Maschinelles Lernen 02

Prof. Dr. Christoph Böhm

Hochschule München

3. Januar 2024

Lineare Regression

Lineare Regression

Lineare Regression

Lineare Regression im Eindimensionalen

Bei der linearen Regression im Eindimensionalen nehmen wir an, dass die Funktion f beschrieben werden kann durch

$$f : \mathbb{R} \to \mathbb{R}$$

mit

$$f_{\mathbf{w}}(x) = \mathbf{w}_1 x + \mathbf{w}_0.$$

Wir nennen $\mathbf{w} = (\mathbf{w}_0, \mathbf{w}_1)^T \in \mathbb{R}^2$ Parameter des Modells.

Lineare Regression

Lineare Regression im Eindimensionalen

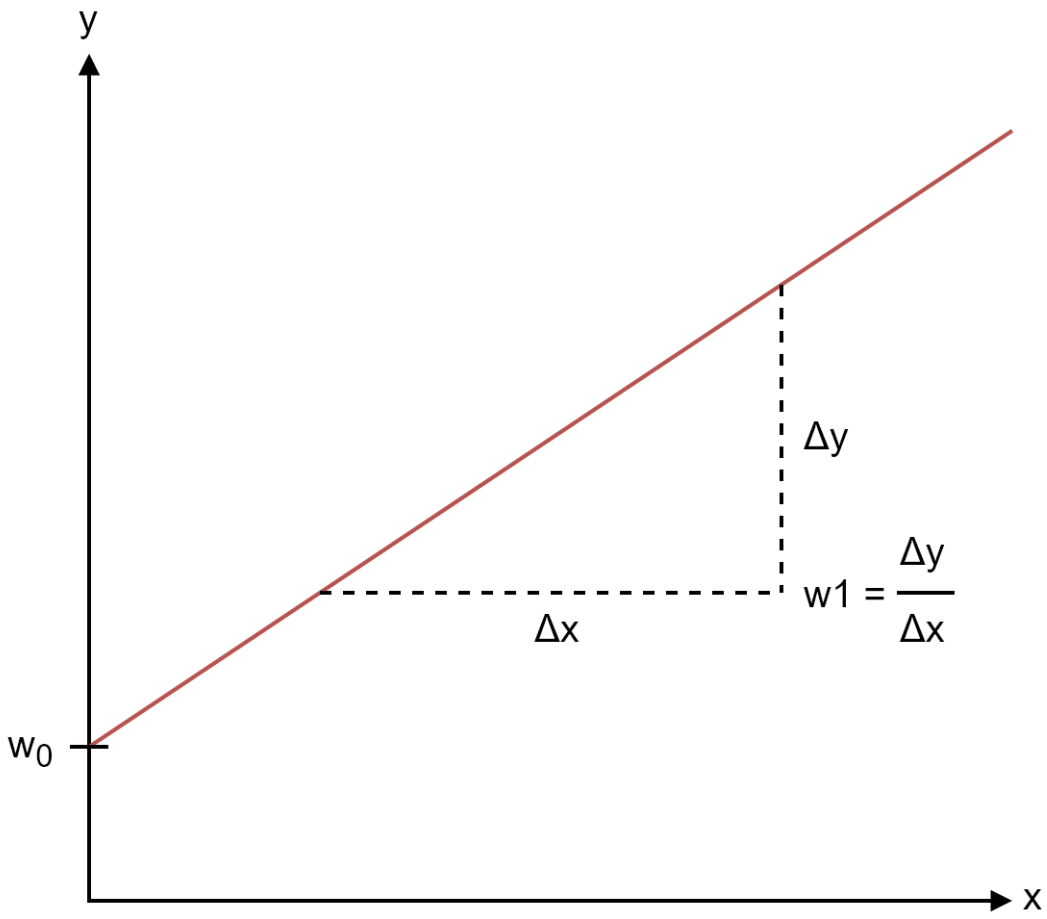


Abbildung 1: Plot einer linearen Funktion mit Parametern $\mathbf{w}_0$ und $\mathbf{w}_1$.

Lineare Regression

Lineare Regression im Eindimensionalen

Gegeben

$$\mathcal{D} = \{(x^{(i)}, y^{(i)}) \in \mathbb{R}^2 \mid 1 \leq i \leq n\},$$

wie bestimmen wir die *besten* Parameter von f?

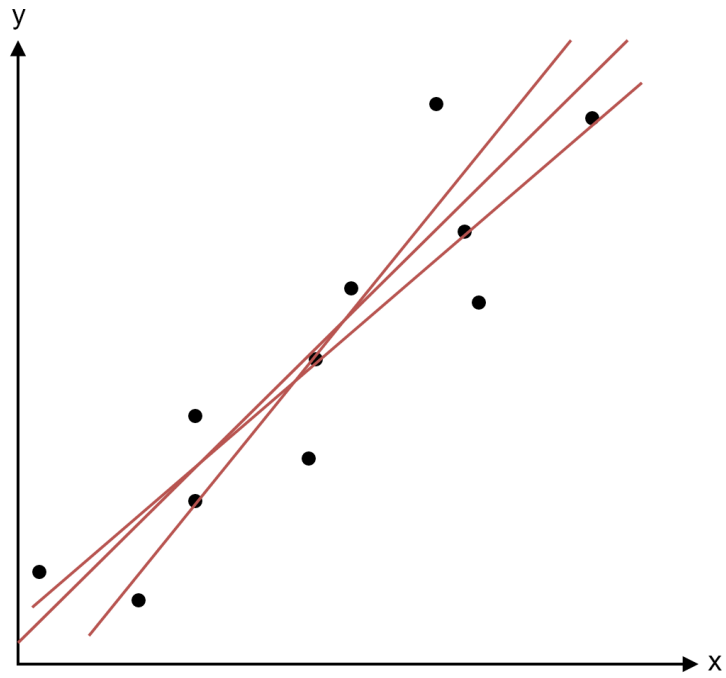


Abbildung 2: Es gibt unendlich viele Wahlmöglichkeiten für jeden der beiden Parameter.

Lineare Regression

Lineare Regression im Eindimensionalen

Wir bestimmen den quadratischen Fehler (Residual Sum of Squares, RSS) der parametrisierten Funktion mit Hilfe der Formel

$$\text{RSS}(\mathbf{w}) = \sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right)^2$$

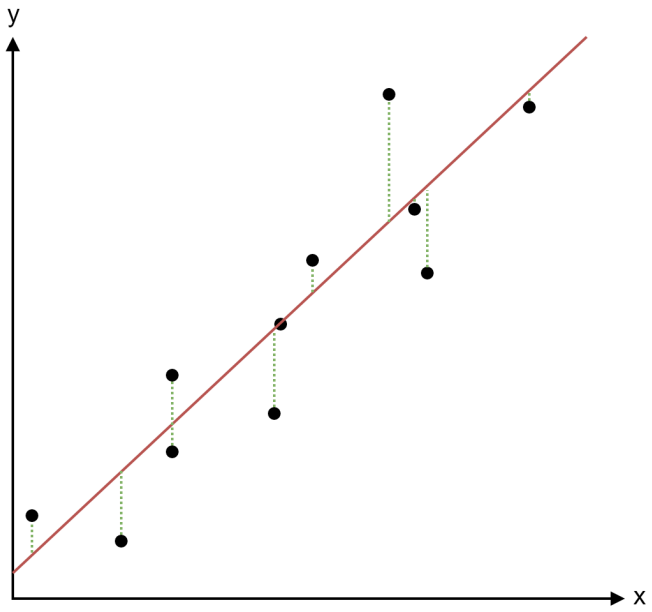


Abbildung 3: Bei der RSS werden die Abstände quadriert und summiert.

Lineare Regression

Lineare Regression im Eindimensionalen

Sollen Modelle mit unterschiedlicher Anzahl von Trainingsdatenpunkte verglichen werden, so verwendet man häufig eine normalisierte Variante der RSS, den mittleren quadratischen Fehler (Mean Squared Error, MSE) definiert als

$$\text{MSE}(\mathbf{w}) = \frac{1}{n}\text{RSS}(\mathbf{w})$$

Im Folgenden jedoch wollen wir der Einfachheit halber den Fehler

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}\left(y^{(i)} - f_{\mathbf{w}}(x^{(i)})\right)^2$$

minimieren, um die beste Funktion $f_{\mathbf{w}}$ zu finden.

Lineare Regression

Lineare Regression im Eindimensionalen

Formal suchen wir also

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}) = \arg\min_{\mathbf{w}} \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right)^2$$

Das Minimum einer Funktion finden wir normalerweise durch

1. Ableiten der Funktion
2. Setzen der Ableitung auf Null
3. Lösen des entstandenen Gleichungssystems
4. Untersuchung der gefundenen Lösungen (Hochpunkt, Tiefpunkt, Sattelpunkt, etc.)

Lineare Regression

Lineare Regression im Eindimensionalen

Glücklicherweise können wir uns in diesem Fall den letzten Punkt sparen, da $E(\mathbf{w})$ eine konvexe Funktion ist und genau ein Minimum besitzt – sofern das Problem wohldefiniert ist.
Wir setzen also

$$\nabla E(\mathbf{w}) = \left(\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_0}, \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_1} \right)^T = \mathbf{0}$$

Lineare Regression

Lineare Regression im Eindimensionalen

$$
\begin{aligned}
\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_0} &= \frac{\partial}{\partial \mathbf{w}_0} \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right)^2 \\
&= \frac{1}{2} \sum_{i=1}^{n} 2 \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right) \frac{\partial}{\partial \mathbf{w}_0} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right) \\
&= \sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right) \left(-\frac{\partial}{\partial \mathbf{w}_0} f(x^{(i)}) \right) \\
&= -\sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right) \frac{\partial}{\partial \mathbf{w}_0} \left(\mathbf{w}_1 x^{(i)} + \mathbf{w}_0 \right) \\
&= -\sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right) \\
&= -\sum_{i=1}^{n} y^{(i)} + \mathbf{w}_1 \sum_{i=1}^{n} x^{(i)} + n\mathbf{w}_0
\end{aligned}
$$

Lineare Regression

Lineare Regression im Eindimensionalen

$$
\begin{aligned}
\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_1} &= \frac{\partial}{\partial \mathbf{w}_1} \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right)^2 \\
&= -\sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right) \frac{\partial}{\partial \mathbf{w}_1} \left(\mathbf{w}_1 x^{(i)} + \mathbf{w}_0 \right) \\
&= -\sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right) x^{(i)} \\
&= -\sum_{i=1}^{n} x^{(i)} y^{(i)} + \mathbf{w}_1 \sum_{i=1}^{n} x^{(i)} x^{(i)} + \mathbf{w}_0 \sum_{i=1}^{n} x^{(i)}
\end{aligned}
$$

Lineare Regression

Lineare Regression im Eindimensionalen

Wir erhalten also mit

$$\left(\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_0}, \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_1} \right)^T = \mathbf{0}$$

ein lineares Gleichungssystem mit zwei Gleichungen und zwei Unbekannten $(\mathbf{w}_0, \mathbf{w}_1)$, welches prinzipiell eindeutig lösbar ist. So ein Gleichungssystem könnten wir direkt mit Hilfe eines entsprechenden Algorithmus wie z.B. dem *Gausschen Eliminationsverfahren* lösen. Im Bereich des maschinellen Lernens kann es uns jedoch schnell passieren, dass die entstehenden Gleichungssysteme sehr groß oder nicht eindeutig lösbar werden. Daher werden meist iterative Verfahren, wie das Gradientenabstiegsverfahren verwendet.

Lineare Regression

Gradientenabstiegsverfahren

Beispiel

Die quadratische Funktion $f(x) = x(x-2)$ besitzt ein Minimum bei $x = 1$, da $f'(x) = 2x - 2 = 0 \Leftrightarrow x = 1$ und $f''(x) = 2 > 0$.

