Woche 06 Daten: Unter- und Überanpassung

Skript

Erarbeitet von
Ann-Kathrin Selker

Lernziele

- Unter- und Überanpassung mit Beispielen erklären können
- Unter- und Überanpassung mithilfe einer grafischen Darstellung identifizieren können
- Das Prinzip von Ausreißern mit Beispielen erläutern können

Inhalt

Einstieg

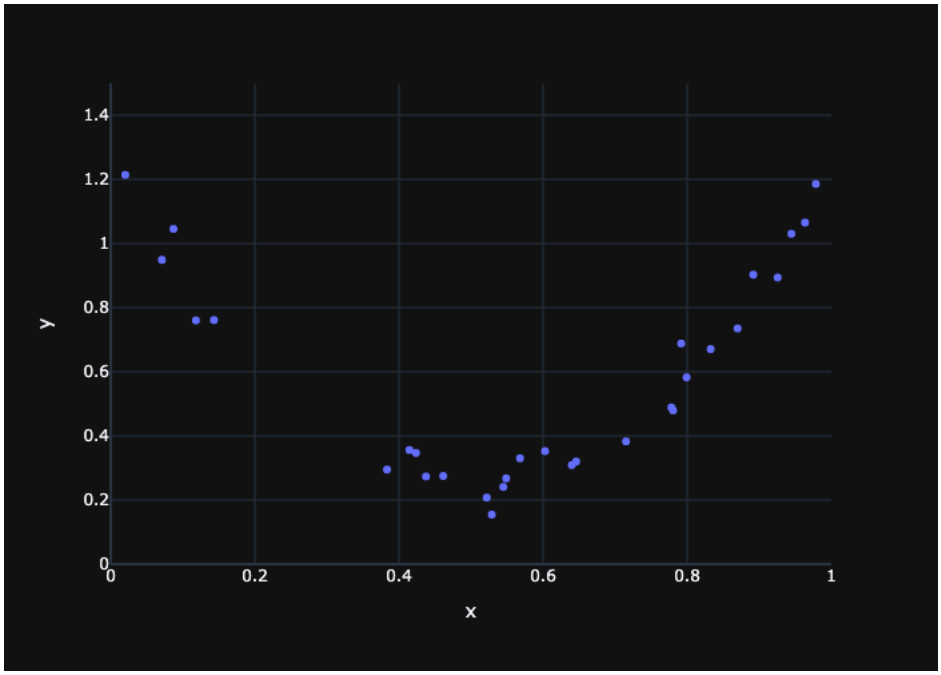
Um ein Maschine-Learning-Modell zu trainieren, „füttern“ wir es mit Trainingsdaten, also Daten, bei denen auch bereits die Ausgabe vorhanden ist. Mithilfe dieser Zuordnung von Eingabe zu gewünschter Ausgabe lernt das Modell einen Zusammenhang zwischen beiden.

Projekt „KI für alle“
HeiCAD – Heine Center for Artificial Intelligence and Data Science

Wenn das Modell dann neue Daten erhält, also neue Eingaben ohne die dazugehörigen Ausgaben, kann es aufgrund des Trainings dann Vorhersagen über die richtigen Ausgaben treffen. Doch damit dies gut funktioniert, müssen wir vorher Annahmen über die Komplexität des Zusammenhangs zwischen Ein- und Ausgabe stellen. Welche Auswirkungen kann es haben, wenn diese gestellten Annahmen falsch sind? Und was passiert, wenn wir mit zu wenigen, falsch zusammengesetzten oder fehlerhaften Trainingsdaten arbeiten?

