

Maschinelles Lernen

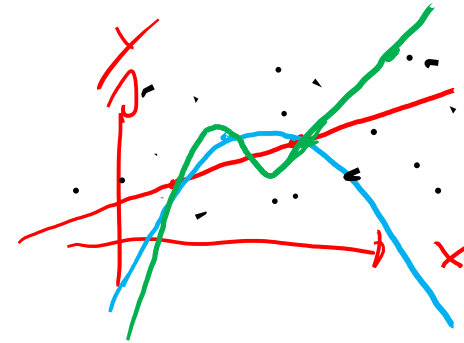
Modellselektion und -validierung

Prof. Dr. Rainer Stollhoff

Übersicht

- Motivation
- Bias-Varianz-Zerlegung
- Regularisierung in der Modellselektion
- Resampling in der Modellvalidierung

Motivation: Polynominterpolation



Aufgabe: Regression, d.h. Vorhersage $\hat{y} = \hat{y}(x) = f(x)$

Erfahrung: Datensatz $(x_i, y_i)_{i=1}^n$

Qualität: Verlustfunktion: $L(y, \hat{y}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - f(x_i; \theta))^2 = L(\theta)$

Maschine: Regression mit $f(x; \theta_1, \dots, \theta_m) = \theta_0 + \theta_1 \cdot x + \theta_2 \cdot x^2 + \dots + \theta_m \cdot x^m$

Lernen: Finde Werte für $\theta = (\theta_1, \dots, \theta_m)$, die die quadratische Verlustfunktion minimieren

Polynominterpolation:

Falls $m \geq n$ gibt es immer Werte für θ so dass $f(x_i; \theta) = y_i$ und damit $L(\theta) = 0$

$$(y_i - f(x_i; \theta)) = 0$$

mit $\ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j} = \frac{(x - x_0) \cdot \dots \cdot (x - x_{i-1}) \cdot (x - x_{i+1}) \cdot \dots \cdot (x - x_n)}{(x_i - x_0) \cdot \dots \cdot (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdot \dots \cdot (x_i - x_n)}$

