Clusteranalyse

Aufgabe 1:
Es seien folgende Medikamente mit ihrem *weight index* und ihrem *pH index* vorgegeben:

Objekt	weight index	pH index
Medizin A	1	1
Medizin B	2	1
Medizin C	4	3
Medizin D	5	4

Gruppieren Sie die Medikamente mithilfe des k-means Clusterverfahrens in 2 Cluster. Medikamente A und B sollen dabei die Startpartition für je eine Klasse sein. Als Distanzmaß soll die euklidische Distanz verwendet werden.

Aufgabe 2: (Fortsetzung zu Aufgabe 2 von Blatt 6)
Der Datensatz `europa.txt` enthält Daten zu $n = 24$ europäischen Ländern. Folgende Variablen wurden erhoben: `ober` (Oberfläche in km^2), `einw` (Einwohner in Millionen), `brut` (BIP pro Kopf in \$) und `arbl` (Arbeitslosenquote in %).

a) Lesen Sie den Datensatz in R ein, standardisieren Sie die Daten und laden Sie das Paket `cluster`.

b) Führen Sie ein k-means Clustering mit Hilfe der Funktion `kmeans` durch. Wählen Sie dazu k=4. Wiederholen Sie es dann nur mit den beiden Variablen `arbl` und `brut`. Plotten Sie die 4 Cluster in 4 verschiedenen Farben im zweidimensionalen Raum.

c) Führen Sie nun, ähnlich wie in Aufgabe 2 von Blatt 6, eine hierarchische Klassifikation mithilfe der Funktion `hclust` mit dem Single-Linkage Verfahren und dem Zentroid-Verfahren nur unter Einbeziehung der beiden Variablen `arbl` und `brut` durch. Vergleichen Sie die Ergebnisse für k=4 mit denen aus Teilaufgabe b) graphisch.

Aufgabe 3:
Der Datensatz `geyser` aus dem R-Paket `MASS` beinhaltet für 299 Eruptionen des berühmten *Old Faithful* Geysirs im Yellowstone Nationalpark die Wartezeit seit der vorangegangenen Eruption (in Minuten) sowie die Eruptionsdauer (in Minuten). Im Folgenden soll dieser Datensatz mit Hilfe eines Mischmodellansatzes untersucht werden. Für die Verteilung in den Klassen wird eine bivariate Normalverteilung angenommen.

a) Skizzieren Sie kurz die verwendeten Modellannahmen des hier betrachteten Mischmodellansatzes. Wie kann das Mischmodell geschätzt werden und wie erhält man hieraus eine Partitionierung der Daten?

b) Plotten Sie die Daten und bestimmen Sie visuell eine geeignete Anzahl an Klassen für den Mischmodellansatz.

c) Clustern Sie die Geysir-Daten mit Hilfe der Funktion `Mclust` aus dem Package `mclust`. Verwenden Sie für Kovarianzmatrizen in den Klassen die Annahme $\boldsymbol{\Sigma}_r = \sigma^2 \boldsymbol{I}$. Visualisieren Sie die gefundene Partitionierung.

d) Verwenden Sie flexiblere Annahmen für die Kovarianzstruktur und vergleichen Sie die sich daraus ergebenden Modelle und Partitionierungen.