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

Knowledge Discovery in Databases II

WS 2015/2016

Übungsblatt 3: Feature Reduction

Aufgabe 3-1 Relevant Component Analysis

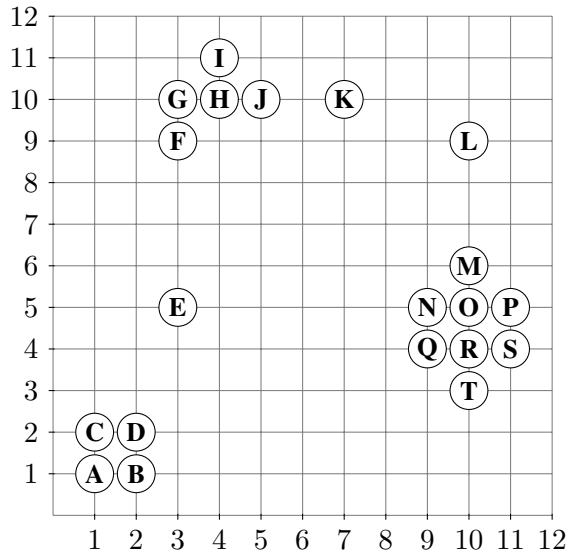
In this exercise, you will compare the result of PCA to the result of RCA as preprocessing step for a kNN classifier. To help you with the implementation download the template *RCA_template.py* from the homepage. The template reads an ARFF-file and performs a cross-validation test for a kNN classifier on the original data. Additionally, it applies PCA and RCA to the data set for all subspaces.

(a) Implement the method pca to reduce the data set to its i principal components.

(b) Implement the method rca to reduce the data set to its i most separating dimensions.

(c) Run the template and compare the resulting classification accuracies.

Aufgabe 3-2 Recapitulating DBSCAN



Compute DBSCAN on the dataset above using Manhattan distance. Indicate core points, border points, and noise points. Use the following parameters:

- Radius $\varepsilon = 1.1$ and *minPts* $= 3$
- Radius $\varepsilon = 1.1$ and *minPts* $= 4$
- Radius $\varepsilon = 2.1$ and *minPts* $= 4$

Aufgabe 3-3 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 $|\mathcal{N}_\epsilon^S(p)| := \{q \in D \mid L_P(\pi_S(p), \pi_S(q)) \leq \epsilon\}$.