Polynom x^n von Grad n

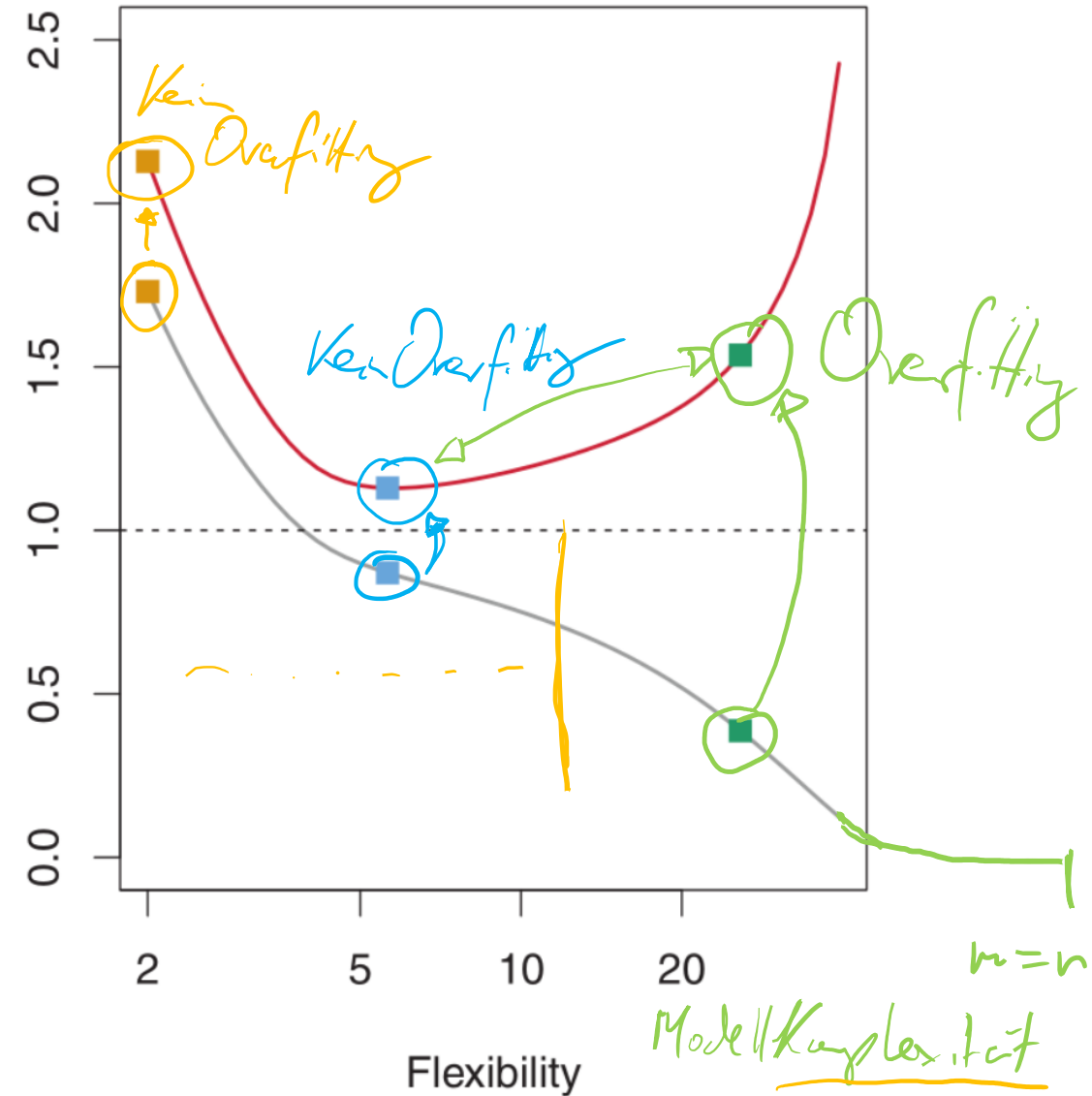
