Machine Learning

Prof. Dr. Fabian Brunner

<fa.brunner@oth-aw.de>

Amberg, 25. Oktober 2021

Übersicht

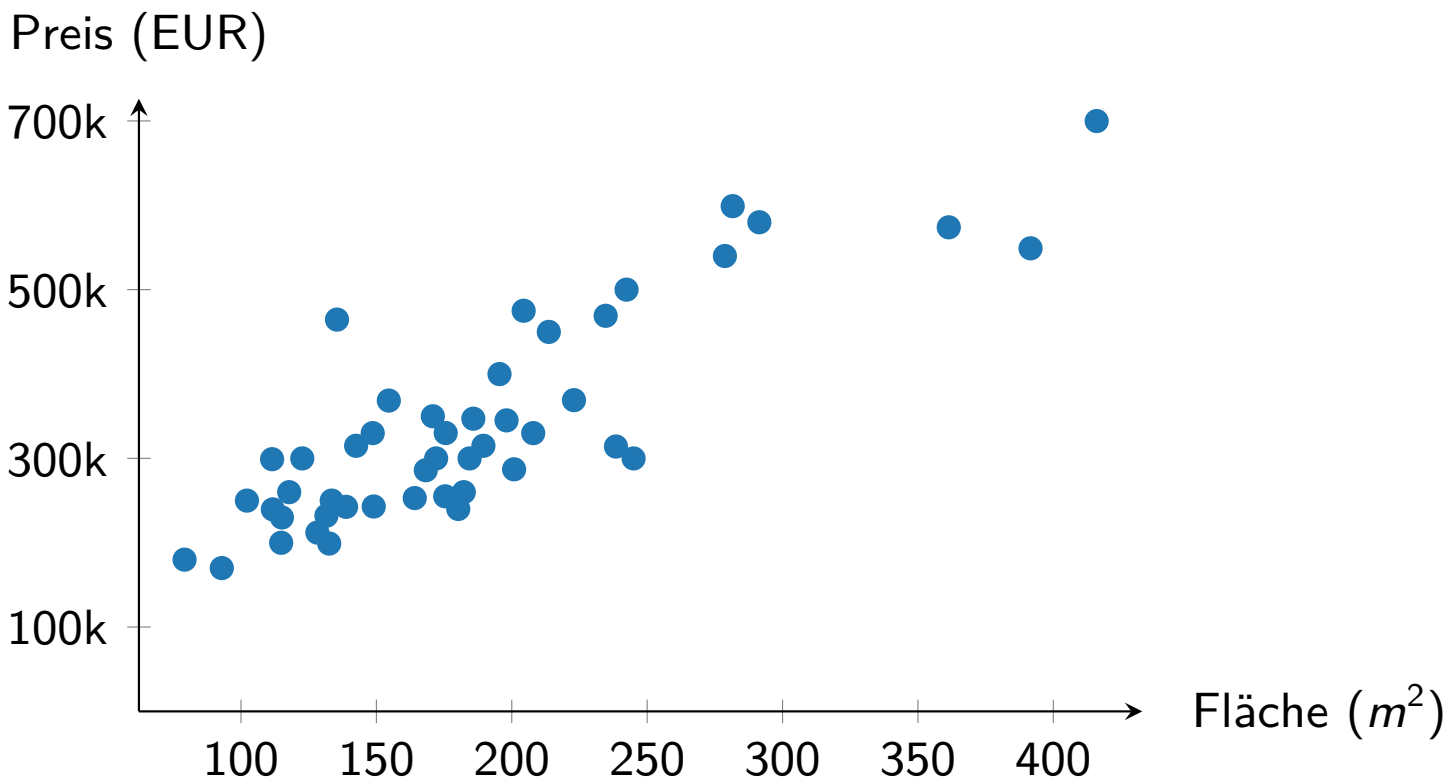
Ostbayerische Technische Hochschule
Amberg-Weiden

Thema heute: Lineare Regression

- Grundidee
- Least-Squares-Funktional
- Mathematische Grundlagen: unrestringierte Optimierung im $\mathbb{R}^n$.
- Normalgleichungen
- Gradientenverfahren
- Modellbewertung

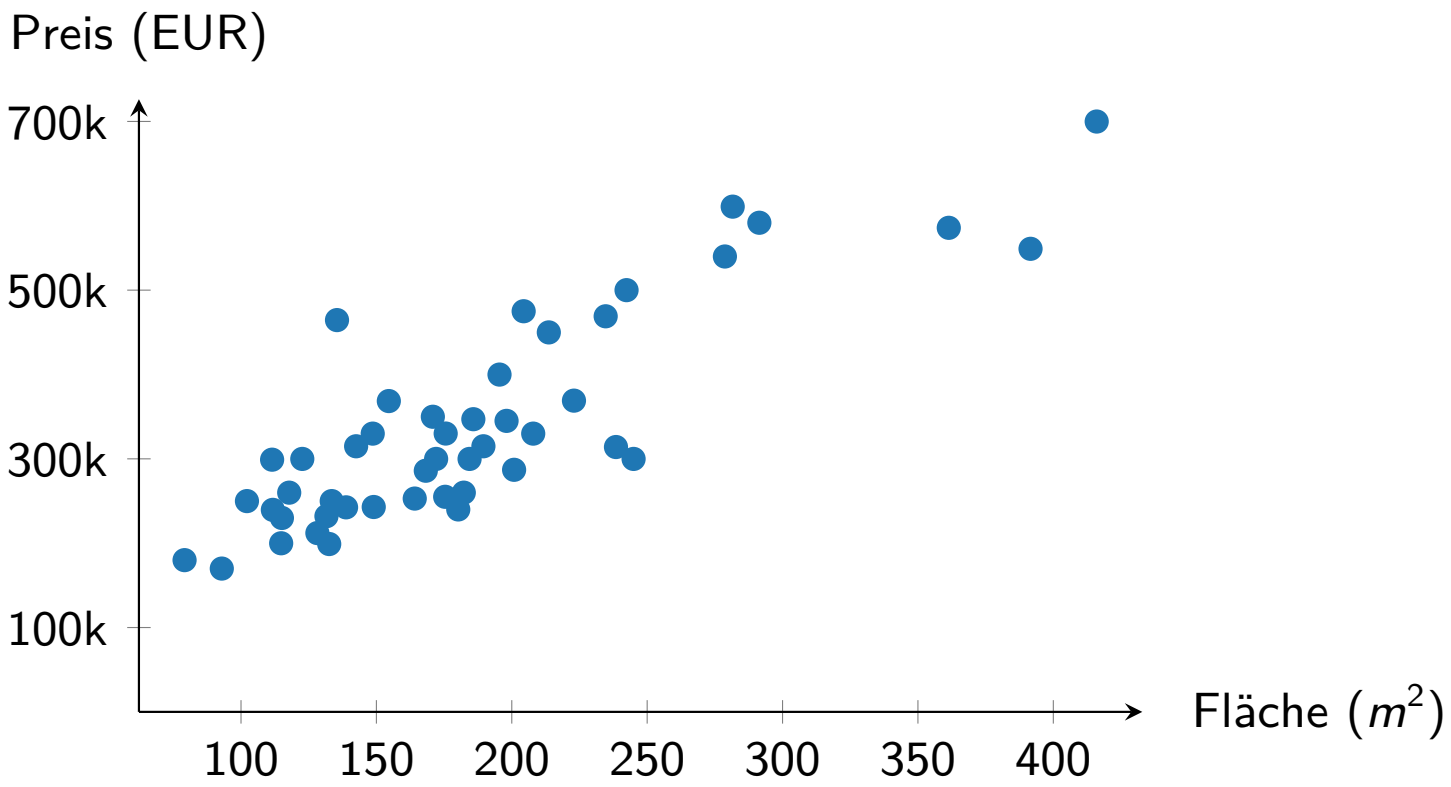
Beispiel: Modellierung von Häuserpreisen mit Linearer Regression

Das folgende Streudiagramm zeigt die Preise von 48 Häusern in Abhängigkeit von der Wohnfläche:



