1 Überanpassung

Die Überanpassung ist ein Phänomen, welches häufig bei Machine Learning auftritt, und beschreibt extreme Gewichtung von Parametern. Diese extreme Gewichtung kann zu einer Überanpassung, also zu einer sehr großen Varianz der Daten führen. Diese Varianz bezeichnet also auch das Problem, dass das zugrundeliegende Modell für die Daten zu komplex ist. Es gibt aber auch ebenfalls die Unteranpassung die zu einem großen Bias führen und somit bei einer Entscheidung immer in eine bestimmte Richtung weisen. Dies ist ein Zeichen für ein unterkomplexes Modell.

Dieser Überanpassung wird nun durch einen sogenannten L2-Regulierungsterm mit dem Regulierungsparameter λ ausgeglichen.

$$\frac{\lambda}{2}||\omega||^2 = \frac{\lambda}{2}\sum_{j=1}^{m}\omega_j^2 \tag{1}$$

Diesen Term addieren wir also zu der Gewichtsänderung dazu.

$$\Delta\omega = \nabla(p + \frac{\lambda}{2}||\omega||^2) \tag{2}$$

Durch die Absicht diese Funktion zu minimieren ist sie kleiner, wenn auch die Gewichte kleiner sind. Der Parameter λ ist hierbei die Stärke der Regulierung. Ist dieser Parameter größer, ändert sich auch das Gewicht stärker.

2 Maximum-Margin-Klassifizierung

Um die Fehlklassifizierungen des Modells zu veringern wird auf die Maximum-Margin-Klassifizierung zurückgegriffen. Diese Klassifizierung hat es zum Ziel den Abstand zwischen den Punkten der Datenmengen und der Hyperebene zu maximieren um die Fehlklassifizierungen zu minimieren. Durch einen größeren Abstand der Punkte zur Hyperebene werden neue Daten eher im richtigen Feld als im falschen Feld einsortiert. Dabei konstruieren wir zwei Hyperebenen, die parallel zur trennenden Hyperbene verlaufen. Dabei nehmen wir an die trennende Hyperebene sitzt im Nullpnunkt des Koordinatensystems. Diese lassen sich durch den Parameter angeben, der nicht mit einer Komponente multipliziert wurde, um zwei Ebenen auf einer Geraden durch einen Punkt vergleochen zu können.

$$\omega_0 + \omega \cdot x_{pos} = 1 \tag{3}$$

Dies ist die Ebene mit dem positiven Versatz.

$$\omega_0 + \omega \cdot x_{neg} = -1 \tag{4}$$

Subtrahiert man nun diese beiden Parameter der Ebenen voneinander gilt:

$$\omega(x_{pos} - x_{neg}) = 2 \tag{5}$$

Dies bezeichnet den Abstand der beiden Ebenen. Normiert man nun diese Parameter gilt:

$$\frac{\omega(x_{pos} - x_{neg})}{||\omega||} = \frac{2}{||\omega||} \tag{6}$$

Diesen Abstand kann man nun mit verschiedenen Verfahren maximieren.

3 Der nicht-linear-trennbare Fall

Nun gibt es Fälle, in denen die Existenz einer Hyperebene nicht gegeben ist. Dort nutzt man nun die Transformation des Raumes der Daten in einen höherdimensionalen Raum, in dem sie durch eine Hyperebene trennbar sind und transformiert dann diese Hyperbene wieder in den ursprünglichen Raum zurück. So erhält man eine nichtlineare Entscheidungsgrenze und kann trotzdem das Modell der logistischen Regression benutzen.
Um nun die Daten zu transformieren nutzen wir den Kerneltrick. Dafür ersetzen wir das Skalarprodukt der Gewichte mit den Features durch das Skalarprodukt der Zuordnungsfunktion α von dem ursprünglichen Skalarpdodukt. Dafür muss dann gelten. Dies erleichtert die Berechnung des Skalarproduktes im höheren Raum.

$$x^{(i)T} \cdot x^j = \alpha(x^{(i)T}) \cdot \alpha(x^j) \tag{7}$$

Nun wird ebenfalls die Kernelfunktion definiert. Diese Kernelfunktion gilt als das Maß der Ähnlichkeit der beiden Koordinaten.

$$k(x^{(i)}, x^{(j)}) = e^{-\frac{||x^{(i)} - x^{(j)}||^2}{2\sigma^2}} \tag{8}$$

Diese Funktion wird also, durch die oben-vorrausgesetzte Transformation der Zurodnungsfunktion, das Skalarprodukt ersetzen. Das Skalarprodukt gibt nämlich ebenfalls, solange es normiert ist, die Ähnlichkeit zwischen zwei Punkten an, da es entweder gleich dem Betrag oder kleiner oder größer ist.