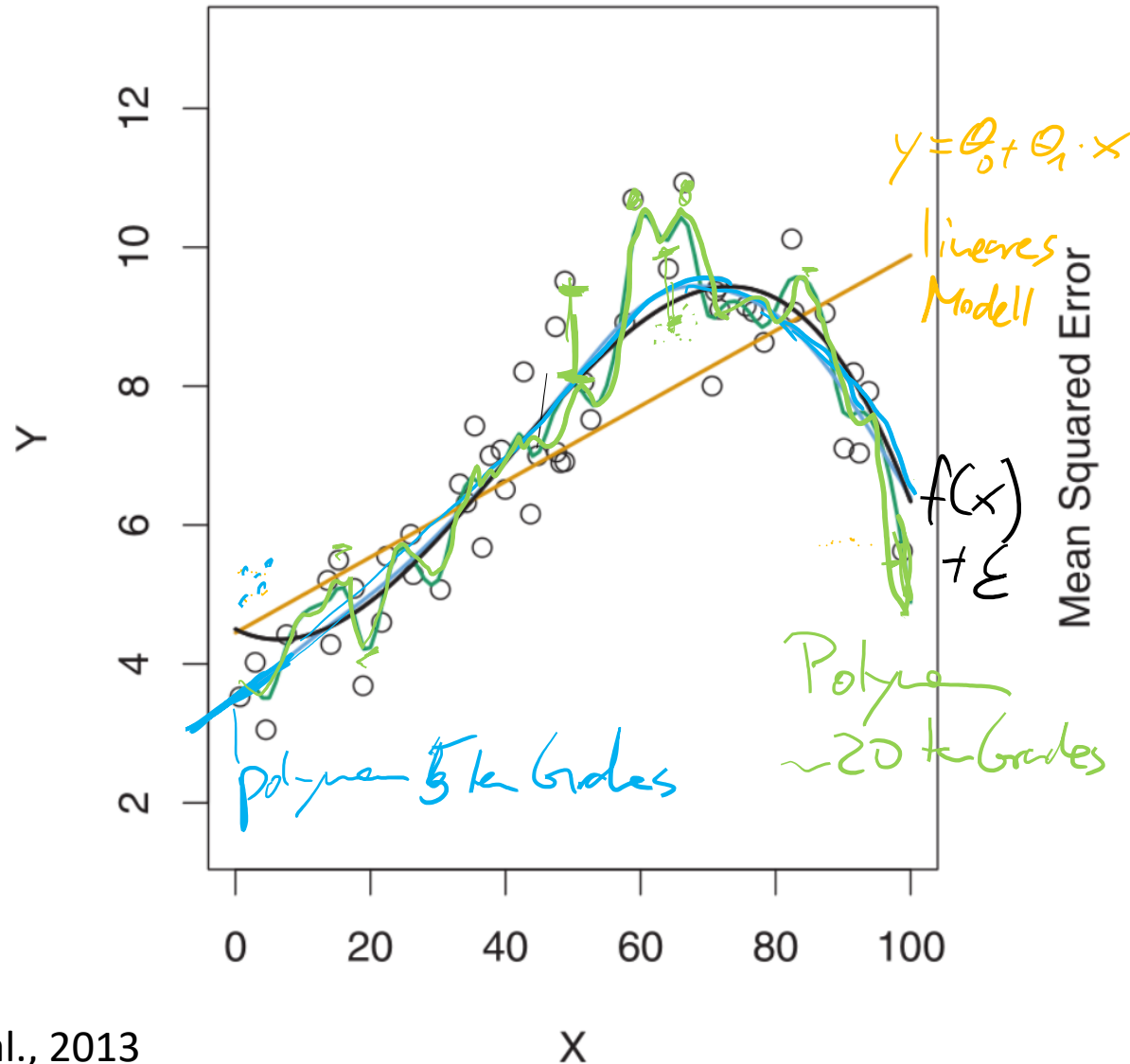
Zahl > 0