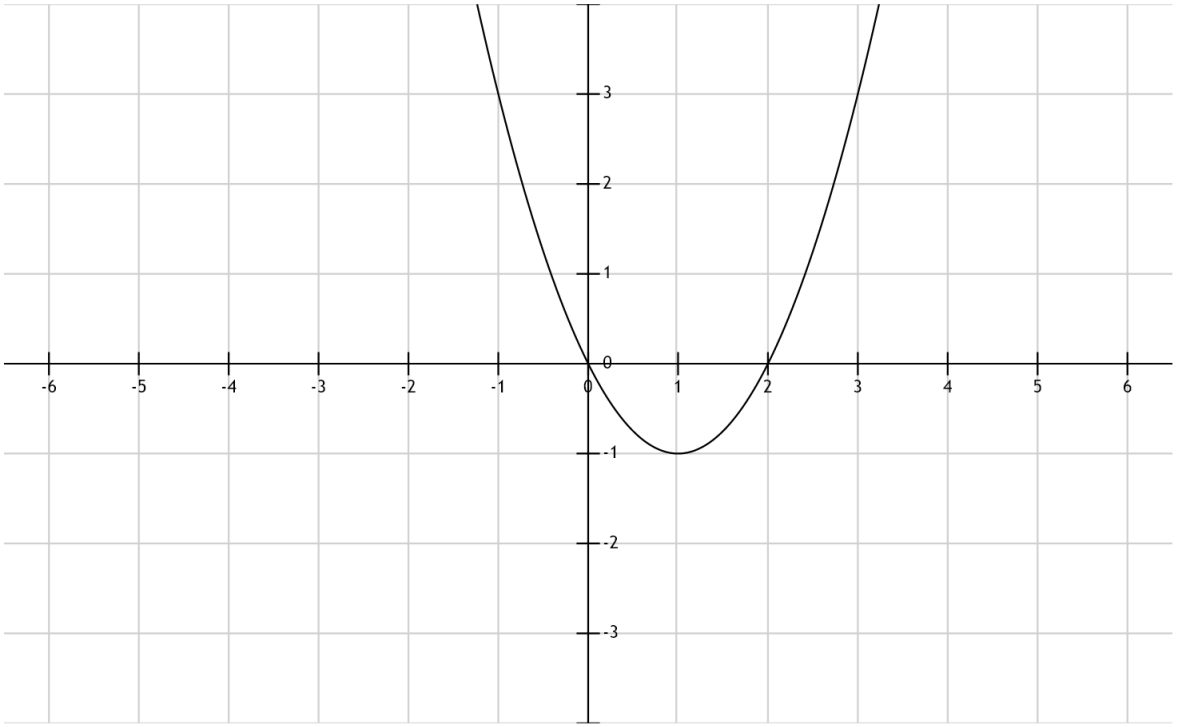


Abbildung 4: Plot der Funktion $f(x) = x(x-2)$.

Lineare Regression

Gradientenabstiegsverfahren

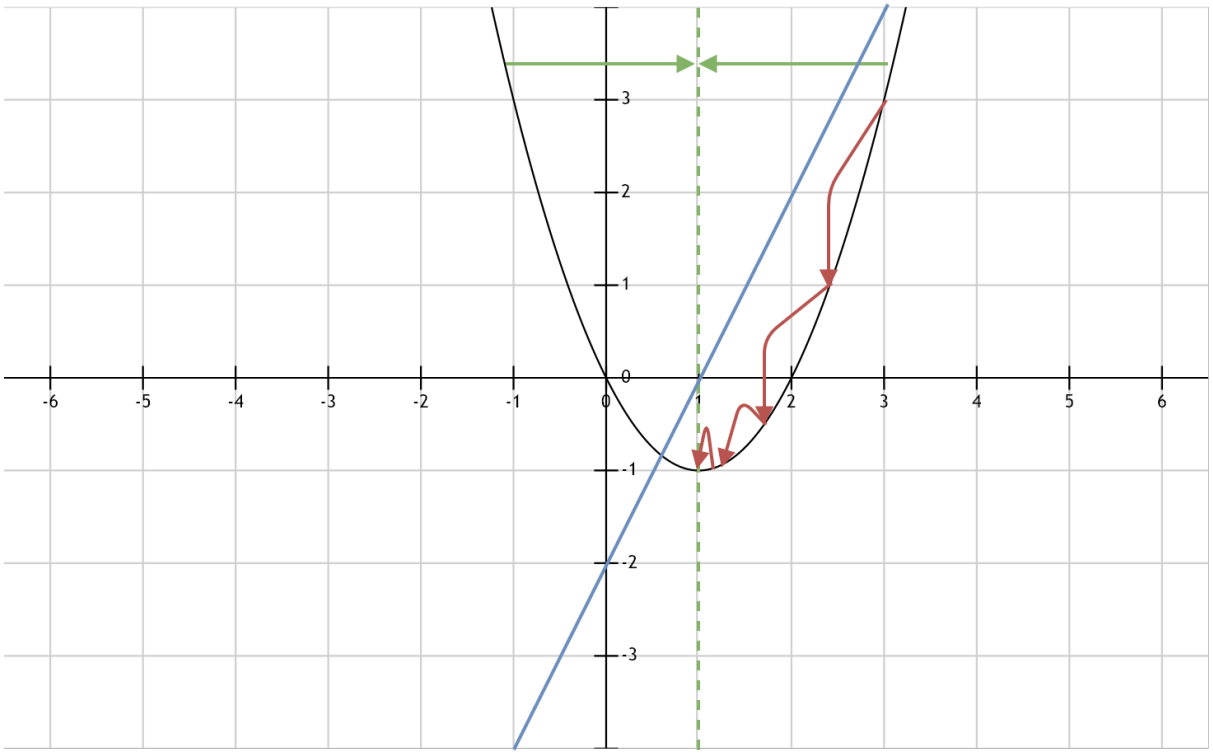


Abbildung 5: Gradientenabstiegsverfahren auf $f(x) = x(x-2)$.

Folgt man iterativ einem Bruchteil η der negativen ersten Ableitung, also $-\eta f'(x) = \eta(2-2x)$ bringt einen dies näher und näher an das Minimum.

Lineare Regression

Gradientenabstiegsverfahren

Algorithm 1 gradient_descent_1D($\mathcal{D}$, η, steps)

$\mathbf{w}_0 = 0$, $\mathbf{w}_1 = 0$
for step $= 1 \ldots$ steps **do**
    $\Delta\mathbf{w}_0 = 0$, $\Delta\mathbf{w}_1 = 0$
    for $(x, y) \in \mathcal{D}$ **do**
        $\Delta\mathbf{w}_0 = \Delta\mathbf{w}_0 - y + \mathbf{w}_1 x + \mathbf{w}_0$
        $\Delta\mathbf{w}_1 = \Delta\mathbf{w}_1 - xy + \mathbf{w}_1 xx + \mathbf{w}_0 x$
    end for
    $\mathbf{w}_0 = \mathbf{w}_0 - \eta\Delta\mathbf{w}_0$
    $\mathbf{w}_1 = \mathbf{w}_1 - \eta\Delta\mathbf{w}_1$
end for
return $\mathbf{w}_0, \mathbf{w}_1$

Lineare Regression

Gradientenabstiegsverfahren

Das Gradientenabstiegsverfahren ist eine Ausprägung von Liniensuchverfahren, bei denen eine Funktion $f : \mathbb{R}^d \to \mathbb{R}$ entlang eines Richtungsvektors (in diesem Fall dem Gradienten) optimiert wird. Der Hyperparameter $\eta \in \mathbb{R}_{>0}$ im Gradientenabstiegsverfahren wird auch Lernrate genannt. Er hat direkten Einfluss auf die Geschwindigkeit, in der sich das Verfahren dem Minimum / der Konvergenz nähert. Üblicherweise beobachtet man den zu minimierenden Fehler $E(\mathbf{w})$ während der Laufzeit, um die Anzahl der Iterationen zu bestimmen.

Lineare Regression

Gradientenabstiegsverfahren

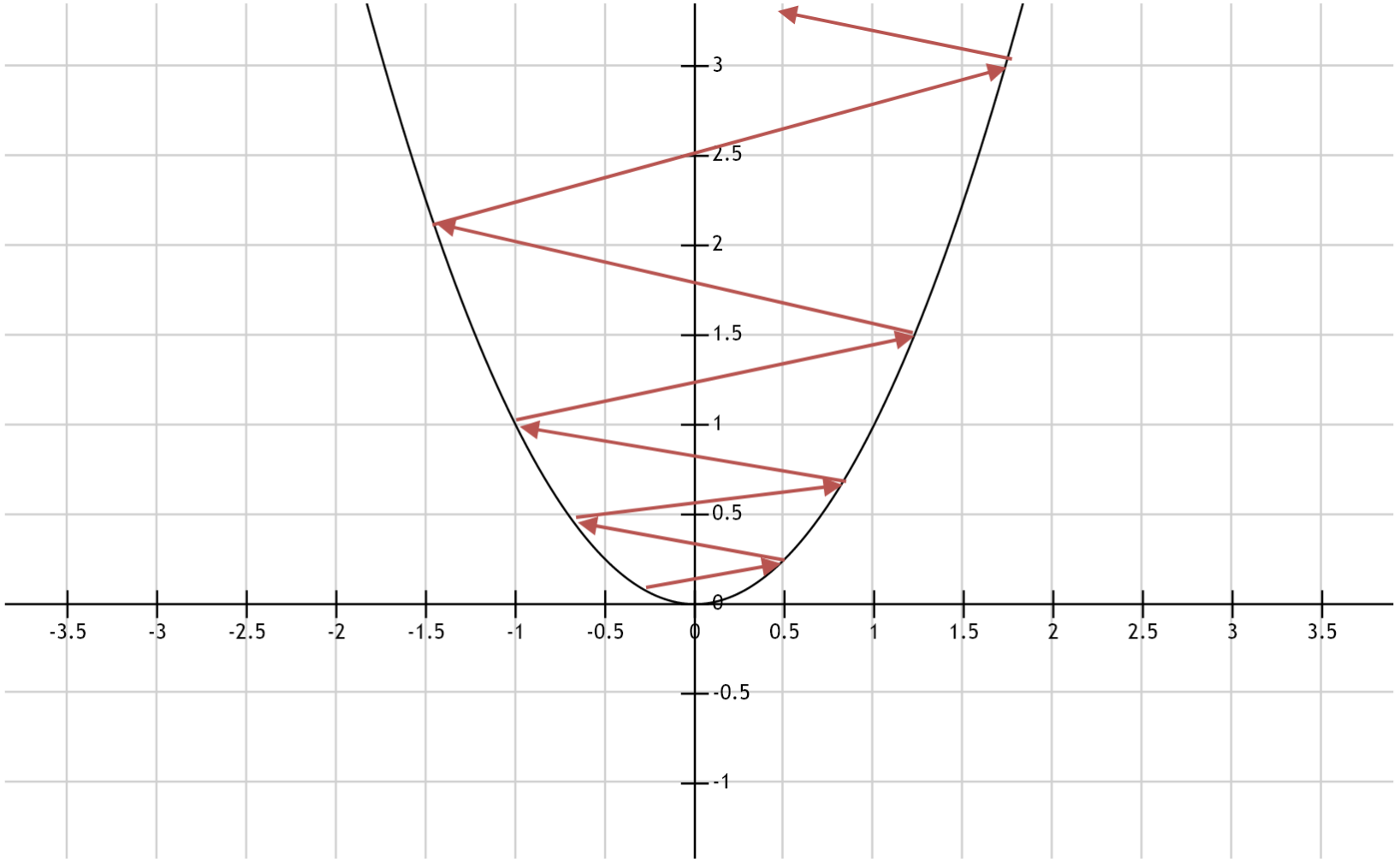


Abbildung 6: Ist die Lernrate η zu groß, kann es zu Oszillationen kommen und das Verfahren konvergiert nicht.