Beispiel: Modellierung von Häuserpreisen mit Linearer Regression

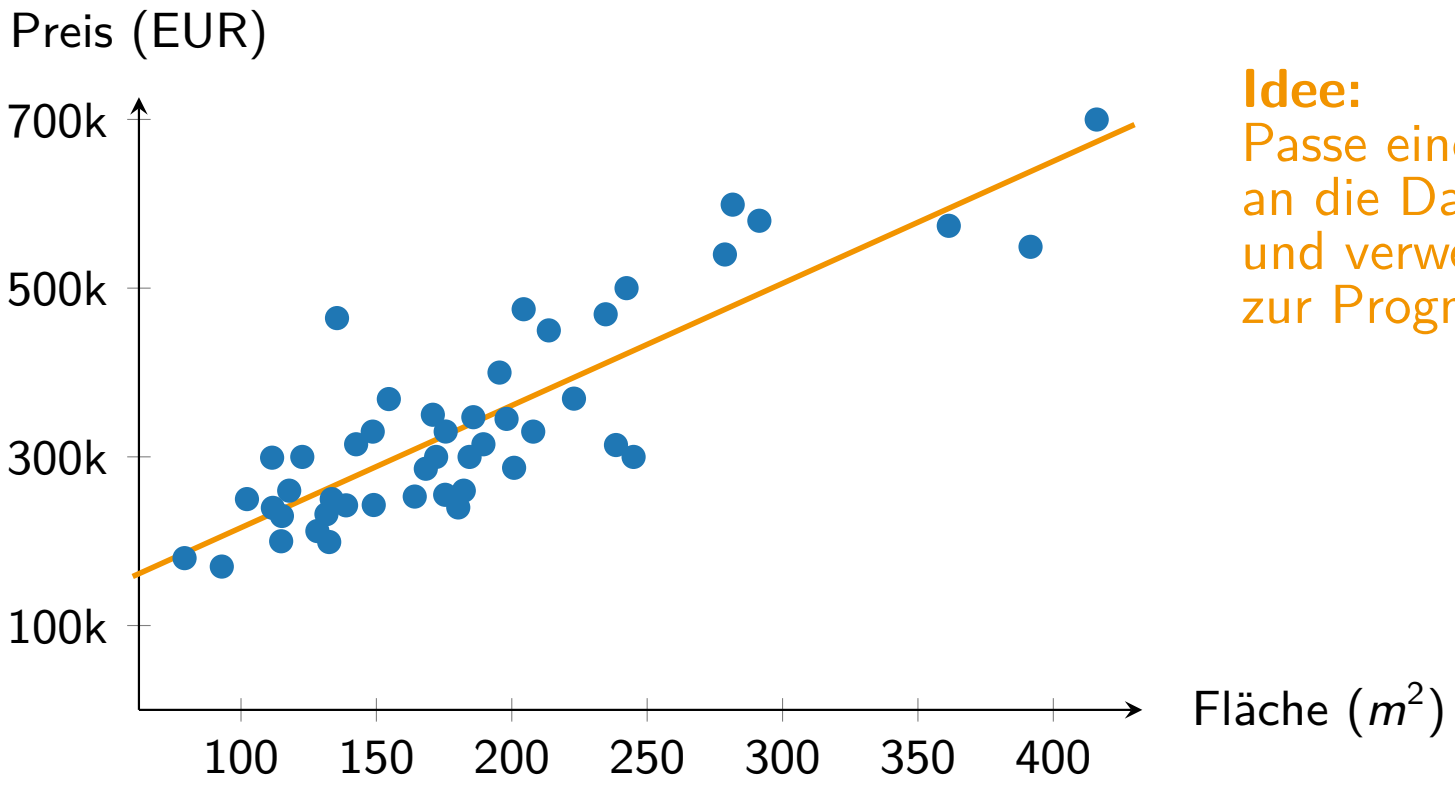
Das folgende Streudiagramm zeigt die Preise von 48 Häusern in Abhängigkeit von der Wohnfläche:



Frage: Wie teuer wäre ein Haus mit $330m^2$?

Beispiel: Modellierung von Häuserpreisen mit Linearer Regression

Das folgende Streudiagramm zeigt die Preise von 48 Häusern in Abhängigkeit von der Wohnfläche:

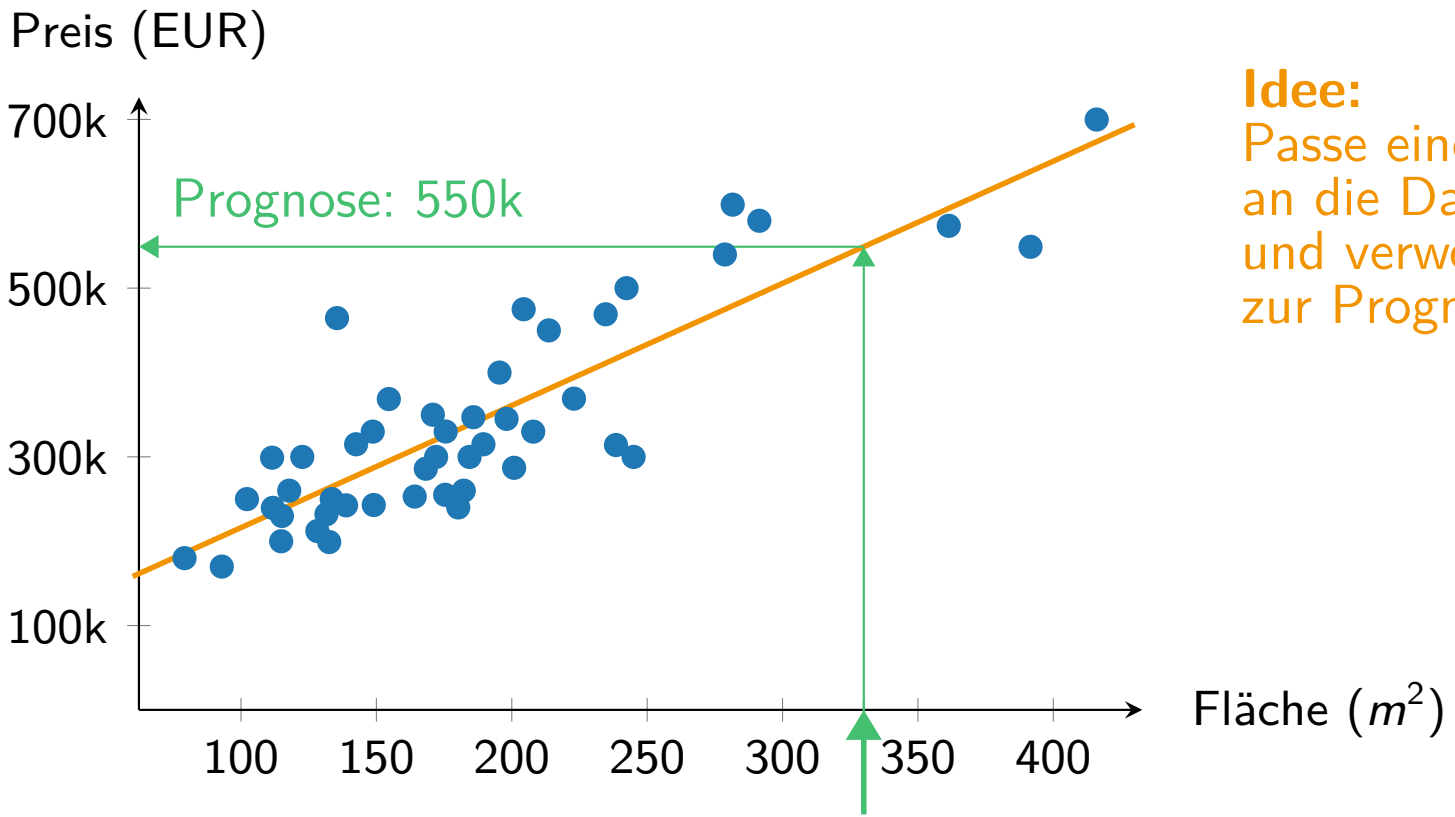


Idee:
Passe eine Gerade an die Daten an und verwende diese zur Prognose!

Frage: Wie teuer wäre ein Haus mit $330m^2$?

Beispiel: Modellierung von Häuserpreisen mit Linearer Regression

Das folgende Streudiagramm zeigt die Preise von 48 Häusern in Abhängigkeit von der Wohnfläche:



Idee:
Passe eine Gerade an die Daten an und verwende diese zur Prognose!

Frage: Wie teuer wäre ein Haus mit $330m^2$?

Beispiel: Modellierung von Häuserpreisen mit Linearer Regression

Ostbayerische Technische Hochschule Amberg-Weiden

Trainingsdaten:

Fläche in m^2 (x)	Hauspreis (y)
195	399900
149	329900
223	369000
132	232000
279	539900
184	299900
⋮	⋮

m Trainingsdatensätze

Feature | Zielvariable

Notation:

x	:	Input-Variable/Feature/unabhängige Größe
y	:	Ausgabe-Variable/Zielvariable/abhängige Größe
m	:	Anzahl der Trainingsdatensätze
$(x^{(i)}, y^{(i)})$	:	i-ter Trainingsdatensatz

Beispiel: Modellierung von Häuserpreisen mit Linearer Regression

Ostbayerische Technische Hochschule Amberg-Weiden

Aufgabenstellung bei der univariaten linearen Regression

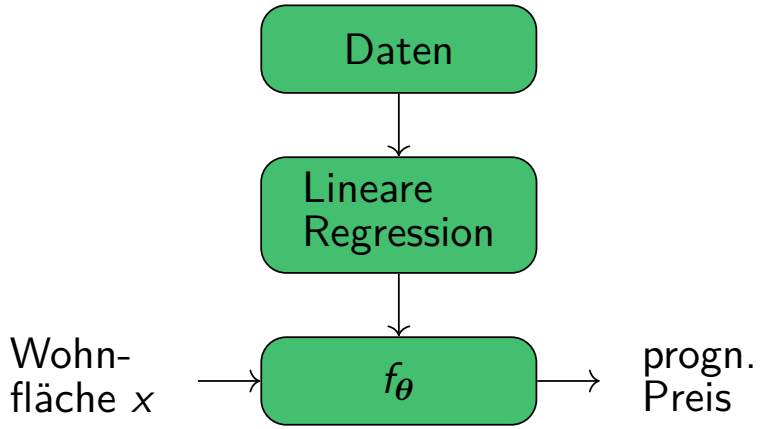
Bestimme die Paramter $\boldsymbol{\theta} = (\theta_0, \theta_1)$ einer linearen Funktion

$$f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x \ ,$$

sodass diese möglichst gut zu den Trainingsdaten „passt“.

