Woche 07 Daten: Dimensionsreduktion

Skript

Erarbeitet von
Ann-Kathrin Selker

Lernziele

- exponentielles Wachstum veranschaulichen und mit linearem und quadratischem Wachstum vergleichen können
- Fluch der Dimensionalität erklären können
- kausale und zufällige Korrelation erklären

Inhalt

Einstieg

Daten sammeln ist häufig teuer und zeitaufwändig, erst recht, wenn man die Datensammlung wiederholen muss, weil man wichtige Features vergessen hat. Manchmal ist es sogar nicht möglich, ein gewisses Feature nachträglich zu sammeln, z. B. bei anonymen

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

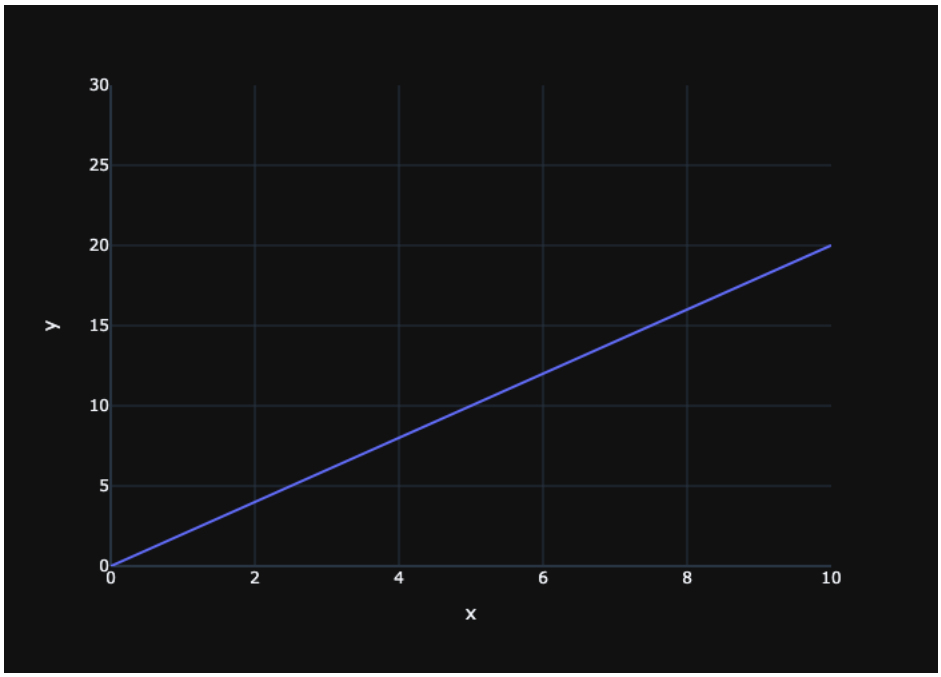
Umfragen. Daher wird in diesen Fällen häufig alles gesammelt, was man in die Finger bekommen kann. Oft ist auch vorher gar nicht klar, welche Features eigentlich eine wichtige Rolle beim späteren Modell spielen werden, sodass man dem Modell lieber mehr Features und damit auch mehr Information zur Verfügung stellen möchte. Doch was für Probleme treten auf, wenn unsere Daten sehr viele Features haben?

Fluch der Dimensionalität

Je nachdem, mit welcher Funktion das Machine-Learning-Modell arbeitet, werden alle Features als gleichwertig angesehen. Ein Beispiel dafür ist das Messen der Entfernung, was du schon bei der Regression kennengelernt hast. Hier wird der Distanz von jedem Feature die gleiche Bedeutung zugemessen, was dann zu einem schlechten Ergebnis führt, wenn irrelevante oder unwichtige Features vorhanden sind. Werden diese Features vorher nicht entfernt, verschlechtert sich also die Qualität des Modells stark.
Wenn wir uns ein Datenobjekt als Vektor vorstellen, ist jeder Eintrag des Vektors der Wert eines Features. Dementsprechend wächst auch die Dimension des Vektors mit der Anzahl an Features. Um den Zusammenhang zwischen den Features der Datenobjekte und der gewünschten Ausgaben zu berechnen, wird eine gewisse Minimalanzahl an Trainingsdaten benötigt. Diese Minimalanzahl ist abhängig von der Dimension des Vektors und wächst exponentiell. Insgesamt führt dies also dazu, dass bei dem Hinzufügen eines Features plötzlich exponentiell mehr Datenobjekte als zuvor benötigt werden, damit das Modell sinnvoll trainiert werden kann. Der Mathematiker Richard Bellman hat dafür den Begriff „Fluch der Dimensionalität“ eingeführt. Falls du dich für Programmierung interessierst, hast du von Bellman sicher schon als Erfinder der dynamischen Programmierung gehört.