Lineare Regression

Gradientenabstiegsverfahren

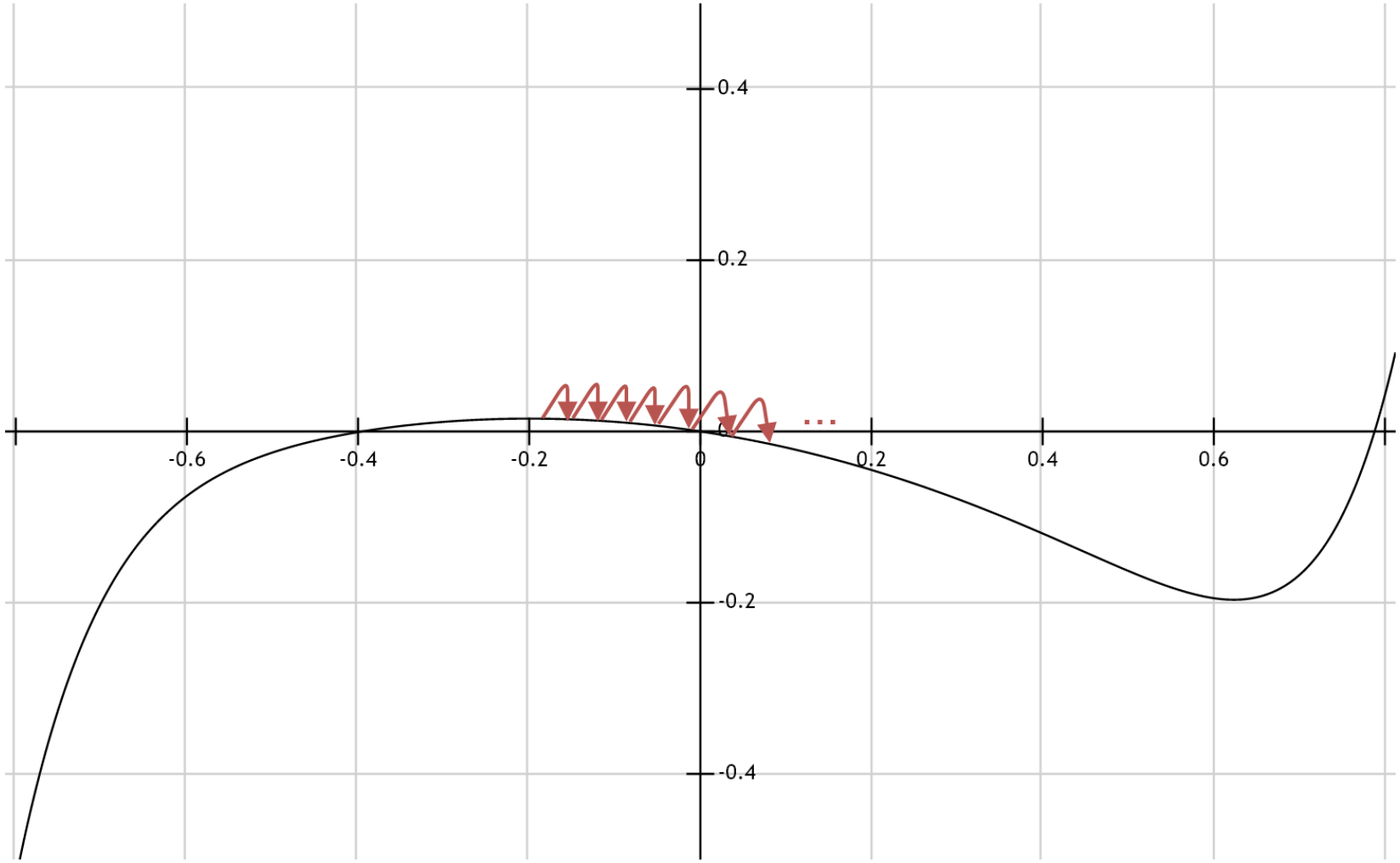


Abbildung 7: Ist die Lernrate η zu klein, werden sehr viele Schritte bis zur Konvergenz benötigt.

Lineare Regression

Mehrdimensionale Lineare Regression

Bei der linearen Regression im Mehrdimensionalen nehmen wir an, dass die Eingabemenge mehrdimensional ist, also $\mathcal{X} = \mathbb{R}^d$ und somit die Funktion f beschrieben werden kann durch

$$f : \mathbb{R}^d \to \mathbb{R}$$

mit

$$f_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{d} \mathbf{w}_i \mathbf{x}_i + \mathbf{w}_0.$$

Die Parameter des Modells sind $\mathbf{w} = (\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_d)^T \in \mathbb{R}^{d+1}$.

Lineare Regression

Mehrdimensionale Lineare Regression

Um eine kompaktere Darstellung zu erreichen, verwenden wir einen Trick. Wir nehmen implizit an, dass wir die Eingabe die Form

$$\mathbf{x} = (1, \mathbf{x}_1, \ldots, \mathbf{x}_d)^T$$

mit $\mathbf{x}_0 = 1$ hat und erhalten schließlich

$$f_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{d} \mathbf{w}_i \mathbf{x}_i + \mathbf{w}_0 = \mathbf{w}^T \mathbf{x}.$$

Lineare Regression

Mehrdimensionale Lineare Regression

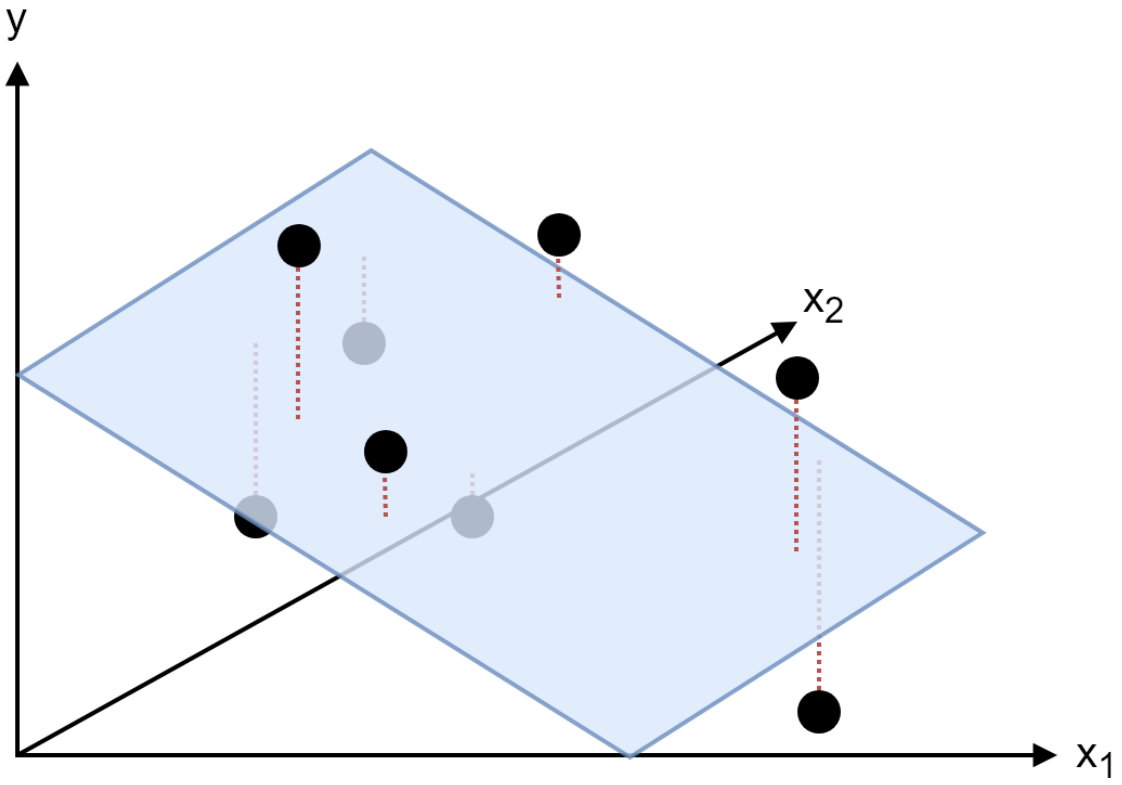


Abbildung 8: Bei der mehrdimensionalen Regression wird eine Verallgemeinerung der Gerade, allgemein eine Hyperebene, im dreidimensionalen Raum wie hier eine normale Ebene, so im Raum positioniert, dass der Abstand zu den Datenpunkten minimiert wird.

Lineare Regression

Mehrdimensionale Lineare Regression

Auch für die mehrdimensionale lineare Regression sind die bekannten Definitionen der RSS und des MSE gültig. Abermals verwenden wir die leicht abgewandelte Fehlermetrik

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - f(\mathbf{x}^{(i)}) \right)^2 .$$

Wir folgen auch wieder dem negativen Gradienten

$$\nabla E(\mathbf{w}) = \left(\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_0}, \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_1}, \ldots, \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_d} \right)^T$$

um das Minimum zu finden.

Lineare Regression

Gradientenabstiegsverfahren

Algorithm 2 gradient_descent($\mathcal{D}$, η, steps)

1: $\mathbf{w} = \mathbf{0}$
2: **for** step $= 1 \ldots$ steps **do**
3: $\quad \Delta\mathbf{w} = \mathbf{0}$
4: $\quad$ **for** $(\mathbf{x}, y) \in \mathcal{D}$ **do**
5: $\quad\quad \Delta\mathbf{w} = \Delta\mathbf{w} - (y - f(\mathbf{x}))\,\nabla f(\mathbf{x})$
6: $\quad$ **end for**
7: $\quad \mathbf{w} = \mathbf{w} - \eta\Delta\mathbf{w}$
8: **end for**
9: **return** $\mathbf{w}$

Hinweis:

$$\nabla f(\mathbf{x}) = (1, \mathbf{x}_1, \ldots, \mathbf{x}_d)^T$$

Lineare Regression

Gradientenabstiegsverfahren

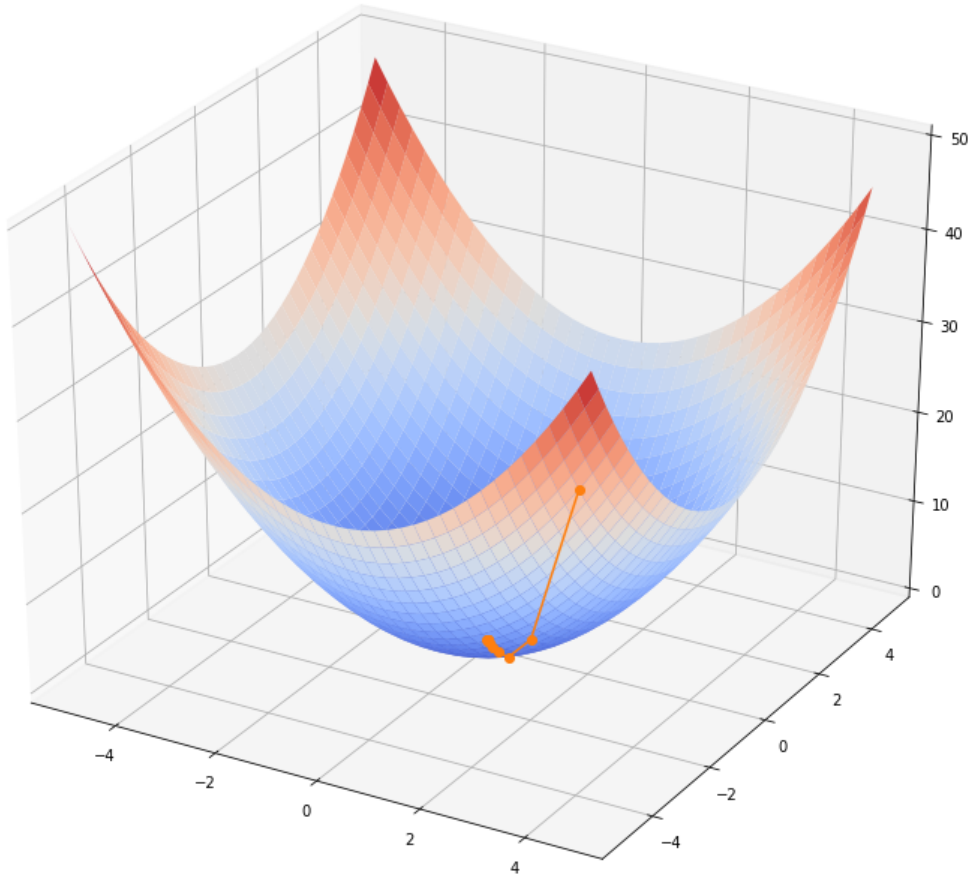


Abbildung 9: Gradientenabstiegsverfahren im mehrdimensionalen Raum bei der Funktion $f(\mathbf{x}) = \mathbf{x}_1^2 + \mathbf{x}_2^2$.