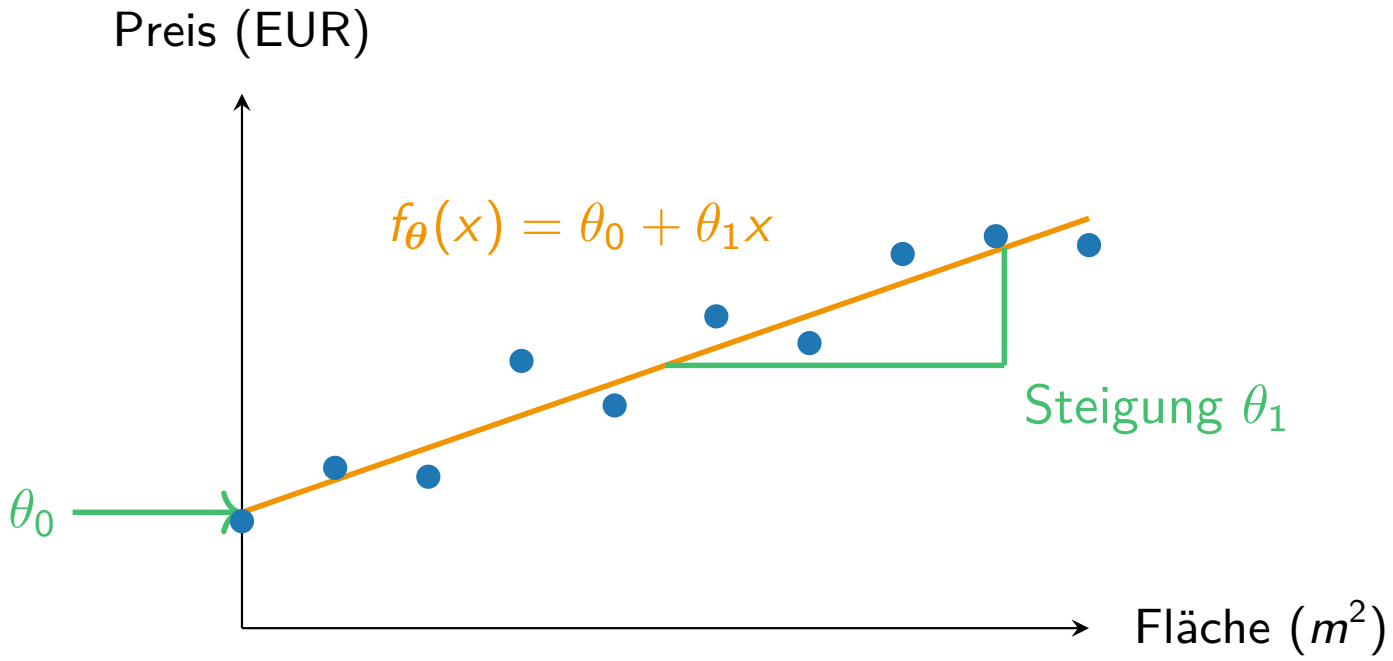
Bemerkungen:

- Es ist noch zu definieren, was „passt“ bedeutet.
- Der Parameter θ_0 wird auch als „Offset“ bezeichnet.
- Die lineare Regression ist ein parametrisiertes ML-Verfahren.
- Man spricht von **linearer** Regression, da die Modellfunktion als Linearkombination der Gewichte darstellbar ist.

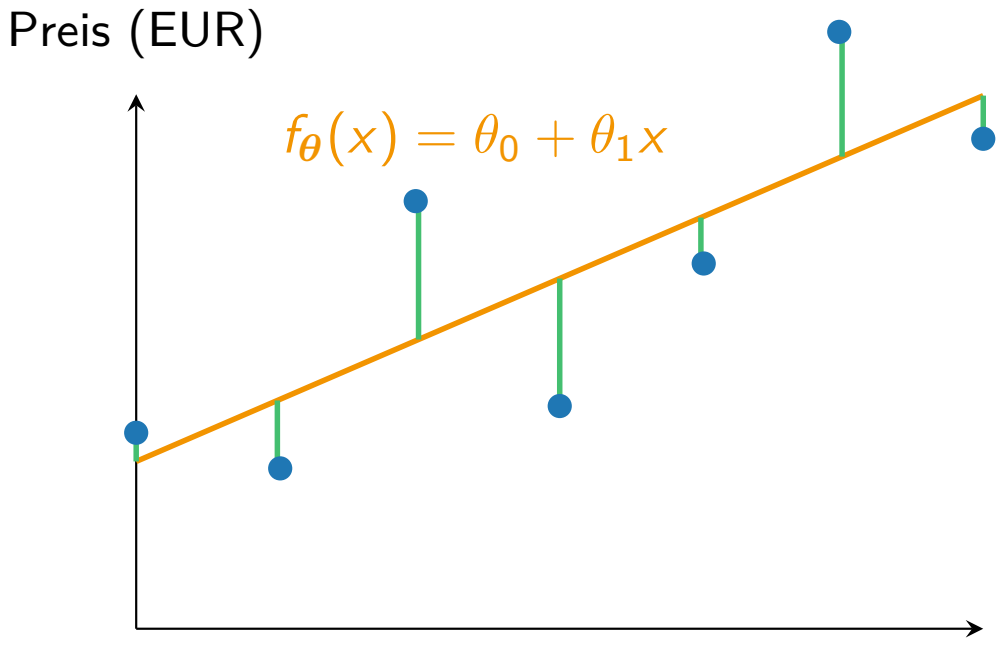


Lineare Regression in einer Variable

Die Parameter θ_0 und θ_1 haben eine geometrische Interpretation:



Least-Squares-Funktional für die Lineare Regression



Idee:
Minimiere die Summe der quadratischen Abweichungen zwischen f_θ und den Trainingsdatenpunkten

Least-Squares-Funktional für die Lineare Regression mit einer Variable

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

Minimierungsaufgabe bei der univariaten Linearen Regression

Minimierungs-Problem

Gesucht sind diejenigen Parameter $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1)^T \in \mathbb{R}^2$, für die das Optimierungsproblem

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

gelöst wird, d.h.

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^2} J(\boldsymbol{\theta}) \ .$$

Fragen:

- Unter welchen Bedingungen ist das Optimierungsproblem lösbar?
- Ist die Lösung eindeutig?

Problemdarstellung in Matrix-Form

$$J(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}\left(f_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)}\right)^2 = \frac{1}{2m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$$

$$= \frac{1}{2m}\left[\begin{pmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(m)} \end{pmatrix} \cdot \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} - \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix}\right]^T \cdot$$

$$\left[\underbrace{\begin{pmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(m)} \end{pmatrix}}_{=:X} \cdot \underbrace{\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}}_{=:\boldsymbol{\theta}} - \underbrace{\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix}}_{\boldsymbol{y}}\right] = \frac{1}{2m}(X\boldsymbol{\theta} - \boldsymbol{y})^T(X\boldsymbol{\theta} - \boldsymbol{y}) \,.$$

Ziel: Wie bestimmt man nun ein Minimum von J?
$\rightarrow$ Gradient berechnen und „gleich Null setzen“.

Partielle Ableitung und Gradient

Ostbayerische Technische Hochschule Amberg-Weiden

Definition 1 (Partielle Ableitung)

Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine Funktion von n Variablen und sei $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ gegeben. Falls der Grenzwert

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}) := \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_n)}{h}$$

existiert, so nennt man ihn die partielle Ableitung von f nach x_i an der Stelle $\boldsymbol{x}$.

Definition 2 (Gradient)

