Überanpassung

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Überanpassung (Overfitting) ist die Fehlerart einer maschinelles Lernen (ML)-Methode, die sich zu eng an ihre Trainingsdatensatz anpasst: Die gelernte Hypothese weist auf der Trainingsdatensatz ein kleines empirisches Risiko auf, aber ein großes Risiko, d. h. einen großen erwarteten Verlust auf Datenpunkte, die aus der zugrunde liegenden Wahrscheinlichkeitsverteilung gezogen werden. Folglich hat die gelernte Hypothese eine große Generalisierungsabstand. Die Validierung erkennt eine Überanpassung durch den Vergleich des Trainingsfehlers mit dem auf einer Validierungsdatensatz erzielten Validierungsfehler. Überanpassung tritt typischerweise auf, wenn der Hypothesenraum im Verhältnis zur Anzahl der Datenpunkte in der Trainingsdatensatz zu groß ist; die Vapnik-Chervonenkis-Dimension (VC-Dimension) und die Rademacher-Komplexität messen diese Größe. Die Regularisierung wirkt der Überanpassung entgegen, indem das Modell beschnitten, dem empirisches Risiko ein Strafterm hinzugefügt oder die Trainingsdatensatz erweitert wird.

Definition

Ein Bild-Klassifikator, der jedes Foto in seiner Trainingsdatensatz auswendig lernt, aber Bilder außerhalb seiner Trainingsdatensatz falsch klassifiziert, weist eine Überanpassung auf. Allgemein sei eine maschinelles Lernen (ML)-Methode betrachtet, die empirische Risikominimierung (ERM) verwendet, um auf einer gegebenen Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ eine Hypothese $\widehat{h} \in \mathcal{H}$ mit Minimumem empirisches Risiko $\widehat{L}\big(\widehat{h}|\mathcal{D}^{(\text{train})}\big)$ zu lernen. Die Methode überpasst sich an $\mathcal{D}^{(\text{train})}$ an, wenn das empirisches Risiko $\widehat{L}\big(\widehat{h}|\mathcal{D}^{(\text{train})}\big)$ klein ist, während das Risiko $\bar{L}\big(\widehat{h}\big)$, d. h. der erwartete Verlust auf unabhängig von $\mathcal{D}^{(\text{train})}$ aus der zugrunde liegenden Wahrscheinlichkeitsverteilung gezogenen Datenpunkte, groß ist. Folglich lernt eine zur Überanpassung neigende ML-Methode eine Hypothese mit einer großen Generalisierungsabstand (Hastie et al., 2009, Sect. 7.2; Jung, 2022).

Eine große Generalisierungsabstand allein impliziert noch keine Überanpassung. Man betrachte ein lineares Modell, das an eine kleine Trainingsdatensatz angepasst wird, deren Labels verrauscht sind und nichtlinear von den Merkmale abhängen. Es kann einen großen Trainingsfehler und ein noch größeres Risiko aufweisen, sodass die Generalisierungsabstand groß ist, während der

Trainingsfehler nicht klein ist. Dieses Beispiel stellt eine Underfitting dar, das gegenteilige Fehlverhalten: Ein Hypothesenraum, der keine Hypothese enthält, die den Zusammenhang zwischen den Merkmale und dem Label eines Datenpunkts abbildet, lässt beide Größen groß werden.

Abb. 1 zeigt, warum ein kleiner Trainingsfehler für sich genommen kein Beleg gegen eine Überanpassung ist. Drei Datenpunkte liegen auf einer Geraden. Keine der drei Kurven ist von Hand gezeichnet: Jede ist eine Hypothese $\widehat{h}$, die von der ERM mit dem quadratischer Verlust auf genau diesen drei Datenpunkte geliefert wird. Was sie unterscheidet, ist der Hypothesenraum, den die ERM durchsucht hat, und die drei sind ineinander geschachtelt, $\mathcal{H}^{(1)} \subset \mathcal{H}^{(2)} \subset \mathcal{H}^{(3)}$: $\mathcal{H}^{(1)}$ sind die Konstanten $h(x) = b$, $\mathcal{H}^{(2)}$ das lineares Modell, dessen Hypothesen $h(x) = wx + b$ sind, und $\mathcal{H}^{(3)}$ alle stetigen Funktionen auf $\mathbb{R}$.