Lineare Regression

Genauigkeit

Wenn wir nun ein Regressionsmodell z.B. durch Anwendung des Gradientenabstiegsverfahrens gefunden haben, sollten wir uns fragen, wie genau, also wie gut, unser Modell eigentlich ist. Hier könnten wir prinzipiell den quadratischen Fehler

$$\text{RSS}(\mathbf{w}) = \sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right)^2$$

oder noch besser den mittleren quadratischen Fehler

$$\text{MSE}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(x^{(i)}) \right)^2$$

verwenden, welcher unabhängig von der Anzahl der Trainingsdatenpunkte ist.

Lineare Regression

Genauigkeit

Für den MSE können wir jedoch keine allgemein gültige Skala angeben, da die Ausmaße abhängig vom Wertebereich des Problems, also der y-Werte ist. Hier hilft uns die R^2-Statistik (sog. Bestimmtheitsmaß), definiert über den quadratischen Gesamtfehler (Total Sum of Squares, TSS)

$$TSS = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y} \right)^2$$

mit

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

als

$$R^2(\mathbf{w}) = \frac{TSS - RSS(\mathbf{w})}{TSS} = 1 - \frac{RSS(\mathbf{w})}{TSS}.$$

Lineare Regression

Genauigkeit

Die TSS misst die komplette Varianz in den Ausgabedaten $y^{(i)}$ und somit misst

$$TSS - RSS(\mathbf{w})$$

die Varianz, die durch das Regressionsmodell mit den Parametern $\mathbf{w}$ erklärt wird. Die R^2-Statistik misst daher den Anteil der kompletten Varianz, der durch das Modell erklärt wird und nimmt normalerweise Werte im Intervall $[0, 1]$ an. Im Prinzip können auch negative Werte entstehen, wenn das Modell nicht zumindest dem linearen Trend der Daten folgt.

- Ein R^2-Wert nahe 1 zeugt von einem relativ passenden Modell, da die Daten sehr gut durch das Modell erklärt werden.
- Ein R^2-Wert nahe 0 bedeutet, dass das Modell die Daten nur relativ schlecht erklären kann.

Lineare Regression

Genauigkeit

In der Praxis wird die R^2-Statistik sehr oft für die Beurteilung von Modellen und dem Vergleich von Modellen untereinander verwendet, da sie unabhängig von der Anzahl der Trainingsdaten und dem Wertebereich ist.
Jedoch im Allgemeinen zu bestimmen, ab welchem Wert ein Modell gut ist, ist nicht zielführend.

- In empirischen Wissenschaften, wie Psychologie, Biologie oder Medizin, ist es oft schwierig das perfekte erklärende Modell zu finden und oft liegen die Daten mit hohem Rauschen vor. Ein Modell mit einem guten jedoch durchaus von 1 weiter entfernten R^2-Wert kann oft schon sinnvoll sein.
- Bei manchen Zusammenhängen z.B. in der Physik weiß man, welcher Natur das dahinterliegende Modell ist. Auch die Messungenauigkeit kann man minimieren. Hier werden oft sehr hohe R^2-Werte angestrebt.

Lineare Regression

Genauigkeit

Meist kann die Leistungsfähigkeit eine ML-Modells nicht beliebig gesteigert werden, denn es muss eine möglichst gute Balance zwischen den zwei Fehlerarten Verzerrung (Bias) und Varianz (Variance) gefunden werden.

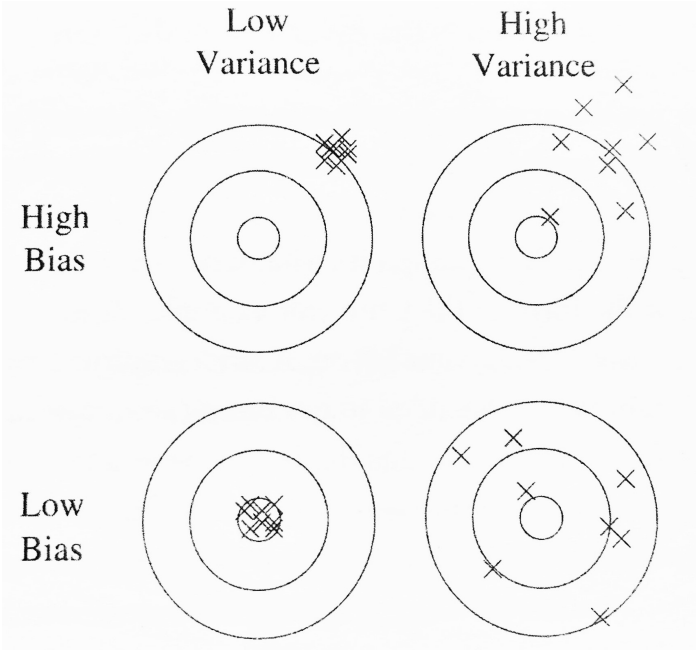


Abbildung 10: Bias-Variance-Tradeoff, Abbildung übernommen aus *Pedro Domingos, The Master Algorithm, Penguin Books, 2017 (S. 79)*.

Lineare Regression

Genauigkeit

No Free Lunch Theorem

Ein Modell basiert meist auf einer Vereinfachung der echten Welt, um den Fokus auf die wirklich wichtigen Aspekte zu lenken. Jedoch müssen hierfür Annahmen getroffen werden, weshalb diese Annahmen in manchen/vielen Situationen richtig sind, in anderen jedoch potentiell falsch, was zu Fehlern bei der Vorhersage führt. Das No Free Lunch Theorem besagt, dass es kein Modell geben kann, welches für alle Probleme am besten funktioniert, da die Annahmen des einen Modells eben nicht unbedingt für das andere Problem gelten müssen.

Lineare Regression

Interpretierbarkeit

Modelle basierend auf linearer Regression besitzen den großen Vorteil, dass die Parameter $\mathbf{w}$ vom Menschen interpretierbar sind. Dies erleichtert u.a. die Sicherstellung der Korrektheit.

$\mathbf{w}_i > 0$	positiver Zusammenhang, d.h. steigt $\mathbf{x}_i$ um m, so steigt y um $m\lvert\mathbf{w}_i\rvert$
$\mathbf{w}_i \approx 0$	(fast) kein (linearer) Zusammenhang zwischen $\mathbf{x}_i$ und y
$\mathbf{w}_i < 0$	negativer Zusammenhang, d.h. steigt $\mathbf{x}_i$ um m, so sinkt y um $m\lvert\mathbf{w}_i\rvert$

jeweils unter der Annahme, dass die anderen Einflussgrößen gleich bleiben.

Lineare Regression

Nichtlineare Zusammenhänge

Mit Hilfe der mehrdimensionalen linearen Regression und einem Trick können wir auch nichtlineare Zusammenhänge lernen. Hierfür benötigen wir eine Funktion $\phi : \mathbb{R} \to \mathbb{R}^d$, mit welcher wir einen Basiswechsel vollziehen können. Beispielsweise erlaubt uns die Funktion

$$\phi : \mathbb{R} \to \mathbb{R}^2, \phi(x) = \left(x, x^2\right)^T$$

und die Funktionskonkatenation $f \circ \phi$ mit der linearen Funktion $f : \mathbb{R}^2 \to \mathbb{R}, f(\mathbf{x}) = \mathbf{w}_2\mathbf{x}_2 + \mathbf{w}_1\mathbf{x}_1 + \mathbf{w}_0$ die Darstellung der quadratischen Funktion

$$(f \circ \phi)(x) = \mathbf{w}_2 x^2 + \mathbf{w}_1 x + \mathbf{w}_0.$$

Lineare Regression

Nichtlineare Zusammenhänge

Wir sind dabei nicht auf eindimensionale Eingabegrößen beschränkt. Mit Hilfe von

$$\phi : \mathbb{R}^2 \to \mathbb{R}^5, \phi(\mathbf{x}) = \left(\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_1\mathbf{x}_2, \mathbf{x}_2^2, \mathbf{x}_1^2\right)^T$$

und der linearen Funktion

$$f : \mathbb{R}^5 \to \mathbb{R}, f(\mathbf{x}) = \sum_{i=1}^{5} \mathbf{w}_i \mathbf{x}_i + \mathbf{w}_0$$

erzeugen wir die nichtlineare Funktion

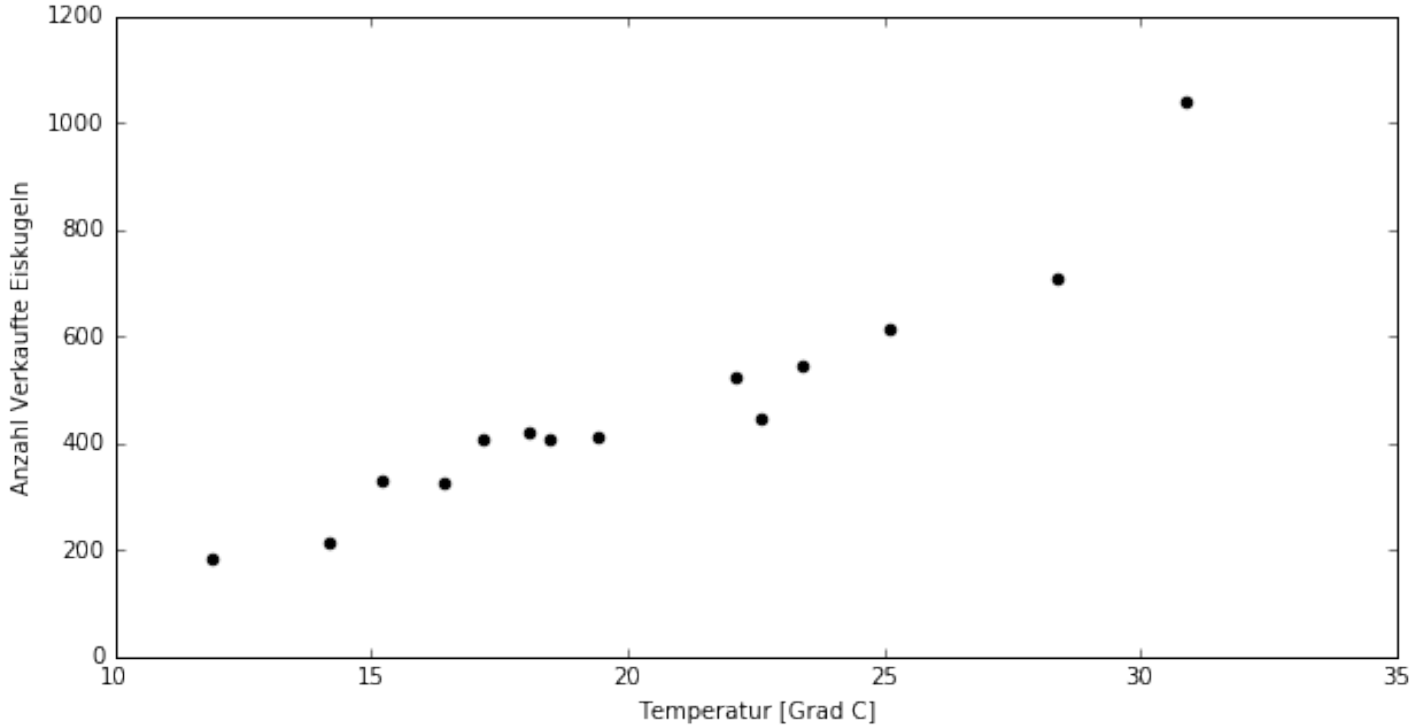
$$(f \circ \phi)(\mathbf{x}) = \mathbf{w}_5\mathbf{x}_1^2 + \mathbf{w}_4\mathbf{x}_2^2 + \mathbf{w}_3\mathbf{x}_1\mathbf{x}_2 + \mathbf{w}_2\mathbf{x}_1 + \mathbf{w}_1\mathbf{x}_2 + \mathbf{w}_0.$$

Lineare Regression

Nichtlineare Zusammenhänge

Predictive Analytics im Eisdielenbusiness

Klar: Je schöner das Wetter desto mehr Eis wird verkauft. Aber wie ist der genaue Zusammenhang?

