Machine Learning

Prof. Dr. Fabian Brunner

<fa.brunner@oth-aw.de>

Amberg, 9. November 2021

Übersicht

Themen heute:

- Probabilistische Sicht auf die Least-Squares-Regression
- Bedingter Erwartungswert als optimaler Regressor
- Verzerrung-Varianz-Zerlegung
- Bias-variance tradeoff
- Kreuzvalidierung mittels Holdout-Methode
- k-fache Kreuzvalidierung
- Regularisierung

Bedingter Erwartungswert

Bedingter Erwartungswert, stetiger Fall

Sei (Ω, Σ, P) ein Wahrscheinlichkeitsraum und seien $X, Y : \Omega \to \mathbb{R}$ zwei stetige Zufallsvariablen mit gemeinsamer Dichte $f(x, y)$ sodass $\int_{-\infty}^{\infty} |y| f_Y(y) dy < \infty$.
Sei ferner $f_X(x) = \int_{\mathbb{R}} f(x, y)\, dy$ die Randdichte von X und

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

die bedingte Dichte von Y gegeben $X = x$.
Dann definiert der Ausdruck

$$E(Y|X = x) := \int_{\mathbb{R}} y f(y|x)\, dy$$

den bedingten Erwartungswert von Y gegeben $X = x$.

Für den diskreten Fall erfolgt die Definition analog mit der Wahrscheinlichkeitsfunktion.

Beispiel zum bedingten Erwartungswert (vgl. letzte Stunde)

Es wird das (unabhängige) Werfen zweier idealer Würfel betrachtet. Sei X die Augenzahl des ersten Würfels und Y die Augensumme beider Würfel.

a) Man bestimme $E(Y)$.

b) Man bestimme $E(Y|X = 1)$.

Lösung:

a) $E(Y) = 7$.

b) $E(Y|X = 1) = 4.5$.

Least-Squares-Regression

Setting:

- Seien X und Y zwei Zufallsvariablen mit gemeinsamer Dichte $f(x, y)$.
- Wir fassen X als Eingabegröße und Y als Ausgabegröße auf, die wir mit Hilfe von X vorhersagen möchten.
- Im besten Fall finden wir eine Funktion f, sodass $Y = f(X)$ gilt.
- Da eine solche Funktion meist nicht existiert, weicht man auf die Frage aus, für welche Funktion der Fehler klein wird. Genauer:

Allgemeines Least-Squares-Regressionsproblem

Gesucht ist eine Funktion f, für die der Ausdruck

$$E((Y - f(X))^2)$$

minimal wird.

Least-Squares-Regression

Bedingter Erwartungswert als optimaler Regressor

Die Lösung des o.g. Minimierungsproblems ist gegeben durch

$$\mu(x) = E(Y|X = x) \ .$$

Bemerkungen:

- Falls es einen funktionellen Zusammenhang der Form $Y = g(X)$ gibt, so gilt
$$\mu(x) = E(Y|X = x) = g(x) \ .$$
- Für den Fehler $e = Y - \mu(X)$ gilt
$$E(Y - \mu(X)|X = x) = 0 \ .$$

Beweis: Sei $\pi(x)$ eine beliebige Funktion. Dann gilt

$$
\begin{aligned}
E\left[(Y-\pi(X))^2\right] &= E\left[((Y-\mu(X))+(\mu(X)-\pi(X)))^2\right] \\
&= E\left[(Y-\mu(X))^2\right] + 2\underbrace{E\left[(Y-\mu(X))(\mu(X)-\pi(X))\right]}_{=0} \\
&\quad + E\left[(\mu(X)-\pi(X))^2\right] \\
&= E\left[(Y-\mu(X))^2\right] + E\left[(\mu(X)-\pi(X))^2\right] \\
&\geq E\left[(Y-\mu(X))^2\right]
\end{aligned}
$$

Nebenrechnung

$$
\begin{aligned}
E\left[(Y-\mu(X))(\mu(X)-\pi(X))\right] &= \int_{\mathbb{R}}\int_{\mathbb{R}}(y-\mu(x))(\mu(x)-\pi(x))\, f(x,y)dy\, dx \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}(y-\mu(x))f(x,y)\, dy\, (\mu(x)-\pi(x))\, dx \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}(y-\mu(x))f(y|x)f_X(x)\, dy(\mu(x)-\pi(x))\, dx \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}(y-\mu(x))f(y|x)\, dy\cdot f_X(x)(\mu(x)-\pi(x))\, dx \\
&= 0\ .
\end{aligned}
$$

Verzerrung-Varianz-Zerlegung

- Seien zunächst x und y zwei beliebige Datenpunkte und sei $\hat{y}_m(x)$ eine Modellschätzung, die aus einem Trainingsdatensatz der Länge m hervorgeht.
- Wir untersuchen nun den Fehler, der durch die zufällige Datenauswahl entsteht, d.h. wir betrachten $\hat{y}_m(x)$ als Realisierung einer Zufallsvariable $\hat{Y}_m(x)$.
- Mittelung (Erwartungswertbildung) über alle möglichen Trainingsdatensätze ergibt:

$$
\begin{aligned}
E\left[(y - \hat{Y}_m(x))^2\right] &= E\left[\left(y - E(\hat{Y}_m(x)) + E(\hat{Y}_m(x)) - \hat{Y}_m(x)\right)^2\right] \\
&= (y - E(\hat{Y}_m(x)))^2 + 2\underbrace{E\left[(y - E(\hat{Y}_m(x)))(E(\hat{Y}_m(x)) - \hat{Y}_m(x))\right]}_{=0} \\
&\quad + E\left[(E(\hat{Y}_m(x)) - \hat{Y}_m(x))^2\right] \\
&= (y - E(\hat{Y}_m(x)))^2 + \mathrm{Var}(\hat{Y}_m(x)) \; .
\end{aligned}
$$

Verzerrung-Varianz-Zerlegung

Nun mitteln wir noch über alle y für gegebenes x (d.h. wir bilden den bedingten Erwartungswert über Y unter der Bedingung $X = x$):

$$
\begin{aligned}
E\left[(Y - \hat{Y}_m(x))^2\right] &= E\left[(Y - E(\hat{Y}_m(x)))^2\right] + \mathrm{Var}(\hat{Y}_m(x)) \\
&= E\left[Y - \mu(x) + \mu(x) - E(\hat{Y}_m(x))^2\right] + \mathrm{Var}(\hat{Y}_m(x)) \\
&= E\left[(Y - \mu(x))^2\right] + 2\underbrace{E\left[(Y - \mu(x))(\mu(x) - E(\hat{Y}_m(x)))\right]}_{=0} \\
&\quad + (\mu(x) - E(\hat{Y}_m(x)))^2 + \mathrm{Var}(\hat{Y}_m(x)) \\
&= E\left[(Y - \mu(x))^2\right] + (\mu(x) - E(\hat{Y}_m(x)))^2 + \mathrm{Var}(\hat{Y}_m(x))
\end{aligned}
$$

