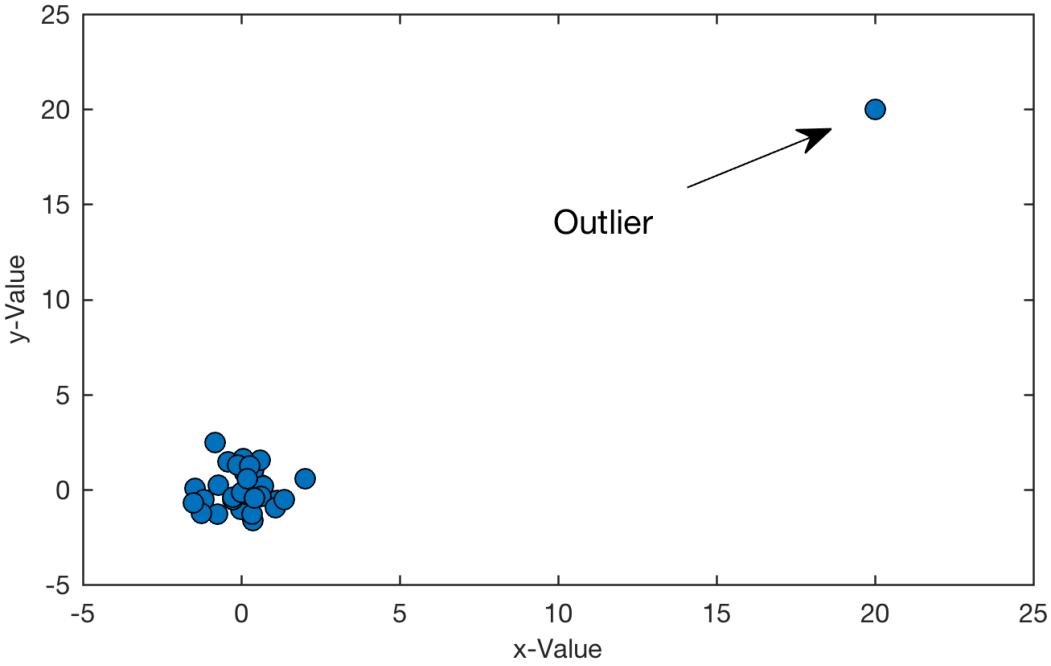
Handbuch: 7.1.7. Vorverarbeitung

Bevor Rohmessdaten im Zusammenhang mit maschinellem Lernen eingesetzt werden, durchlaufen diese gewöhnlich mehrere Bearbeitungsstufen. Diese Methoden, die zusammen als Vorverarbeitung (engl. pre-processing) bezeichnet werden, sind von essentieller Wichtigkeit, um gute Ergebnisse zu erzielen.

Unplausible Werte werden vom Datensatz entfernt, und zwar entsprechend der erlaubten Schwankungsbereiche, wie sie in den Metadaten spezifiziert sind. Ausreißer werden als nächstes entfernt. Das sind Datenpunkte, die höchst irreguläre oder gar unmögliche Eigenschaften aufweisen wie etwa sehr große Gradienten. Aus diesen Gründen entfernte Werte werden ersetzt durch möglichst plausible Messdaten.



Werte werden in ein Koordinatensystem normalisiert, damit jede Variable einen Durchschnitt von Null und eine Standardabweichung von 1 aufweist. Dadurch werden natürliche Skalierungen ebenso entfernt wie auch Inkompatibilitäten von Daten, die es im Zusammenhang mit großem Datenumfang immer mal wieder gibt. Beispielsweise könnten Sie ein Gewicht in Kilogramm oder in Tonnen berechnen. Für unser menschliches Verständnis macht es keinen Unterschied, welche Maßeinheit Sie benutzen, aber für einen Computer sind Angaben in Kilogramm sehr viel größer als in Tonnen, so dass sie im Vergleich zu anderen Größen – etwa Druck – überproportionalen Einfluss nehmen können. Die Normalisierung entfernt solche Besonderheiten und fokussiert sich nur auf die relativen Veränderungen.