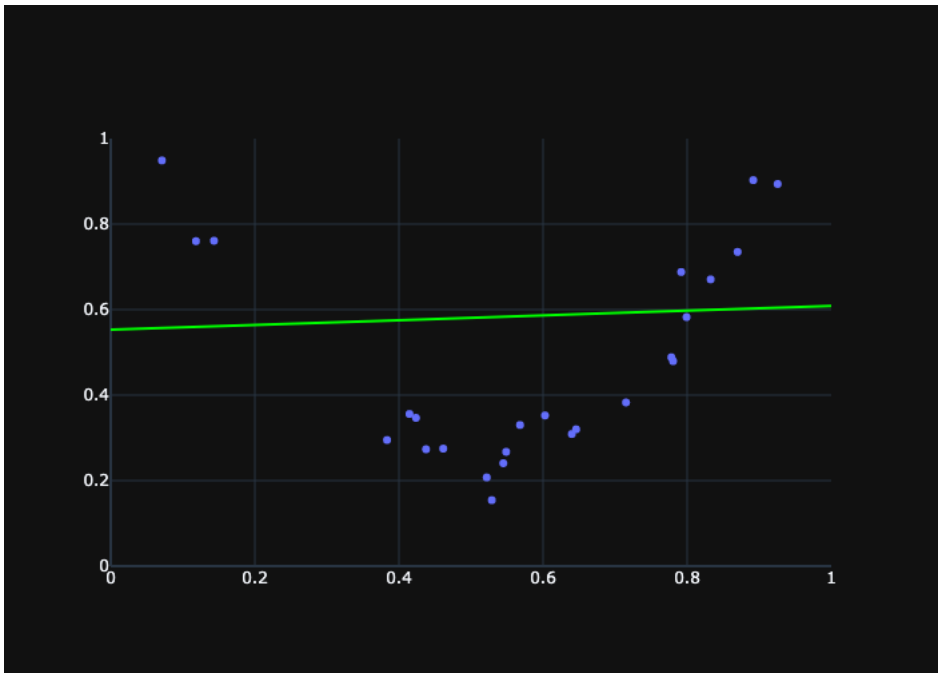
Unteranpassung

Betrachten wir zuerst den Fall, dass wir die Komplexität des Zusammenhangs zwischen Ein- und Ausgabe unterschätzt haben. Angenommen, wir haben die folgenden Trainingsdaten:



Einblendung Trainingsdaten

Auf der x-Achse ist unser einziges vorhandenes Feature aufgetragen, die y-Achse gibt die Zielgröße an, also die gewünschte Ausgabe. Da es sich um Trainingsdaten handelt, kennen wir die Zielgröße für unsere Daten bereits. Man sieht sofort, dass es hier wohl keinen linearen Zusammenhang zwischen Feature und Zielgröße gibt, lass uns aber trotzdem diese Annahme machen. Als Verfahren benutzen wir die lineare Regression. Unser Modell berechnet anhand der Trainingsdaten die folgende Gerade als Vorhersage.



