

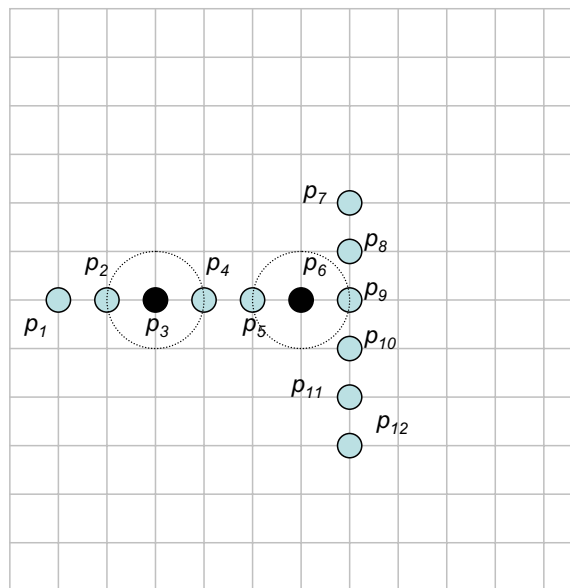
Knowledge Discovery in Databases II  
 WS 2014/2015

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

**Aufgabe 4-1 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  $\epsilon$ -neighborhood, each dimension whose variance is below  $\delta$  is weighted more heavily ( $\kappa$ ).

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  $\epsilon$ -neighborhood.



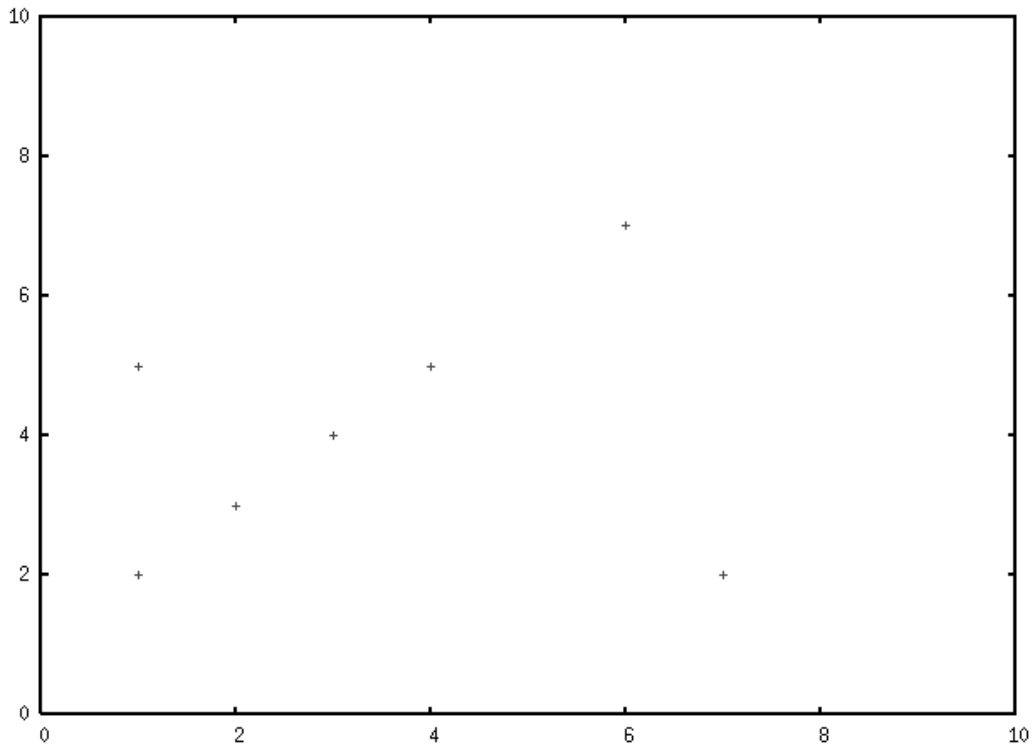
Calculate, if  $p_3$  and  $p_6$  are core points. Assume the following parameter values:  $minPts = 3, \epsilon = 1, \delta = 0.25, \lambda = 1, \kappa = 100$

### 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:

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

(Note:  $\alpha_d = 0$ ).

Visualize the parameter functions. Where are dense regions located?