**Data Mining WS 2012/13**
**Institut für Neuroinformatik**
Dr. F. Schwenker

2. Aufgabenblatt (Abgabe bis zum 27.11.2012 in der Vorlesung)

**5. Aufgabe (6): Agglomerative Clusterung**
Gegeben sei die folgende Datenmatrix von 6 Vektoren aus $\mathbb{R}^3$

$$X = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 0 & 1 & 3 \\ 3 & 4 & 3 \\ 0 & 3 & 4 \\ 2 & 3 & 2 \end{bmatrix}$$

Als Abstandmaß $d$ soll der quadrierte Euklidische Abstand benutzt werden

$$d(x,y) = \sum_{i=1}^{p} (x_i - y_i)^2 \quad x, y \in \mathbb{R}^p.$$

1. Stellen sie die Distanzmatrix $D$ auf.
2. Führen sie die folgenden agglomerativen Clusterverfahren durch. Stellen sie jeweils die Folge der Clusterung als Dendrogramm dar:
   (a) Single-Linkage-Verfahren
   (b) Complete-Linkage-Verfahren
   (c) Group-Average-Verfahren
   (d) Centroid-Verfahren
   (e) Ward-Verfahren

Hinweis: Verwenden sie z.B. die `Statistics Toolbox` von `matlab` oder auch `R`

**6. Aufgabe (6): Optimale Clusterung**
Es sei $k \in \mathbb{N}$ fest gewählt und sei $\mathcal{C} = \{C_1, \ldots, C_k\}$ eine Partition von $k$ Teilmengen einer Objektmenge, deren Objekte durch reelle Merkmalsvektoren $x_\mu \in \mathbb{R}^p$ gegeben seien.
Für eine (endliche) Menge $C \in \mathcal{C}$ ist der *Durchmesser* definiert durch

$$\operatorname{diam}(C) = \max\{\|x - y\|_2 \ : \ x, y \in C\}.$$

Die Bewertung einer Clusterung $\mathcal{C} = \{C_1, \ldots, C_k\}$ ist dann definiert durch

$$D_{\text{diam}}(\mathcal{C}) = \frac{1}{k} \sum_{j=1}^{k} \operatorname{diam}(C_j)$$

Für die optimale Clusterung $\mathcal{C}^*$ (mit $k$ Clustern) gilt:

$$S(\mathcal{C}^*) = \min_{\mathcal{C}} D_{\text{diam}}\,(\mathcal{C})$$

Bestimmen sie die optimale Clusterung mit $k = 2$ Clustern für die Punkte

$$\{(1,0), (2,0), (0,1), (2,2), (2,1)\} \subset \mathbb{R}^2$$