

**Knowledge Discovery in Databases II**  
SoSe 2009

**Übungsblatt 3: Feature Reduktion und Clustering in hochdimensionalen Daten**

Besprechung am Donnerstag, 14.5.2009

**Aufgabe 3-1**     *Singular Value Decomposition*

Ein weiteres zentrales Konzept in der Feature Reduktion ist die Singular Value Decomposition. Gegeben sei eine Matrix  $M$  und ihre SVD-Zerlegung:

$$M = T * S * D'$$

wobei

$$M = \begin{bmatrix} 1 & 2 \\ 6 & 3 \\ 0 & 2 \end{bmatrix}$$

$$T = \begin{bmatrix} -0.2707 & 0.5458 \\ -0.9509 & -0.2797 \\ -0.1497 & 0.7899 \end{bmatrix}$$

$$S = \begin{bmatrix} 7.0257 & 0 \\ 0 & 2.1539 \end{bmatrix}$$

$$D = \begin{bmatrix} -0.8507 & -0.5257 \\ -0.5257 & 0.8507 \end{bmatrix}$$

Führen Sie nun nach dem in der Vorlesung beschriebenen Verfahren eine Reduktion auf 1 Feature durch.

**Aufgabe 3-2**     *Dichte-basiertes Subspace-Clustering (SubClu)*

Beweisen Sie die folgende Aussage (Monotonie der Kernpunkt-Eigenschaft):

Sei  $D$  eine Menge von  $d$ -dimensionalen Featurevektoren,  $\mathcal{A}$  die Menge aller Attribute (Dimensionen/Feature).

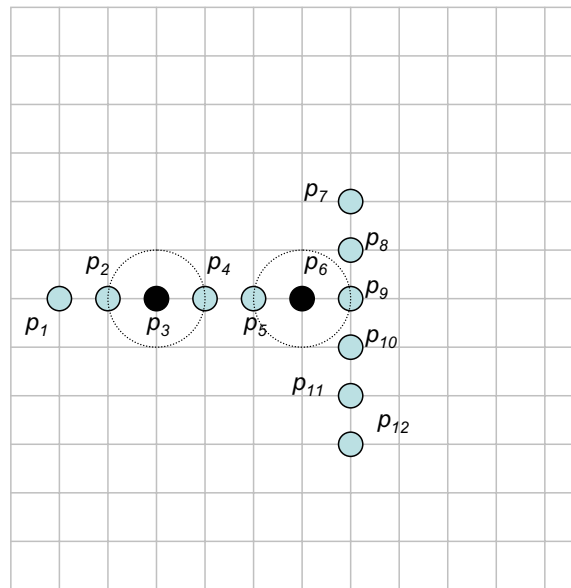
Sei weiter  $p \in D$  und  $S \subseteq \mathcal{A}$  ein Unterraum (Attribut-Teilmenge).

Dann gilt für beliebige  $\epsilon \in \mathbb{R}^+$  und  $minPts \in \mathbb{N}$ :

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

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

**Aufgabe 3-3** Dichte-basiertes Projected-Clustering (PreDeCon)



Gegeben sei obige 2D Datenmenge (der Abstand zwischen den Gitterlinien beträgt 1), die mit euklidischer Distanz verglichen werden soll. Berechnen Sie, ob  $p_3$  und  $p_6$  Kernpunkte im Algorithmus PreDeCon wären. Nehmen Sie hierzu folgende Parameterwerte an:  $minPts = 3$ ,  $\epsilon = 1$ ,  $\delta = 0.25$ ,  $\lambda = 1$ ,  $\kappa = 100$