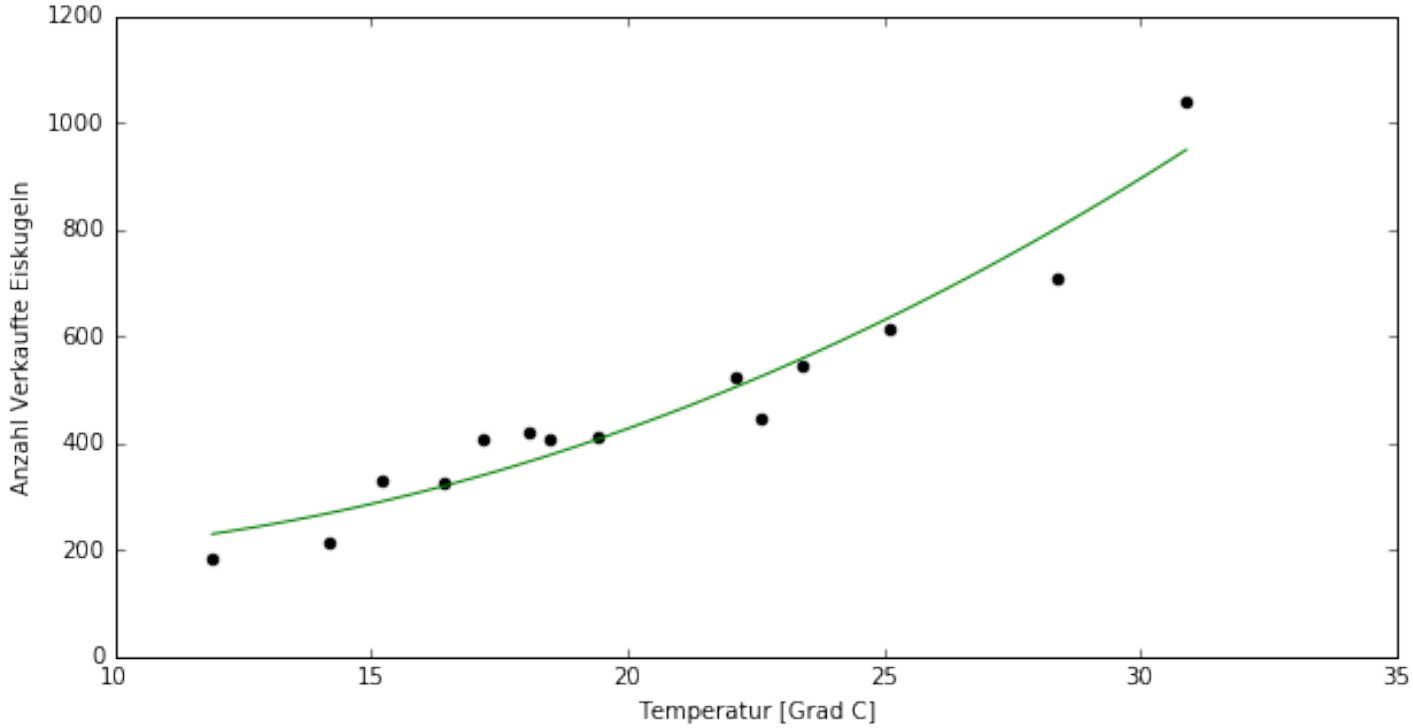


Lineare Regression

Nichtlineare Zusammenhänge

Anwendungsbeispiel

Wir nehmen einen quadratischen Zusammenhang $f(x) = \mathbf{w}_2 \cdot x^2 + \mathbf{w}_1 \cdot x + \mathbf{w}_0$ zwischen der Temperatur und der Anzahl der verkauften Kugeln Eis an.



Lineare Regression

Nichtlineare Zusammenhänge

Prinzipiell hätten wir nun mit

- der mehrdimensionalen linearen Regression,
- dem Basiswechseltrick und
- dem Gradientenabstiegsverfahren

alle nötigen Werkzeuge um ein Polynom n-ten Grades perfekt an unsere n Datenpunkte $\mathcal{D}$ zu fitten. Aber ist das immer eine gute Idee?

Lineare Regression

Nichtlineare Zusammenhänge

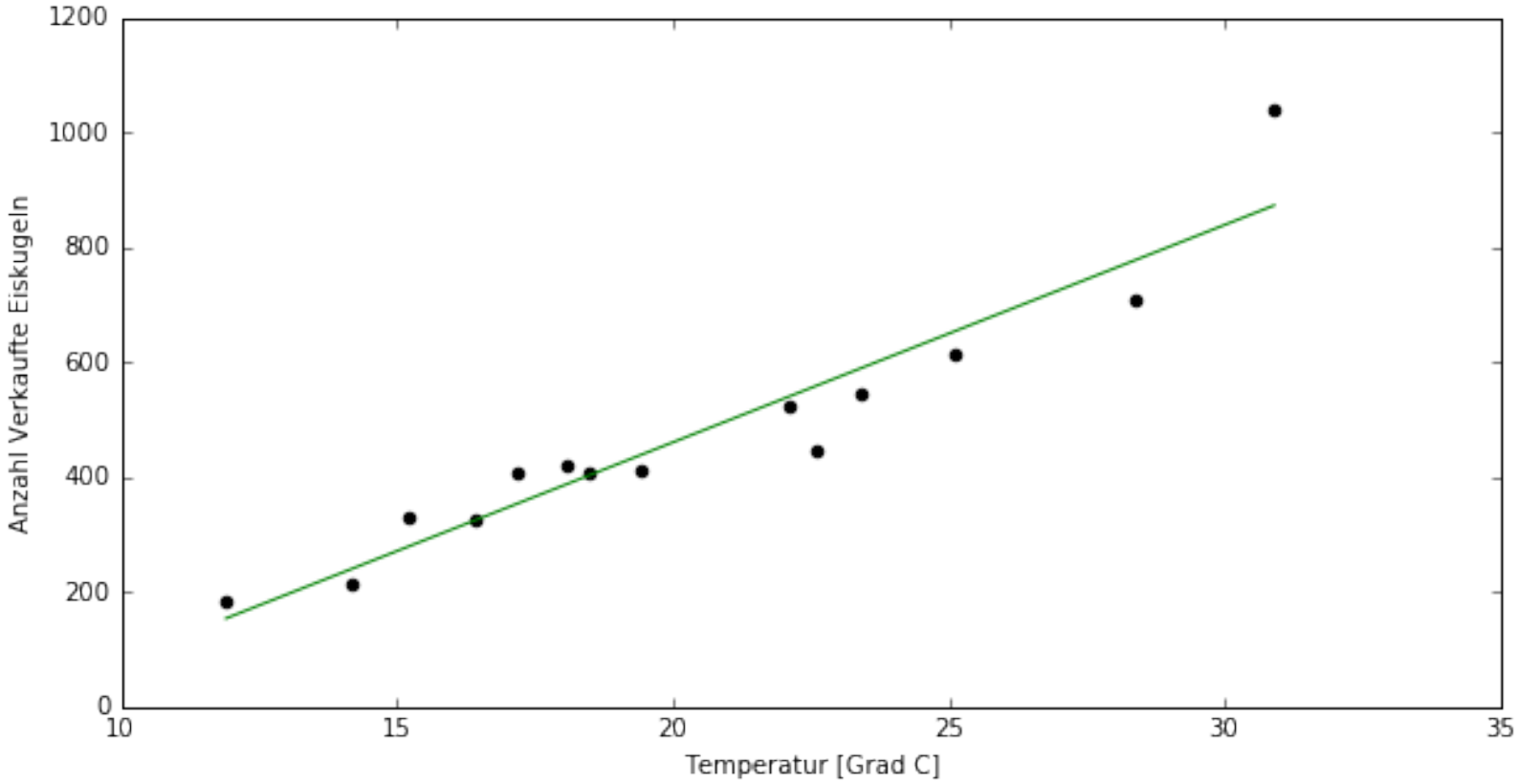


Abbildung 11: Lineare Regression eines Polynoms 1-ten Grades an die Eisverkaufdaten durch Basiserweiterung. Gewichte $\mathbf{w} \approx (-296, 37.8)^T$

Lineare Regression

Nichtlineare Zusammenhänge

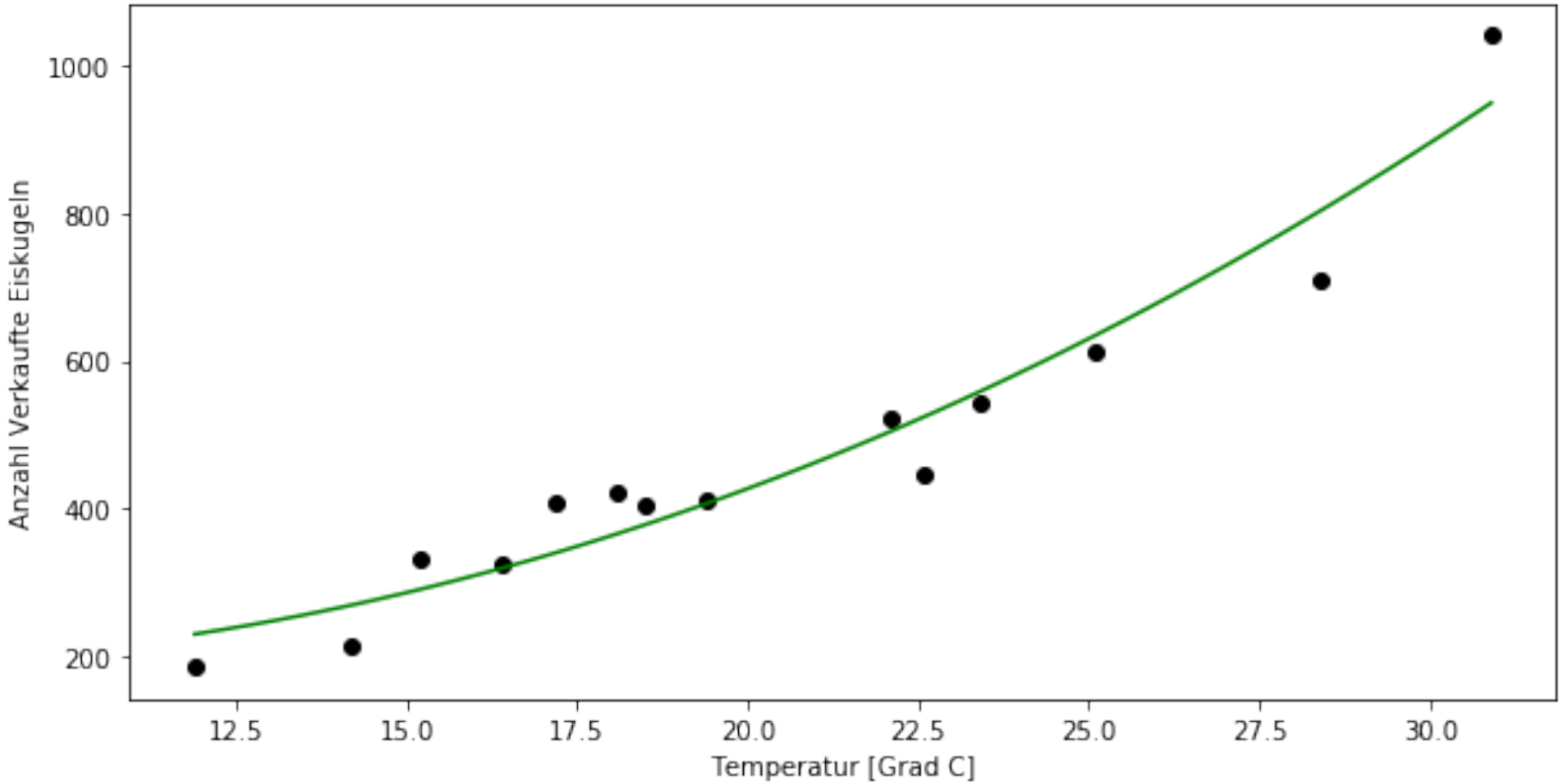


Abbildung 12: Lineare Regression eines Polynoms 2-ten Grades an die Eisverkaufdaten durch Basiserweiterung. Gewichte
$\mathbf{w} \approx (237, -15.3, 1.24)^T$

