Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Yifeng Lu

Knowledge Discovery in Databases II

SS 2018

Exercise 5: High Dimensional Data Clustering

Exercise 5-1 Subspace vs Projected Clustering

Download the package 'subspace' in R and compare the results of CLIQUE, ProClus, SubClu with the given dataset provided in the package. You can also try out the package orclus.

Exercise 5-2 ProClus

V1	V2	V3	V4	V5
45	651	308	543	246
51	649	496	536	25
50	655	578	535	253
46	657	228	533	251
53	653	617	535	244
46	646	516	531	253
48	650	679	540	249
41	648	86	536	253
51	645	718	547	248
54	653	548	528	250

Try to find two 3-dim Clusters using Proclus algorithm.

Exercise 5-3 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 $minPts \in \mathbb{N}$:

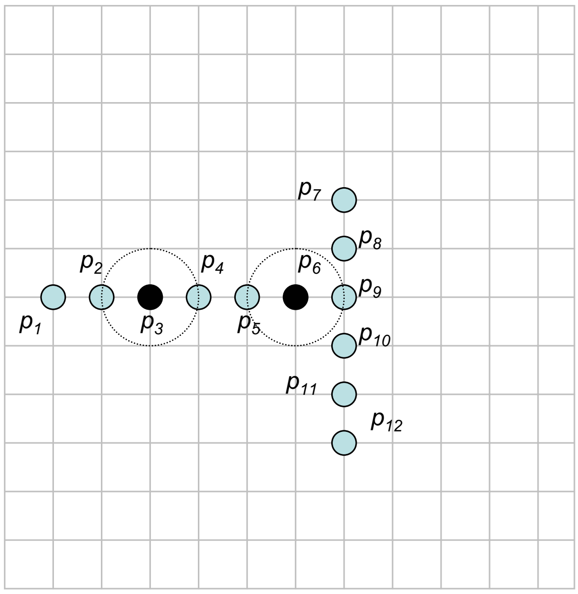
$$\forall T \subseteq S \ : \ |\mathcal{N}_\epsilon^S(p)| \geq minPts \ \Rightarrow \ |\mathcal{N}_\epsilon^T(p)| \geq minPts$$

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

Exercise 5-4 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.



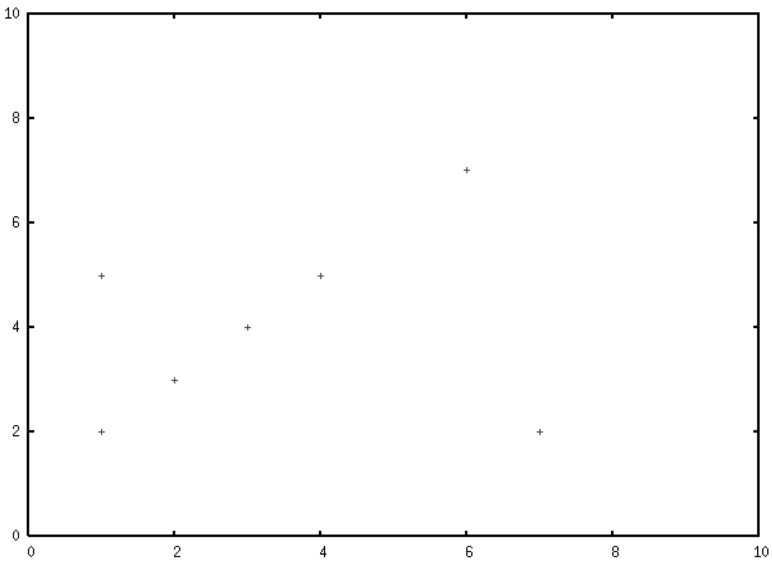
Calculate, if p_3 and p_6 are core points. Assume the following parameter values: $minPts = 3, \epsilon = 1, \delta = 0.25, \lambda = 1, \kappa = 100$

Exercise 5-5 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:

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

(Note: $\alpha_d = 0$).

Visualize the parameter functions. Where are dense regions located?