Handbuch: 7.2.5. Variablenreduktion

Viele der Daten, die in einer typischen Datenhistorie einer Industrieanlage abgespeichert sind, werden nicht benötigt, um zuverlässige Aussagen über den Zustand der Maschinerie oder über die Optimierung des Verarbeitungsprozesses zu machen. Gewöhnlich sind es weniger als 10% aller Messungen, die tatsächlich gebraucht werden. Und unter diesen Daten gibt es im Allgemeinen noch erhebliche Doppelungen an Informationen.

Jedesmal, wenn wir ein System modellieren wollen, ist es unsere erste Aufgabe, die notwendigen Daten auf intelligente Weise auszuwählen. Dabei besteht allerdings die Gefahr, dass man der Analyse keine ausreichenden Daten zur Verfügung stellt, weshalb wir eher vorsichtig sein müssen und auch solche Messungen einbeziehen sollten, von denen wir nicht sicher sind, ob sie wirklich einen informativen Beitrag leisten können oder nicht. Um zu vermeiden, die Analyse mit unnötigen Daten zu überlasten, müssen wir festlegen, welche Messungen wirklich wichtig sind und welche nicht.

Hinzu kommt, dass manche Messungen mit anderen zusammenhängen, so dass sie zwar Informationen liefern, aber doch mit vergleichsweise wenig Mehrwert. Nehmen Sie beispielsweise Ihr Körpergewicht, Ihre Körpergröße, Ihr Geschlecht und das Ausmaß Ihrer Körperbetätigungen. Allein auf der Basis Ihrer Körpergröße, Ihres Geschlechts und Ihrer Körperbetätigung kann man Ihr Gewicht zwar nicht sicher wissen, aber dazu doch eine vernünftige Vorhersage machen. Wenn wir dann doch noch Ihr Gewicht tatsächlich messen, trägt das zwar zur Gesamtinformation bei, aber doch keineswegs so viel wie die ersten drei Werte, da Ihr Gewicht ja eng mit diesen zusammenhängt. Je nachdem, was man über eine Person herauszufinden wünscht, kann man das mit Hilfe der drei ursprünglichen Werte tun, ohne auf die zusätzliche Information (in diesem Fall: das gemessene Gewicht) unbedingt angewiesen zu sein.

Auf der Grundlage dieser Erkenntnisse können wir zwei unterschiedliche Ideen verfolgen, um die Zahl der Variablen eines Modells zu minimieren. Wir könnten solche Variablen identifizieren, die überhaupt nicht benötigt werden und sie völlig vom Modell ausschließen. Oder wir verwandeln den Datensatz in ein anderes Koordinatensystem mit weniger Dimensionen, bei dem jede neue Dimension allerdings als Kombination der ursprünglichen Dimensionen gedacht wird. Diese zweite Idee basiert auf der soeben beispielhaft erläuterten Erkenntnis, dass manche Messungen zwar zusätzliche Informationen liefern und deshalb nicht gänzlich ausgeschlossen werden sollen, sie aber keine zusätzliche Dimension rechtfertigen.

Diese Idee führt zu der Hauptkomponentenanalyse, die man am besten visuell durch die nachfolgende Grafik erklärt:

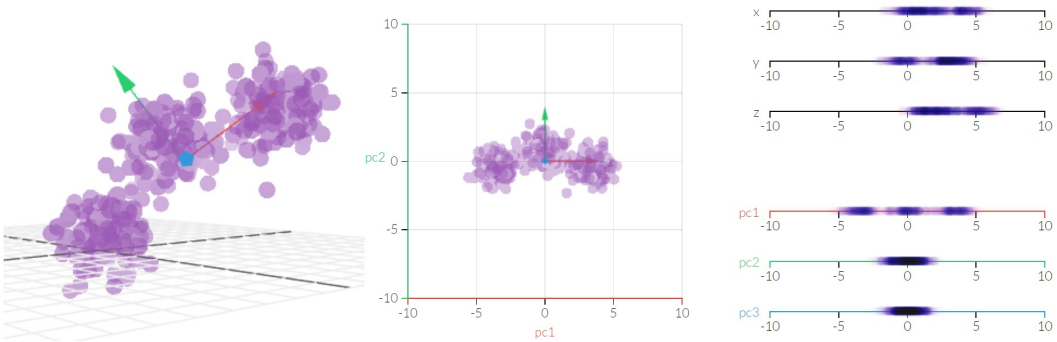

Abb. Links sehen wir den ursprünglichen dreidimensionalen Datensatz, dargestellt im ursprünglichen Koordinatensystem. Es war festzustellen, dass die größten Datenvariationen sich entlang des roten Pfeils im linken Schaubild zeigten. Die zweit- und dritthäufigsten Variationen befinden sich entlang des grünen und des (kaum sichtbaren) blauen Pfeils. Beim mittleren Bild wurden die Daten in einen zweidimensionalen Raum umgewandelt, bei dem der rote und der grüne Pfeil als die neuen Koordinaten angenommen wurden. Wir können leicht erkennen, dass die blaue Richtung nicht genug Informationen beiträgt, um die drei sichtbaren Datengruppen zu trennen, so dass wir diese Dimension ruhig ignorieren können. Das Bild rechts zeigt die Daten entlang der jeweils ursprünglichen, aber nun gedrehten drei Dimensionen. Wir können deutlich sehen, dass die erste Hauptkomponente (der rote Pfeil) ausreicht, um die drei vorhandenen Datengruppen zu trennen. Wir dürfen schlussfolgern, dass die ursprünglich dreidimensionalen Daten durch diese Umwandlungen in einen eindimensionalen Datensatz verschlichtet werden können.

Die Entscheidung darüber, wie viele Hauptkomponenten nötig sind, um die Informationen eines Datensatzes zu erklären, kann mit Hilfe von statistischen Informationen genau getroffen werden. In diesem Fall wird es für den menschlichen Betrachter schnell klar, dass eine Dimension ausreicht, um die drei Klassen von Datenpunkten zu trennen. Wenn es aber darum geht, 10.000 Dimensionen auf 500 zu reduzieren, wird das für den menschlichen Nutzer zwar nicht mehr sichtbar sein, kann aber dennoch präzise ausgeführt werden.

Wir verwenden diese Methode, um die Dimensionen eines Datensatzes zu verringern und dennoch den darin enthaltenen Informationsgehalt aufrecht zu erhalten. Der Vorteil des Ganzen ist, dass wir den Datensatz mit vergleichsweise wenigen Parametern modellieren können, ohne die Genauigkeit zu opfern. Das spart nicht nur Zeit, sondern verbessert auch die Allgemeingültigkeit des Modells; denn es gilt beim maschinellen Lernen als generell wahr, dass ein Modell mit wenigen Parametern ein größeres Potenzial zur allgemeinen Anwendung bietet.