1 Datenvorverarbeitung

Bevor überhaupt ein Maschinelles Lernen stattfinden kann müssen die Daten erst einmal vorverarbeitet werden. Diese Vorverarbeitung beinhaltet das ergänzen fehlender Daten und die Reduktion der Menge an Features, die es zu verarbeiten gilt.
Die fehlenden Daten in einem Datensatz können nicht einfach ignoriert werden, somit müssen die fehlenden Daten entweder ersetzt oder dieser Datenpunkt aus dem Datensatz genommen werden. Die Herausnahme des Datenpunktes kann die Datenmenge um viele Punkte verkürzen und somit die Präzision des Modells um einiges verringern. Daher gibt es auch die Möglichkeit diese Daten aus den übrigen Datenpunkten auszuwerten, indem man Interpolations- verfahren anwendet. Besonders geeignet sind hierfür der Mittelwert der anderen Daten dieses Features oder das Einsetzen des Wertes, welcher am häufigsten vorkommt.
Bei kategoralen Daten gibt es Unterschiede zwischen ordinalen und nominalen Merkmalen. Ordinale Merkmale sind zwar keine Zahlen, haben aber eine Ordnung untereinander und können so durch natürliche Zahlen, samt ihrer Ordnung, dargestellt werden. Bei nominalen Merkmalen wird hier einfach eine Liste mit binären Werten verwendet um die Klasse zu beschreiben.
Ebenfalls gehört die Auffteilung in Testdaten und Trainingsdaten zur Datenvorverarbeitung. Eine geeignete Auswahl stellt die Minimierung an Informationsverlust durch Verlust der Trainingsdaten, also auch die Maximierung der Fehlerabschätzung durch Testdaten da. Danach werden die Daten entweder normiert

$$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

oder standardisiert.

$$x_i^* = \frac{\mu_i - x_i}{\sigma_i} \tag{2}$$

Normierte Daten werden immer auf das Intervall von [0;1] begrenzt, während standatisierte Daten um einen Mittelwert herum, gestaucht mit der Standardabweichung, ein besseres Maß darstellt, da die standartisierten Daten bessere Übersicht über ungewöhnlich hohe oder kleine Daten geben und bei Auftreten dieser auch nicht so empfindlich ist, was die Werte der anderen Daten angeht.

Die Auswahl aussagekräftiger Merkmale reduziert die Berechnungslast und ist deswegen, besonders bei sehr großen Datenmengen sehr wichtig.

Zur Komplexitätsreduktion dient hier die L2- und L1-Regulierung.

$$L1 : ||\omega||_1 = \sum_{j=1}^{n} \omega_j \tag{3}$$

$$L2 : ||\omega||_2^2 = \sum_{j=1}^{n} \omega_j^2 \tag{4}$$

Beide Regulierungen werden als Terme zur Funktion, die es zu optimieren gilt, addiert und fördern Merkmale, die keinen Einfluss auf das Ergebnis haben heraus. Diese Merkmale können dann herausgenommen werden, indem man schaut, welche Gewichtungen im finalen Gewichtungsvektor Null sind und damit keinen Einfluss auf das Ergebnis haben.

Ein weiteres Verfahren zur Merkmalsreduktion ist die sequenzielle Rückwärtsauswahl von Merkmalen. Diesem Algorithmus liegt die Idee zugrunde, aus einem k-dimensionalem Raum einen d-dimensionalem Raum zu machen, indem wir (k-d)-mal das Modell mit einer kleineren Datenmenge trainieren und dabei jeweils ein Feature auslassen. Die Leistung des Modells wird dann von der ursprünglichen Leistung des Modells subtrahiert und es wird das Merkmal entfernt, welches die geringste Leistungseinbuße vorweißt. Dieser Algorithmus schaut also auch nur auf Einzelmerkmale, vernachlässigt aber das Gesamtbild und Features, die auf Langzeit in Kombinationen mit anderen Merkmalen sinnvolle Ergebnisse liefern. Allerdings ist dies ein kleiner Kritikpunkt und der Algorithmus funktioniert immer noch sehr gut.

Mit der Random-Forests-Methode lassen sich die Merkmale ebenfalls reduzieren, indem man einen Random-Forest mit diesen Daten trainiert und dabei schaut, auf welche Features dieser besonderen Wert legt. Die Features, die keine besondere Bedeutung haben, können somit aus der originalen Trainingsdatenmenge herausgenommen werden.