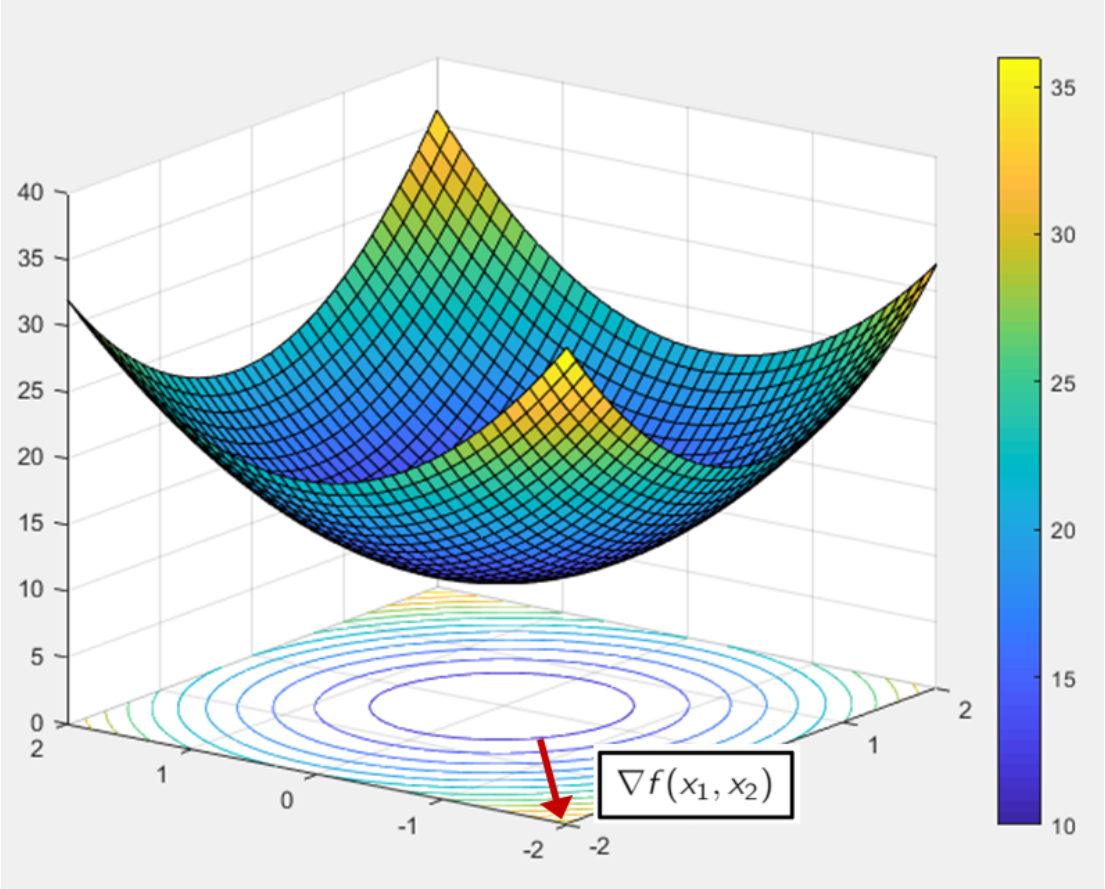
Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine reelle Funktion, deren partielle Ableitungen in einem Punkt $\boldsymbol{x} \in \mathbb{R}^n$ existieren. Dann heißt

$$\nabla f(\boldsymbol{x}) := \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \right)^T$$

der Gradient von f an der Stelle $\boldsymbol{x}$.

Interpretation des Gradienten

Der Gradient an einer Stelle $\boldsymbol{x} = (x_1, x_2)$ zeigt stets in die Richtung des steilsten Anstiegs der Funktion f:



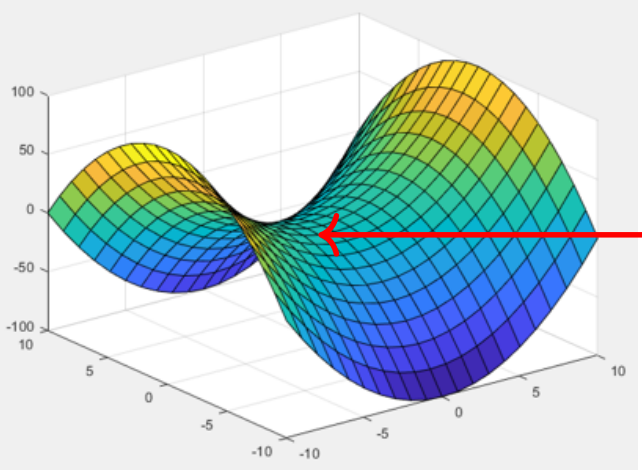
Notwendige Optimalitätsbedingung

Theorem 3 (Notwendige Optimalitätsbedingung)

Ist $f : \mathbb{R}^n \to \mathbb{R}$ *differenzierbar und ist* $\boldsymbol{a} \in \mathbb{R}^n$ *eine Extremstelle von* f*, dann gilt*

$$\nabla f(\boldsymbol{a}) = \mathbf{0} = (0, \ldots, 0)^T \ .$$

- Punkte, bei denen der Gradient verschwindet, nennt man kritische Punkte.
- Ein kritischer Punkt muss noch kein Minimum oder Maximum sein, z.B. ist bei der Funktion $f(x_1, x_2) = x_1^2 - x_2^2$ die Stelle $\boldsymbol{x} = (0, 0)^T$ ein kritischer Punkt, es ist aber weder ein Maximum sondern ein Minimum (sondern ein Sattelpunkt).



Der Nullpunkt ist hier zwar ein kritischer Punkt, aber weder ein lokales Maximum noch ein lokales Minimum. Es handelt sich vielmehr um einen Sattelpunkt.

Positiv definite Matrizen

Definition 4 (SPD-Matrizen)

Sei $A \in \mathbb{R}^{(n,n)}$ eine symmetrische Matrix. Dann heißt A

positiv definit,	falls $\boldsymbol{x}^T A\boldsymbol{x} > 0$ für alle $\boldsymbol{x} \in \mathbb{R}^n$,
positiv semidefinit,	falls $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ für alle $\boldsymbol{x} \in \mathbb{R}^n$,
negativ definit,	falls $\boldsymbol{x}^T A\boldsymbol{x} < 0$ für alle $\boldsymbol{x} \in \mathbb{R}^n$,
negativ semidefinit,	falls $\boldsymbol{x}^T A\boldsymbol{x} \leq 0$ für alle $\boldsymbol{x} \in \mathbb{R}^n$.

Ist A weder positiv noch negativ semidefinit, so nennt man sie **indefinit**.

Theorem 5 (Definitheit und Eigenwerte)

Eine symmetrische Matrix $A \in \mathbb{R}^{(n,n)}$ ist genau dann

positiv (semi-)definit,	falls alle Eigenwerte positiv (nicht-negativ) sind ,
negativ (semi-)definit,	falls alle Eigenwerte negativ (nicht-positiv) sind ,
indefinit,	falls es positive und negative Eigenwerte gibt .

Hinreichendes Optimalitätskriterium

Ostbayerische Technische Hochschule Amberg-Weiden

Theorem 6 (Hinreichende Optimalitätsbedingung)

Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei zweimal stetig differenzierbar und $\boldsymbol{a} \in \mathbb{R}^n$ sei ein kritischer Punkt von f (d.h. $\nabla f(\boldsymbol{a}) = \mathbf{0}$) und sei $H_f(\boldsymbol{a})$ die Hesse-Matrix von f in $\boldsymbol{a}$. Ist

$H_f(\boldsymbol{a})$	positiv definit, so ist $\boldsymbol{a}$ ein striktes lokales Minimum von f ,
$H_f(\boldsymbol{a})$	negativ definit, so ist $\boldsymbol{a}$ ein striktes lokales Maximum von f ,
$H_f(\boldsymbol{a})$	indefinit, so ist $\boldsymbol{a}$ kein lokales Extremum von f .

Übung

Gegeben sei die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ durch

$$f(x, y) = y^2(x - 1) + x^2(x + 1) \ .$$

Bestimmen Sie alle kritischen Punkte und untersuchen Sie, ob ein lokales Maximum, ein lokales Minimum oder ein Sattelpunkt vorliegt.

Übung

Gegeben sei die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ durch

$$f(x, y) = y^2(x - 1) + x^2(x + 1) \ .$$

Bestimmen Sie alle kritischen Punkte und untersuchen Sie, ob ein lokales Maximum, ein lokales Minimum oder ein Sattelpunkt vorliegt.

$$\nabla f(x, y) = \begin{pmatrix} y^2 + 3x^2 + 2x \\ 2y(x - 1) \end{pmatrix} \ ,$$

$$H_f(x, y) = \begin{pmatrix} 6x + 2 & 2y \\ 2y & 2(x - 1) \end{pmatrix} \ .$$

Aus $\nabla f(x, y) = \mathbf{0}$ ergeben sich die kritischen Punkte $(x_1, y_1) = (0, 0)^T$ und $(x_2, y_2) = (-2/3, 0)^T$. Für die Hesse-Matrizen erhält man

$$H_f(x_1, y_1) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \ , \quad H_f(x_2, y_2) = \begin{pmatrix} -2 & 0 \\ 0 & -10/3 \end{pmatrix} \ .$$

Also ist $(x_1, y_1)^T$ ein Sattelpunkt (Hesse-Matrix indefinit) und $(x_2, y_2)^T$ ein striktes lokales Maximum (Hesse-Matrix negativ definit).

Normalgleichungen

Man kann zeigen, dass der Gradient von J wie folgt lautet:

$$\nabla J(\boldsymbol{\theta}) = \frac{1}{m} \left(X^T X \boldsymbol{\theta} - X^T \boldsymbol{y} \right) \ .$$

Der Gradient muss in einem Minimum von J verschwinden. Daraus ergibt sich die folgende notwendige Optimalitätsbedingung:

Normalgleichungen für das lineare Least-Squares-Funktional

$$X^T X \boldsymbol{\theta} = X^T \boldsymbol{y} \ .$$

Bemerkungen:

- Die Bedingung ist ein lineares Gleichungssystem in den Unbekannten $\boldsymbol{\theta}$.
- Man kann zeigen, dass dieses immer lösbar ist und dass jede Lösung ein globales Minimum von J ist.
- Wenn die Matrix $X^T X$ invertierbar ist, erhält man eine eindeutige Lösung:

$$\hat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \boldsymbol{y} \ .$$

- Die Invertierbarkeit der Matrix $X^T X$ wird im Folgenden untersucht.