$\ell_i(x_k) = \begin{cases} 0 & i \neq k \\ 1 & i = k \end{cases}$

definiere $f(x) = \sum_{i=1}^n y_i \ell_i(x)$

$$f(x_k) = \sum_{i=1}^n y_i \cdot \ell_i(x_k) = y_k \cdot 1 = y_k$$

Motivation: Overfitting der Trainingsdaten



Bias-Varianz-Zerlegung des Vorhersagefehlers

- Bestmögliche Vorhersage

- Vorhersage $y = f(x) + \epsilon$ dabei $f(x)$ als bestmögliche Vorhersage und ϵ als echter Zufallswert

- Unvermeidbarer Fehler $E_{y,x}[(y - f(x))^2]$ $E_{y,x}$ Erwartungswert über alle möglichen (x, y)

- Bestmögliches geschätztes Modell

- Vorhersage $\hat{f}(x)$ minimiert $(f(x) - \hat{f}(x))^2$ über x

- Bias des Modells $E_x[(f(x) - \hat{f}(x))^2]$

- Bestmögliches auf einem Trainingsdatensatz geschätztes Modell

- Trainingsdatensatz $T = (x_i, y_i)$ für $i = 1, \dots, n$

- Vorhersage $\hat{y} = \hat{f}(x; T)$ minimiert $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$

- Varianz der Vorhersagen $E_{x,T}[(\hat{f}(x) - \hat{f}(x; T))^2]$

- Bias-Varianz-Zerlegung

$$E_{y,x}[(y - \hat{f}(x))^2] = E_{y,x}[(y - f(x))^2] + E_x[(f(x) - \hat{f}(x))^2] + E_{x,T}[(\hat{f}(x) - \hat{f}(x; T))^2]$$