Auf $\mathcal{H}^{(1)}$ hat die ERM eine eindeutige Lösung: die Konstante, deren Wert der Durchschnitt der drei Labels ist. Diese Konstante trifft den mittleren Datenpunkt und verfehlt die beiden anderen, und kein Element aus $\mathcal{H}^{(1)}$ schneidet besser ab. Der Hypothesenraum ist hier zu klein, und die Methode unterpasst sich an. Auf $\mathcal{H}^{(2)}$ ist die Lösung ebenfalls eindeutig und erreicht einen Trainingsfehler von null; der Hypothesenraum passt zum Zusammenhang zwischen Merkmal und Label, und die Methode weist weder eine Unter- noch eine Überanpassung auf. Auf $\mathcal{H}^{(3)}$ erreicht jede stetige Funktion durch die drei Datenpunkte einen Trainingsfehler von null, und es gibt unendlich viele davon: Die ERM hat dort unendlich viele Lösungen, und Abb. 1 zeigt eine davon.

Der Trainingsfehler kann diese Kurve nicht von der Lösung auf $\mathcal{H}^{(2)}$ unterscheiden, also der linearen Hypothese, die durch die drei Datenpunkte verläuft, und er kann auch keine Auswahl unter den Lösungen auf $\mathcal{H}^{(3)}$ treffen. Die beiden Hypothesen stimmen auf der Trainingsdatensatz überein, unterscheiden sich jedoch bei Merkmalswerten zwischen denen der Trainings-Datenpunkte, und das Risiko unterscheidet sich zwischen ihnen, da aus der Wahrscheinlichkeitsverteilung gezogene Datenpunkte ebenfalls in diesen Bereich fallen. Der Hypothesenraum ist hier zu groß, und die Methode überpasst sich an.

