h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Antje Jahn
Multivariate Statistik
Sommersemester 2019

Arbeitsblatt 8

A 1 Der Datensatz diab (moodle) enthält Daten von 724 indischen Frauen zu Diabeteserkrankungen. Neben dem Erkrankungsstatus (diabetes: 1=erkrankt, 0=nicht erkrankt) sind Ergebnisse diagnostischer Tests (pgc = Plasma Glukose Konzentration, insulin = 2-Stunden-Seruminsulin) und potentielle prognostischen Faktoren enthalten (dbp=diastolischer Blutdruck, bmi=body mass index, pregn=Anzahl an Schwangerschaften, hfd = Hautfaltendicke).

Führen Sie eine Diskriminanzanalyse (Funktionen lda und qda im Paket MASS) durch, um die Diabeteserkrankung aus dem Ergebnis des Glukosetoleranztest (pgc) und dem BMI vorherzusagen.

a) Erstellen Sie zunächst einen Plot in dem BMI und Glukosekonzentration dargestellt werden mit verschiedenen Farben für die Erkrankten und Gesunden
b) Führen Sie eine lineare Diskriminanzanalyse durch. Welche Annahmen treffen Sie hier?
c) Erstellen Sie für jede Gruppe separat ein Histogramm für die Werte, die die lineare Diskriminanzfunktion in dieser Gruppe annimmt.
d) Welcher Trainingsfehler ergibt sich aus der Diskriminanzanalyse?
e) Wie würde eine neue Person von mit einer Glukosekonzentration von 120 und einem BMI von 27 klassifiziert werden? Wie würde die posterior-Wahrscheinlichkeit für eine Diabetes-Erkrankung geschätzt werden? (Funktion predict)
f) Erstellen Sie einen Plot der Klassifikationsgrenzen. Erstellen Sie dazu z.B. Vektoren x, y über den Wertebereich von bmi und pgc, den Vektor z als Klassifikationen zu x,y (predict) und plotten Sie die Höhenlinien von x,y,z in das Streudiagramm aus a).
g) Führen Sie einen Test auf Homogenität der Kovarianzen durch. Welches Ergebnis liefert dieser Test? (Funktion boxM im Pakte biotools)
h) Führen Sie eine quadratische Diskriminanzanalyse durch, schätzen Sie den Klassifikationsfehler und plotten Sie auch hier die Klassifikationsgrenze.

A 2 Berechnen Sie den Trainingsfehler aus einer Klassifikationsregel, die alle Personen unabhängig von ihren Einflussvariablen als gesund klassifiziert? Vergleichen Sie den Fehler mit dem Fehler aus Aufgabe 1.

A 3 Beweisen Sie, dass das Ergebnis einer linearen Diskriminanzanalyse unabhängig davon ist, ob die Prädiktoren zunächst standardisiert werden, d.h. statt X die p-dimensionale Zufallsvariable

$$\tilde{X} := \hat{D}^{-1}(X - \hat{\mu})$$

mit $\hat{\mu}$ der geschätzte Erwartungswertvektor und $\hat{D}$ die Diagonalmatrix mit den geschätzten Standardabweichungen $\hat{\sigma}_i$ der Prädiktoren auf der Diagonalen.

Hinweis: Berechnen Sie für die transformierten Beobachtungen $\tilde{x}$ die geschätzte Kovarianzmatrix, $\hat{\Sigma}$, die geschätzten Erwartungswertvektoren pro Gruppe, $\hat{\mu}_k$, und daraus die lineare Diskriminanzfunktion $h(\tilde{x})$ und vergleichen Sie diese mit der Diskrimininanzfunktion für die Original-Beobachtungen, $h(x)$. Nutzen Sie dabei bekannte Rechenregeln für (Kovarianz-)Matrizen.