Vertiefende statistische Verfahren

4. Übungsblatt SS 2024

Stefan Kolb, Joachim Waltl

Allgemeine Information

Alle Aufgaben sind mit R zu lössen, wenn nicht explizit anders angegeben. Die Berechnungen sollen nachvollziehbar und dokumentiert sein. Um die vollständige Punktezahl zu erreichen, müssen alle Ergebnisse und Fragen entsprechend interpretiert bzw. beantwortet werden. Code alleine ist nicht ausreichend! Die Abgabe erfolgt über Moodle entsprechend der Abgaberichtlinien als pdf und Rmd File. Bitte inkludieren Sie namentlich alle beteiligten Gruppenmitglieder sowohl im Bericht als auch im Source Code. Die jeweiligen Datensätze die für diese Übung relevant sind finden Sie ebenfalls in Moodle.

Datensatz: Reaven and Miller Diabetes Daten

Verwenden Sie den Datensatz `diabetes_RM.csv`. Der Datensatz enthält fünf Messungen, die an 145 nicht adipösen erwachsenen Patienten durchgeführt wurden, die in drei Gruppen eingeteilt wurden.

Die drei primären Variablen sind die Glukoseintoleranz, die Insulinantwort auf orale Glukose und die Insulinresistenz (gemessen durch die Steady-State-Plasmaglukose, die nach chemischer Suppression der endogenen Insulinsekretion bestimmt wird). Zwei zusätzliche Variablen, das relative Gewicht und die Nüchternplasmaglukose, sind ebenfalls enthalten. Zusammengefasst ergeben sich folgende Prädiktorvariablen:

- `rw`: relatives Gewicht, Verhältnis zwischen aktuellem Gewicht und zu erwartendem Gewicht bei der Körpergröße
- `fpg`: Nüchternglukoselevel im Plasma in mg/dl
- `glucose`: Fläche unter Glukose-Antwort (mg/dl*h) nach 3h oralem Glukosetoleranztest (OGTT)
- `insulin`: Fläche unter der Insulin-Antwort (mg/dl*h) nach OGTT
- `sspg`: Steady-State-Plasmaglukose (mg/dl) als Maß für die Insulinresistenz

Reaven und Miller [ref] wendeten in Anlehnung an Friedman und Rubin (1967) eine Clusteranalyse auf die drei primären Variablen an und identifizierten drei Cluster: “normal”, “chemical” und “overt” diabetsiche Probanden. Die Variable `group` enthält die Klassifizierungen der Probanden in diese drei Gruppen.

1 Diskriminanzanalyse [5P]

Führen Sie eine Diskriminanzanalyse unter Berücksichtigung folgender Punkte durch:

i) Explorative Analyse der Prädiktoren mit Hilfe von Histogrammen. Gibt es Prädiktoren, die bereits eine gute Trennung zwischen den Klassen erlauben?

ii) Überprüfen Sie ob die Vorraussetzungen für eine LDA gegeben sind und führen Sie eine Standardisierung der Daten durch.

iii) Unterteilen sie die gesamten Daten in Trainings- und Test-Daten und führen Sie in der weiteren Folge eine Klassifizierung mit einer LDA durch. Evaluieren Sie die Perfomance der Klassifizierung und stellen Sie die Ergebnisse graphisch dar (Darstellung der Projektionen, Partition Plot).

iv) Vergleichen Sie unterschiedliche Varianten der Diskriminanzanalyse (QDA, MDA, FDA) hinsichtlich ihrer Klassifikationsgenauigkeit.

2 Principal Component Analyse [5P]

Führen Sie eine PCA unter Berücksichtigung folgender Punkte durch:

i) Überprüfung der paarweisen Kovarianzen / Korrelationen

ii) Berechnen der PCA und Beurteilung wie viele PCs sinnvoll sind (Eigenwerte und Screeplot).

iii) Transformation der ursprünglichen Daten in ein PC-Koordinatensystem mit zwei PCs. Vergleichen Sie die Darstellung mit dem Ergebnis der LDA (LD1 und LD2 Projektionen).

iv) Stellen Sie den Correlation Circle und Biplot graphisch dar. Welche Information liefert diese Darstellung?

v) Beurteilen Sie die Qualität und Beiträge der Variablen auf die PCs.

vi) Wiederholen Sie die LDA von Aufgabe 1 unter Verwendung der PCs zur Klassifizierung. Achten Sie auf die Verwendung der gleichen Trainings- und Test-Daten und vergleichen Sie die Performance. Vergleichen Sie weiters die Performance der LDA mit den Variablen `glucose` und `fpg` mit der PCA+LDA mit zwei PCs.