Verzerrung-Varianz-Zerlegung

Verzerrung-Varianz-Zerlegung bei der Regressionsanalyse

Für festes x hat der erwartete Fehler die folgende Zerlegung:

$$E\left[(Y - \hat{Y}_m(x))^2\right] = \underbrace{E\left[(Y - \mu(x))^2\right]}_{\text{irreduzibler Fehler}} + \underbrace{(\mu(x) - E(\hat{Y}_m(x)))^2}_{\text{Verzerrung}^2} + \underbrace{\mathrm{Var}(\hat{Y}_m(x))}_{\text{Varianz}} .$$

Bemerkung: Die Erwartungswertbildung erfolgt bezüglich Y (gegeben $X = x$) und bezüglich der möglichen Trainingsdatensätze.

Verzerrung-Varianz-Zerlegung

Verzerrung-Varianz-Zerlegung bei der Regressionsanalyse

Für festes x hat der erwartete Fehler die folgende Zerlegung:

$$E\left[(Y - \hat{Y}_m(x))^2\right] = \underbrace{E\left[(Y - \mu(x))^2\right]}_{\text{irreduzibler Fehler}} + \underbrace{(\mu(x) - E(\hat{Y}_m(x)))^2}_{\text{Verzerrung}^2} + \underbrace{\mathrm{Var}(\hat{Y}_m(x))}_{\text{Varianz}} .$$

Bemerkung: Die Erwartungswertbildung erfolgt bezüglich Y (gegeben $X = x$) und bezüglich der möglichen Trainingsdatensätze.

- Der **irreduzible Fehler** gibt an, wie schwer es (intrinsisch) ist, Y an der Stelle $X = x$ vorherzusagen. Er kann nicht beeinflusst werden.

Verzerrung-Varianz-Zerlegung

Verzerrung-Varianz-Zerlegung bei der Regressionsanalyse

Für festes x hat der erwartete Fehler die folgende Zerlegung:

$$E\left[(Y - \hat{Y}_m(x))^2\right] = \underbrace{E\left[(Y - \mu(x))^2\right]}_{\text{irreduzibler Fehler}} + \underbrace{(\mu(x) - E(\hat{Y}_m(x)))^2}_{\text{Verzerrung}^2} + \underbrace{\mathrm{Var}(\hat{Y}_m(x))}_{\text{Varianz}} .$$

Bemerkung: Die Erwartungswertbildung erfolgt bezüglich Y (gegeben $X = x$) und bezüglich der möglichen Trainingsdatensätze.

- Der **irreduzible Fehler** gibt an, wie schwer es (intrinsisch) ist, Y an der Stelle $X = x$ vorherzusagen. Er kann nicht beeinflusst werden.
- Die sog. **Verzerrung** (engl.: **bias**) bzw. der (systematische) **Approximationsfehler** gibt an, wie gut sich μ durch die Modellfunktion approximieren lässt.

Verzerrung-Varianz-Zerlegung

Verzerrung-Varianz-Zerlegung bei der Regressionsanalyse

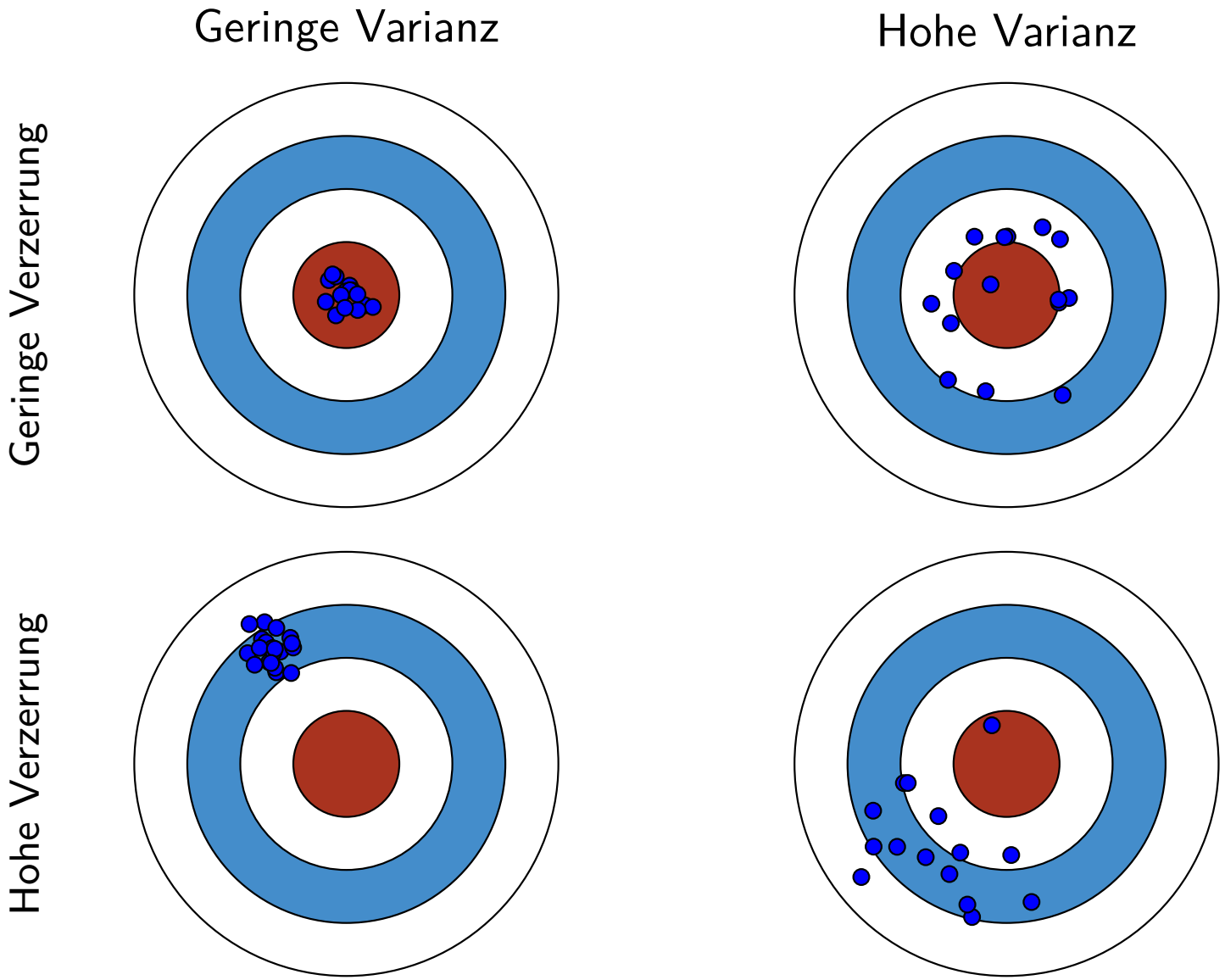
Für festes x hat der erwartete Fehler die folgende Zerlegung:

$$E\left[(Y-\hat{Y}_m(x))^2\right] = \underbrace{E\left[(Y-\mu(x))^2\right]}_{\text{irreduzibler Fehler}} + \underbrace{(\mu(x)-E(\hat{Y}_m(x)))^2}_{\text{Verzerrung}^2} + \underbrace{\mathrm{Var}(\hat{Y}_m(x))}_{\text{Varianz}} .$$

Bemerkung: Die Erwartungswertbildung erfolgt bezüglich Y (gegeben $X = x$) und bezüglich der möglichen Trainingsdatensätze.

- Der **irreduzible Fehler** gibt an, wie schwer es (intrinsisch) ist, Y an der Stelle $X = x$ vorherzusagen. Er kann nicht beeinflusst werden.
- Die sog. **Verzerrung** (engl.: **bias**) bzw. der (systematische) **Approximationsfehler** gibt an, wie gut sich μ durch die Modellfunktion approximieren lässt.
- Der dritte Term gibt die **Varianz** der Schätzung der Regressionsfunktion bei Verwendung unterschiedlicher Trainingsdatensätze an.

Verzerrung und Varianz - Illustration



Bias variance tradeoff

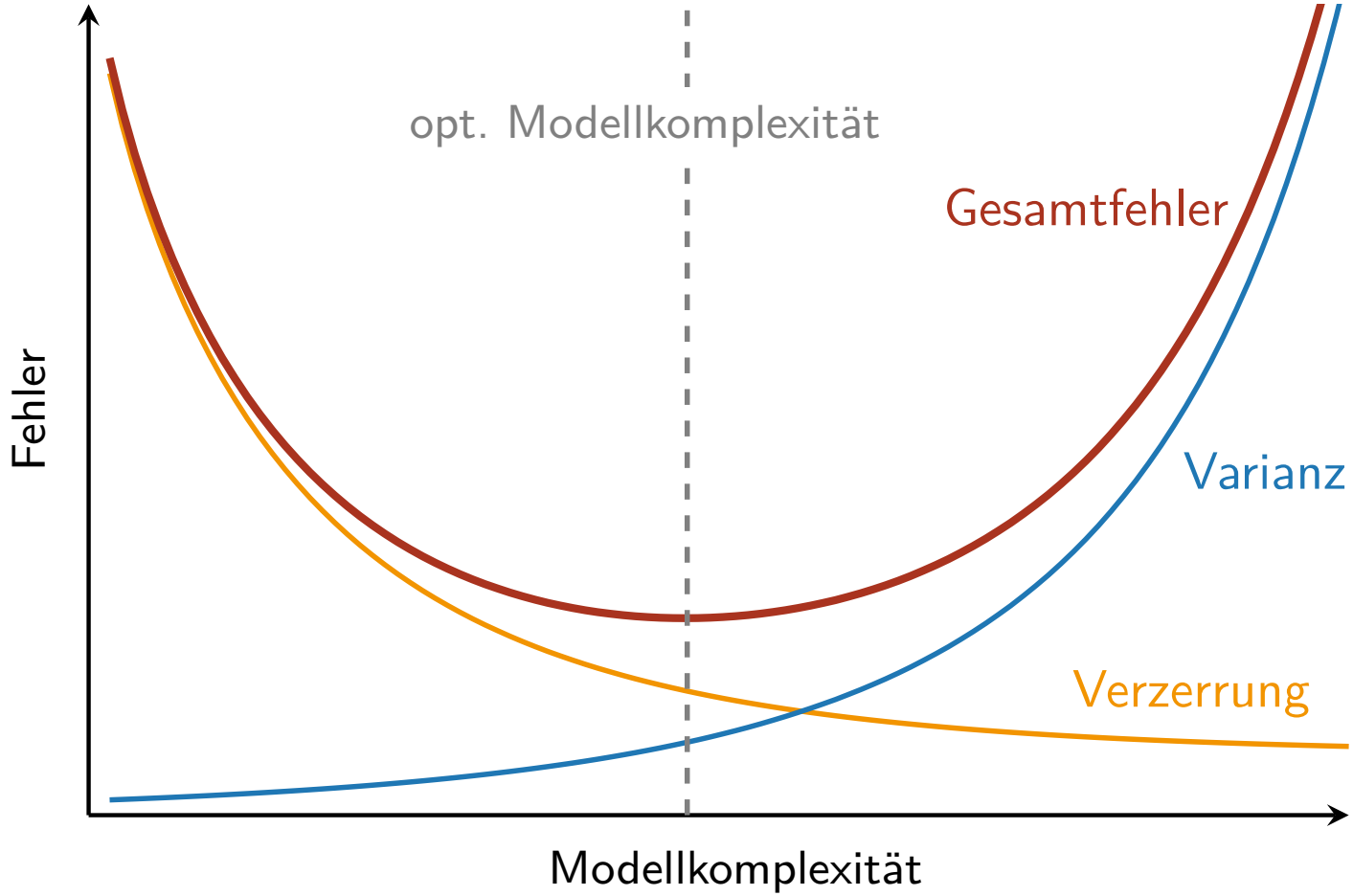
- Modelle mit hoher Varianz haben die Eigenschaft, dass sie schlecht auf neue, ungesehene Daten verallgemeinern (hoher **generalization error**).
- Dies ist beispielsweise dann der Fall, wenn das verwendete Modell zu komplex ist, d.h. wenn es zu viele erklärende Variablen enthält.
- Dann passt sich das Modell zu gut an die Trainingsdaten an (**Übertraining** bzw. **Overfitting**) und erfasst z.B. auch unerwünschtes Rauschen.
- Ist die gewählte Modellkomplexität hingegen nicht hinreichend, so äußert sich dies in einem großen Approximationsfehler, da relevante Zusammenhänge nicht abgebildet werden (**Untertraining** bzw. **Underfitting**).