Aufgaben und Verständnisfragen

Ostbayerische Technische Hochschule Amberg-Weiden

Wir können uns nun der Lösbarkeit der Normalgleichungen zuwenden:

- Welche Eigenschaften hat die Matrix $X^T X$?
- Wann ist sie positiv semidefinit, wann positiv definit?
- Welche Bedingung muss erfüllt sein, damit die Spalten von X linear unabhängig sind?
- Warum ist die Matrix $X^T X$ invertierbar, wenn die Spalten von X linear unabhängig sind?
- Wie sieht anschaulich eine Situation aus, in der keine eindeutige Lösung des eindimensionalen linearen Regressionsproblems existiert?
- Kann eine Situation eintreten, in der $J(\boldsymbol{\theta}) = 0$ ist?
- Zeigen Sie, dass die Normalgleichungen stets lösbar sind. Verwenden Sie dazu die Beziehungen $\mathrm{Kern}(A) = \mathrm{Bild}(A^T)^\perp$ und $\mathrm{Kern}(A^T A) = \mathrm{Kern}(A)$, die für jede reelle Matrix A gelten.

Bemerkungen zur Lösbarkeit der Normalgleichungen

- Das Funktional $J(\boldsymbol{\theta})$ ist zweimal stetig differenzierbar und die Hesse-Matrix ist positiv semidefinit. Daraus folgt, dass J konvex ist.
- Bei konvexen, zweimal stetig differenzierbaren Funktionalen ist jeder kritische Punkt ein globales Minimum.
- Die Normalgleichungen sind immer lösbar, d.h. es gibt immer einen kritischen Punkt und damit ein globales Minimum.
- Sind die Normalgleichungen eindeutig lösbar, dann gibt es ein eindeutiges Minimum.
- Die Normalgleichungen sind eindeutig lösbar, wenn nicht alle $x^{(i)}$ gleich sind (=degenerierter Fall).

Multivariate lineare Regression

- Bisher haben wir den Einfluss einer Eingangsgröße auf eine Ausgangsgröße durch lineare Regression modelliert.
- Häufig sind aber mehrere Eingangsgrößen relevant, deren Einfluss auf eine Zielgröße modelliert werden soll.
- Angenommen, wir hätten den folgenden Datensatz gegeben:

Wohnfläche $[m^2]$	Alter [Jahre]	Entfernung Zentrum $[km]$	Hauspreis
195	4	4.1	399900
149	5	2.8	329900
223	10	5.2	369000
132	15	1.5	232000
279	8	6.7	539900
184	25	12.0	299900
⋮			⋮
Feature x_1	Feature x_2	Feature x_3	Zielvariable y

m Trainingsdatensätze

- Der multivariate lineare Regressionsansatz lautet in diesem Fall:

$$y = f_{\boldsymbol{\theta}}(\boldsymbol{x}) = f_{\boldsymbol{\theta}}(x_1, x_2, x_3) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \ .$$

Multivariate lineare Regression

Ostbayerische Technische Hochschule Amberg-Weiden

Univariate lineare Regression

Bisher hatten wir den Fall **einer** Eingangsgröße x (z.B. Wohnfläche) und einer Zielvariable y (z.B. Hauspreis) betrachtet und ein Modell der Form

$$y = f_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$$

hergeleitet.

Multivariate lineare Regression

Bei der multivariaten linearen Regression wird der allgemeine Fall von p Features $\boldsymbol{x} = (x_1, x_2, \ldots x_p)$ betrachtet und ein Zusammenhang der Form

$$y = f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \ldots + \theta_p x_p$$

modelliert. Die Aufgabe besteht also darin, anhand der Trainingsdaten die Parameter $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_p)$ zu bestimmen.

Parameterfitting durch Least-Squares-Ansatz

Gegeben seien nun Trainingsdaten $(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)})$, wobei $\boldsymbol{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_p^{(i)})$. Unter Verwendung der Konvention $x_0^{(i)} := 1$ lautet das Least-Squares-Funktional wie folgt:

Least-Squares-Funktional der multivariaten linearen Regression

$$J(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)} \right)^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(\sum_{j=0}^{p} \theta_j x_j^{(i)} - y^{(i)} \right)^2 .$$

Minimierungs-Problem

Gesucht sind diejenigen Parameter $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \ldots, \hat{\theta}_p)$, für die das Minimierungsproblem

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

gelöst wird.

Least-Squares-Funktional in Matrix-Form

$$J(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} \left(\sum_{j=0}^{p} \theta_j x_j^{(i)} - y^{(i)} \right)^2$$

$$= \frac{1}{2m} \left[\begin{pmatrix} 1 & x_1^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & \dots & x_p^{(m)} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} - \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} \right]^T .$$

$$\left[\underbrace{\begin{pmatrix} 1 & x_1^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & \dots & x_p^{(m)} \end{pmatrix}}_{=:X} \underbrace{\begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}}_{=:\boldsymbol{\theta}} - \underbrace{\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix}}_{=:\boldsymbol{y}} \right]$$

$$= \frac{1}{2m} (X\boldsymbol{\theta} - \boldsymbol{y})^T (X\boldsymbol{\theta} - \boldsymbol{y}) .$$

Fazit - multivariate Lineare Regression

- Das multivariate lineare Regressionsproblem hat dieselbe Struktur wie das univariate. Letzteres erhält man für den Spezialfall $p = 1$.
- Eine Lösung des Minimierungsproblems ist gegeben durch die Lösung der Normalgleichungen

Normalgleichungen

$$X^T X\boldsymbol{\theta} = X^T \boldsymbol{y}$$

mit

$$X = \begin{pmatrix} 1 & x_1^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & \dots & x_p^{(m)} \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix}.$$

- Frage: Unter welchen Bedingungen ist die Lösung eindeutig?

Algebraische Eigenschaften der Least-Squares-Lösung bei der linearen Regression

Theorem 7

Sei $\boldsymbol{\theta}$ die Least Squares-Lösung der multivariaten linearen Regression und seien $e^{(i)} := f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)}$ die Residuen auf den Trainingsdaten. Dann gilt:

$$\sum_{i=1}^{m} e^{(i)} = 0 \ , \tag{1}$$

$$\sum_{i=1}^{m} e^{(i)} x_j^{(i)} = 0 \text{ für alle } j \in \{1, \ldots, p\} \ . \tag{2}$$

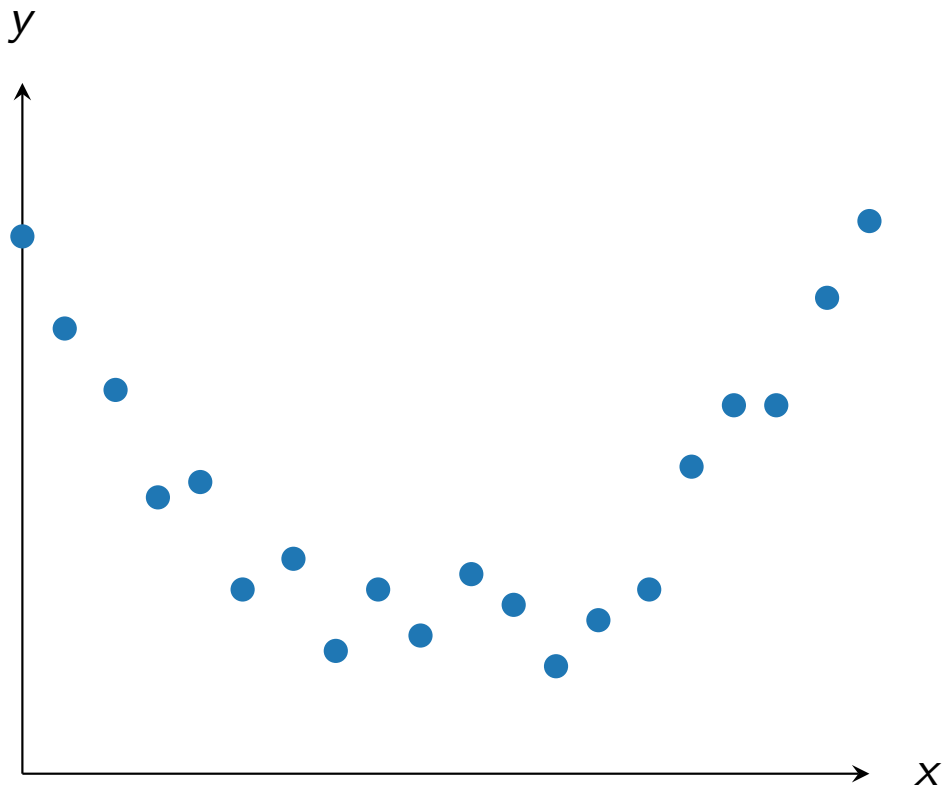
Bemerkung:
Aus den beiden Eigenschaften folgt für $\hat{y}^{(i)} := f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})$ auch

$$\sum_{i=1}^{m} e^{(i)} \hat{y}^{(i)} = 0 \ , \tag{3}$$

d.h. Residuen und Vorhersagen sind unkorreliert.

