Aufgaben zur Hausarbeit 2

Aufgabe 1

Datensatz

Laden Sie den Datensatz "glass.txt" herunter. Weitere Informationen zum Datensatz und allen Variablen finden Sie auf **UCI Machine Learning Repository**.
Lesen Sie den Datensatz in R ein. Führen Sie eine Hauptkomponentenanalyse auf einer Korrelationsmatrix der Daten ohne Variable `Type` durch. Die ersten zwei empirischen Hauptkomponenten werden nun für die Diskriminanzanalyse benutzt.

Aufgaben

1. Man möchte eine Entscheidungsregel finden, die anhand der ersten zwei empirischen Hauptkomponenten bestimmt, ob das Glas ein Fensterglas (`Type` 1 bis 3) ist oder nicht (`Type` 5 bis 7). Plotten Sie die beiden empirischen Hauptkomponenten gegeneinander, markieren Sie jeden Punkt mit dem Glastyp (Fensterglas oder nicht) und beschreiben Sie das Ergebnis. Sind die empirischen Hauptkomponenten dazu geeignet, die beide Glastypen zu unterscheiden? Anschließend testen Sie, ob die Erwartungswerte der beiden Klassen sich signifikant unterscheiden. (3 Punkte)

2. Führen Sie eine lineare Diskriminanzanalyse mit den Bayes- und Maximum-Likelihood-Entscheidungsregeln durch und rechnen Sie die entsprechenden Resubstitutionsfehler aus. Welche Entscheidungsregel hat den kleineren Resubstitutionsfehler? Stellen Sie das Ergebnis der Diskriminanzanalyse nach den beiden Methoden graphisch dar, vergleichen und interpretieren Sie das Ergebnis. (4 Punkte)

3. Bestimmen Sie nun eine Entscheidungsregel, die berücksichtigt, dass $C(\epsilon_{1|2})/C(\epsilon_{2|1}) = 1/4$, wobei $C(\epsilon_{i|j})$ die Kosten des Ereignisses $\epsilon_{i|j}$, $i, j = 1, 2$ sind. Klasse 1 entspricht dem Fensterglas. Rechnen Sie den entsprechenden Resubstitutionsfehler aus und vergleichen Sie diesen mit den Resubstitutionsfehlern aus Punkt 2. Beschreiben Sie die Ereignisse $\epsilon_{i|j}$, $i, j = 1, 2$ und interpretieren Sie $C(\epsilon_{1|2})/C(\epsilon_{2|1}) = 1/4$. (3 Punkte)

4. Bei der linearen Diskriminanzanalyse wird angenommen, dass die Kovarianzmatrizen der beiden Klassen identisch sind. Ist diese Annahme für die vorliegenden Daten sinnvoll? Führen Sie die quadratische Diskriminanzanalyse mit der Bayes-Entscheidungsregel durch, rechnen Sie den Resubstitutionsfehler aus und stellen sie das Ergebnis graphisch dar. Interpretieren Sie das Ergebnis. (3 Punkte)

5. Nun möchte man eine Entscheidungsregel finden, die anhand der ersten zwei empirischen Hauptkomponenten zwischen Fensterglas (`Type` 1 bis 3), Geschirrglas (`Type` 5 und 6) und Scheinwerferglas (`Type` 7) unterscheidet. Plotten Sie die beiden empirischen Hauptkomponenten gegeneinander und markieren Sie jeden Punkt mit dem Glastyp (Fensterglas, Geschirrglas und Scheinwerferglas). Kommentieren Sie

das Ergebnis. Führen Sie eine lineare Diskriminanzanalyse mit der Bayes-Entscheidungsregel durch, rechnen Sie den entsprechenden Resubstitutionsfehler aus und stellen das Ergebnis graphisch dar. Interpretieren Sie das Ergebnis. Erklären Sie was die Stichproben-Diskriminanten sind und plotten Sie diese indem Sie auch die drei Glastypen auf dem Plot markieren. Vergleichen Sie die Resubstitutionsfehler der quadratischen und linearen Diskriminanzanalyse mit der Bayes-Entscheidungsregel. (5 Punkte)

Aufgabe 2

Datensatz

In dieser Aufgabe wird weiter mit dem Datensatz `glass.txt` aus Aufgabe 1 gearbeitet. Die Klassen (Cluster) sollen anhand der ersten zwei empirischen Hauptkomponenten bestimmt werden.

Aufgaben

1. Wenden Sie den k-Means-Algorithmus auf die Daten an um diese in 2 Klassen aufzuteilen. Erstellen Sie einen Plot, der die Daten und die geschätzten Klassen zeigt. Inwieweit stimmen die Klassen, die Sie mit dem k-Means-Algorithmus bekommen haben mit den beiden Glastypen aus Aufgabe 1, Punkt 1 (Fensternglas und alle andere Glastypen) überein? Welchen Resubstitutionsfehler macht der k-Means-Algorithmus? Erstellen Sie auch den Silhouettenplot und kommentieren Sie diesen. (4 Punkte)

2. Wenden Sie die drei hierarchischen Verfahren (Single, Complete und Average Linkage) auf die Daten an. Plotten Sie die entsprechenden Dendrogramme. Für jede Methode erstellen Sie für 2 Klassen die Silhouettenplots and Plots, die die Daten und die beiden Klassen zeigen. Beschreiben und interpretieren Sie die Ergebnisse. Wie gut können die drei Methoden das Fensterglas unterscheiden? Wie unterscheiden sich die Ergebnisse von denen aus Punkt 1? (5 Punkte)

3. Nehmen Sie nun an, dass die Daten einer Mischung von p zweidimensionalen Normalverteilungen folgen. Benutzen Sie die Funktion `Mclust` um die Daten anzupassen. Welches Modell und mit wie vielen Klassen p wurde gewählt? Stellen Sie das Ergebnis graphisch dar und beschreiben Sie es. Schätzen Sie auch ein Modell mit zwei Klassen, stellen Sie dieses graphisch dar und vergleichen Sie das Ergebnis mit dem aus Punkt 1. Ist die Annahme der Normalverteilung für diese Daten sinnvoll? (3 Punkte)