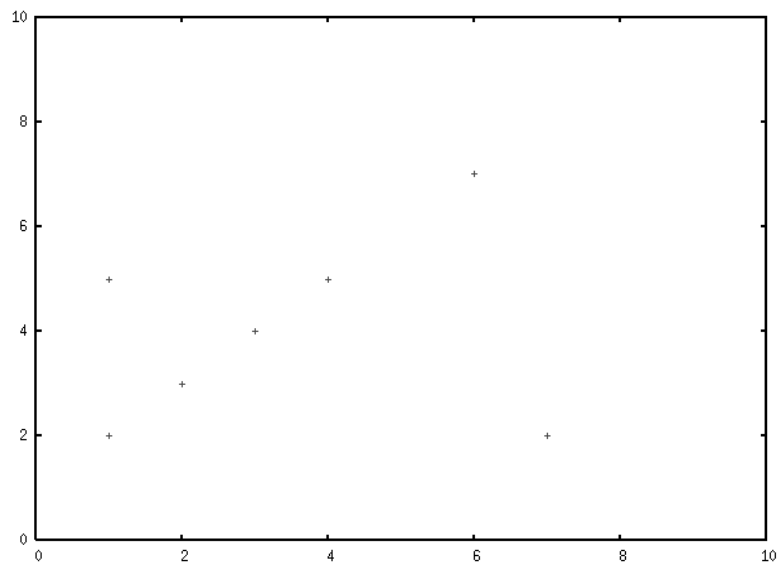
- (f) Try out several parameters for δ to find the two linear correlation clusters using the 4C algorithm.

Exercise 4-4 CASH: Hough-Transform

Consider the data set “cashDaten.txt”.

(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:

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

(Note: $\alpha_d = 0$).

Visualize the parameter functions. Where are dense regions located?