Lineare Regression

Nichtlineare Zusammenhänge

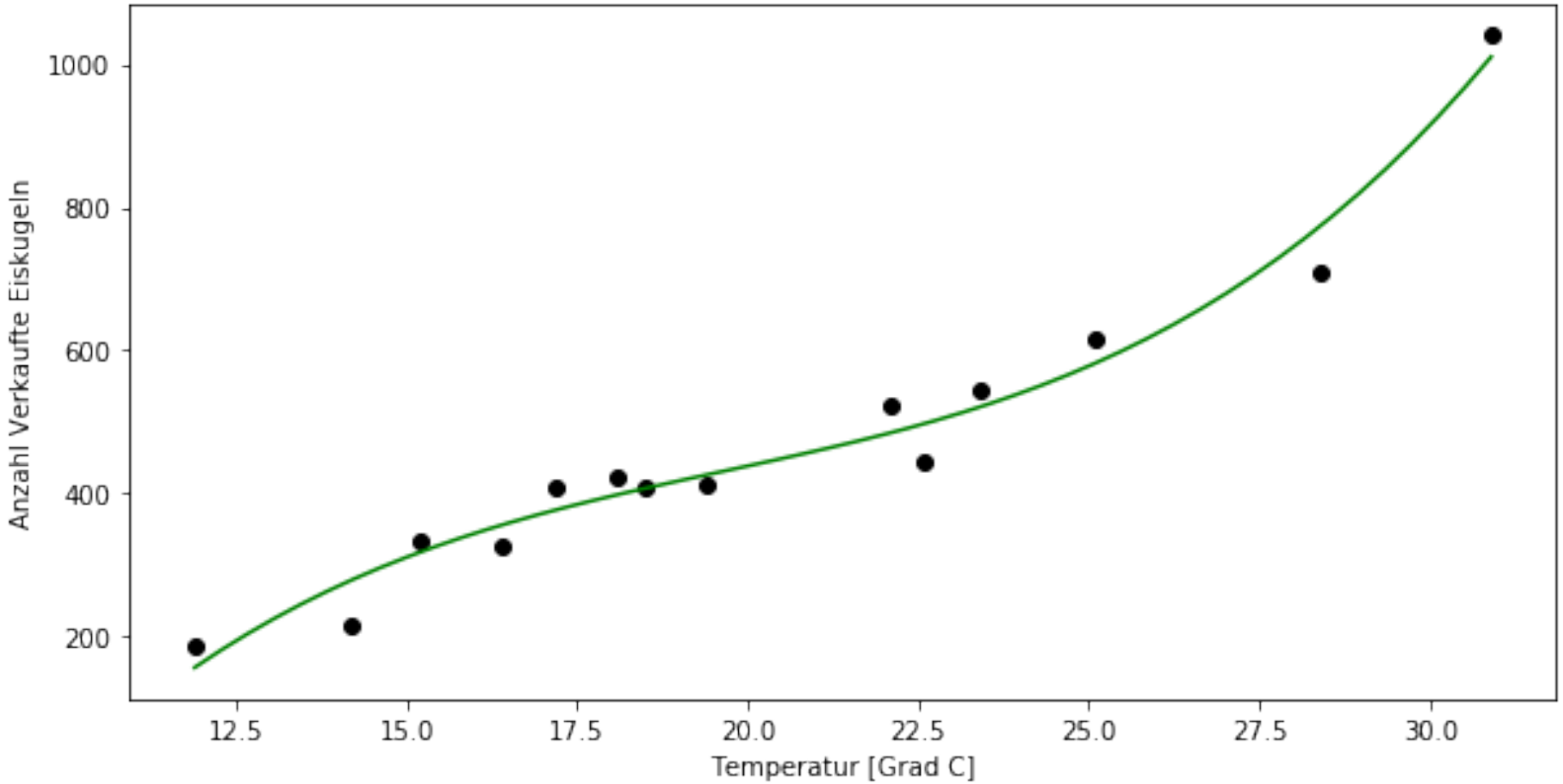


Abbildung 13: Lineare Regression eines Polynoms 3-ten Grades an die Eisverkaufdaten durch Basiserweiterung. Gewichte
$\mathbf{w} \approx (-1853, 307, -14.6, 0.247)^T$

Lineare Regression

Nichtlineare Zusammenhänge

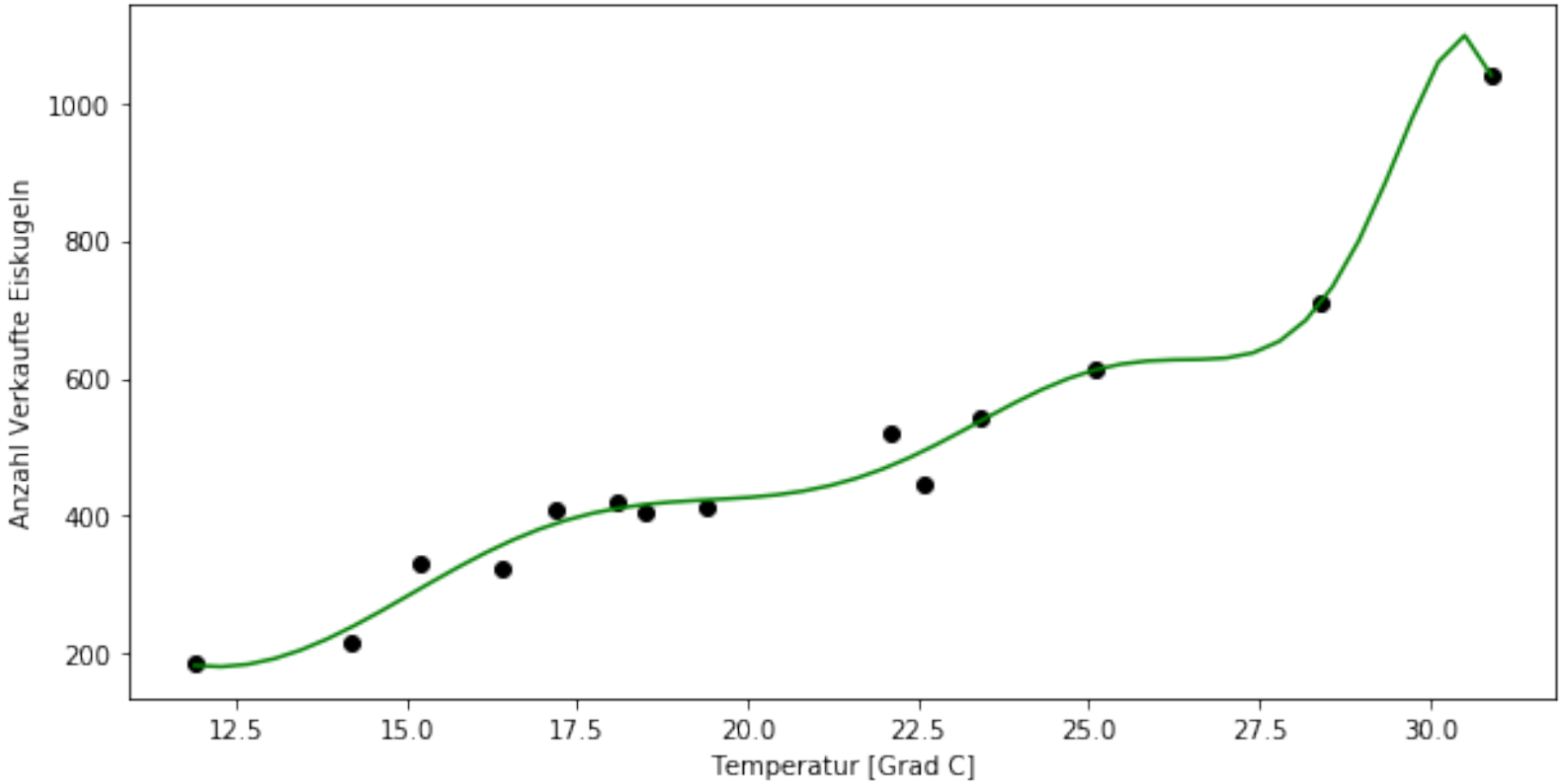


Abbildung 14: Lineare Regression eines Polynoms 12-ten Grades an die Eisverkaufdaten durch Basiserweiterung. Gewichte $\mathbf{w} = (0, -0.0000457, -0.00000496, -0.0000570, -0.000489, -0.00297, -0.00977, 0.00256, -0.000271, 0.0000152, -0.000000471, 0.00000000777, -0.0000000000530)^T$

Lineare Regression

Nichtlineare Zusammenhänge

Die Probleme, die durch unkontrollierte Modellkomplexität, in diesem Fall dem Grad und damit auch die Koeffizienten des zu lernenden Polynoms, sind

- numerische Probleme
- Überanpassung (Overfitting), d.h. das Modell passt sich zu sehr an die Daten $\mathcal{D}$ und damit auch an dort enthaltenes Rauschen oder eine systematische Auswahl an und ist nicht in der Lage zu generalisieren, was meist in einer guten Performance unter Laborbedingungen aber einer schlechten Leistung in der Praxis resultiert

Lineare Regression

Trainings- und Testdaten

Woher sollen wir wissen, ob unser Modell zu komplex ist? Dazu verwenden wir wieder einen Trick, wir zerteilen den kompletten Datensatz $\mathcal{D}$ in zwei Teile

$$\mathcal{D} = \mathcal{T} \dot{\cup} \mathcal{V}$$

- den Trainingsdatensatz $\mathcal{T}$, welcher für das Training/Lernen verwendet wird und
- den Testdatensatz $\mathcal{V}$, welcher *ungesehene* Daten für die Validierung der Praxistauglichkeit dient.

Lineare Regression

Trainings- und Testdaten

Prinzipiell können nun folgende Situationen auftreten:

- Idealerweise hat man geringen Fehler sowohl auf $\mathcal{T}$ und $\mathcal{V}$.
- Ein hoher Fehler auf $\mathcal{T}$ lässt auf Unteranpassung schließen, beispielsweise durch zu wenig Daten bzw. Trainingsschritte oder eine zu niedrige Modellkomplexität.
- Ein geringer Fehler auf $\mathcal{T}$ aber hoher Fehler auf $\mathcal{V}$ ist oft ein Resultat von Überanpassung, was unter Umständen durch eine verringerte Modellkomplexität korrigiert werden kann.

Lineare Regression

Optimierung von Hyperparametern

Meist besitzt ein Modell Hyperparameter, welche die Komplexität beeinflußen.

