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

München, den 21.6.2010

Knowledge Discovery in Databases II

SoSe 2010

Übungsblatt 7: Multirepräsentiertes Data Mining

Besprechung am Donnerstag, 24.6.2010

Aufgabe 7-1 *Komplementarität von Klassifikatoren*

Gegeben seien zwei binäre Klassifikatoren f_1 und f_2, die auf je einer Repräsentation der Objekte eines Datensatzes D mit Klassen $\{0, 1\}$ arbeiten. Entscheiden Sie, ob in den folgenden Fällen eine Kombination der Klassifikatoren sinnvoll ist:

(a) $f_1(x) = f_2(x)$ für alle $x \in D$

(b) $f_1(x) = 1 - f_2(x)$ für alle $x \in D$

Aufgabe 7-2 *Abhängigkeitsmaß*

Gegeben sei ein Maß h, welches die Abhängigkeit zwischen zwei Kernelmatrizen K und K' misst. Anschaulich heißt das, dass $h(K, K')$ groß ist, wenn die zugehörigen Kernels k und k' dieselben Objekte als ähnlich und als unähnlich betrachten. Wenn sie die Ähnlichkeit derselben Objekte unterschiedlich bewerten, sei $h(K, K')$ niedrig.

Seien nun ein Datensatz D mit einem Klassenlabel und r Repräsentationen pro Objekt gegeben. Wir berechnen eine Kernelmatrix K_i für jede der r Repräsentationen und eine Kernelmatrix L auf den Klassenlabels. Überlegen Sie sich, wie man mittels h eine Linearkombination der K_i bestimmen kann, die die Ähnlichkeit der Klassenlabels möglicht gut widerspiegelt.

Aufgabe 7-3 *Multirepräsentiertes Clustering*

Gegeben sei ein Datensatz X, so dass jeder Punkt durch 2 zweidimensionale Vektoren repräsentiert wird.

```
A = ( 0, 1); (  3,  0) B = (-1,-1); (  2,  0) C = ( 0, 0); (  3,
1) D = ( 0,-3); ( -2,  2) E = ( 2, 1); ( -2, -3)
```

Wir wollen auf diesem Datensatz multirepräsentiertes Clustering mittels DBSCAN durchführen.

(a) Wie unterscheidet sich multirepräsentiertes Clustering von gewöhnlichem Clustering? Welche besonderen Schwierigkeiten sind damit verbunden?

(b) Es sei $MinPoints = 3$. Für welche Werte von ϵ_1, ϵ_2 sind die Objekte C und D Kernobjekte nach

- der Vereinigungsmethode?
- der Schnittmethode?