Bias variance tradeoff

- Verzerrung und Varianz können nicht unabhängig voneinander eliminiert werden.
- Sie stehen typischerweise in einer wechselseitigen Beziehung.
- Es muss daher ein „Mittelweg“ zwischen einem zu einfachen und einem zu komplexen Modell gewählt werden („bias variance tradeoff“).

Bias variance tradeoff

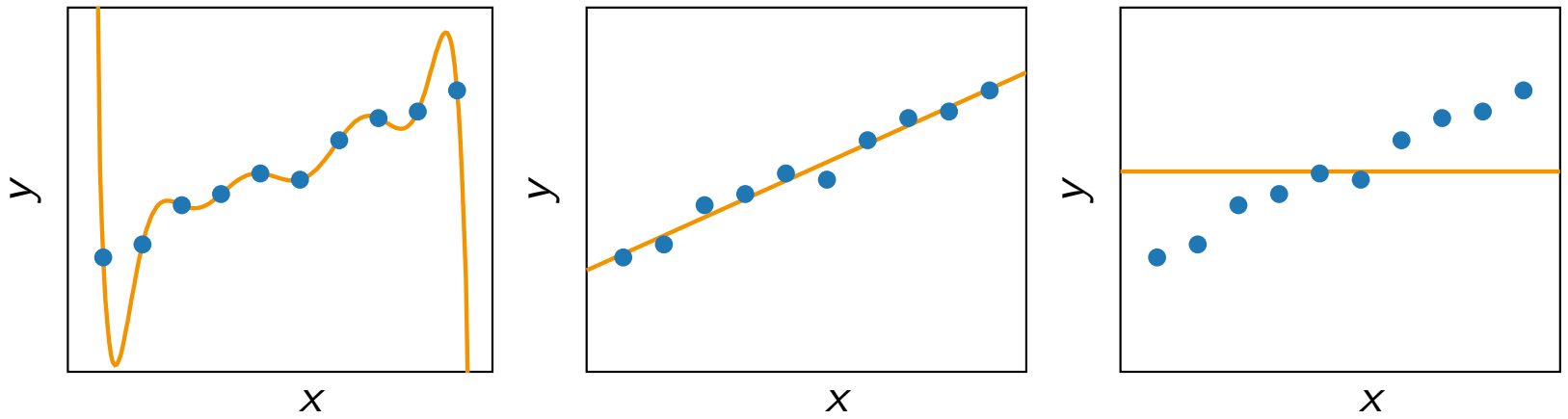
Verzerrung und Varianz werden durch die Modellkomplexität beeinflusst:



Overfitting und Underfitting

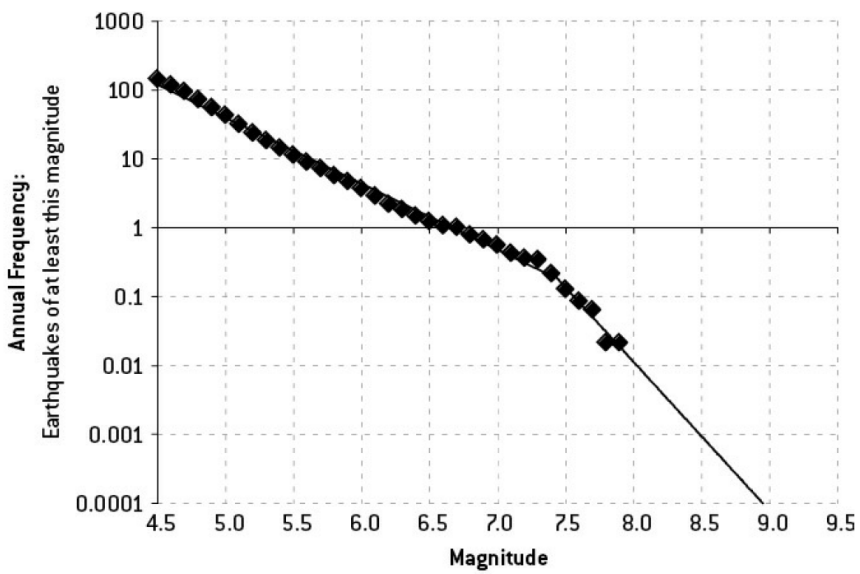
Verständnisfrage:

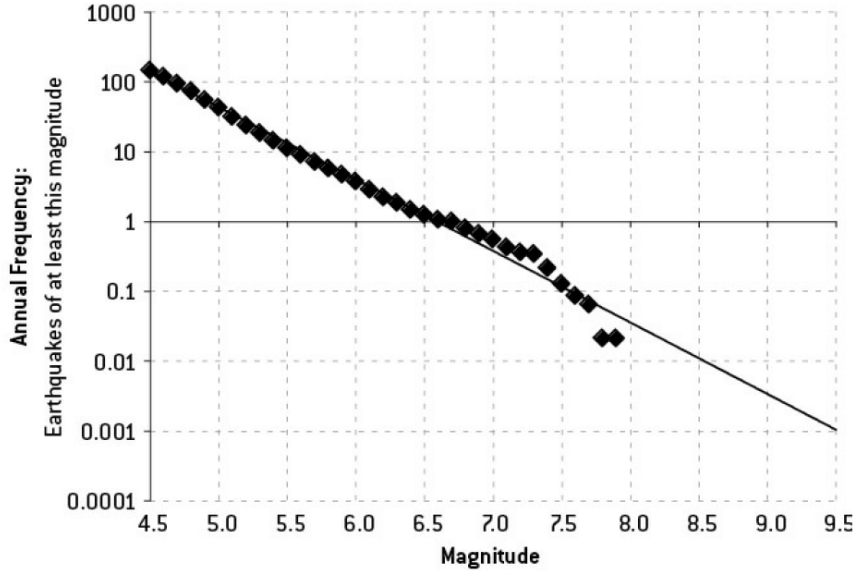
Bei welchem der folgenden Modelle liegt Underfitting vor, bei welchem Overfitting?