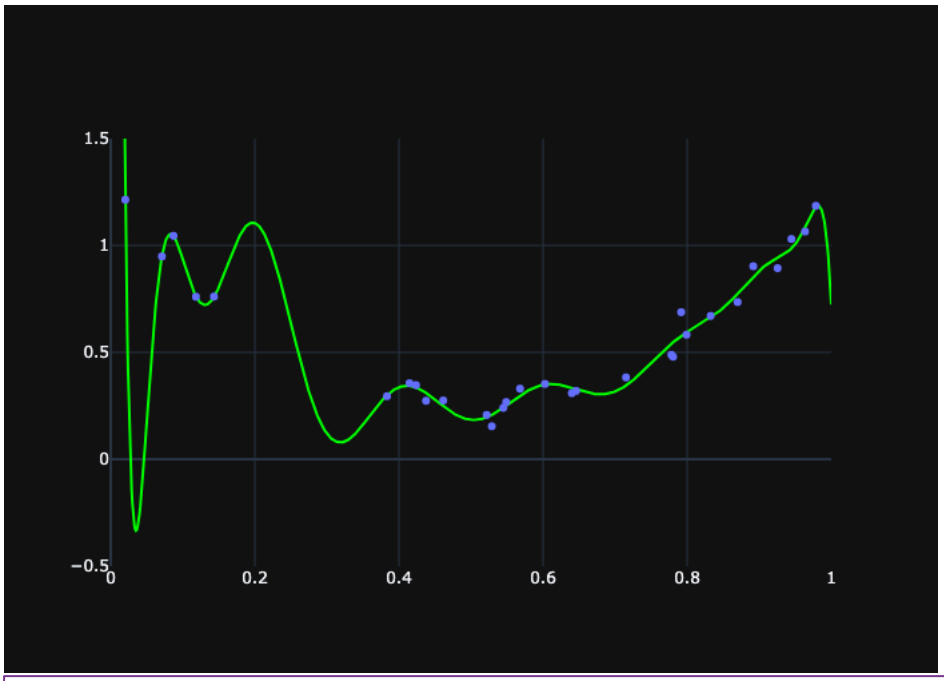
Einblendung berechnete Gerade

Man sieht gleich, dass die Summe der Fehler, also die Summe der Abstände der Datenpunkte zur Gerade, hier sehr hoch ist. Eine Gerade war für die Vorhersage der Zielgröße also keine gute Wahl.
Das Prinzip der unterschätzten Komplexität eines Modells nennen wir Unteranpassung, im englischen Underfitting. Unser Beispiel zeigt auch gleich, wie du in der Praxis erkennst, dass es bei deinem Modell zu einer Unteranpassung gekommen ist. Bei einer Unteranpassung ist der Fehler hoch, sowohl bei den Testdaten als auch schon bei den Trainingsdaten. Zur Erinnerung: Der Fehler sollte bei den Trainingsdaten klein sein, da das Machine-Learning-Modell ja anhand dieser Daten gelernt hat. Bei einer Unteranpassung hilft es nicht, das Modell mit mehr Daten zu trainieren, da das gewählte Modell das Problem darstellt. Auch die Regression an sich ist nicht das Problem und Unteranpassung tritt auch bei anderen Machine-Learning-Verfahren auf.

Überanpassung

Als Nächstes betrachten wir den Fall, dass wir die Komplexität des Zusammenhangs überschätzt haben. Hierfür benutzen wir als Beispiel wieder eine Regression und dieselben Trainingsdaten. Da wir - hoffentlich - aus dem letzten Beispiel gelernt haben, setzen wir die Komplexität dieses Mal höher an und wählen als Verfahren eine Regression, die uns eine Kurve zurückgibt. Die Komplexität einer Kurve bestimmt sich dabei aus der Anzahl der Richtungswechsel, die möglich sind. Bei einer Geraden gibt es keine Richtungswechsel. Wir lassen jetzt im Beispiel bis zu 14 Richtungswechsel zu. Falls du schon einen mathematischen Hintergrund hast: Dies heißt, dass das Polynom, das die Kurve beschreibt, einen Grad von 15 hat.

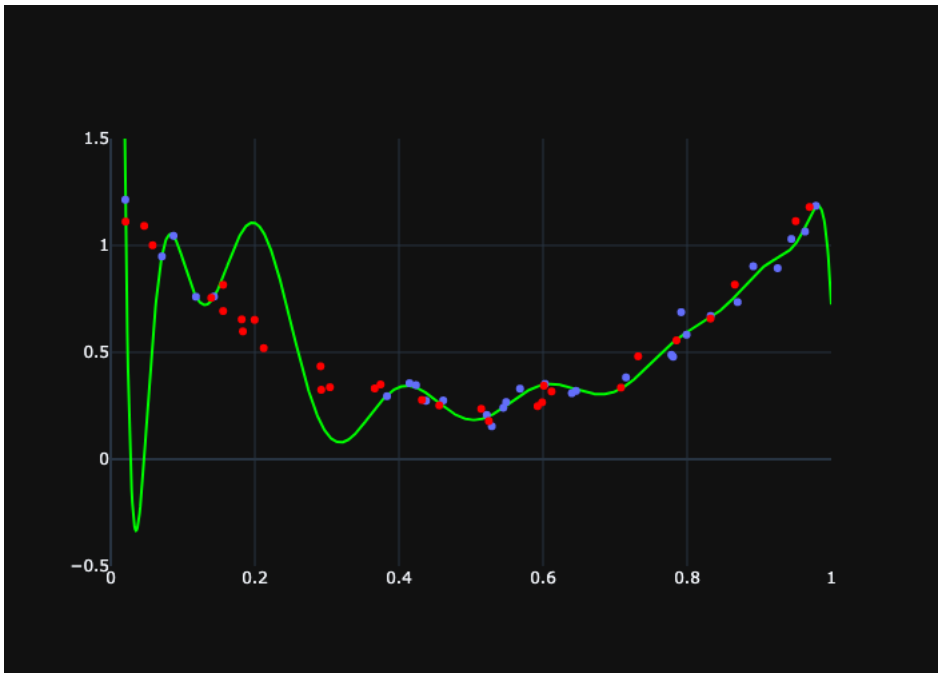
Wenn wir uns die Vorhersage des Modells angucken, dann sehen wir, dass der Fehler bei den Trainingsdaten sehr klein ist. Die Punkte liegen dicht an der Kurve, und viele, zum Beispiel ganz links, liegen sogar genau auf der Kurve und haben daher einen Fehler von 0.