Polynomiale Regression

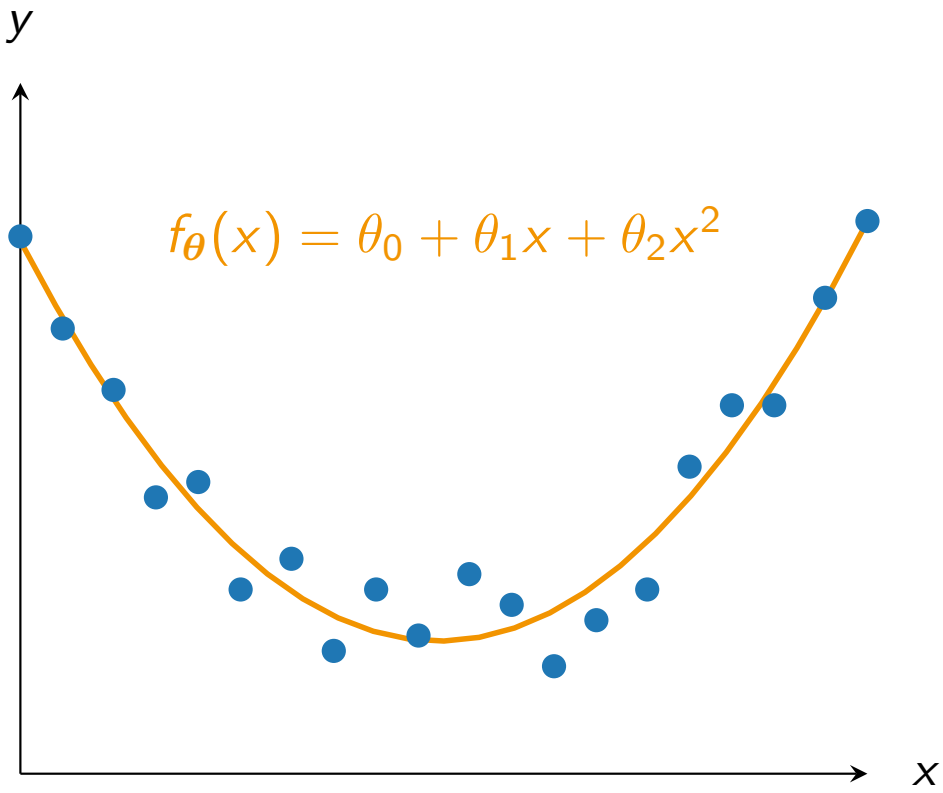
- Lineare Ansätze sind nicht immer geeignet, um Zusammenhänge in den Daten zu modellieren.
- Durch welche Funktionsklasse könnte man die folgenden Daten besser fitten als durch einer Gerade?



- Frage: Wie kann man anhand gegebener Daten untersuchen, ob die Annahme eines linearen Zusammenhangs zwischen zwei Merkmalen

Polynomiale Regression

- Lineare Ansätze sind nicht immer geeignet, um Zusammenhänge in den Daten zu modellieren.
- Durch welche Funktionsklasse könnte man die folgenden Daten besser fitten als durch einer Gerade?



- Frage: Wie kann man anhand gegebener Daten untersuchen, ob die Annahme eines linearen Zusammenhangs zwischen zwei Merkmalen

Polynomiale Regression

Modell bei der polynomialen Regression

Der Modellansatz bei der (univariaten) polynomialen Regression lautet

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_d x^d \ ,$$

wobei d den Polyomgrad bezeichnet.

Sind $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$ die gegebenen Trainingsdaten, so ist die Least-Squares-Lösung $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_d)^T$ durch die Lösung der Normalgleichungen

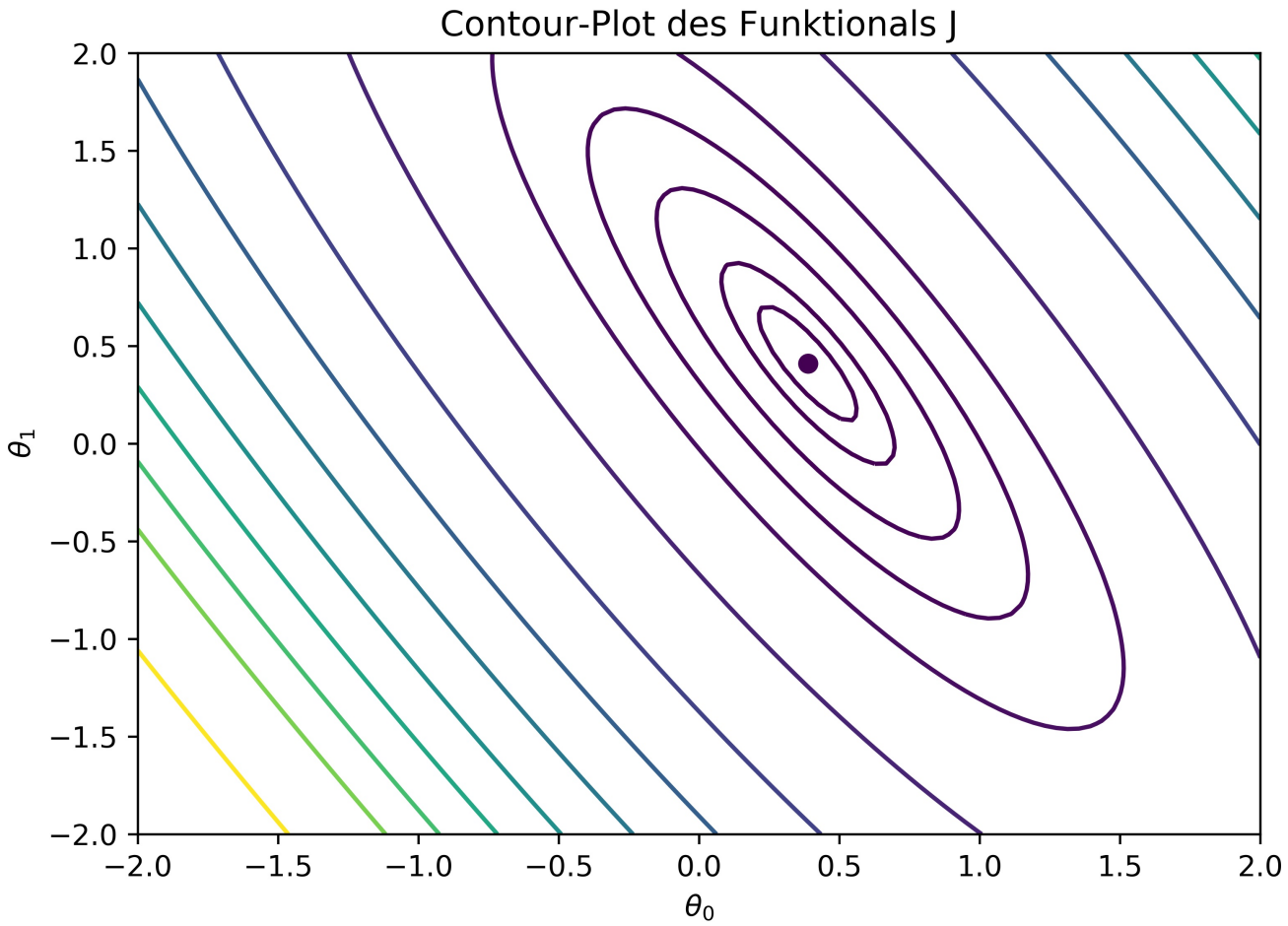
$$X^T X \boldsymbol{\theta} = X^T \boldsymbol{y}$$

gegeben, wobei

$$X = \begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & \ldots & (x^{(1)})^d \\ 1 & x^{(2)} & \ldots & \ldots & (x^{(2)})^d \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & x^{(m)} & (x^{(m)})^2 & \ldots & (x^{(m)})^d \end{pmatrix} \ , \quad \boldsymbol{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} \ .$$

Contour-Plot des Least-Squares-Funktionals

- Das Least-Squares-Funktional ist konvex und hat immer ein Minimum.
- Ist dieses eindeutig? -> Mathematische Analyse des Funktionals