Vorhersagefehler = Unvermeidbarer Fehler + Bias + Varianz

Aufgabenstellung

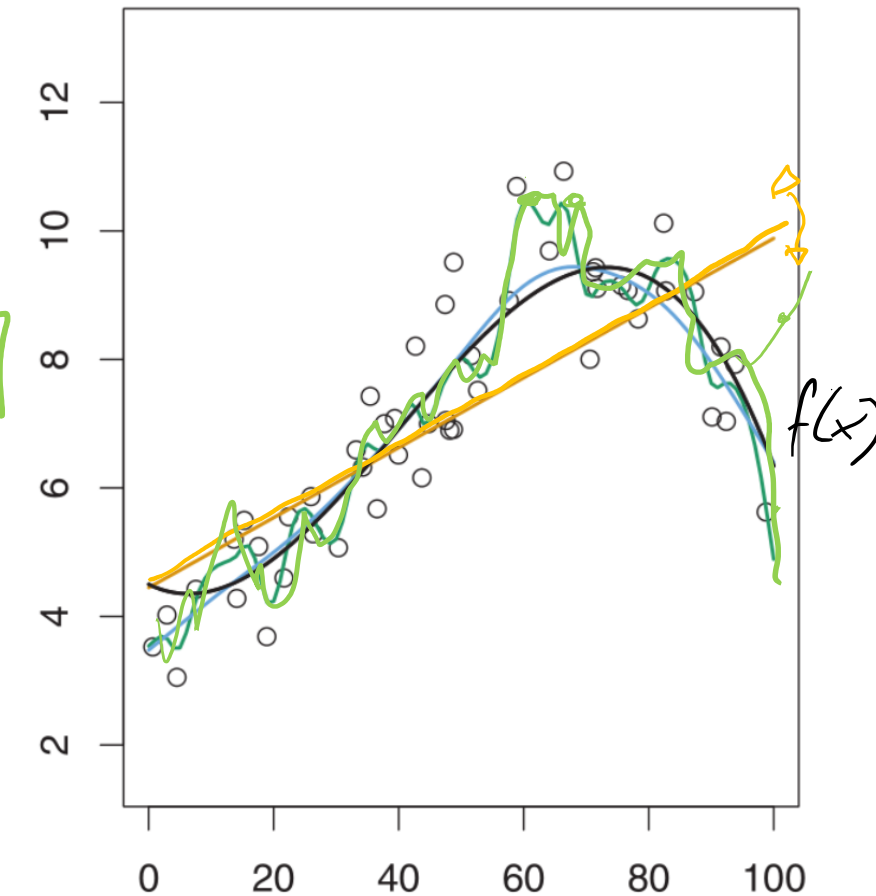
Modellkürze

Übersicht

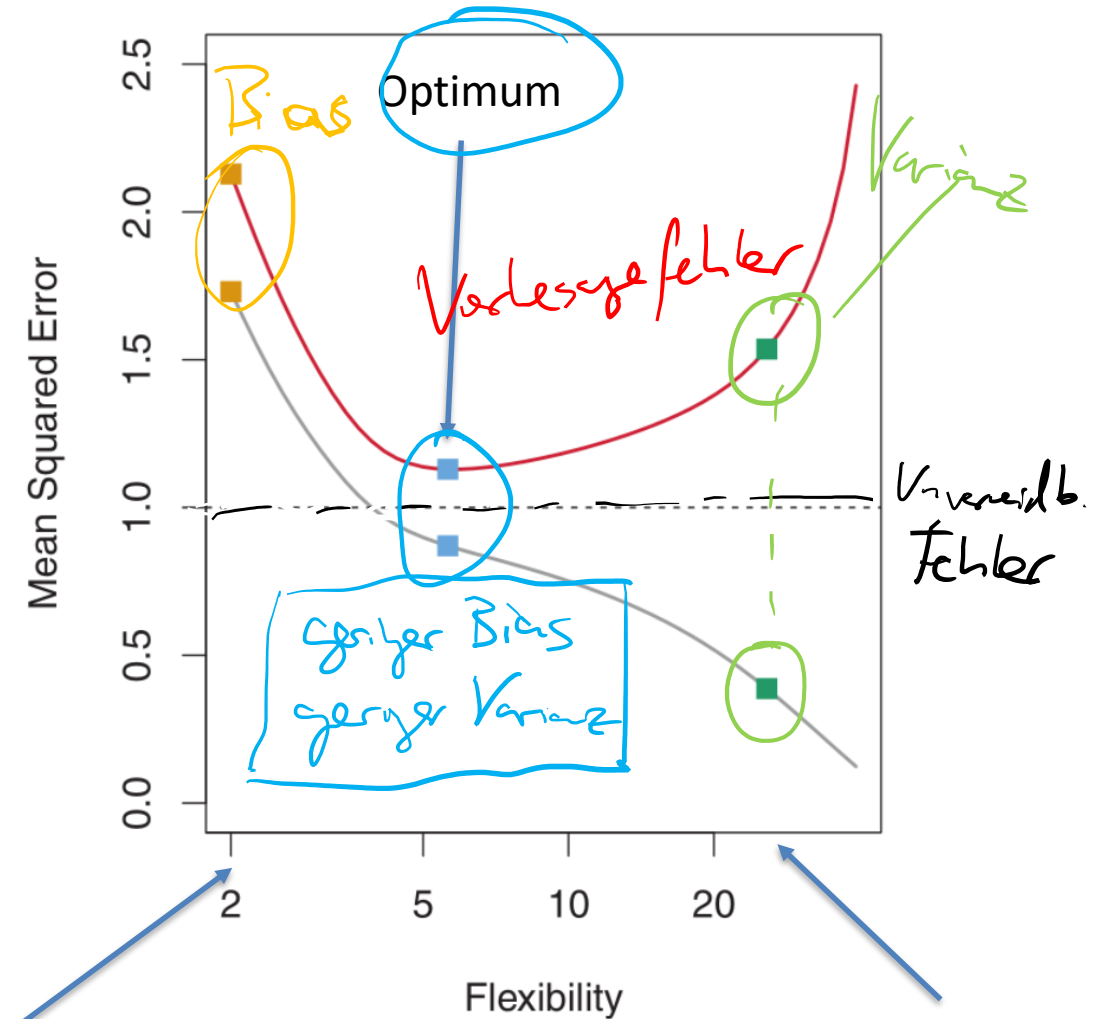
- Motivation
- Bias-Varianz-Zerlegung
- **Regularisierung in der Modellselektion**
 - Ziel: Reduktion der Modellkomplexität und damit Verringerung der Varianz
- Resampling in der Modellvalidierung

Bias-Varianz-Zerlegung des Vorhersagefehlers

- Unvermeidbarer Fehler
 $E_x[(y - f(x))^2]$
- Bias des Modells
 $E_x[(f(x) - \hat{f}(x))^2]$
- Varianz der Vorhersagen
 $E_{x,T}[(\hat{f}(x) - \hat{f}(x; T))^2]$
- Vorhersagefehler
 $E_x[(y - \hat{f}(x))^2]$