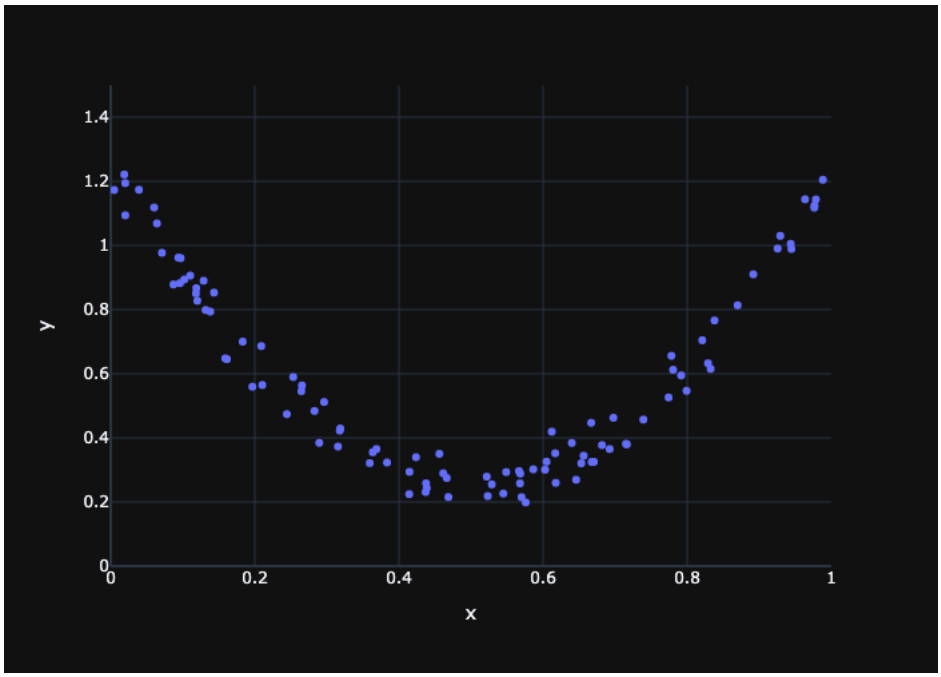
Einblendung berechnete Kurve

Wenn wir aber jetzt die Vorhersage mit den rot markierten Testdaten testen, fällt auf, dass die Kurve überhaupt nicht mehr gut passt.