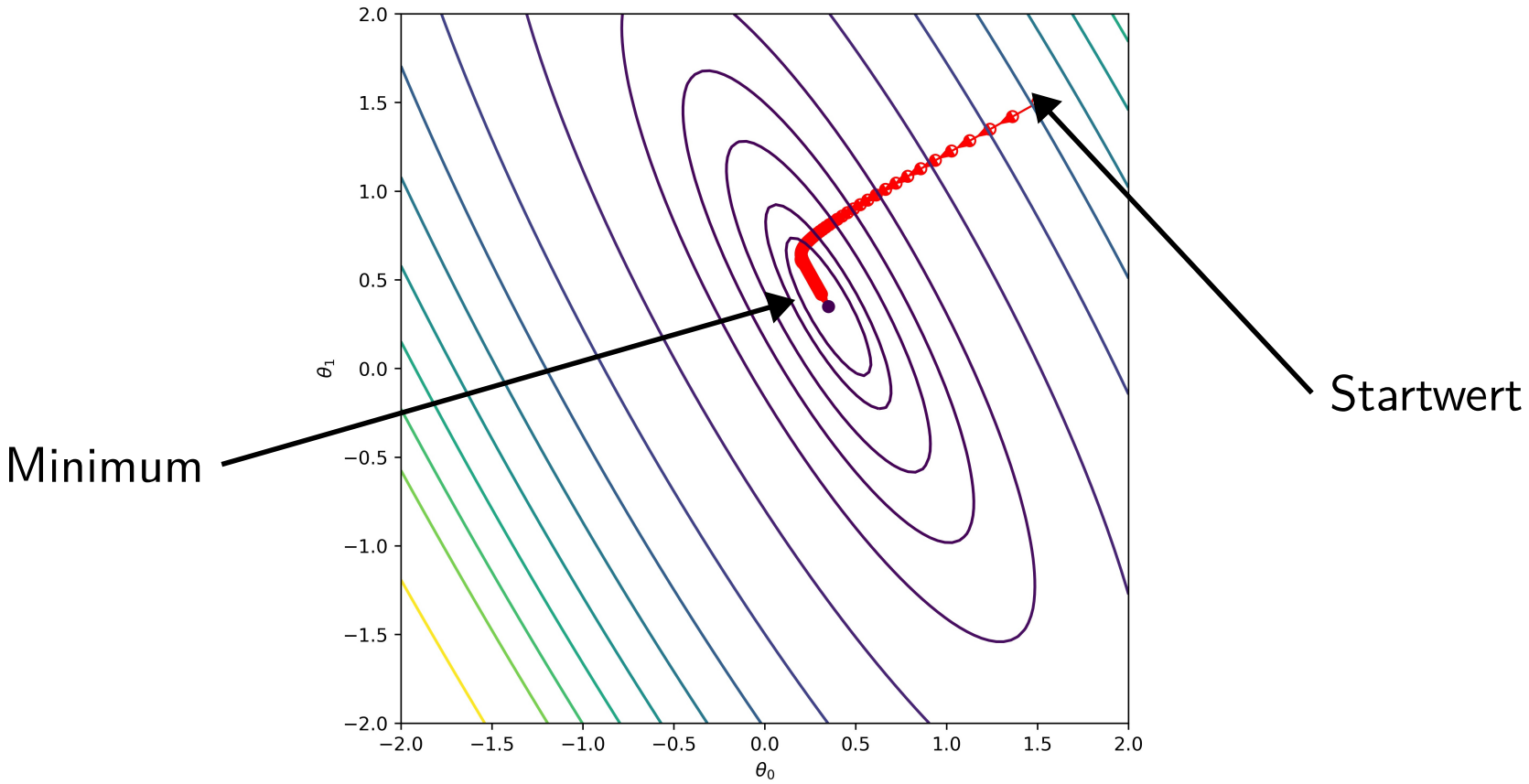


Das Gradientenverfahren

- Im Fall der linearen Regression konnten wir das Minimum des Least-Squares-Fehlerfunktionals analytisch berechnen.
- Diest ist häufig nicht möglich. Dann behilft man sich mit Näherungsverfahren.
- Eines der wichtigsten Näherungsverfahren zur Approximation von Minimalstellen reeller Funktionen ist das sog. **Gradientenverfahren** bzw. **Gradienten-Abstiegsverfahren**.
- Es beruht auf der Tatsache, dass der Gradient einer reellen Funktion in einem Punkt stets in die Richtung des stärksten Anstiegs der Funktion in diesem Punkt zeigt.
- Ausgehend von einem Startwert wird iterativ ein Minimum bestimmt, indem in jedem Iterationsschritt ein Teil (der durch die sog. **Lernrate** spezifiziert wird) des negativen Gradienten als Update addiert wird.

Approximation eines Minimums mit dem Gradientenverfahren

Das folgende Bild zeigt exemplarisch, wie ausgehend vom Startwert $\boldsymbol{\theta}^0 = (1.5, 1.5)^T$ mit Hilfe des Gradientenverfahrens das Minimum im Punkt $(0.34, 0.34)^T$ approximiert wird. Die Ellipsen stellen die Niveaulinien der Funktion dar.



Gradientenverfahren

Gradientenverfahren

1. Wähle einen Startwert $\boldsymbol{x}^0$ und eine Lernrate $\alpha > 0$.
2. Solange das Abbruchkriterium nicht erfüllt ist, addiere im k-ten Schritt wie folgt ein Update:

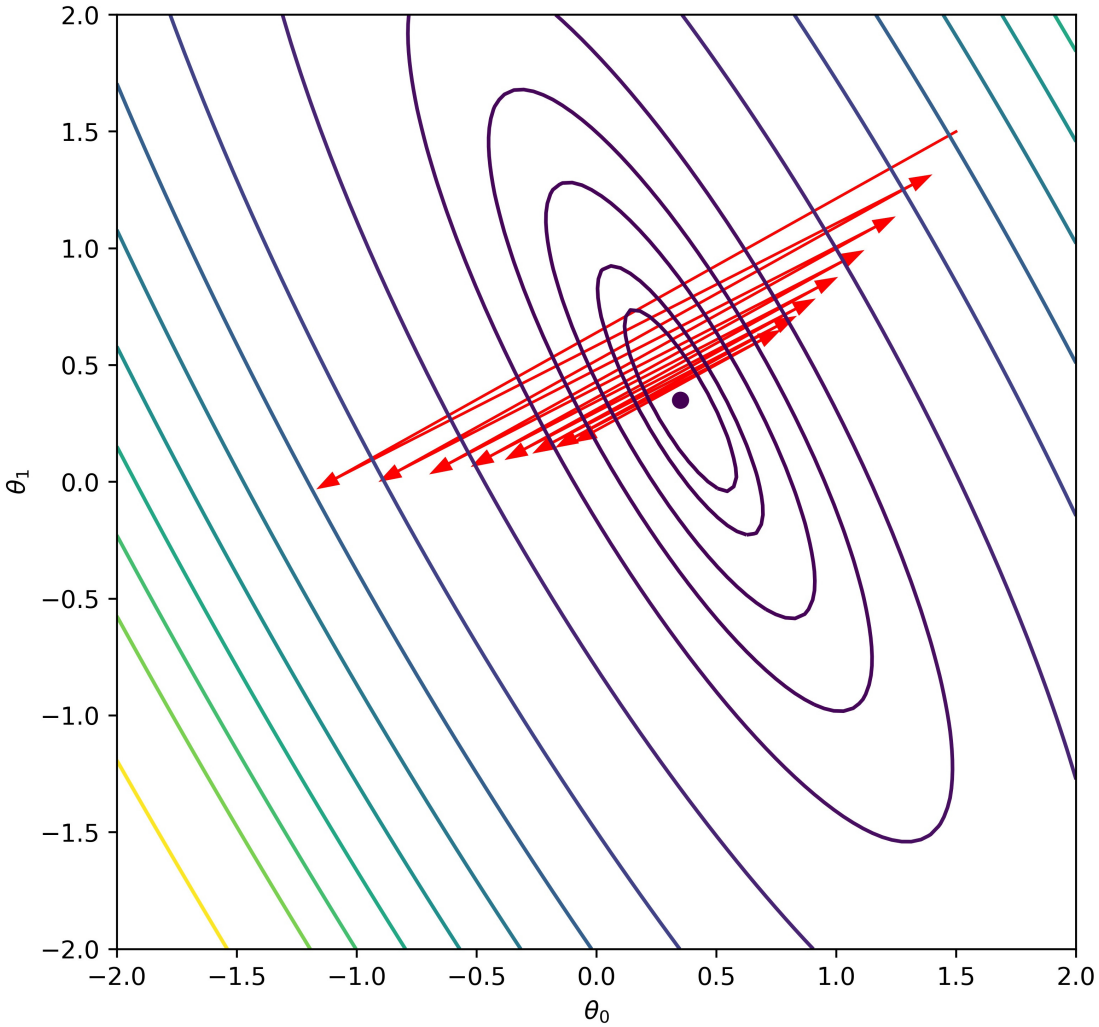
$$\boldsymbol{x}^k := \boldsymbol{x}^{k-1} - \alpha \nabla f(\boldsymbol{x}^{k-1}) \ .$$

Bemerkungen:

- **Abbruchkriterium:** die Iteration wird abgebrochen, wenn eine vorgegebene Anzahl an Iterationen erreicht wird oder wenn das Update „hinreichend klein“ ist (z.B.: $\|\Delta \boldsymbol{x}^k\| < 10^{-6}$).
- **Lernrate**: eine zu kleine Lernrate führt dazu, dass potentiell viele Iterationen nötig sind, um das Minimum zu erreichen. Eine zu große Lernrate kann dazu führen, dass man „am Ziel vorbei schießt“. Oft hilft nur Ausprobieren, welche Lernrate am besten funktioniert.

Überschießen

Bei zu großer Lernrate kann man „über das Ziel hinaus“ schießen:



Formulierung des Gradientenverfahrens für die lineare Regression

Im Fall der linearen Regression hatten wir den Gradienten des Least-Squares-Funktionals bereits berechnet:

$$\nabla J(\boldsymbol{\theta}) = \frac{1}{m}(X^T X\boldsymbol{\theta} - X^T \boldsymbol{y})$$

Die Update-Regel des Gradientenverfahrens lautet für diesen Fall also

Gradientenverfahren für das LS-Funktional der linearen Regression

$$\boldsymbol{\theta}^k = \boldsymbol{\theta}^{k-1} - \frac{\alpha}{m}(X^T X\boldsymbol{\theta}^{k-1} - X^T \boldsymbol{y}) \ .$$

Verständnisfrage: warum kann es trotzdem sinnvoll sein, das Minimum mit dem Gradientenverfahren approximativ zu bestimmen, obwohl durch die Normalgleichungen die analytische Lösung bestimmt werden kann?

