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

Antje Jahn
Multivariate Statistik
Sommersemester 2019

Arbeitsblatt 9

A 1 Beschreiben Sie die Unterschiede/Gemeinsamkeiten zwischen den drei Klassifikationsverfahren 1. Diskriminanzanalyse, 2. logistische Regressionsanalyse und 3. K-Nächste-Nachbarn-Klassifikation hinsichtlich

- den Voraussetzungen an die Zielvariable
- den Voraussetzungen an die Einflussvariablen
- der resultierenden Klassifikationsgrenzen
- den Konsequenzen einer Standardisierung von Einflussvariablen

A 2 Führen Sie mit dem Datensatz diab (s. Praktikum 8) eine logistische Regression für die abhängige Variable diabetes mit den unabhängigen Variablen pgc und bmi durch (Funktionen `glm` und `predict.glm`).

a) Wie würde anhand dieser Ergebnisse für eine neue Person mit einem BMI von 27 und einer Glukosekonzentration (pgc) von 120 die Wahrscheinlichkeit für eine Diabetes-Erkrankung prognostiziert werden? Leiten Sie die Prognose aus den Schätzern der Regressionskoeffizienten ab und vergleichen Sie Ihr Ergebnis mit dem Ergebnis einer predict-Funktion. Vergleichen Sie das Ergebnis mit den Ergebnissen aus der linearen Diskriminanzanalyse (s. Praktikum 8, A1 e))
b) Steigt oder fällt das Risiko für eine Diabetes-Erkrankung bei höherer Glukosekonzentration (pgc)? Woran lässt sich dies ablesen?
c) Um wieviel % ist die Odds für eine Diabeteserkrankung bei einer Frau erhöht/reduziert gegenüber einer Frau mit gleicher Glukosekonzentration aber einem um 5 Einheiten erhöhtem BMI?
d) Plotten Sie eine ROC-Kurve (Funktion roc im Paket pROC)
e) Schätzen Sie aus den Daten die Misklassifikationswahrscheinlichkeit, die Sensitivität und die Spezifität für die Klassifikationsregel $\delta(x) = 1 \Leftrightarrow \hat{P}(Y = 1|X = x) > 0.5$, wobei $\hat{P}$ die Schätzung aus der logistischen Regression sei. Wo in der ROC Kurve liegt dieser Punkt?
f) Berechnen Sie eine Kalibrierungsgerade. Teilen Sie dazu die Beobachtungen in die Dezile der vorhergesagten Wahrscheinlichkeiten. Berechnen Sie pro Dezil die durchschnittliche

geschätzte Wahrscheinlichkeit für eine Diabeteserkrankung, daraus die erwartete Anzahl an Frauen mit Diabetes und plotten Sie diese gegen die beobachtete Anzahl an Frauen mit Diabetes. Zeichnen Sie in den Plot zum Vergleich auch eine Gerade ein, die einer perfekten Kalibrierung entspräche.

g) Welchen zusätzlichen Einfluss auf die Prognose hat das Ergebnis des zweiten diagnostischen Tests (insulin)? Schauen Sie sich dazu zunächst die Verteilung der Variablen an, transformieren Sie die Variable ggf. und schätzen Sie ein Modell mit dieser zusätzlichen Einflussvariable. Beschreiben Sie die Veränderungen im erweiterten Modell hinsichtlich Anpassungsgüte, Diskriminierung und Kalibrierung.

A 3 Teilen Sie den Datensatz zufällig in einen Test- und einen Validierungsdatensatz gleicher Größe auf. Führen Sie eine K-Nächste-Nachbarn-Klassifikation mit K=1, K=5, K=10 durch, berechnen Sie jeweils den Klassifikationsfehler und vergleichen Sie das Ergebnis mit den Misklassifikationswahrscheinlichkeiten aus Diskriminanzanalyse und logistischer Regression (Funktion knn im Paket class).