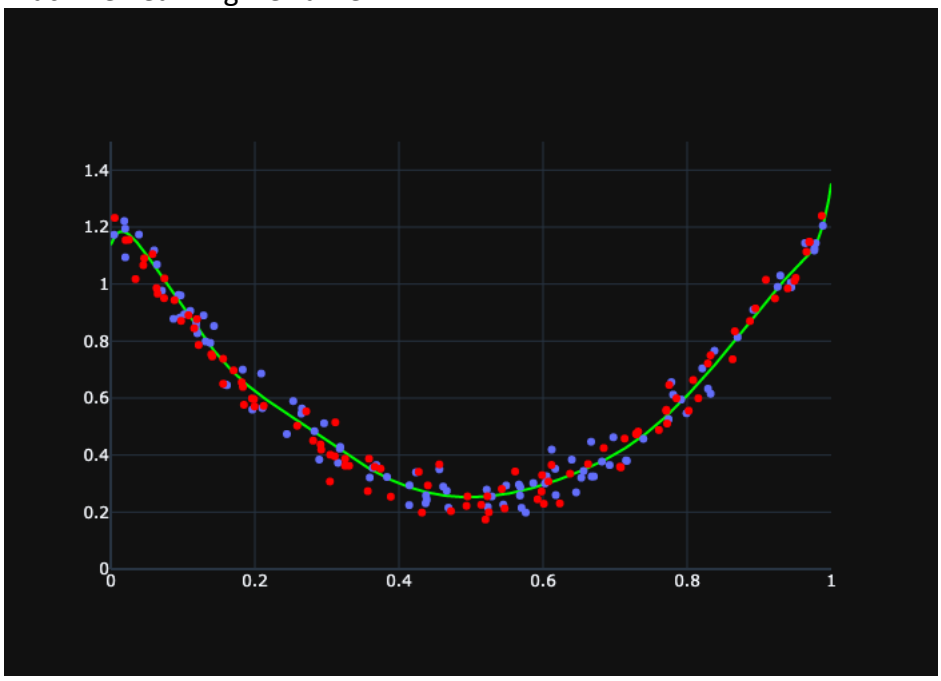
Einblendung berechnete Kurve inkl. Testdaten



Im eben gezeigten Beispiel ist es zu einer Überanpassung gekommen, im englischen Overfitting. Dies geschieht unter anderem, weil die Komplexität des Zusammenhangs überschätzt wurde. Kleinen Abweichungen in den Daten werden zu große Bedeutung beigemessen und in die Vorhersage mit eingefügt. Das Modell passt sich also zu stark an die Trainingsdaten an. Eine Überanpassung erkennst du daran, dass der Fehler bei den Trainingsdaten extrem klein ist, der Fehler bei den Testdaten aber sehr hoch. Das Problem der Überanpassung wird durch zu wenige Trainingsdaten noch verstärkt. Wenn nur wenige Daten vorhanden sind, wird jeder kleinen Abweichung von der Norm automatisch eine größere Bedeutung zugemessen. Daher hilft es, die Zahl der Trainingsdaten deutlich zu erhöhen, wenn du bei dir eine Überanpassung bemerkst.



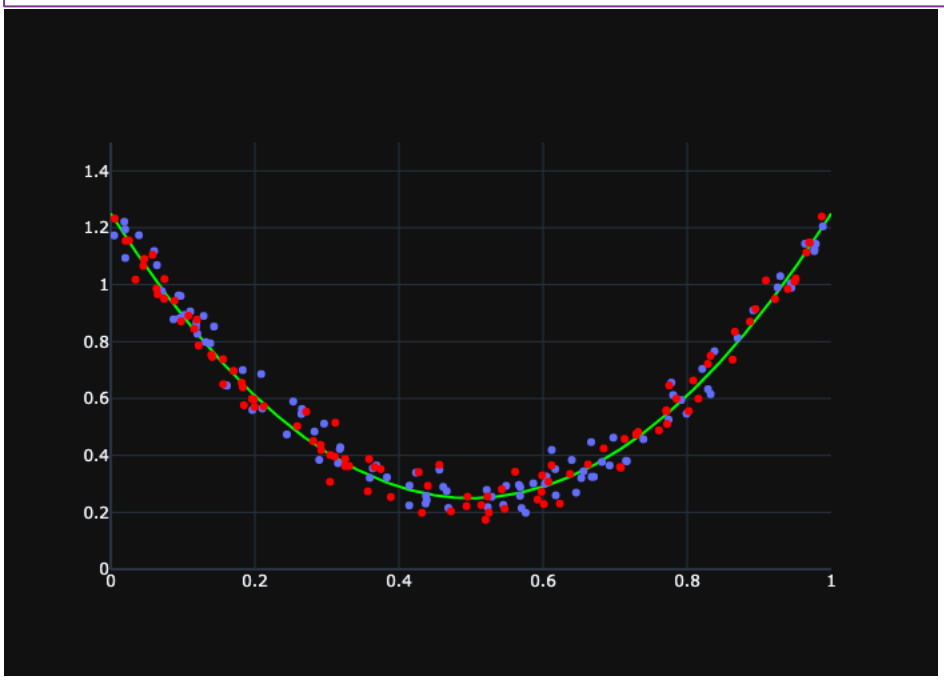
Wenn wir in unserem Beispiel anstelle von 30 Datenobjekten 100 Datenobjekte zum Training verwenden, berechnet unser Machine-Learning-Modell folgende Kurve als Vorhersage des Zusammenhangs zwischen Daten und Zielgröße. Wie du sehen kannst, hat der Fehler bei den roten Testdaten jetzt deutlich abgenommen.
Auch die Überanpassung tritt nicht nur bei Regression auf, sondern auch bei anderen Machine-Learning-Verfahren.