Beispiel für Overfitting

- Modellierung der Erdbebenhäufigkeit in der Region Fukushima anhand historischer Daten der letzten 400 Jahre (s. unten).
- Gutenberg-Richter-Gesetz: lineare Beziehung zwischen der Stärke eines Erdbebens und der (logarithmierten) Häufigkeit des Auftretens.
- Modellierung der Ingenieure mit überangepasstem Modell (linke Grafik): ein Erdbeben mit Stärke >9 alle 13.000 Jahre.
- Auslegung des Atomkraftwerks auf Stärke 8.6
- Ergebnis mit linearem Modell (rechte Grafik): alle 300 Jahre.
- 11.03.2011: Erdbeben der Stärke 9.1





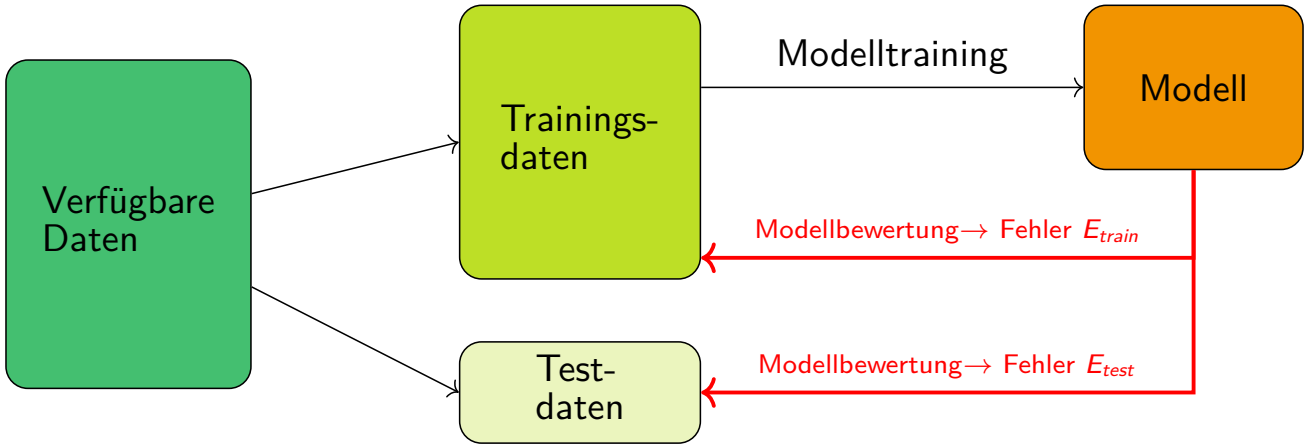
Quelle: N. Silver: The signal and the noise (2012)

Modellvalidierung mittels Holdout-Methode

Bewertung von Verzerrung und Varianz

Um Varianz und Verzerrung für ein vorgegebenes Modell zu bewerten, wird der gegebene Datensatz vorab gesplittet und das Modelltraining nur auf einem Teil des Datensatzes ausgeführt. Anschließend kann man die Performance auf beiden Datensätzen auswerten und vergleichen.

Schematischer Ablauf der Holdout-Methode



Vertändnisfrage: Warum sollte der Split in Trainings- und Testdaten zufällig erfolgen?

Modellvalidierung mittels k-facher Kreuzvalidierung

Ostbayerische Technische Hochschule Amberg-Weiden

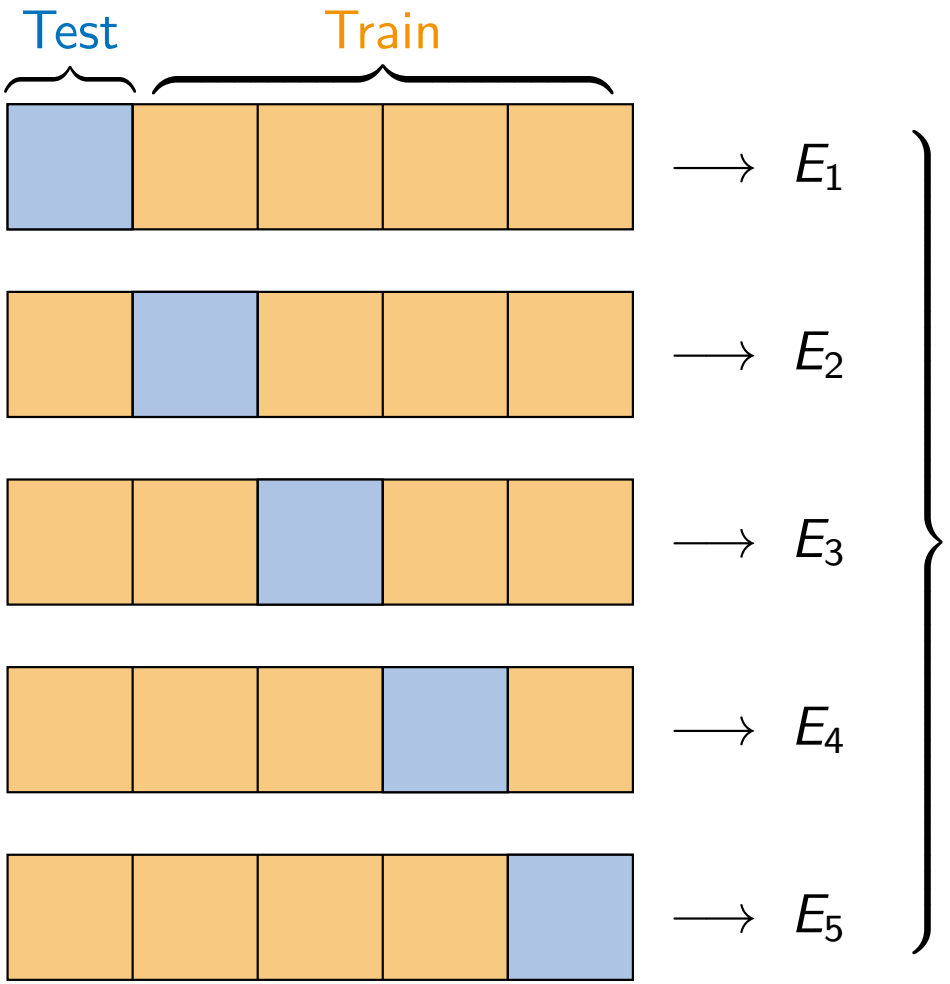
Holdout-Methode

- Modelltraining auf Trainingsdatensatz, Modellbewertung auf dem Testdatensatz.
- Durch ungünstige (zufällige) Auswahl der Testdaten kann die Varianz des Modells als zu hoch bzw. zu gering eingeschätzt werden.