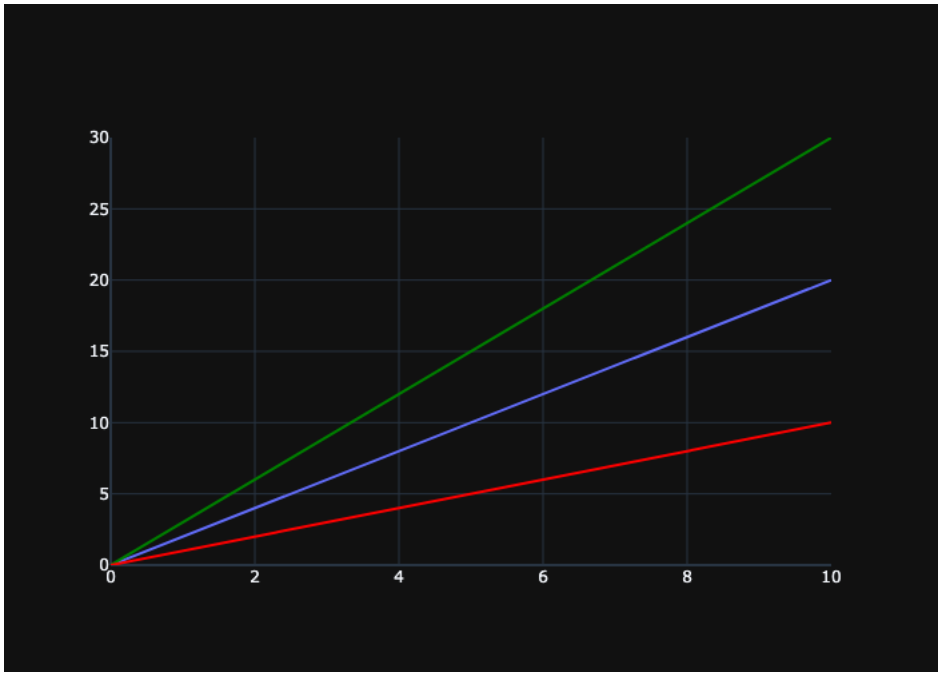
Exponentielles Wachstum

Doch was bedeutet exponentielles Wachstum überhaupt? Betrachten wir zuerst noch mal zur Wiederholung das lineare und quadratische Wachstum.
Es handelt sich um lineares Wachstum, wenn in jedem Schritt eine konstante Anzahl zu unserem vorherigen Ergebnis hinzugefügt wird. Ein Beispiel dafür ist das Gehalt. Angenommen, du verdienst 20 Euro pro Stunde. Nach 0 Stunden Arbeit hast du 0 Euro verdient, nach einer Stunde 20, nach zwei Stunden 40, nach drei Stunden 60 usw. Den genauen Verlauf des verdienten Gehalts kannst du in diesem Diagramm sehen. (Einheit 10€, d.h. ein Wert von 20 steht für 20*10€ = 200€)