Einblendung Trainingsdaten + neue Kurve

Und wie sieht eine Vorhersage mit der richtigen Komplexität aus?

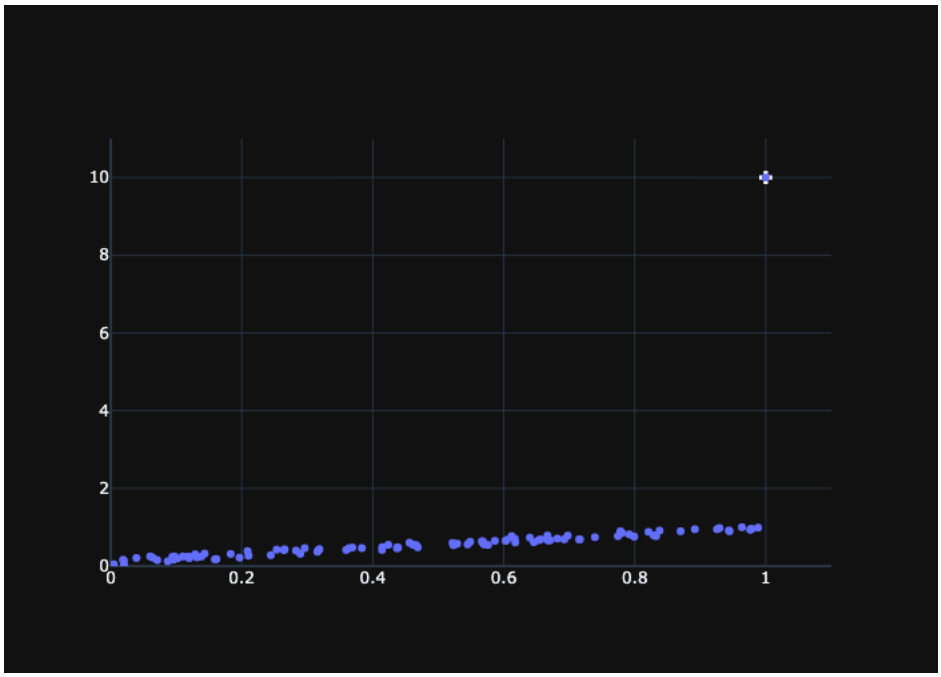
Einblendung optimale Kurve



So, mit einem Richtungswechsel. Sowohl der Abstand zu den blauen Trainingspunkten als auch der Abstand zu den roten Testpunkten ist klein. Das bedeutet, dass unser trainiertes Modell wahrscheinlich auch bei zukünftigen Eingaben gute Vorhersagen treffen wird.