k-fache Kreuzvalidierung

- Bei der k-fachen Kreuzvalidierung wird das Vorgehen k-mal wiederholt und anschließend der Mittelwert über die Leistungsmaße E_i, $i \in \{1, \ldots, k\}$ gebildet. Dadurch ist ein weniger verzerrtes Ergebnis zu erwarten.
- Zu diesem Zweck wird der Datensatz in k etwa gleich große Datensätze zerlegt, von denen jeder genau einmal als Testdatensatz verwendet wird.
- Am Ende wurden alle Datenbeispiele $k-1$-mal zum Training und genau einmal zum Testen verwendet.

k-fache Kreuzvalidierung zur Modellbewertung



Gesamt-Leistungsbewertung:

$$E = \frac{1}{5} \sum_{i=1}^{5} E_i$$

Modellauswahl

Ostbayerische Technische Hochschule Amberg-Weiden

Eine wichtige Aufgabe bei der Modellerstellung ist die Auswahl eines geeigneten Modells (engl. „model selection“). Diese erfordert verschiedene Festlegungen:

- Auswahl der Modellklasse (z.B. KNN, Entscheidungsbaum, Neuronales Netz etc.)
- Auswahl und ggf. Erzeugung geeigneter Features (engl. „feature selection“, „feature engineering“)
- Festlegung der Modellparametrisierung anhand der (modellspezifischen) Hyperparameter (z.B. Tiefe des Baums, Regularisierungsparameter etc.)
- Festlegung der Trainingsmethode (z.B. Fehlerfunktional, Optimierungsverfahren)

Diese Festlegungen sollten so getroffen werden, dass der Gesamtfehler möglichst gering wird, d.h. dass weder Overfitting noch Underfitting vorliegt.

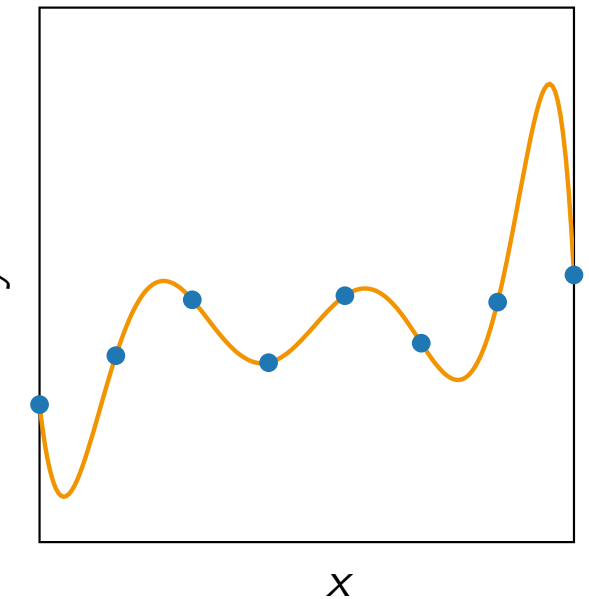
Beispiel zur Modellauswahl bei der polynomialen Regression

Angenommen, Sie möchten ein polynomiales Regressionsmodell auf dem nebenstehend abgebildeten Datensatz erstellen. Dann gehört die Festlegung des Polynomgrads zur Modellauswahl.

- Bei Wahl des Polynomgrads 7 und Schätzung der Parameter $\theta_0, \ldots, \theta_7$ verschwindet der Fehler auf dem Trainingsdatensatz:

$$J(\boldsymbol{\theta}) = \frac{1}{16} \sum_{i=1}^{8} (f_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)}) = 0 \ .$$

- Auf einem unabhängigen Testdatensatz wäre dies nicht der Fall (Overfitting).
- Der Polynomgrad wurde zu hoch gewählt.



$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \ldots + \theta_7 x^7$$

Frage: wie würden Sie vorgehen, um den Polynomgrad geeignet zu bestimmen?

Naives Vorgehen bei der Modellauswahl

Hypothetisches Vorgehen zur Auswahl des Polynomgrads in der obigen Situation:

- Train-Test-Split der verfügbaren Daten
- Training mehrerer Regressionsmodelle für die Polynomgrade $d = 0, \ldots, 7$ auf dem Trainingsdatensatz
- Auswertung des MSE für jedes Modell auf dem Testdatensatz
- Auswahl des Modells, bei dem der Fehler auf dem Testdatensatz am geringsten war.

Arbeitsauftrag
Diskutieren Sie dieses Vorgehen.

Modellauswahl mittels Holdout-Methode

Ostbayerische Technische Hochschule Amberg-Weiden

Holdout-Methode: Train-Test-Validate-Split

Schritt 1 Split der Daten in drei Teile: einen Trainings-, einen Validierungs- und einen Testdatensatz (z.B. 70%, 20%, 10%)

Schritt 2 Hyperparameter-Optimierung: Trainiere das Modell für mehrere verschiedene Werte der Hyperparameter auf dem Trainingsdatensatz.

Schritt 3 Bewerte die Leistung der Modelle auf dem Validierungsdatensatz und wähle diejenigen Werte der Hyperparameter, für die die beste Performance gemessen wurde.

Schritt 4 Füge Trainings- und Validierungsdatensatz zusammen und trainiere ein Modell darauf. Verwende dazu die Hyperparameter, die im vorherigen Schritt bestimmt wurden.

Schritt 5 Bewerte die Übertragbarkeit auf unbekannte Daten durch Auswertung auf dem (bisher unbenutzten) Testdatensatz.

