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

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

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, \( A \) the set of all attributes (dimensions/features). Further let \( p \in D \) and \( S \subseteq A \) be a subspace (attribute subset).
Then the following holds for arbitrary \( \epsilon \in \mathbb{R}^+ \) and \( \text{minPts} \in \mathbb{N} \):
\[
\forall T \subseteq S : |N^S_\epsilon(p)| \geq \text{minPts} \Rightarrow |N^T_\epsilon(p)| \geq \text{minPts}
\]
with \( |N^S_\epsilon(p)| := \{ q \in D | L_P(\pi_S(p), \pi_S(q)) \leq \epsilon \} \).

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 \( \epsilon \)-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.

![Diagram of 2D data set]

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

**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 ''
```

![Plot of data space]

Determine the parameter space associated with this data space, i.e. for each point a parameter function of the following form:
\[ f_p(\alpha_1, \ldots, \alpha_{d-1}) = \sum_{i=1}^{d} p_i \cdot \left( \prod_{j=1}^{i-1} \sin(\alpha_j) \right) \cdot \cos(\alpha_i) \]

(Note: \( \alpha_d = 0 \)).

Visualize the parameter functions. Where are dense regions located?