Methoden der Datenrepräsentation und Klassifikation
Aufgabenblatt 5: Bildung von Partitionen

Checkliste Lernziele:

- Wie kann man mit `R` „zufällige“ Zahlen erzeugen?
- Wie lässt sich eine *k-means* Clusteranalyse durchführen?
- Wie lassen sich die Vor- und Nachteile von Verfahren durch künstliche Daten verdeutlichen?

Erinnerung: Die Normalverteilung folgt der Dichtefunktion

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

wobei μ der Erwartungswert und σ die Standardabweichung ist. Ist $\mu = 0$ und $\sigma = 1$ spricht man von einer Standardnormalverteilung. Um anzuzeigen, dass eine Zufallsvariable X einer Normalverteilung mit Parametern μ und σ folgt, verwenden wir die Schreibweise $X \sim \mathcal{N}(\mu, \sigma)$.

Aufgaben:

1. Betrachten Sie die Hilfe zum Befehl `rnorm`. Verwenden Sie diesen Befehl um einen Datensatz zu erstellen, der aus insgesamt 100 Fällen besteht, die in zwei Gruppen zu jeweils 50 Fällen aufgeteilt sind; erzeugen Sie für jeden Fall zwei Werte zweier normalverteilter Zufallsvariablen X und Y, wobei für die erste Gruppe $X \sim \mathcal{N}(3, 1)$ und $Y \sim \mathcal{N}(3, 1)$ und für die zweite Gruppe $X \sim \mathcal{N}(6, 1)$ und $Y \sim \mathcal{N}(5, 1)$ gelten soll. Weitere nützliche Befehle hierfür sind `cbind`, `rbind` und `as.data.frame`.

2. Stellen Sie die in der vorherigen Aufgabe erstellten „Zufallsvariablen“ grafisch dar. Verändern Sie die Parameter der Verteilungen und betrachten Sie die Effekte.

3. Verwenden Sie den Befehl `kmeans`, um eine 2-Cluster-Lösung für die künstlich erzeugten Daten aus Aufgabe 1 zu berechnen. Stellen Sie dieses Ergebnis grafisch dar.

4. Erzeugen Sie zudem Lösungen mit 3, 4 und 5 Clustern. Nutzen Sie eine grafische Darstellung, um die Ergebnisse zu vergleichen.

5. Verändern Sie Ihre Daten so, dass fünf Fälle „klare Ausreißer“sind. Berechnen Sie erneut eine 2-Cluster-Lösung. Interpretieren Sie das Ergebnis!

6. Erzeugen Sie einen neuen Datensatz, der aus 100 Fällen und zwei standardnormalverteilten Zufallsvariablen besteht. Berechnen Sie Lösungen mithilfe des `kmeans`-Befehls für 2, 3, 4 und 5 Cluster.

7. Erstellen Sie für die Daten aus der letzten Aufgabe abermals 2-Cluster-Lösungen mittels `kmeans`. Variieren Sie hierbei den Parameter `nstart` mit Werten von 1 bis 5. Stellen Sie die Ergebnisse grafisch dar. Interpretieren Sie die Ergenisse!

8. Laden Sie den Datensatz `bs1.dat` und erstellen Sie aus diesem die (transponierte) Tabelle 2.3-2.

9. Benutzen Sie die Datenmatrix aus der letzten Aufgabe, um mittels `kmeans` Partitionen der Größen 2, 3 und 4 zu bilden. Vergleichen Sie Ihre Ergebnisse mit den im Skript in Abschnitt 5.2.5 angegebenen Resultaten.

10. Benutzen Sie den Befehl `pam` aus dem Paket `cluster` um das *k-medoids* Verfahren auf die Abstandsmatrix aus der Datei `auto.dat` anzuwenden. Die Zahl der Cluster können Sie frei wählen. Begründen Sie Ihre Entscheidung!