Beispiel: Polynome

Lineare Regression auf Polynomen mit Hilfe des Gradientenabstiegsverfahren besitzt normalerweise drei Hyperparameter:

- Lernrate η: Einfluss auf Modellkomplexität relativ komplex, sollte nicht zu hoch oder niedrig sein
- Anzahl der Lernschritte: Je geringer die Anzahl der Schritte, desto mehr wird die Überanpassung verhindert, kann jedoch schnell zur Unteranpassung führen
- Polynomgrad: Je höher, desto komplexer das Modell – zu hoch: Überanpassung, zu niedrig: Unteranpassung

Lineare Regression

Optimierung von Hyperparametern

Die Hyperparameter lassen sich durch Überlegung und manuelle Justierung optimieren als auch automatisiert über eine weitere Schleife.

Algorithm 3 optimal_polynome($\mathcal{D}$, η, steps, max_d)

$\mathcal{T}, \mathcal{V} = \text{split}(\mathcal{D})$
$\mathbf{w}^* = \mathbf{0}$, $\text{MSE}^* = \infty$
for d $= 0 \ldots$ max_d **do**
    $\mathbf{w} = \text{gradient\_descent}_d(\mathcal{T}, \eta, \text{steps})$
    $\text{MSE} = \text{MSE}_{\mathbf{w}}(\mathcal{V})$
    if $\text{MSE} < \text{MSE}^*$ **then**
        $\mathbf{w}^* = \mathbf{w}$, $\text{MSE}^* = \text{MSE}$
    end if
end for
return $\mathbf{w}$

Lineare Regression

Optimierung von Hyperparametern

Mehrere Hyperparameter werden durch Rastersuche optimiert.

- Hier wird der Hyperparameterraum entlang eines regelmäßigen Rasters
- meist in einer linearen oder logarithmischen Skala exploriert.
- Dies kann auch rekursiv wiederholt werden (Binärsuche).

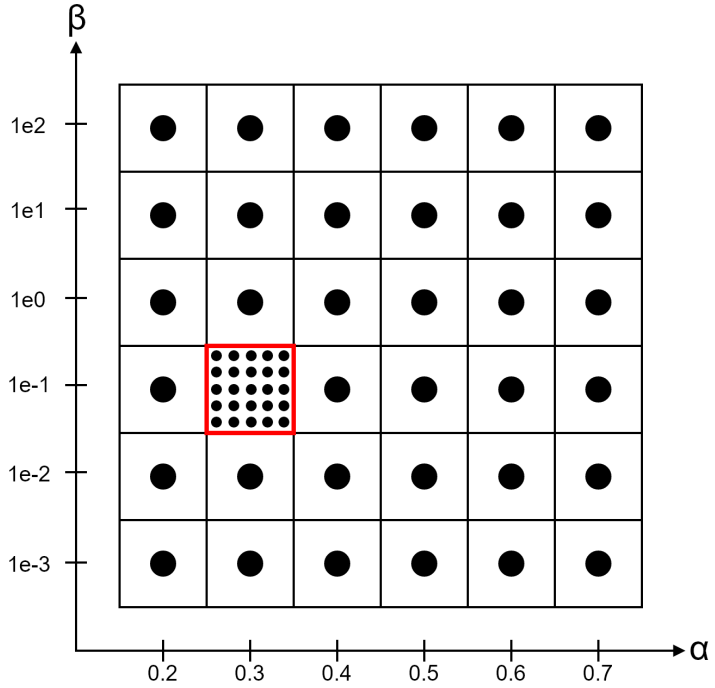


Abbildung 15: Muster der Rastersuche.

Lineare Regression

Optimierung von Hyperparametern

Eine weitere Möglichkeit ist es

- den Hyperparameter entlang eines zufälligen Rasters
- mit in einer uniformen oder logarithmischen Verteilung zu explorieren.
- Dies kann ebenso rekursiv wiederholt werden.

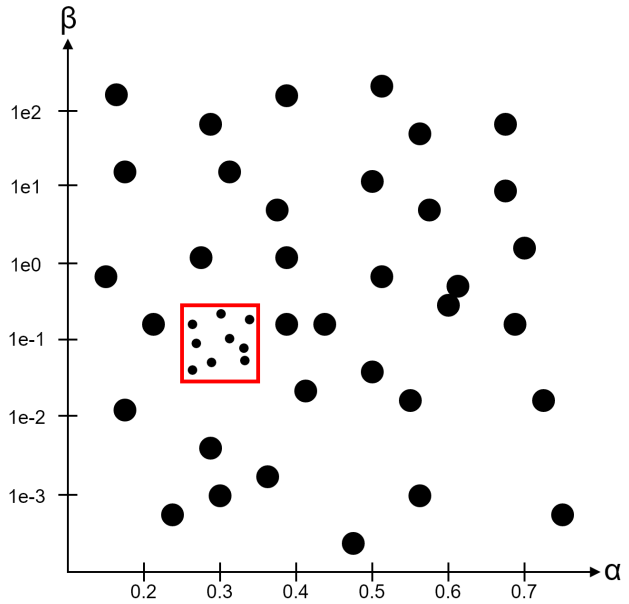
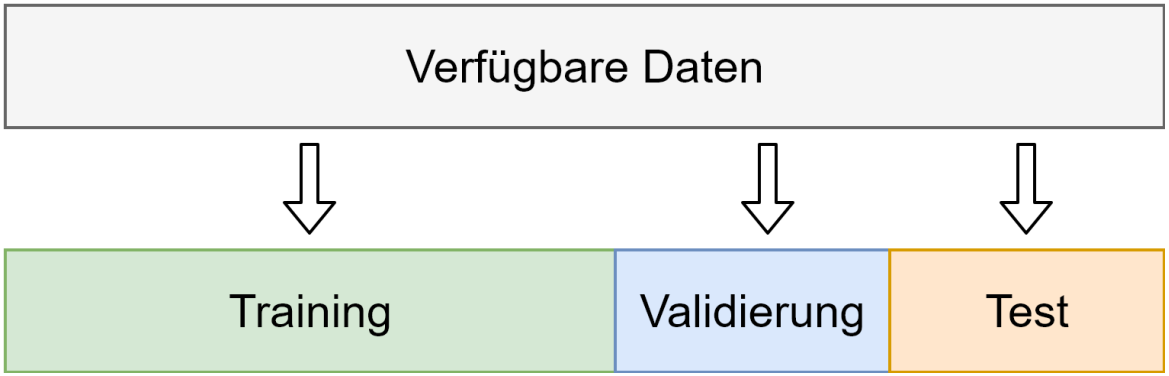


Abbildung 16: Zufälliges Suchraster.

Lineare Regression

Optimierung von Hyperparametern

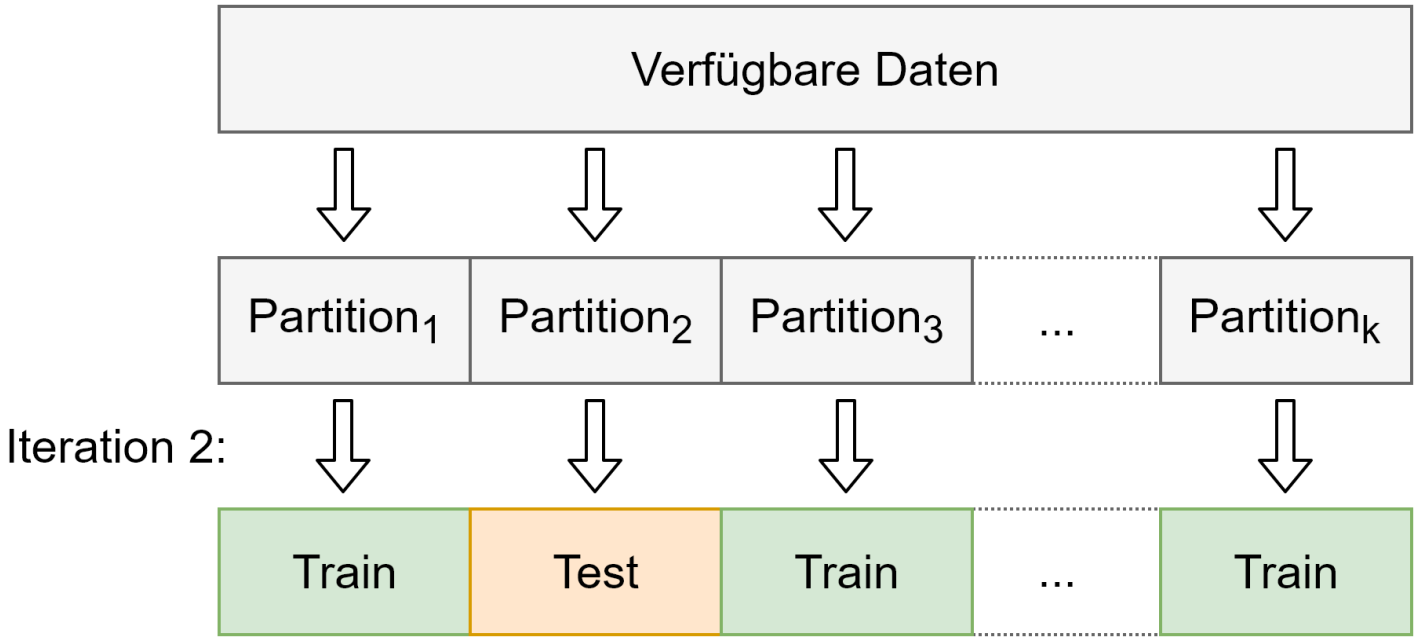
Müssen für ein Modell auch dessen Hyperparameter optimiert werden, so wird für diesen Prozess die verfügbare Datenmenge weiter unterteilt. Die Modellparameters werden durch Trainingsdaten bestimmt. Die Hyperparameter werden auf den Validierungsdaten getestet und die engültige Modellperformance wird auf den Testdaten bestimmt.



Lineare Regression

Optimierung von Hyperparametern

Um eine verlässlichere Schätzung für die Güte eines Modells zu bekommen, kann die Kreuzvalidierung verwendet werden. Hier wird der gesamte Datensatz in k Partitionen zerteilt. Das Training findet nun in k Iterationen statt. In Iteration i wird auf Partition i getestet. Der Rest wird für das Training verwendet. Die Leistungsmetrik wird schließlich über die k Iterationen gemittelt.



Lineare Regression

Ridge Regression

Wir können auch direkt versuchen eine Überanpassung zu verhindern, indem wir exzessive Werte für die Parameter $\mathbf{w}$ *bestrafen*. Dieses Verfahren nennt sich Ridge Regression und beruht auf der angepassten Fehlerfunktion

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - f_{\mathbf{w}}(\mathbf{x}^{(i)}) \right)^2 + \alpha ||\mathbf{w}||^2.$$

Wir bekommen durch den Hyperparameter $\alpha \in \mathbb{R}_{\geq 0}$ einen weiteren Freiheitsgrad durch welchen wir intuitiv den tatsächlichen Polynomgrad stufenlos einstellen können.