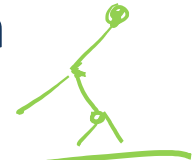
James et al., 2013 X



Hoher Bias
Geringe Varianz

Geringer Bias
Hohe Varianz

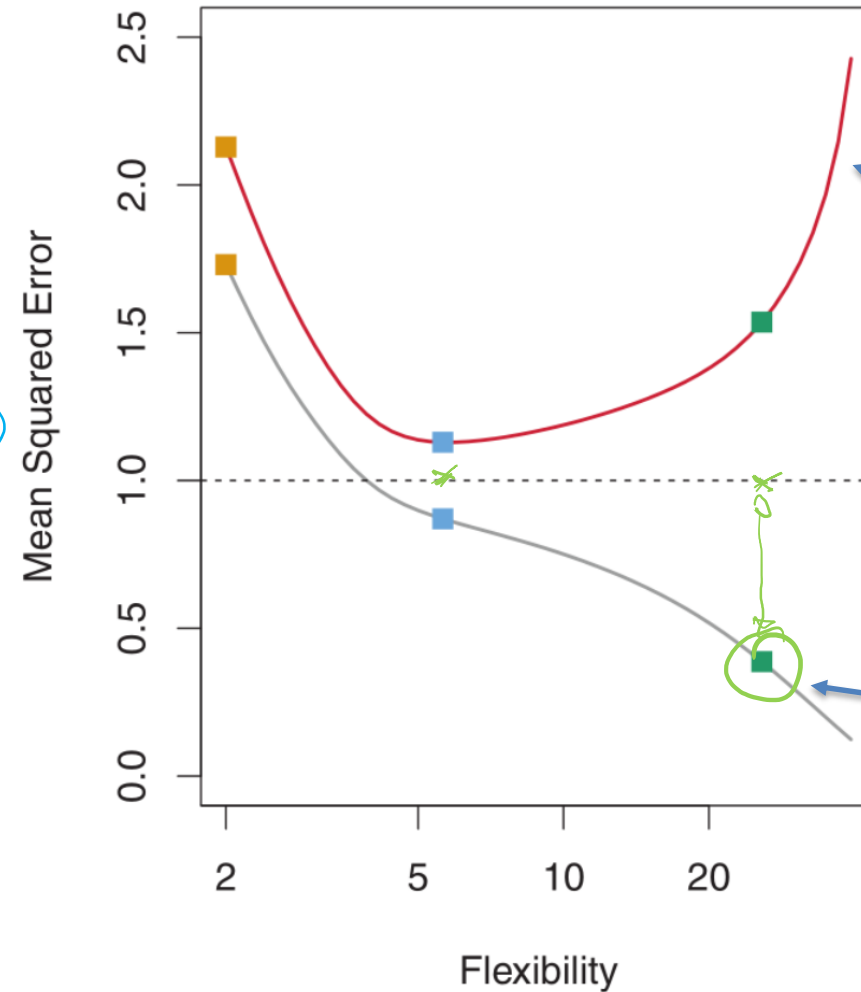
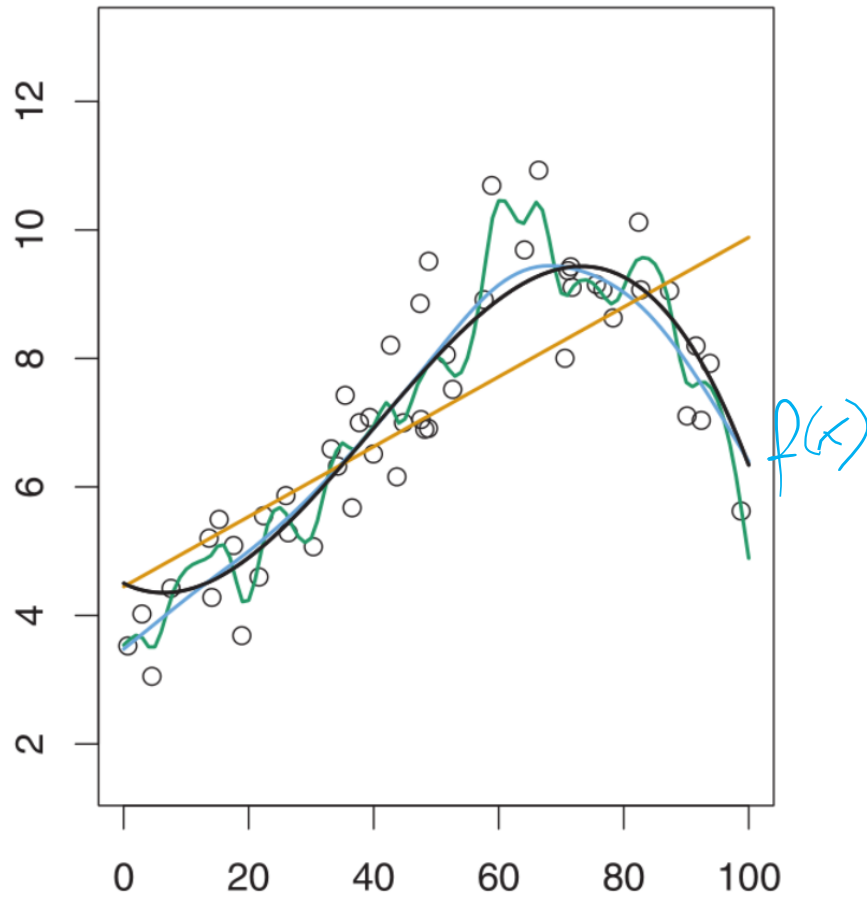
Methoden der Regularisierung in der Modellselektion