Beispiel zur Modellauswahl mittels Holdout-Methode

Hyperparameter-Optimierung: welcher Polynomgrad ist geeignet?

$$
\begin{aligned}
f_{\boldsymbol{\theta}}(x) &= \theta_0 && \longrightarrow \boldsymbol{\theta}_0 && \longrightarrow E_{val}(\boldsymbol{\theta}_0) \\
f_{\boldsymbol{\theta}}(x) &= \theta_0 + \theta_1 x && \longrightarrow \boldsymbol{\theta}_1 && \longrightarrow E_{val}(\boldsymbol{\theta}_1) \\
f_{\boldsymbol{\theta}}(x) &= \theta_0 + \theta_1 x + \theta_2 x^2 && \longrightarrow \boldsymbol{\theta}_2 && \longrightarrow E_{val}(\boldsymbol{\theta}_2) \\
f_{\boldsymbol{\theta}}(x) &= \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 && \longrightarrow \boldsymbol{\theta}_3 && \longrightarrow E_{val}(\boldsymbol{\theta}_3) \\
&\vdots \\
f_{\boldsymbol{\theta}}(x) &= \theta_0 + \theta_1 x + \ldots + \theta_7 x^7 && \longrightarrow \boldsymbol{\theta}_7 && \longrightarrow E_{val}(\boldsymbol{\theta}_7)
\end{aligned}
$$

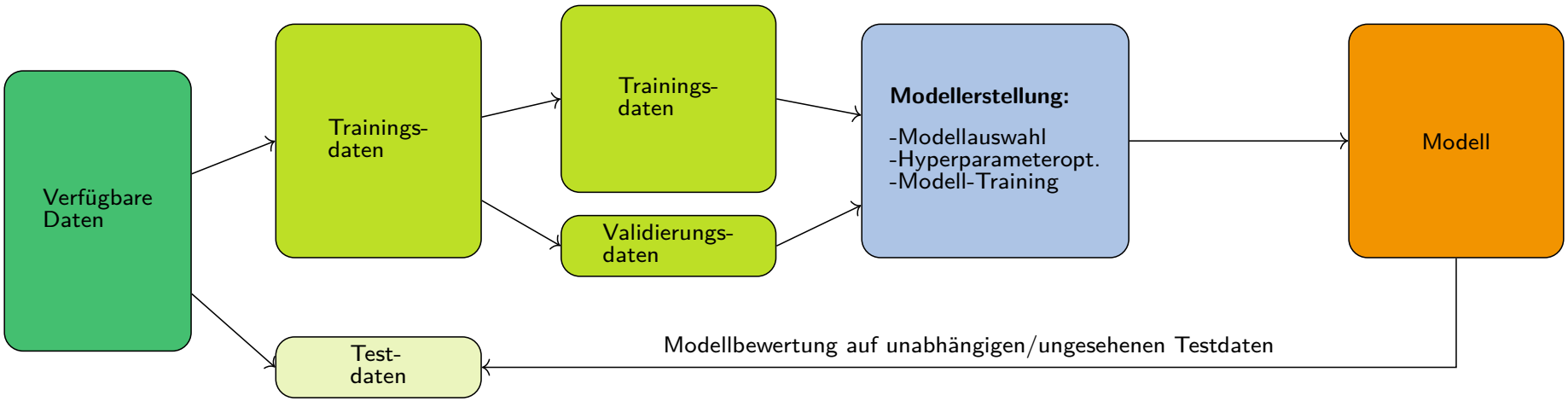
Vorgehen:

- Berechne für verschiedene Polynomgrade den Modell-Fit $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_7$
- Werte die Fehler $E_{val}(\boldsymbol{\theta}_0), \ldots, E_{val}(\boldsymbol{\theta}_7)$ auf dem Validierungsdatensatz aus.
- Ermittle den Polynomgrad, für den der Fehler auf dem Validierungsdatensatz am geringsten ist.
- (optional: Trainiere das Modell auf dem Trainings- und Validierungsdatensatz und verwende den oben ermittelten Polynomgrad)
- Bewerte das resultierende Modell auf dem Testdatensatz.

Modellerstellung mit der Holdout-Methode

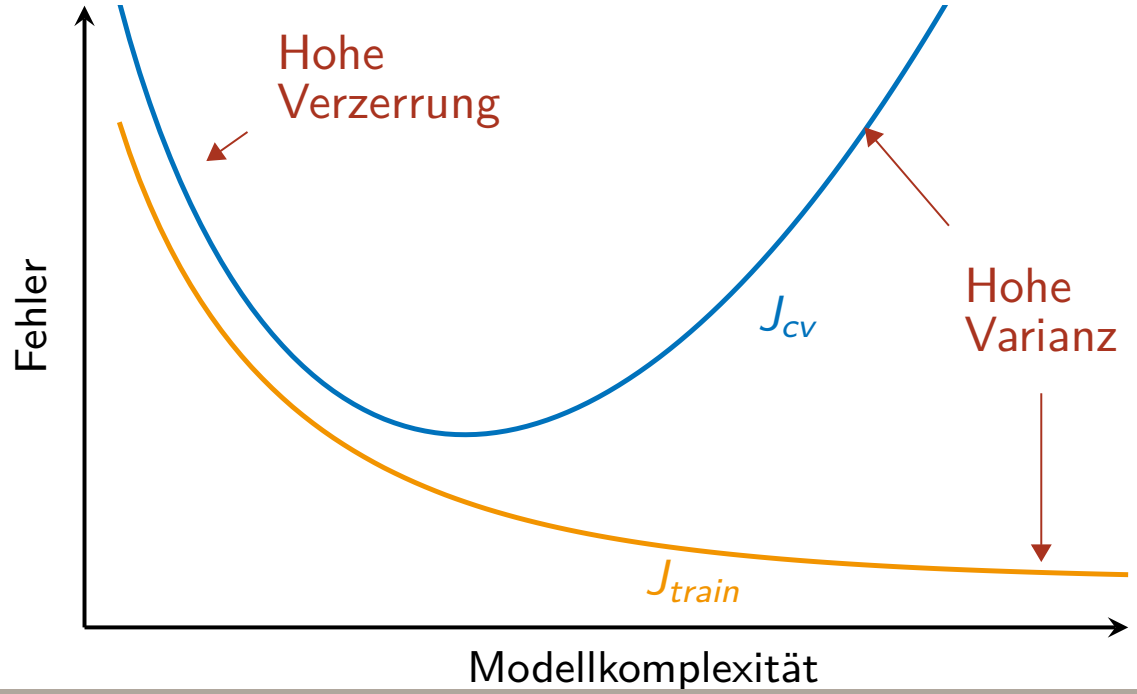
Schematischer Ablauf der Modellerstellung mittels Holdout-Methode:

- Das Modelltraining erfolgt ausschließlich auf dem Trainingsdatensatz.
- Die Bewertung verschiedener Modellansätze oder Hyperparameter im Zuge des Modelltrainings erfolgt auf dem Validierungsdatensatz
- Die abschließende Modellbewertung erfolgt auf dem Testdatensatz.