Lineare Regression

Ridge Regression

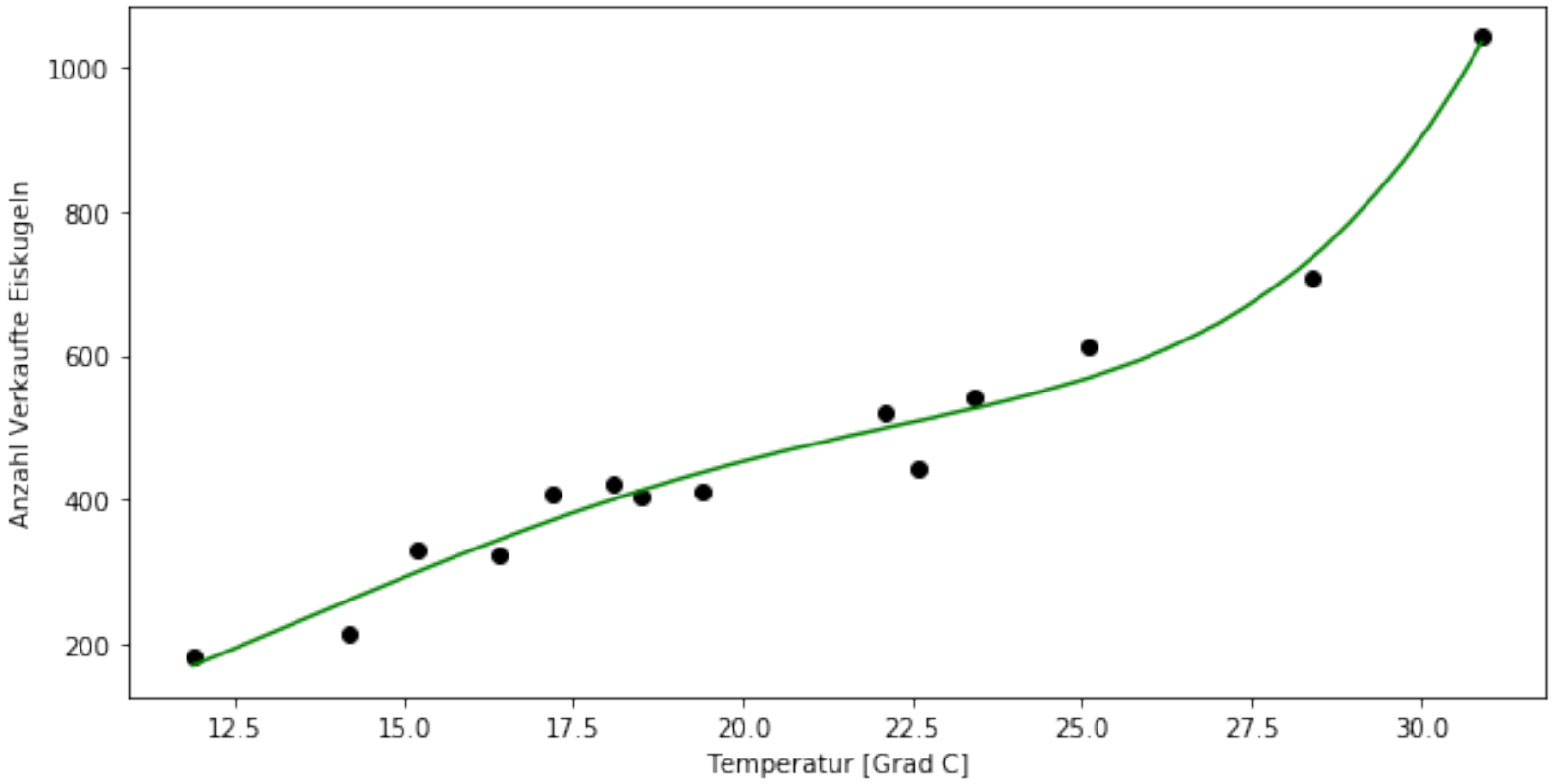


Abbildung 17: Ridge Regression eines Polynoms 5-ten Grades an die Eisverkaufdaten durch Basiserweiterung, $\alpha = 10$.

Lineare Regression

Ridge Regression

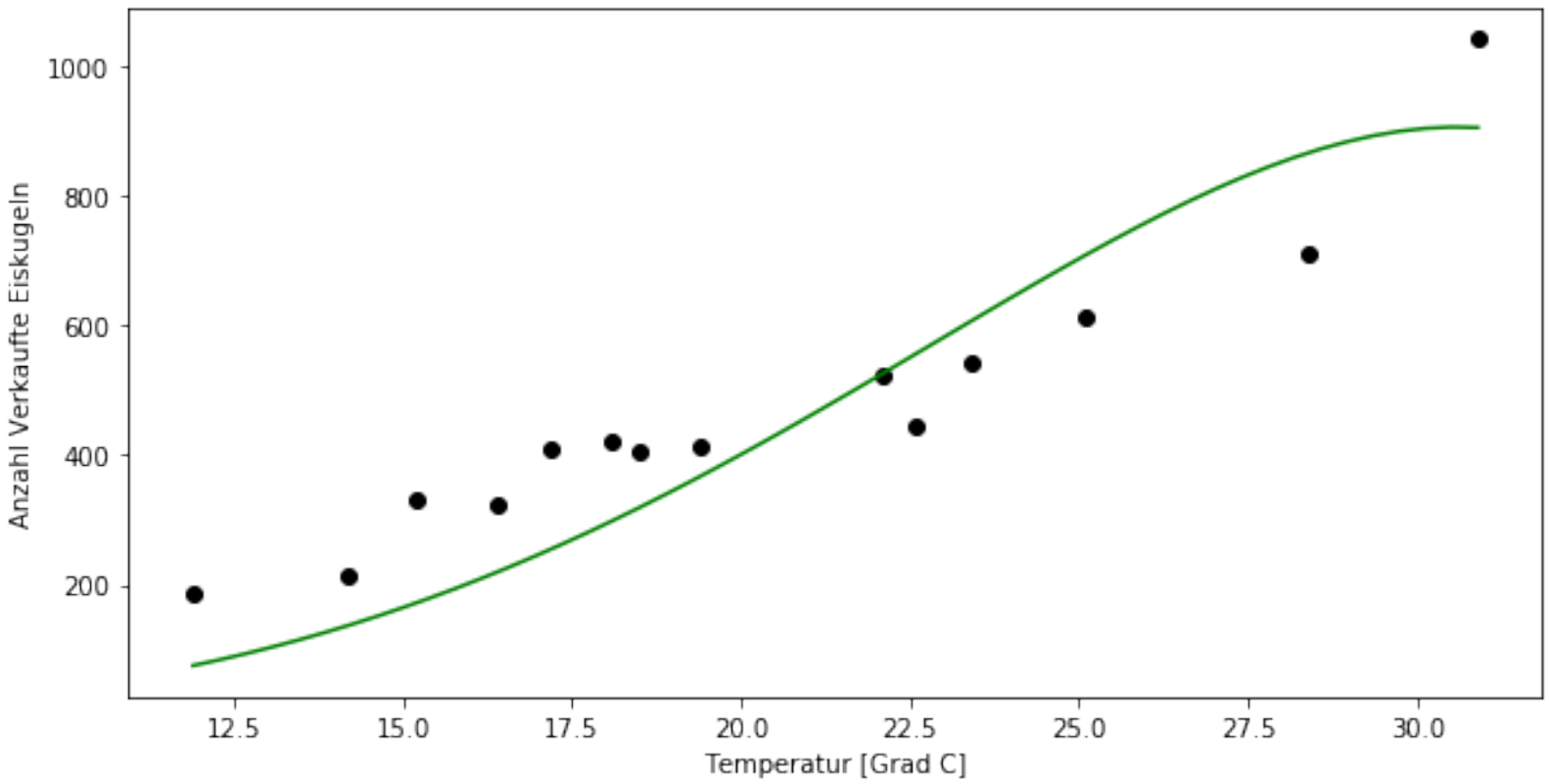


Abbildung 18: Ridge Regression eines Polynoms 5-ten Grades an die Eisverkaufdaten durch Basiserweiterung, $\alpha = 10^{10}$.

Lineare Regression
Ridge Regression

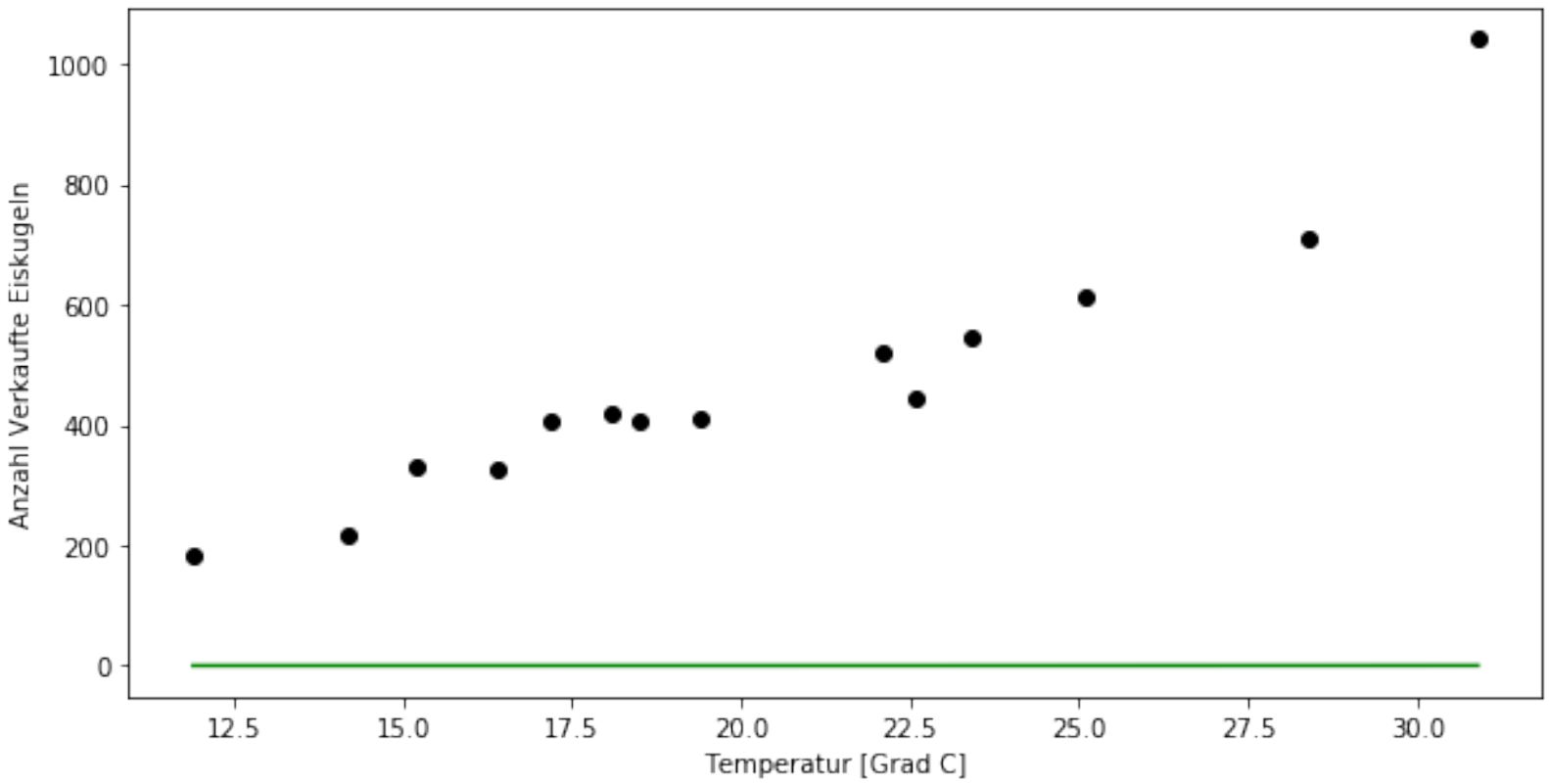


Abbildung 19: Ridge Regression eines Polynoms 5-ten Grades an die Eisverkaufdaten durch Basiserweiterung, $\alpha = 10^{20}$.

Lineare Regression

Ridge Regression

Die Bedeutung des Hyperparameters lässt sich somit wie folgt intuitiv charakterisieren:

- $\alpha = 0$: klassische Regression
- $\alpha > 0$: normaler Wirkungsbereich, mit wachsendem α werden die Parameter $\mathbf{w}$ immer weiter eingeschränkt und der effektive Polynomgrad sinkt
- $\lim \alpha \to \infty$: $f(\mathbf{x}) = 0$, da die Parameter $\lim \mathbf{w} \to \mathbf{0}$