- Regularisierungsterm in der Verlustminimierung für parametrisches Modell mit $\theta = (\theta_1, \dots, \theta_m)$
 - Tikhonov Regularization / Ridge Regression:
Trade-Off
$$L(\theta; \lambda) = \sum_{i=1}^n \underbrace{(y_i - f(x_i; \theta))^2}_{\text{Fehler}} + \lambda \sum_{j=1}^m \underbrace{(\theta_j)^2}_{\text{Modellkomplexität}}$$
 - Akaike Information Criterion:
$$AIC(\theta) = -2 \left(\sum_i \log \hat{p}(y_i; x_i, \theta) \right) + 2 \cdot \underbrace{m}_{\text{Anzahl Parameter}}$$
 - Bayesian Information Criterion:
$$BIC(\theta) = -2 \left(\sum_i \log \hat{p}(y_i; x_i, \theta) \right) + \ln(n) \cdot \underbrace{m}_{\text{Anzahl Datenpunkte}}$$
- Early-Stopping, d.h. vorzeitiges Abbrechen der Optimierungsiterationen zum Beispiel

 - Begrenzen der Anzahl der aufeinanderfolgenden Splits bei Klassifikationsbäumen oder
 - Begrenzen der Anzahl der Boosting-Iterationen
- Feste Einschränkung der Modellkomplexität zum Beispiel
$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$
 - Beschränken des maximalen Grades einer polynomialen Funktion
 - Beschränken auf lineare Funktionen in der multivariaten Regression

Übersicht

- Motivation
- Bias-Varianz-Zerlegung
- Regularisierung in der Modellselektion
- **Resampling in der Modellvalidierung**
 - Ziel: Verlässlichere Schätzung des Vorhersagefehlers