Verbesserung der Konvergenz durch Standardisierung

- In unserem Eingangsbeispiel waren die Variablen von stark unterschiedlicher Größenordnung:

Fläche in m^2 (x)	Hauspreis in EUR (y)
195	399900
149	329900
$\vdots$	$\vdots$

- Dies führt dazu, dass die Niveaulinien des Least-Squares-Funktionals stark entlang einer Achse „verzogen“ sind.
- Das Gradientenverfahren konvergiert dann ggf. nur sehr langsam bzw. mit sehr geringer Lernrate.
- Idee: skaliere die Variablen, damit sie dieselbe Größenordnung haben (betrachte beispielsweise den Hauspreis nicht in Euro, sondern in Tausend Euro).

Standardisierung

Standardisierung

Sei x ein Feature, das im Datensatz durch die Beobachtungen $\boldsymbol{x} = (x^{(1)}, \dots x^{(m)})^T$ repräsentiert wird. Dann geht die standardisierter Größe $\hat{x}$ aus x durch Subraktion des empirischen Mittelwerts und Division der empirischen Standardabweichung hervor:

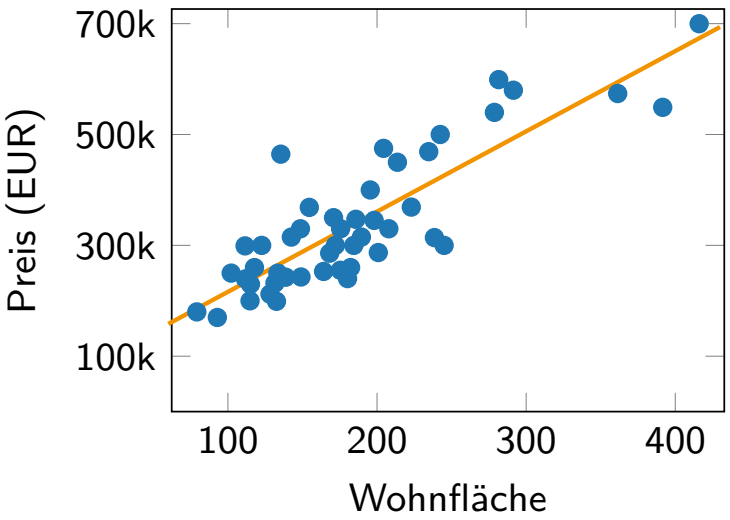
$$\hat{\boldsymbol{x}} := \frac{\boldsymbol{x} - \bar{\boldsymbol{x}}}{s} \ ,$$

wobei

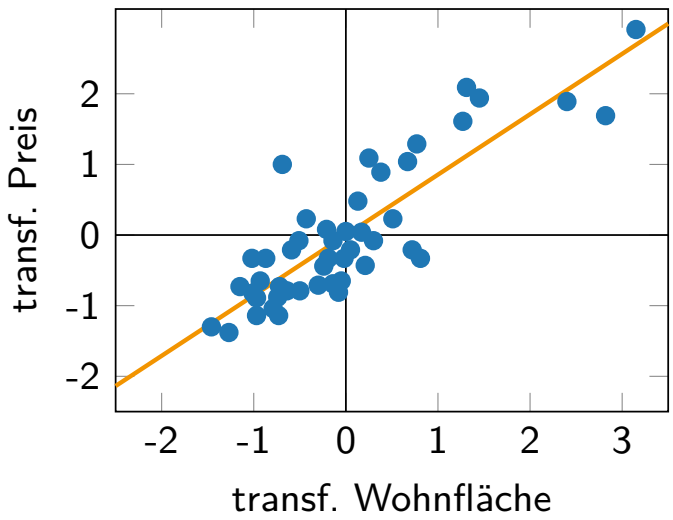
$$\bar{\boldsymbol{x}} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} \ , \quad s = \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} (x^{(i)} - \bar{\boldsymbol{x}})^2} \ .$$

Standardisierung des Häuserpreis-Datensatzes

Unskaliert



Standardisiert



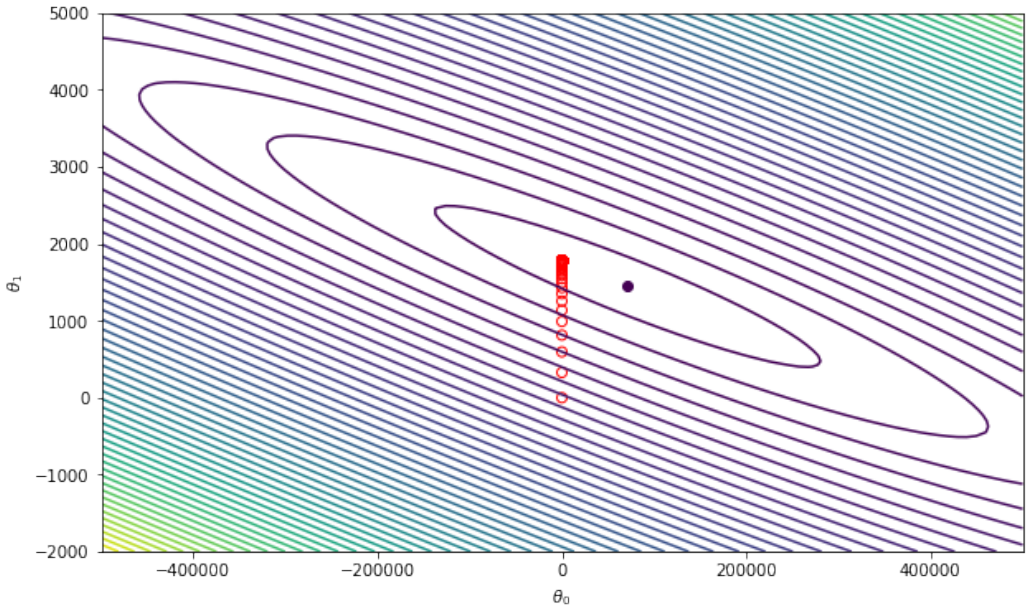
Vorgehen bei der Skalierung

Beachte

- Wird ein skaliertes Modell zu Prognosezwecken auf unbekannte Daten angewendet, müssen diese auf dieselbe Weise transformiert werden wie der Trainingsdatensatz, d.h. mit Hilfe des Mittelwerts und der Standardabweichung, die auf dem Trainingsdatensatz berechnet wurden.
- ein häufig gemachter Fehler ist, den Testdatensatz mit Hilfe des eigenen Mittelwerts und der eigenen Standardabweichung zu standardisieren.
- wurde im Vorfeld des Modelltrainings auch die Zielvariable transformiert, erhält bei der Modellanwendung eine Vorhersage für eine entsprechend transformierte Größe (Beispiel: Transformation der Hauspreise in Tausend Euro führt dazu, dass das Modell Vorhersagen in Tausend Euro liefert.)

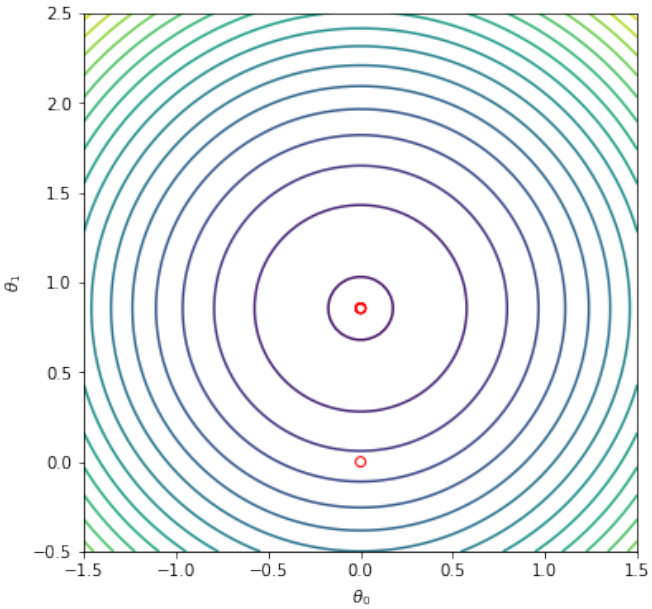
Niveaulinien des Least-Squares-Funktionals beim Häuserpreis-Datensatz

Unskaliertes Problem



Nach 100.000 Iterationen bei einer Lernrate von $\alpha = 10^{-6}$ wurde das Minimum noch nicht annähernd erreicht. Bei größerer Lernrate divergiert das Verfahren.

Nach Standardisierung



Nach nur zwei Iterationen mit Lernrate $\alpha = 1$ befindet man sich sehr nahe am Minimum.

Nach Standardisierung ist

- der empirische Mittelwert der Daten 0
- die empirische Standardabweichung der Daten 1

Zusammenfassung und Ausblick

Ostbayerische Technische Hochschule Amberg-Weiden

Zusammenfassung

- Univariate, multivariate und polynomiale Regression
- Modell-Fitting mit der Least-Squares-Methode
- Analytische Lösung des Optimierungsproblems durch Lösen der Normalgleichungen
- Lösbarkeit der Normalgleichungen
- Gradientenverfahren

Ausblick (nächste Vorlesung)

- Umgang mit Ausreißern
- Fehlermessung
- Signifikanz des Modells
- Probabilistische Sichtweise der linearen Regression
- Verzerrung-Varianz-Zerlegung