Lernkurven

- Zur Festlegung der Modellkomplexität oder Optimierung von Hyperparametern können die Fehler/Bewertungsmaße auf dem Trainingsdatensatz und auf dem Validierungsdatensatz für verschieden komplexe Modelle verglichen werden.
- Hohe Verzerrung: Fehler auf dem Trainingsdatensatz und auf dem Validierungsdatensatz sind groß.
- Hohe Varianz: Fehler auf dem Trainingsdatensatz ist klein und Fehler auf dem Validierungsdatensatz ist groß.



k-fache Kreuzvalidierung zur Hyperparameter-optimierung

Die k-fache Kreuzvalidierung kommt häufig bei der **Hyperparameteroptimierung** zusammen mit **Rastersuche** zum Einsatz, um geeignete Hyperparameter für ein Modell zu bestimmen.

Schritt 1 Split der Daten in zwei Teile: einen Trainingsdatensatz und einen Testdatensatz

Schritt 2 Ermittlung der besten Hyperparameter durch Rastersuche. Für jede Parameter-Konfiguration wird k-fache Kreuzvalidierung auf dem Trainingsdatensatz angewendet. Die Konfiguration der Hyperparameter, für die das (gemittelte) Leistungsmaß das Optimum annimmt, wird anschließend verwendet.

Schritt 3 Modelltraining auf dem gesamten Trainingsdatensatz unter Verwendung der ermittelten optimalen Hyperparameter.

Schritt 4 Modellauswertung auf dem Testdatensatz

Regularisierung

Ansatz bei der polynomialen Regression:

$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \ldots + \theta_d x^d \ .$$

Die Parameter $\boldsymbol{\theta}$ werden durch Minimierung des Kostenfunktionals

$$J(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} (f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2$$

bestimmt (Normalgleichungen oder Gradienten-Verfahrens).

Regularisierung
Durch Regularisierung kann die Varianz des Modells verringert werden, indem ein Regularisierungsterm zum Funktional J addiert wird. Der Parameter $\lambda \geq 0$ heißt **Regularisierungsparameter**.

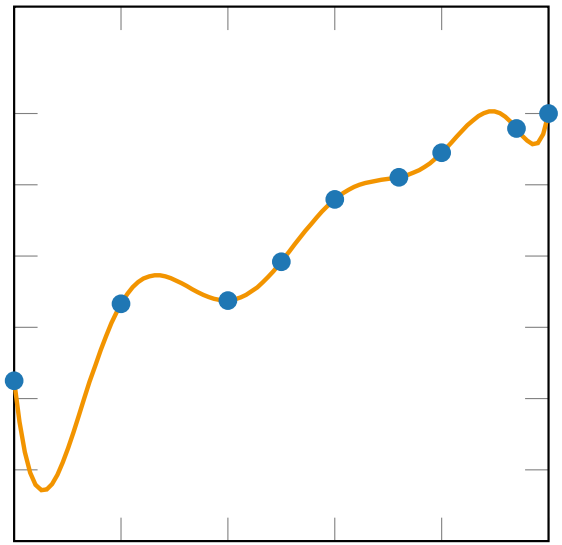
Regularisiertes Least-Squares-Funktional bei der linearen Regression

$$J_\lambda(\boldsymbol{\theta}) = \frac{1}{2m} \left[\sum_{i=1}^{m} (f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{d} \theta_j^2 \right] \ .$$

Beispiel zur Regularisierung

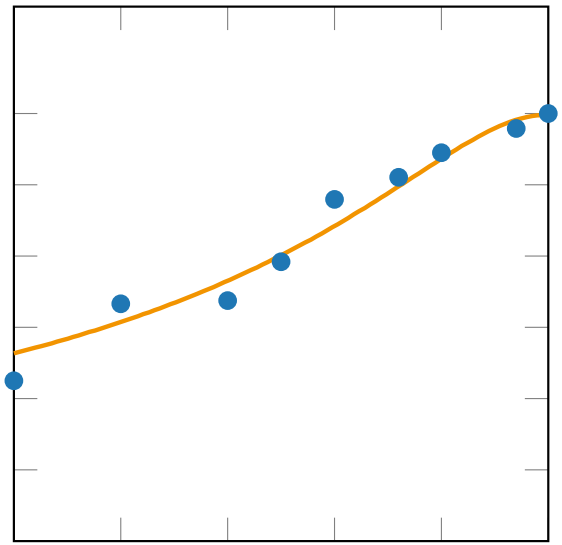
Modellansatz: regularisierte polynomiale Regression mit $d = 8$, d.h.

$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \ldots + \theta_8 x^8 \ .$$



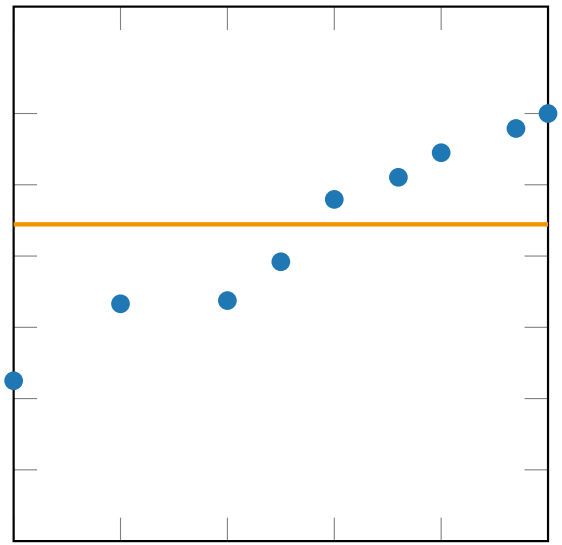
$\lambda = 0$

Hohe Varianz



$\lambda = 0.1$

OK



$\lambda = 1000$

Hoher Bias

Zusammenfassung

- Bei der Regressionsanalyse lässt sich der Fehler in den irreduziblen Fehler, die (quadrierte) Verzerrung und die Varianz zerlegen.
- Verzerrung und Varianz werden durch die Modellkomplexität gegenläufig beeinflusst. Es muss eine „Abwägung“ zwischen beiden gefunden werden.
- Die Bewertung der Güte eines Modells und die Optimierung von Hyperparametern kann mittels Kreuzvalidierung erfolgen. Es wurden der Ansatz der Holdout-Methode (2-fache Kreuzvalidierung) und die k-fache Kreuzvalidierung besprochen.
- Durch Regularisierung kann die Varianz eines Modells reduziert werden.