Technische Universität Dortmund
Fakultät Statistik
Lehrstuhl Computergestützte Statistik
Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

Prof. Dr. Claus Weihs
Dr. Thomas Liebig
Sebastian Buschjäger
Malte Jastrow
Jennifer Neuhaus-Stern

Übung zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2018

Übungsblatt Nr. 9

Der Abgabetermin ist Dienstag der 23.07.2018 bis 10:00 Uhr im moodle-Raum

Aufgabe 1 (Entscheidungsgrenzen) (5 Punkte)

Inzwischen haben Sie in der Vorlesung weitere Klassifikationsverfahren kennen gelernt, sodass sich ein erneuter Vergleich der Entscheidungsgrenzen verschiedener Verfahren anbietet. Betrachten Sie die folgenden Datensätze:

- mlbench.2dnormals(500, 2)
- mlbench.circle(500, 2)
- mlbench.spirals(500, cycles = 2)

a) (2 Punkte) Wenden Sie die folgenden Lernverfahren mit mlr auf die Datensätze an: Naive Bayes, kNN mit $k = 3$ und mit $k = 21$, SVM mit linearem und mit radialem Kern, AdaBoost, Entscheidungsbaum und Random Forest mit 5 und mit 500 Bäumen. Visualisieren Sie die Entscheidungsgrenzen (plotLearnerPrediction)

b) (3 Punkte) Erläutern Sie Unterschiede zwischen den Verfahren, beschreiben Sie vor allem den Effekt der jeweiligen Parameter. Welche Verfahren trennen die Klassen am sinvollsten?

Aufgabe 2 (Hauptkomponentenanalyse: Theorie) (5 Punkte)

Die folgende Aufgabe zum Thema Hauptkomponentenanalyse können Sie sowohl per Hand als auch mithilfe von R lösen.

a) (2 Punkte) Berechnen Sie die Loadings g_1 und g_2 für die Kovarianzmatrix

$$S = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$$

und geben Sie die Hauptkomponenten z_1 und z_2 an.

Berechnen Sie außerdem den Anteil der Gesamtvarianz, der durch die erste Hauptkomponente z_1 erklärt wird.

b) (2 Punkte) Transfomieren Sie die Kovarianzmatrix S aus Teil a) in eine Korrelationsmatrix R. Berechnen Sie auch hier die Loadings g_1^* und g_2^* sowie die Hauptkomponenten z_1^* und z_2^*. Welcher Anteil der Gesamtvarianz wird durch z_1^* erklärt?

c) (1 Punkt) Vergleichen Sie die in den Teilen a) und b) berechneten Hauptkomponenten. Sind sie gleich? Sollten sie gleich sein?

Aufgabe 3 (Hauptkomponentenanalyse: Anwendung) **(5 Punkte)**

In dieser Aufgabe soll der Datensatz bank.txt betrachtet werden. Er enthält diverse Längenmaße von 100 echten und 100 gefälschten schweizer 1000-Franc-Banknoten. Eine genauere Beschreibung finden Sie in der Datei info.txt.

a) (2 Punkte) Führen Sie zwei Hauptkomponentenanalysen für den gesamten Datensatz durch und zwar sowohl auf Basis von Kovarianzen als auch auf Basis von Korrelationen. Dies ist in R mithilfe der Funktionen princomp und prcomp möglich.

b) (1.5 Punkte) Wieviele Hauptkomponenten würden Sie wählen, um eine Dimensionsreduktion durchzuführen? Schauen Sie sich dazu den screeplot und den Prozentsatz der erklärten Gesamtvarianz (z.B. mit summary) an. Für welche der beiden Hauptkomponentenanalysen aus Teil a) sollten die Loadings überhaupt interpretiert werden? Interpretieren Sie die entsprechenden Loadings der ersten Hauptkomponente.

c) (1.5 Punkte) Erstellen Sie für beide Hauptkomponentenanalysen aus a) jeweils einen biplot. Welche Scores-Struktur liegt vor? Vergleichen und interpretieren Sie die Bi-Plots.

Aufgabe 4 (Zeitreihen) **(5 Punkte)**

Betrachten Sie die beiden stochastischen Prozesse

$$y_t = 1 - 0.9y_{t-1} + \epsilon_t \qquad \text{und}$$
$$y_t = -0.2 + 1.25y_{t-1} + \epsilon_t$$

Die ϵ_t seien unabhängig und identisch $N(0, 0.5)$ verteilt.

a) (1 Punkte) Um welche Art von Prozessen handelt es sich? Sind die Prozesse stationär?

b) (2 Punkte) Simulieren Sie 500 Beobachtungen aus den beiden stochastischen Prozessen und plotten Sie die erzeugten Zeitreihen. Berechnen Sie, falls es sich um einen stationären Prozess handelt, jeweils den Erwartungswert und zeichnen Sie ihn in die Grafik mit ein.

c) (2 Punkte) Glätten sie die Zeitreihen mithilfe eines einfachen gleitenden Durchschnitts der Länge 20. Zeichnen Sie die geglätteten Zeitreihen in die bereits erzeugten Plots mit ein. Für die Berechnung gleitender Durchschnitte in R stehen mehrere Funktionen in verschiedenen Paketen zur Verfügung, z. B. die Funktion runmean im Paket caTools.