Ludwig-Maximilians-Universität München
Institut für Informatik
Priv. Doz. Dr. Matthias Schubert

München, den 24.11.2011

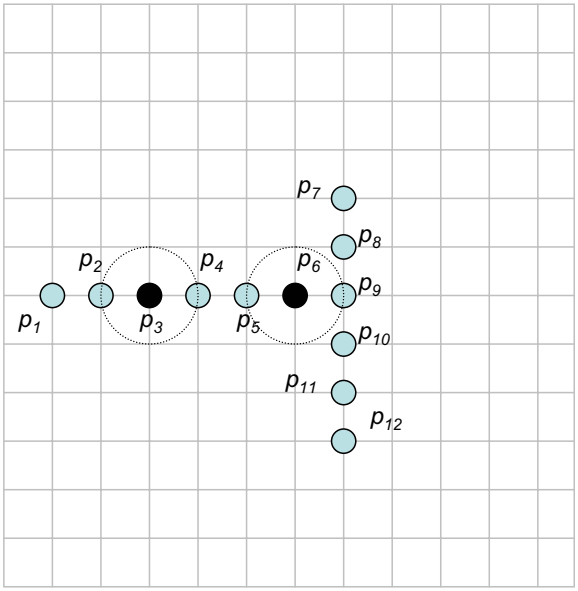
Knowledge Discovery in Databases II

WiSe 2011

Übungsblatt 4: Clustering in hochdimensionalen Räumen

Besprechung am 29.11.2011

Aufgabe 4-1 *Dichte-basiertes Projected-Clustering (PreDeCon)*



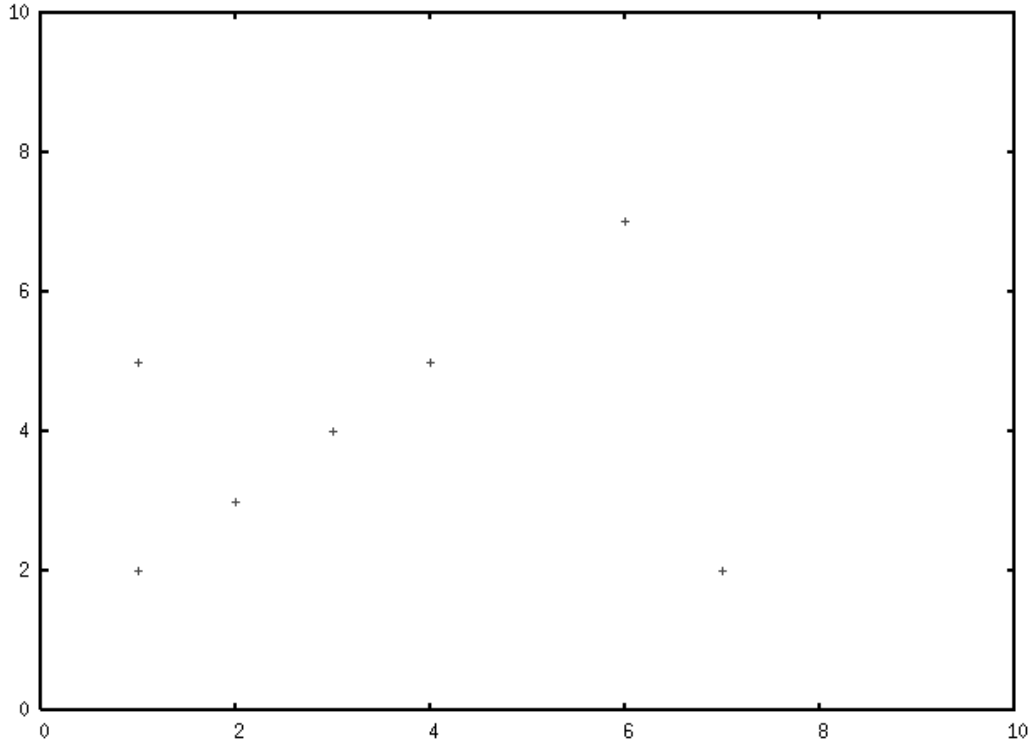
Gegeben sei obige 2D Datenmenge (der Abstand zwischen den Gitterlinien beträgt 1), die mit euklidischer Distanz verglichen werden soll. Berechnen Sie, ob p_3 und p_6 Kernpunkte im Algorithmus PreDeCon wären. Nehmen Sie hierzu folgende Parameterwerte an: $minPts = 3, \epsilon = 1, \delta = 0.25, \lambda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

Aufgabe 4-2 *CASH: Hough-Transformation*

Betrachten Sie den Datensatz "`cashDaten.txt`", den Sie auf der Vorlesungsseite finden.

Zur Visualisierung des Datenraums können Sie folgenden Befehl in gnuplot eingeben:

```
plot [0:10][0:10] "cashDaten.txt" title ''
```



Bestimmen Sie für diesen Datenraum den entsprechenden Parameterraum, d.h. für die im Datensatz enthaltenen Punkte jeweils die Parametrisierungsfunktion nach dem Schema:

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

(Beachten Sie: $\alpha_d = 0$).

Veranschaulichen Sie die Parametrisierungsfunktionen – wo finden sich dichte Regionen?

Aufgabe 4-3 *Clustering-Algorithmen und der "Curse of dimensionality"*

Als Effekte des "Curse of Dimensionality" sind für das Clustering-Problem in hochdimensionalen Vektorräumen vor allem folgende Probleme relevant:

(a) **Komplexität der Mustersuche.** Mehr Attribute bedeuten mehr Variablen für eine Optimierung. Ein vollständige Suche wird mit mehr Dimensionen immer schwieriger oder gar undurchführbar.

(b) **Irrelevanz von Distanz-Unterschieden.** Konzepte wie "Nähe", "Distanz" oder "Nachbarschaft" werden mit zunehmender Dimensionalität immer bedeutungsloser.

(c) **Irrelevante Attribute.** Da immer mehr Attribute "auf verdacht" aufgesammelt werden, sind viele dieser Attribute für bestimmte Muster letztlich wahrscheinlich irrelevant. Die Relevanz ist aber unterschiedlich für unterschiedliche Teilmengen der Datenobjekte.

Obwohl dieses Problem unabhängig vom Problem der Irrelevanz von Distanz-Unterschieden ist, werden auch durch diesen Effekt volldimensionale Distanzen zwischen Objekten fragwürdig.

(d) **Korrelierte Attribute.** Mit der steigenden Anzahl von Attributen wird ebenfalls das Auftauchen von Korrelationen zwischen Attributen immer wahrscheinlicher. Wie das Problem irrelevanter Attribute kann dieses Problem allerdings auch schon bei niedrig-dimensionalen Daten auftauchen.

Korrelationen zwischen Attributen verändern die Gestalt eines Datensatzes und darin verborgener Muster erheblich. Darüberhinaus ist es nicht nur von Interesse, verborgene Muster trotz dieser Korrelationen zu entdecken, sondern auch diese Korrelationen selbst.

Überlegen Sie für die Clustering-Algorithmen CLIQUE, SUBCLU, FIRES, DiSH, PROCLUS, PreDeCon, Cheng&Church, p-cluster, ORCLUS, 4C und CASH, welche dieser Probleme jeweils behandelt werden und welche vernachlässigt oder nicht zufriedenstellend gelöst werden.

	CLIQUE	SUBCLU	FIRES	DiSH	PROCLUS	PreDeCon	Cheng&Church	p-cluster	ORCLUS	4C	CASH
Problem (a)											
Problem (b)											
Problem (c)											
Problem (d)											