Prof. Dr. Andreas Weber
DHBW Karlsruhe
andreas.weber@dhbw-karlsruhe.de

Version 27. Juni 2022

Übungsaufgaben 7

Analyse des `Boston Housing`-Datensatz

Der `BostonHousing`-Datensatz (enthalten im R Paket `mlbench`) enthält Zensus-Daten aus Boston von 1970. Dabei gibt die Variable `medv` (median value) den Median des Werts der Häuser in dem jeweiligen Zensus-Bezirk in 1000 USD an.

Unser Ziel ist es, ein lineares Modell zu berechnen, das den Median des Werts der Häuser in dem jeweiligen Zensus-Bezirk mit Hilfe der anderen Variablen in diesem Datensatz schätzt.

1 Trainings- und Testdaten

Laden Sie den Datensatz mit

```
data("BostonHousing")
```

und teilen Sie den `BostonHousing`-Datensatz in Trainings- und Testdaten auf. Verwenden Sie hierbei ca. 80 % der Daten als Trainingsdaten und die verbleibenden 20 % als Testdaten.

2 Lineare Regression

Berechnen Sie ein Lineares Regressionsmodell für die Output Variable `medv` und verwenden Sie alle anderen Variablen als Input Variablen. Bestimmen Sie Trainings- und Testfehler (mittlere quadratische Abweichung).

3 Bias-Variance, Overfitting oder Underfitting?

Hat Ihr berechnetes Modell ein Problem mit einem hohen Bias oder einer hohen Varianz? Begründen Sie Ihre Antwort.

4 Ridge Regression

Berechnen Sie ein Ridge Regressionsmodell und bestimmen Sie den Testfehler (mittlere quadratische Abweichung). Vgl. Hinweise unten.

5 Lasso Regression

Berechnen Sie ein Lasso Regressionsmodell und bestimmen Sie den Testfehler (mittlere quadratische Abweichung). Geben Sie die Input-Variablen an, deren Koeffizienten in Ihrem finalen Modell mit Null geschätzt werden (Feature Selection). Vgl. Hinweisen unten.

6 Vergleich

Vergleichen Sie die Testfehler der drei berechneten Modelle und begründen Sie, weshalb durch die Regularisierung (Ridge/Lasso Regression) das Modell besser/schlechter wurde.

Hinweise

- Verwenden Sie Paket `glmnet` für Ridge Regression bzw. Lasso Regression. Vgl. [ISLR, 6.6 Lab2: Ridge Regression and the Lasso].
- Die Funktion

  ```
  glmnet(x= ...,y= ...,alpha = [0|1], lambda =...)
  ```

  berechnet Ridge Regression Modelle (`alpha` = 0) bzw. Lasso Regression Modelle (`alpha` = 1) und standardisiert automatisch die Variablen (Default Einstellung).
- Verwenden Sie für die Berechnung eines geeigneten Ridge Regression oder Lasso Regression Modells 10-fold Cross-Validation, um einen geeigneten Wert für den Strafterm-Parameter `lambda` zu bestimmen. Wählen Sie hierzu `lambda` aus dem Grid

  ```
  lambda <- 10^seq( from = 5, to = -3, length = 100)
  ```

  Sie können hierfür die Funktion

  ```
  cv.glmnet(x= ...,y= ...,alpha = [0|1], lambda =...)
  ```

  verwenden.