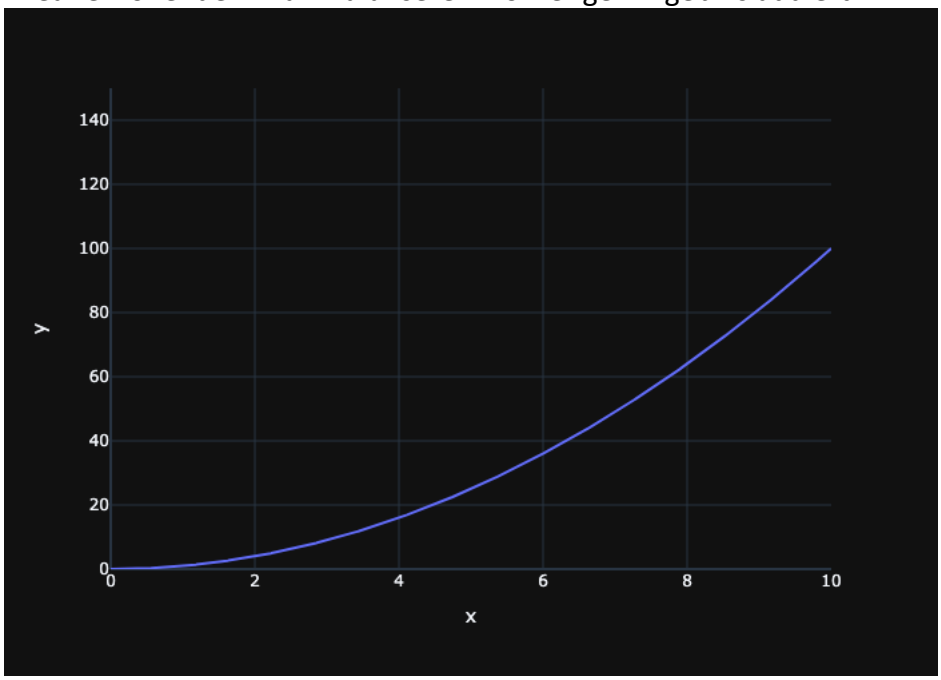
Einblendung Grafik Verlauf

Wenn wir das Gehalt pro Stunde auf 30 erhöhen, ergibt das die Gerade in Grün, und eine Reduzierung auf 10 Euro pro Stunde ergibt die rote Gerade.



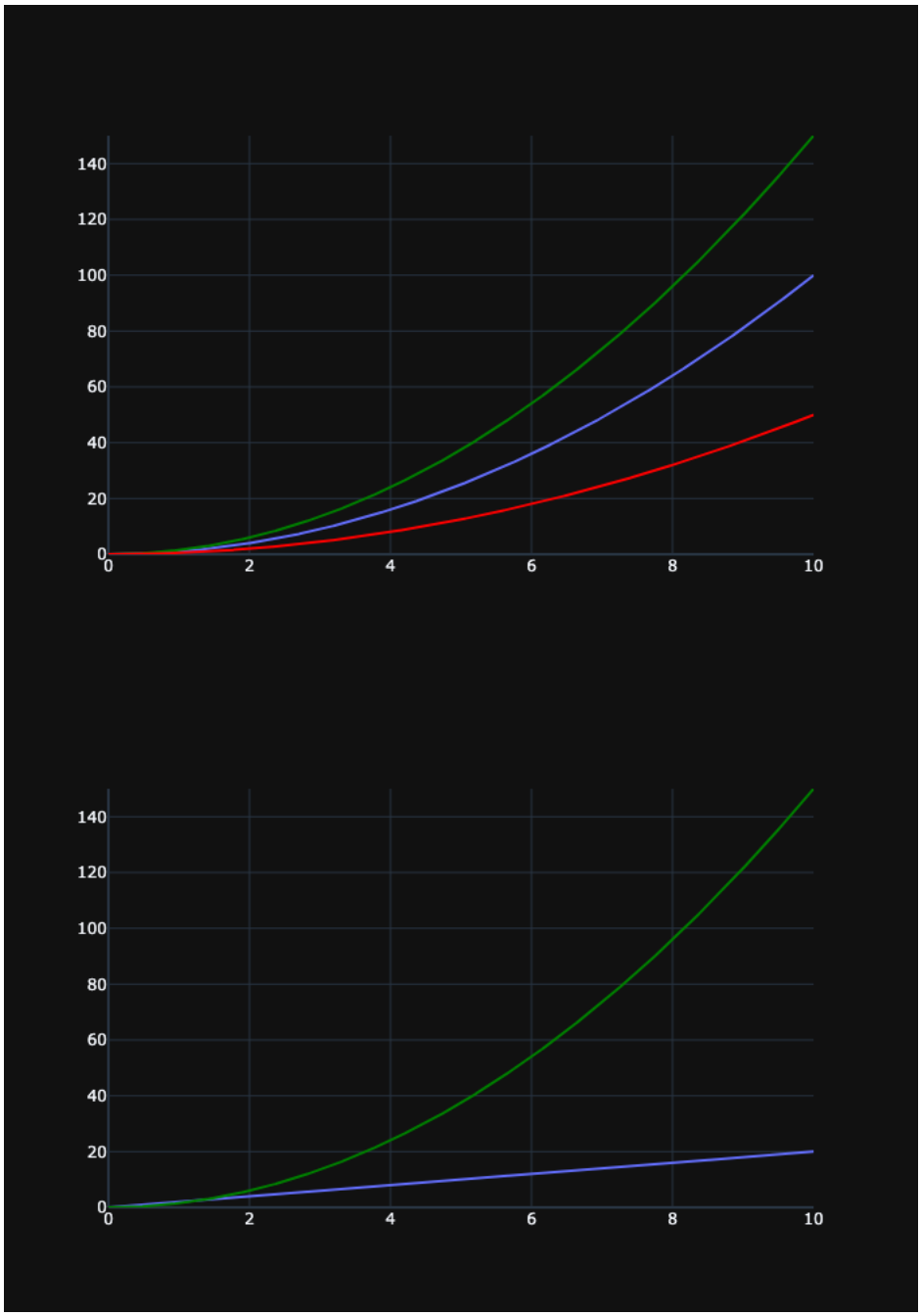
Ein Beispiel für quadratisches Wachstum ist die zurückgelegte Strecke beim Beschleunigen. Die genaue Formel zum Berechnen der Strecke kannst du in einem Physikbuch deines Vertrauens nachschlagen. Angenommen, du fährst mit deinem Fahrrad aus dem Stand an

