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:
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 und standardisieren Sie die Daten.

b) Führen Sie mithilfe der Funktion `hclust()` eine hierarchische Klassifikation unter Einbeziehung aller vier Kovariablen mit dem *Single Linkage* Verfahren durch. Verwenden Sie als zugrundeliegende Distanz zwischen einzelnen Objekten die quadrierte euklidische Distanz.

c) Führen Sie mithilfe der Funktion `hclust()` eine hierarchische Klassifikation unter Einbeziehung aller vier Kovariablen mit dem *Zentroid* Verfahren durch.

d) Führen Sie mithilfe der Funktion `hclust()` eine hierarchische Klassifikation unter Einbeziehung aller vier Kovariablen mit dem *Complete Linkage* Verfahren durch. Verwenden Sie als zugrundeliegende Distanz zwischen einzelnen Objekten die Mahalanobis-Distanz.

Hinweis: Die Funktion `mahalanobis()` könnte hilfreich sein.

e) Visualisieren und vergleichen Sie ihre Ergebnisse der Teilaufgaben b), c) und d) jeweils mithilfe eines Dendrogramms.

f) Führen Sie ein k-means Clustering mit Hilfe der Funktion `kmeans()` aus dem Paket `cluster` durch. Wählen Sie dazu k=4.

g) Wiederholen Sie die Verfahren aus den Teilaufgaben b), c), d) und f) nur unter Einbeziehung der beiden Variablen `arbl` und `brut` durch. Vergleichen Sie die Ergebnisse für k=4, indem Sie jeweils die 4 Cluster in 4 verschiedenen Farben im zweidimensionalen Raum plotten.

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 unterschiedliche Annahmen (z.B. „EII“, „VVI“, „VVV“) für die Kovarianzstruktur in den Klassen. Vergleichen Sie die sich daraus ergebenden Modelle und Partitionierungen.