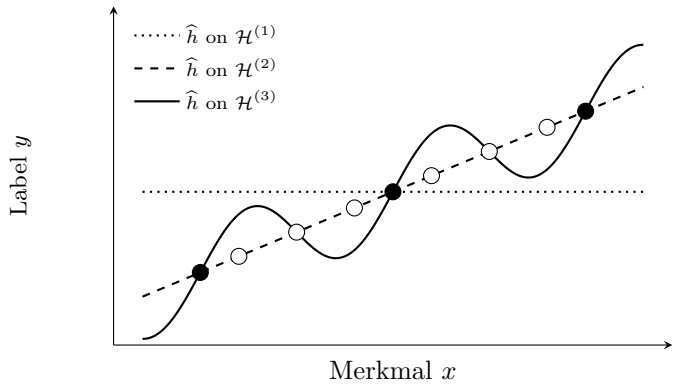


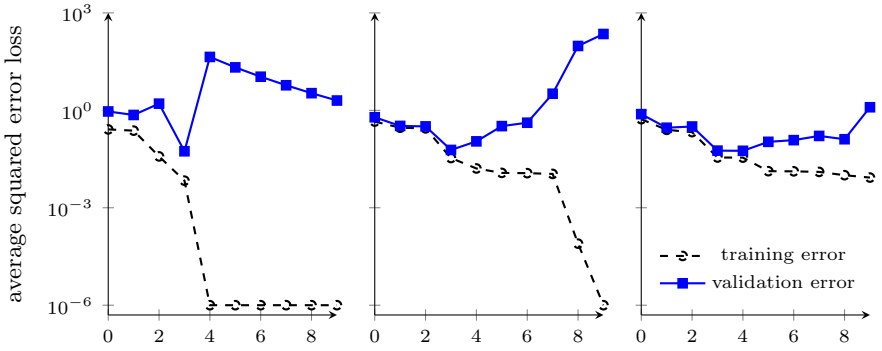
Abbildung 1: Die Hypothesen $\widehat{h}$, welche die ERM mit dem quadratischer Verlust aus drei geschachtelten Hypothesenräume liefert: den Konstanten $\mathcal{H}^{(1)}$ (gepunktete Linie), dem lineares Modell $\mathcal{H}^{(2)}$ (gestrichelte Linie) und allen stetigen Funktionen $\mathcal{H}^{(3)}$ (durchgezogene Kurve). Ausgefüllte Kreise sind die drei Datenpunkte der Trainingsdatensatz, die auf einer Geraden liegen; offene Kreise sind weitere Datenpunkte auf derselben Geraden außerhalb der Trainingsdatensatz. Dieselbe Trainingsdatensatz und dieselbe ERM erzeugen alle drei Kurven, sodass der Hypothesenraum das Einzige ist, was variiert. Auf $\mathcal{H}^{(1)}$ trifft die Konstante nur den mittleren Datenpunkt, was einer Underfitting entspricht. Auf $\mathcal{H}^{(2)}$ stellt die lineare Hypothese den Zusammenhang wieder her, dem die Datenpunkte folgen. Auf $\mathcal{H}^{(3)}$ ist der Trainingsfehler ebenfalls null, sodass der Trainingsfehler die Kurve nicht von der linearen Hypothese unterscheiden kann. Die Kurven unterscheiden sich zwischen den Trainings-Datenpunkte, wo die offenen Kreise liegen und wo sich das Risiko der beiden Kurven unterscheidet: Die Methode überpasst sich an

Die Validierung erkennt eine Überanpassung, da die sie definierende Generalisierungsabstand geschätzt werden kann (Shalev-Shwartz and Ben-David, 2014, Thm. 11.1): Eine ML-Methode, die zur Überanpassung neigt, liefert einen kleinen Trainingsfehler, aber einen großen Validierungsfehler auf einer vom Training zurückgehaltenen Validierungsdatensatz.

Abb. 2 veranschaulicht die Überanpassung bei der polynomiale Regression. Jeder Datenpunkt trägt ein einzelnes Merkmal x und ein verrauschtes Label y, die aus einer in der Bildunterschrift beschriebenen Wahrscheinlichkeitsverteilung gezogen werden. Polynome vom Grad $p = 0, \ldots, 9$ werden mithilfe der ERM mit dem quadratischer Verlust an Trainingsdatensätze von $m = 5$, 10 und 20 Datenpunkte angepasst; jedes Teilbild von Abb. 2 zeigt den Trainingsfehler und den Validierungsfehler für eine Größe. Der Trainingsfehler nimmt in jedem Teilbild mit steigendem Grad ab: Das Polynom vom Grad 9 hat 10 Modellparameter und interpoliert die $m = 10$ Datenpunkte; für $m = 5$ interpoliert jeder Grad ab 4; für $m = 20$ interpoliert keiner der gezeigten Grade. Der Validierungsfehler wird auf einer gemeinsamen Validierungsdatensatz von 100

Datenpunkte ermittelt. Er ist im Gegensatz dazu bei einem moderaten Grad am kleinsten und wächst für größere Grade um mehrere Größenordnungen an. Die Polynome hohen Grades überpassen sich an die Trainingsdatensatz an, und die Differenz zwischen Validierungsfehler und Trainingsfehler schätzt ihre große Generalisierungsabstand.

Ein Vergleich der drei Größen der Trainingsdatensatz zeigt, dass der Grad, ab dem der Validierungsfehler vom Trainingsfehler abweicht, umso niedriger ist, je kleiner die Trainingsdatensatz ist. Für $m = 5$ springt der Validierungsfehler bereits bei Grad 4 um fast drei Größenordnungen nach oben. Grad 4 ist der niedrigste Grad, dessen 5 Modellparameter ausreichen, um die Trainingsdatensatz zu interpolieren. Für $m = 20$ variiert der Validierungsfehler über alle gezeigten Grade hinweg um weniger als eine Größenordnung.