und beschleunigst dabei mit 2 m/s^2. Dann hast du nach einer Sekunde einen Meter zurückgelegt, nach zwei Sekunden 4 m und nach drei Sekunden 9 m. Das liegt daran, dass sich deine Geschwindigkeit bei gleichbleibender Beschleunigung linear erhöht. Du wirst also immer schneller und legst damit pro Sekunde immer mehr Strecke auf einmal zurück. Es gilt also: Bei quadratischem Wachstum wird pro Schritt nicht eine konstante, sondern eine sich linear erhöhende Anzahl zu unserem vorherigen Ergebnis addiert.



Einblendung Grafik Verlauf

Zum Vergleich siehst du in diesem Diagramm noch, wie sich der Verlauf bei einer Beschleunigung von 3 m/s^2 (grün) und einem Meter pro s^2 (rot) ändert. Betrachten wir sowohl lineares als auch quadratisches Wachstum im selben Diagramm, wird deutlich, dass unsere Ausgabe bei quadratischem Wachstum viel schneller wächst als bei linearem Wachstum.

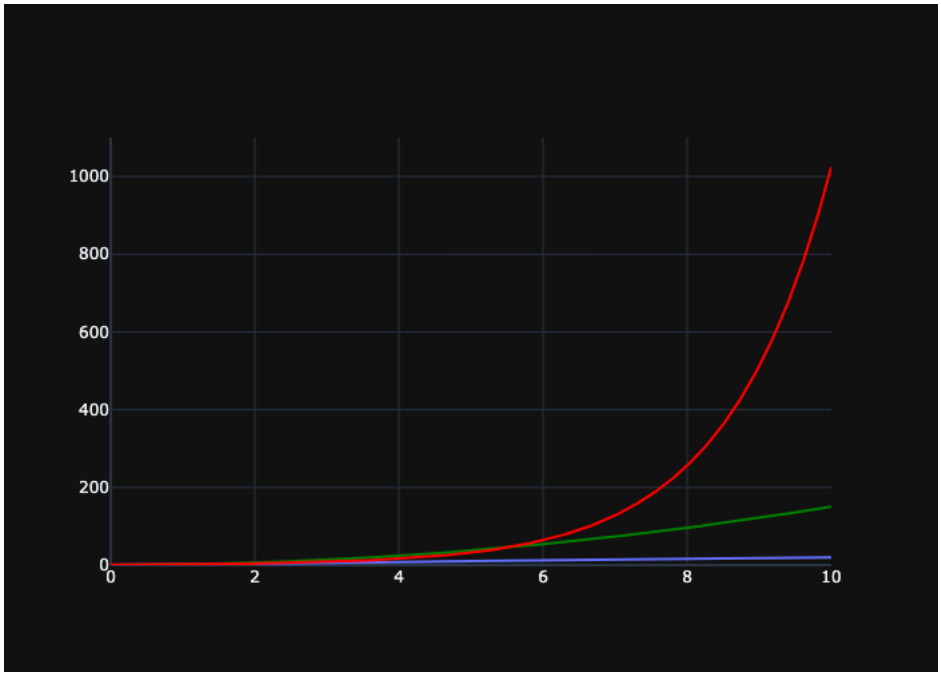


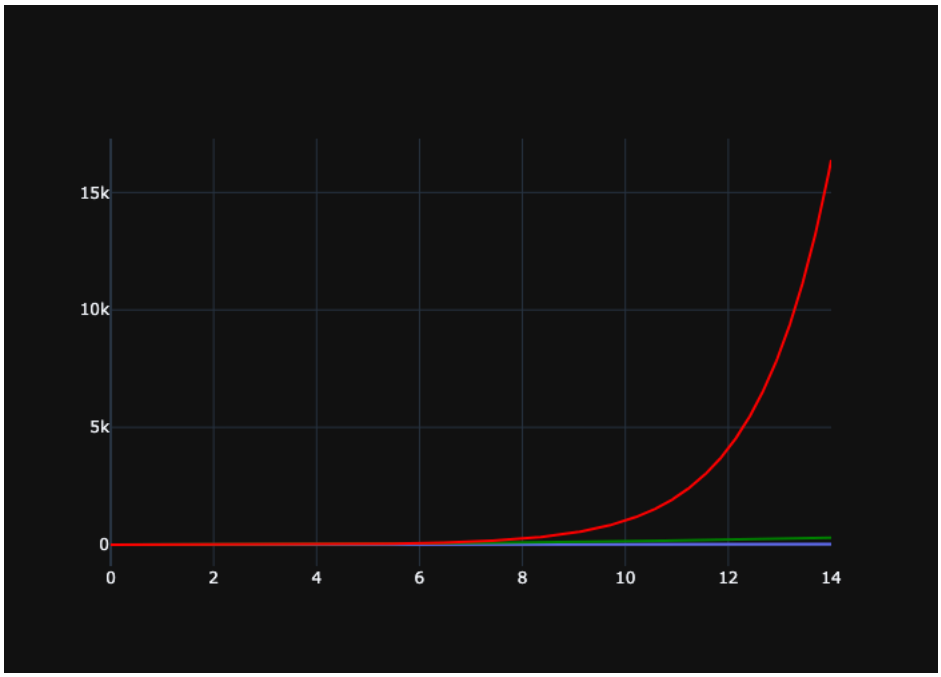
Beim exponentiellen Wachstum verläuft das Wachstum jetzt folgendermaßen:
Nehmen wir zum Beispiel die Vermehrung einer Bakterie. Der Einfachheit halber nehmen wir an, dass wir mit einer Bakterie starten und dass jede Bakterie jeden Schritt eine weitere Bakterie erzeugt. Am Anfang, bei Vermehrungsschritt 0, haben wir also nur eine Bakterie. Diese vermehrt sich aber, sodass wir nach einem Vermehrungsschritt zwei Bakterien haben. Beide vermehren sich, wir sind also nach zwei Schritten bei 4 Bakterien und nach drei Schritten bei 8 angekommen. Das hört sich erstmal noch nicht schlimm an. Doch wenn wir so weitermachen, sind wir nach zehn Schritten bei über tausend Bakterien, nach 20 Schritten bei über einer Million, und nach 30 Schritten bereits bei über einer Milliarde Bakterien angekommen. Ich wiederhole noch einmal: Nach 30 Vermehrungsschritten haben wir bereits eine Milliarde Bakterien erreicht! Beim exponentiellen Wachstum wird also das

aktuelle Ergebnis bei jedem Schritt mit einem konstanten Faktor multipliziert, während wir beim quadratischen Wachstum addieren und nur die Erhöhung mit einem konstanten Faktor multiplizieren.