Ausreißer

Als Nächstes beschäftigen wir uns mit Ausreißern, im Englischen Outlier. Unter Ausreißern versteht man Datenobjekte, bei denen sich der Wert eines Features oder einer Kombination von Features stark von dem typischen Wert bei anderen Datenobjekten unterscheidet. Ein Beispiel dafür kannst du in diesem Diagramm sehen. Der Punkt oben rechts ist weit von den anderen Werten für dieses feature entfernt.

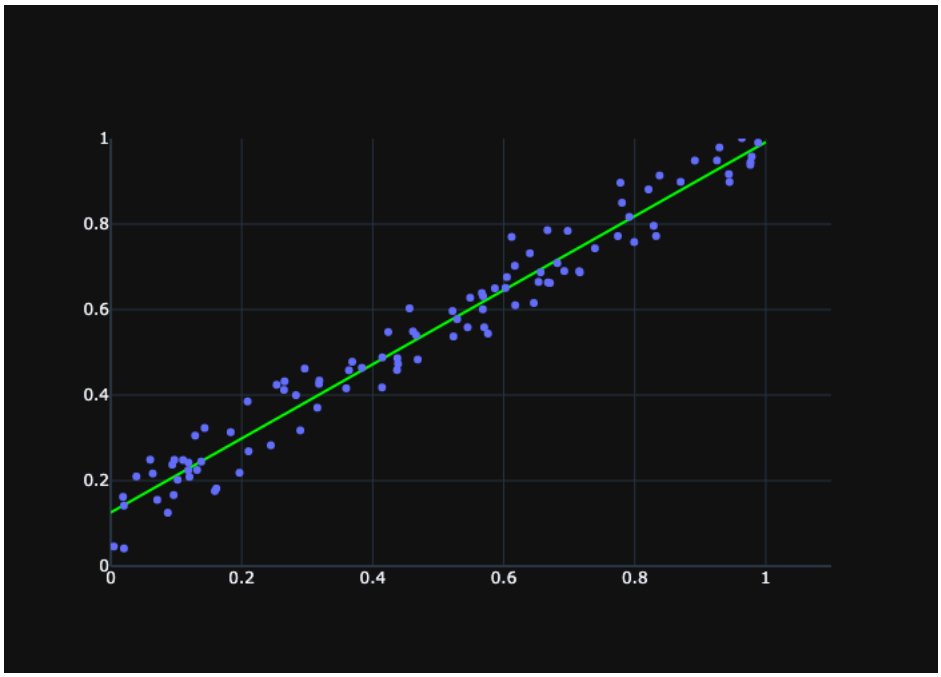


Einblendung Daten mit Ausreißer

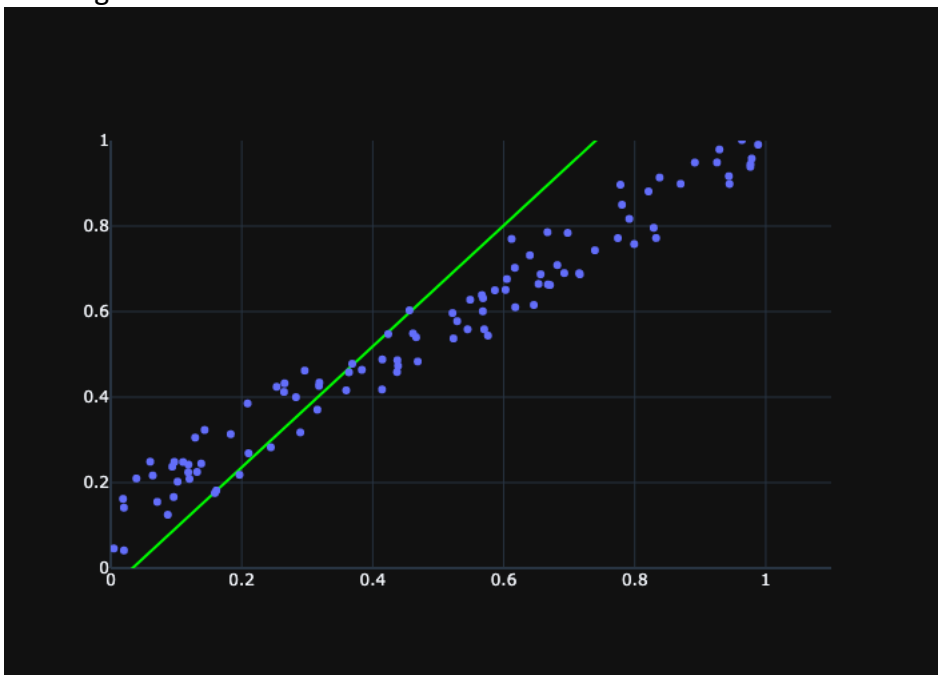
Bei verschiedenen Machine-Learning-Verfahren machen sich Ausreißer unterschiedlich stark bemerkbar. Nehmen wir als Beispiel die lineare Regression.