(a) $m = 5$ (b) $m = 10$ (c) $m = 20$

Abbildung 2: Trainingsfehler und Validierungsfehler der polynomiale Regression über dem Polynomgrad p (gestrichelte Kurve mit offenen Kreisen und durchgezogene Kurve mit ausgefüllten Quadraten); die Teilbilder (a), (b) und (c) entsprechen einer Trainingsdatensatz von $m = 5$, 10 bzw. 20 Datenpunkte. Das Merkmal x wird gleichverteilt aus $[0, 1]$ gezogen, und das Label ist $y = \sin(2\pi x) + \varepsilon$ mit additivem gaußschem Rauschen ε. Alle drei Größen teilen sich eine Validierungsdatensatz von 100 Datenpunkte, die aus derselben Wahrscheinlichkeitsverteilung gezogen wurden. Je kleiner die Trainingsdatensatz, desto niedriger der Grad, ab dem die Überanpassung einsetzt. Daten erzeugt mit `pythondemos/overfitting.py`

Überanpassung tritt typischerweise auf, wenn der Hypothesenraum $\mathcal{H}$ im Verhältnis zur Anzahl der Datenpunkte in der Trainingsdatensatz zu groß ist (wie für $\mathcal{H}^{(3)}$ in Abb. 1). Die Vapnik-Chervonenkis-Dimension (VC-Dimension) (Vapnik, 2000, Sect. 3.6) von $\mathcal{H}$ und seine Rademacher-Komplexität (Shalev-Shwartz and Ben-David, 2014, Sect. 26.1) messen die Größe von $\mathcal{H}$, ohne Modellparameter zu zählen. Das Zählen von Modellparameter ist der naheliegende Weg, diese Größe zu messen, aber nicht immer der richtige: Die Hypothe-

sen $h^{(w)}(x) = \lceil 0.5 \sin(wx) \rceil$ besitzen einen einzigen Modellparameter, bilden aber dennoch einen Hypothesenraum mit unendlicher VC-Dimension (Shalev-Shwartz and Ben-David, 2014, Sect. 6.3.5).

Die Anzahl ist im Gegensatz dazu genau richtig für die Klassifikatoren $h(x) = \text{sign}\left(w_0 + w_1 x + \ldots + w_p x^p\right)$, die durch Schwellenwertbildung der Polynome vom Höchstgrad p in einem einzelnen Merkmal entstehen. Ihr Hypothesenraum hat die VC-Dimension $p+1$, was der Anzahl der Modellparameter entspricht und um eins größer ist als der Höchstgrad. Die Merkmalsabbildung $x \mapsto (x, x^2, \ldots, x^p)$ überführt diese Klassifikatoren in lineare Klassifikatoren auf $\mathbb{R}^p$, deren Hypothesenraum die VC-Dimension $p + 1$ hat (Shalev-Shwartz and Ben-David, 2014, Thm. 9.3). Umgekehrt werden beliebige $p + 1$ Datenpunkte mit verschiedenen Merkmale zerschmettert. Die Klassifikatoren realisieren jede mögliche Zuordnung binärer Labels zu diesen Datenpunkte, da ein Polynom vom Grad höchstens p beliebige Labels auf ihnen interpoliert. Diese schwellenwertbasierten Klassifikatoren betreffen einen anderen Fall als Abb. 2, in der Polynome derselben Grade an eine Regressionsaufgabe angepasst werden.

Beide Maße gehen in die oberen Schranken für die Generalisierungsabstand ein, die mit ihnen wachsen und mit der Anzahl der Datenpunkte in der Trainingsdatensatz schrumpfen (Shalev-Shwartz and Ben-David, 2014, Sect. 26.1; Vapnik, 2000, Sect. 3.7).

Die Regularisierung wirkt der Überanpassung in drei elementaren Formen entgegen: durch Beschneiden des Modells, Hinzufügen eines Strafterms zum empirisches Risiko oder Erweitern der Trainingsdatensatz mittels Datenerweiterung.

Siehe auch: ERM, Generalisierungsabstand, Generalisierung, Regularisierung, Underfitting, Validierung, VC-Dimension.

Querverweise

Literatur

1. Jung (2022). Machine Learning: The Basics. Springer Nature, Singapore, Singapore.

2. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

3. Vapnik (2000). The Nature of Statistical Learning Theory. 2nd ed., Springer, New York, NY, USA.

4. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.