Einblendung Grafik Verlauf

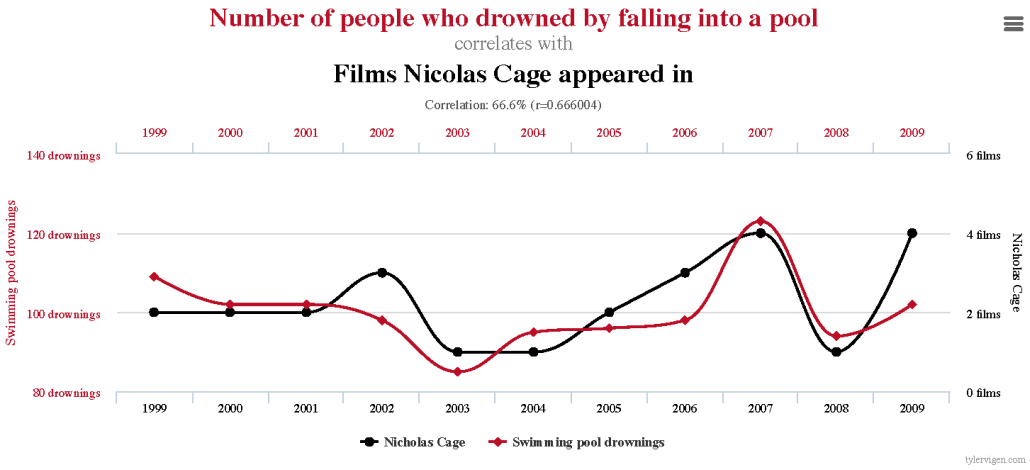
Um den Vergleich noch deutlicher zu sehen, gucken wir uns mal alle drei vorgestellten Wachstumsarten in einem Diagramm an. Nach 10 Schritten sieht man deutlich, wie viel schneller die exponentielle rote Kurve wächst. Nach 14 Schritten kann man bereits die lineare blaue und die quadratische grüne Kurve kaum noch auseinanderhalten, während die exponentielle Kurve rasant wächst.





Korrelation

Wir haben gesehen, dass bei steigender Feature-Anzahl nicht nur die Qualität und Leistung unserer Machine-Learning-Modelle nachlässt, sondern wir zum Trainieren auch exponentiell viele Trainingsdaten abhängig von der Anzahl der Features benötigen und dies schon bei kleiner Feature-Anzahl eine riesige Menge an Daten ausmacht. Um die Anzahl der Features zu verringern, können redundante Features zusammengefasst oder entfernt werden. Dies geschieht als Teil der Vorverarbeitung der Daten. Manche Features sind redundant, weil sie mit einem oder mehreren anderen Features zusammenhängen. Ein Beispiel ist die Angabe von sowohl Geburtsdatum als auch Alter einer Person, bei der eins von beiden Features gestrichen werden kann. Solche zusammenhängenden, also korrelierenden Features können von einem Menschen z. B. durch Visualisierung gefunden werden. Geburtsdatum und Alter hängen auch kausal zusammen, d. h. das Geburtsdatum einer Person bedingt auch ihr Alter. Ein weiteres, nicht so direktes Beispiel hierfür ist der Zusammenhang zwischen Zigarettenkonsum und Lungenkrebs. Allerdings müssen nicht alle Features, die korrelieren, auch kausal zusammenhängen. Ein Beispiel dafür ist das folgende Diagramm von Tyler Vigen, der auf seiner Webseite lustige Korrelationen vorstellt.



Einblendung Grafik (Quelle [1])

Die Anzahl der Filme, in denen Nicolas Cage mitspielt, korreliert mit der Anzahl der Menschen, die in den USA ertrunken sind, weil sie in einen Pool gefallen sind. Dies erkennst du an den ähnlich verlaufenden Kurven. Wir wollen Nicolas Cage aber natürlich keinen kausalen Zusammenhang unterstellen. Es handelt sich also um einen zufälligen Zusammenhang.
Wie genau die Anzahl der Features reduziert werden können, ohne zu viel Verlust bei der Datenqualität zu erleiden, ist ein eigener Forschungsbereich und mathematisch sehr involviert. In Python kann Dimensionsreduktion z. B. mit Machine-Learning-Methoden mit dem Modul Scikit-learn durchgeführt werden.

Abschluss

In diesem Video hast du gelernt, was exponentielles Wachstum ist, und kannst es mit linearem und quadratischem Wachstum vergleichen. Außerdem kannst du jetzt den Fluch der Dimensionalität sowie kausalen und zufälligen Zusammenhang erklären.

Quellen

Quelle [1] Tyler Vigen, https://tylervigen.com/spurious-correlations

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

Weiterführendes Material

Video "Fluch der Dimensionalität Zwei – Georgia Tech – Maschinelles Lernen"
https://studyflix.de/mathematik/exponentielles-wachstum-2029
https://studyflix.de/statistik/korrelation-und-kausalitat-2216
https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

Disclaimer