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

Antje Jahn
Nichtparametrische und nichtlineare Modelle
Sommersemester 2019

Arbeitsblatt 7

A 1 Wir passen ein lineares Regressionsmodell mit den Basisfunktionserweiterungen in X, $h_1(X) = X$ und $h_2(X) = (X-1)^2 \cdot \mathbb{1}_{(X \geq 1)}$ an:

$$Y = b_0 + b_1 h_1(X) + b_2 h_2(X) + \epsilon$$

Als Koeffizientenschätzer erhalten wir $\hat{b}_0 = 1, \hat{b}_1 = 1, \hat{b}_2 = -2$. Skizzieren Sie die geschätzte Kurve zwischen $X = -2$ und $X = 2$.

A 2 Es liegen Stichprobendaten $(x_i, y_i)_{i=1,\ldots,n}$ vor mit $x_i \in [a, b] \;\; \forall i = 1, \ldots, n$. Betrachten Sie zwei Kurven $\hat{g}_1$ und $\hat{g}_2$, die definiert sind durch

$$\hat{g}_1 = \arg\min_{g \in G_3} \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_a^b [g^{(3)}(x)]^2 dx \right)$$

$$\hat{g}_2 = \arg\min_{g \in G_4} \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_a^b [g^{(4)}(x)]^2 dx \right)$$

dabei bezeichnet $g^{(m)}$ die m-te Ableitung von g und G_3 bzw. G_4 bezeichne die Menge aller 3-fach bzw. 4-fach stetig differenzierbaren Funktionen auf $[a, b]$. Betrachten Sie folgende Fälle:

a) Besitzt $\hat{g}_1$ oder $\hat{g}_2$ einen kleineren Trainingsfehler für $\lambda \to \infty$?
b) Besitzt $\hat{g}_1$ oder $\hat{g}_2$ einen kleineren Testfehler für $\lambda \to \infty$?
c) Besitzt $\hat{g}_1$ oder $\hat{g}_2$ einen kleineren Trainingsfehler für $\lambda \to 0$?
d) Besitzt $\hat{g}_1$ oder $\hat{g}_2$ einen kleineren Testfehler für $\lambda \to 0$?

A 3 Betrachten Sie den Datensatz mcycle im Paket MASS. Dieser enthält Daten zu einem simulierten Motorradunfall im Rahmen eines Crash-Tests.

a) Führen Sie ein lineares Regressionmodell mit abhängiger Variable Beschleunigung (accel) und unabhängiger Variable Zeit (times) durch. Inwiefern weisen die diagnostischen Plots auf eine Verletzung der Modellannahmen hin?

b) Führen Sie eine polynomiale Regression von Grad 2 und Grad 3 durch:
   i. Formulieren Sie jeweils das Regressionsmodell.
   ii. Berechnen Sie die Anzahl freier Parameter aus den beiden Modellen.
   iii. Führen Sie die Regressionen in R durch. Welche Beschleunigung erwarten Sie jeweils nach einer Zeit von 10, 20, 30 und 40 msec? Leiten Sie dies aus den Schätzern der Regressionskoeffizienten her und überprüfen Sie Ihre Berechnung mit der predict-Funktion.
   iiii. Führen Sie für beide Modelle einen geeigneten statistischen Test durch, um zu überprüfen, ob auch ein lineares Modell ausreichen würde.

c) Führen Sie eine polynomiale Regression vom Grad p durch, dabei soll ein optimales p mittels Kreuzvalidierung bestimmt werden (Funktion cv.glm aus dem Paket boot). Testen Sie bei der Kreuzvalidierung Polynome vom Grad 1-10.

d) Bestimmen Sie auch für dieses Modell die Anzahl freier Parameter.

e) Vergleichen Sie die drei Modelle (Grad 2, Grad 3, Grad p) hinsichtlich der Anpassungsgüte. Welchen Parameter ziehen Sie zur Modellwahl heran?

f) Erstellen Sie ein Streudiagramm der Daten, in dem auch $\hat{f}$ für jedes der drei geschätzten Modelle dargestellt ist. In welchen Bereichen wirkt die Anpassung jeweils schlecht / gut / zu gut?