Bias-Varianz-Zerlegung des Vorhersagefehlers

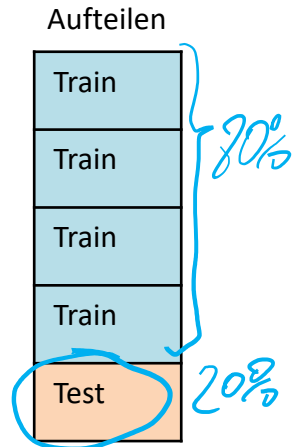


Echter Fehler unbekannt

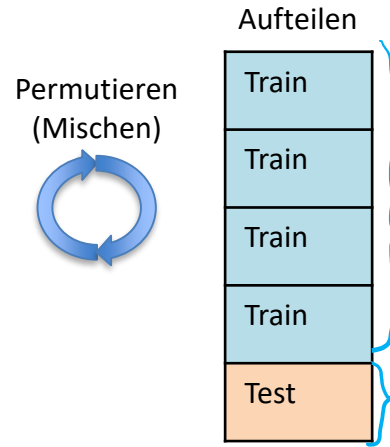
Trainingsfehler
zu optimistisch

Sampling zur Modellvalidierung

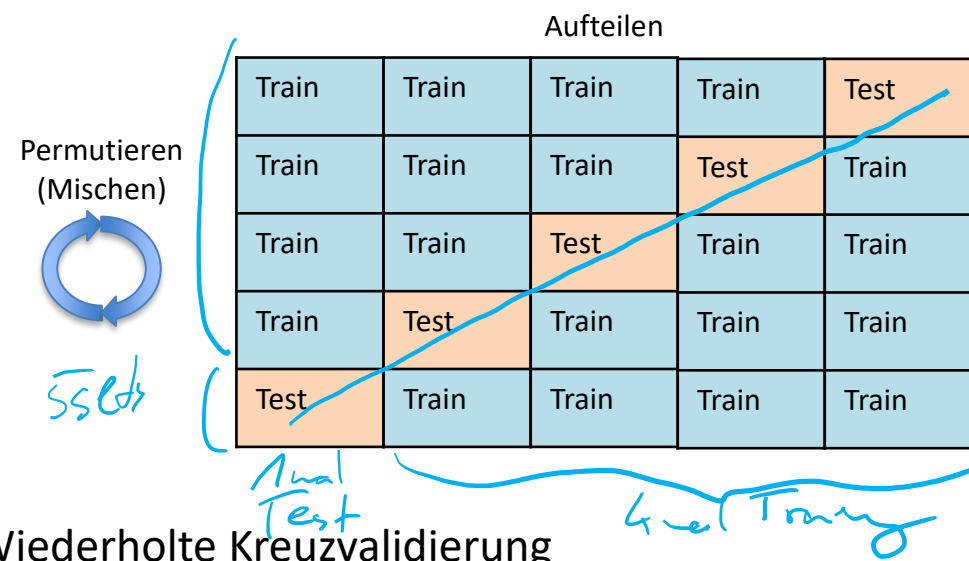
Festes Holdout



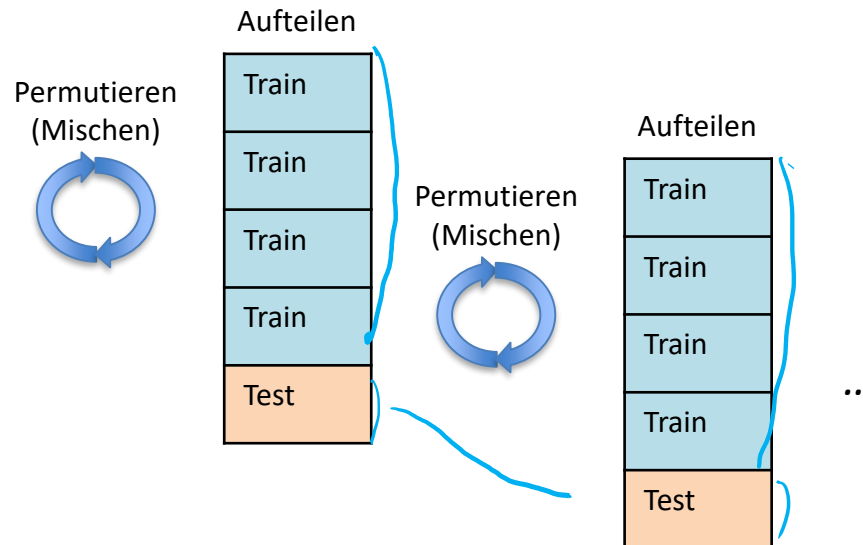
Zufälliges Holdout



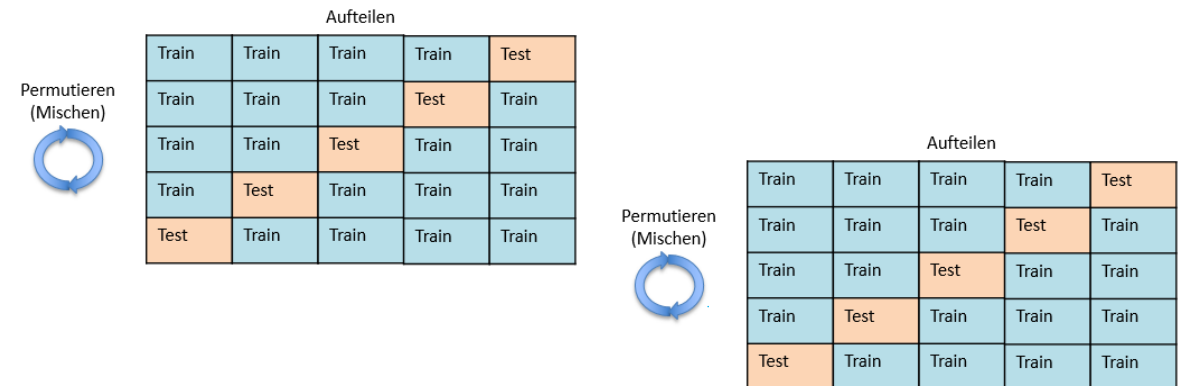
Kreuzvalidierung



Wiederholtes zufälliges Holdout



Wiederholte Kreuzvalidierung



Leave-One-Out Kreuzvalidierung

Training auf $n-1$ Daten, Test auf Restbeobachtung