Einblendung Gerade mit und ohne Ausreißer

Die folgenden Diagramme zeigen den Unterschied der berechneten Geraden, wenn einmal der Ausreißer vor dem Trainieren entfernt wird,



gegenüber der Geraden, wenn mit dem Ausreißer trainiert wird. Die wenigen Trainingsdaten verstärken natürlich den Effekt des Ausreißers.



Ausreißer treten z. B. häufig auf, wenn Daten von Menschen gesammelt werden. Hier kommt es zu Tippfehlern - wie 101,1 anstelle von 10,11 - und Werten in falschen Spalten, wie einer 40 bei Größe anstelle von Alter. Auch technische Fehler führen zu Ausreißern, z. B. bei einem defekten Messgerät. Aber nicht jeder Ausreißer muss Folge eines Fehlers sein, sondern ist vielleicht ein seltener Einzelfall. So war Robert Wadlow tatsächlich 2,72 m groß, auch wenn diese Größe seitdem nicht mehr erreicht wurde.

Bei wenigen Featuren können extreme Ausreißer noch mithilfe von Datenvisualisierung von dir selber gefunden werden. Das funktioniert bei mehr als drei Featuren nicht mehr gut. Während der Datenvorverarbeitung können Ausreißer auch unter anderem mit statistischen Mitteln ermittelt werden. Es gibt aber auch einige Machine-Learning-Algorithmen für die Identifizierung von Ausreißern.

Die Behandlung von Ausreißern solltest du dir vorher gut überlegen. Bei großen Datenmengen zum Trainieren fallen sie kaum ins Gewicht, können aber trotzdem noch eine negative Auswirkung haben, wenn das trainierte Modell nur eine extrem kleine Fehlerquote haben darf. Wenn die Zusammensetzung der Trainingsdaten nicht gut gewählt wurde, können Datenobjekte auch als Ausreißer identifiziert werden, wenn sie eigentlich im echten Leben keine sind. Falls wir Krokodile auf Bildern erkennen wollen, ist ein einziges Bild von einem weißen Krokodil auch ein Ausreißer im echten Leben, da Albino-Krokodile extrem selten sind.

Einblendung Albinokrokodil (Quelle [1])

Falls wir Katzen auf Bildern erkennen wollen, ist ein einzelnes Bild von einer weißen Katze kein Ausreißer im echten Leben, sondern ein Zeichen dafür, dass unsere Trainingsdaten nicht divers genug für die gestellte Aufgabe sind.

Einblendung weiße Katze (Quelle [2])

Abschluss

In diesem Video hast du die Begriffe Über- und Unteranpassung kennengelernt und kannst sie jetzt anhand von Beispielen sowohl erklären als auch identifizieren. Außerdem hast du gelernt, was Ausreißer sind.

Quellen

Quelle [1] Bild von misterfarmer, pixabay.com

Quelle [2] Bild von Pexels, pixabay.com

Weiterführendes Material

https://www.kaggle.com/getting-started/157623
https://scikit-learn.org/stable/modules/outlier_detection.html
https://www.kaggle.com/code/nareshbhat/outlier-the-silent-killer

Disclaimer