Validierungsfehler

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

Der Validierungsfehler einer gelernten Hypothese $\widehat{h}$, die durch eine Methode des maschinelles Lernen (ML) mithilfe eines Trainingsdatensatz gewonnen wurde, ist ihr durchschnittlicher Verlust über einem Validierungsdatensatz. Das Validierungsdatensatz besteht aus Datenpunkte, die nicht zu dem Trainingsdatensatz gehören, das zum Lernen der Hypothese $\widehat{h}$ verwendet wurde. Wenn die Erzeugung der Datenpunkte im Validierungsdatensatz durch unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) modelliert wird, ist der Validierungsfehler ein unverzerrter Schätzer des Risikos. Wird das Trainingsdatensatz als zufällig modelliert, so wird auch die gelernte Hypothese zu einer ZV, und der Validierungsfehler ist erwartungstreu bedingt auf die gelernte Hypothese. Der Validierungsfehler beschreibt die Güte einer gelernten Hypothese auf Datenpunkte außerhalb des für das Training des Modells verwendeten Trainingsdatensatz. Er ist daher ein geeigneteres Kriterium für die Modellauswahl als der Trainingsfehler. Ein Vergleich des Validierungsfehlers mit dem Trainingsfehler einer gelernten Hypothese ermöglicht es, Überanpassung und Underfitting zu diagnostizieren.

Definition

Die Wetterstation Krems an der Donau (Österreich) zeichnete von 2020 bis 2024 für jeden Tag das morgendliche Minimum und das Maximum der Temperatur auf, ein Datensatz von 1824 Tagen, für die beide Messwerte vorliegen. Jeder Tag ist ein Datenpunkt: Sein Merkmal x ist die Lufttemperatur um 06:00 (UTC), sein Label y die Lufttemperatur um 15:00 (UTC) desselben Tages. Abb. 1 zeigt zwanzig dieser Datenpunkte als Streudiagramm. Drei davon, in Schwarz eingezeichnet, bilden ein Trainingsdatensatz, mit dem zwei Hypothesen mittels empirische Risikominimierung (ERM) gelernt werden: die lineare Hypothese $\widehat{h}^{(1)}$ über dem lineares Modell und das Polynom zweiten Grades $\widehat{h}^{(2)}$, das die polynomiale Regression liefert. Das Polynom verläuft durch alle drei Tage und weist einen Trainingsfehler von null auf, weit unter dem Trainingsfehler 2.22 von $\widehat{h}^{(1)}$. Sein durchschnittlicher Verlust über alle 1824 Datenpunkte von 2020 bis 2024 beträgt jedoch 26.93.

Die Brauchbarkeit einer Hypothese $\widehat{h} = \mathcal{A}(\mathcal{D}^{(\text{train})})$, die durch eine Methode des maschinelles Lernen (ML) $\mathcal{A}$ aus einem Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ gelernt

wurde, muss über $\mathcal{D}^{(\text{train})}$ hinaus beurteilt werden. Zu diesem Zweck wird $\widehat{h}$ auf einem Validierungsdatensatz $\mathcal{D}^{(\text{val})}$ ausgewertet, das aus Datenpunkte außerhalb des Trainingsdatensatz besteht. Der Validierungsfehler E_v von $\widehat{h}$ ist ihr durchschnittlicher Verlust auf dem Validierungsdatensatz (Hastie et al., 2009, Abschn. 7.2),

$$E_v = \frac{1}{|\mathcal{D}^{(\text{val})}|} \sum_{\mathbf{z} \in \mathcal{D}^{(\text{val})}} L\left(\mathbf{z}, \widehat{h}\right), \tag{1}$$

wobei $L\left(\mathbf{z}, \widehat{h}\right)$ der Verlust ist, den $\widehat{h}$ auf einem Datenpunkt $\mathbf{z}$ erleidet. Abb. 1 veranschaulicht ein Validierungsdatensatz aus drei Datenpunkte, die als rote Dreiecke eingezeichnet und nach einer festen Regel so gewählt sind, dass sie den Wertebereich des Merkmals abdecken. Die gepunkteten Balken sind die Fehler der Vorhersage des Polynoms auf diesen Punkten. Der Durchschnitt der drei quadrierten Balken ist der Validierungsfehler (1) des Polynoms, 14.73. Er liegt weit über dem Trainingsfehler von null, legt also offen, was der Trainingsfehler verbirgt, und liegt dennoch 45% unter dem durchschnittlichen quadratischer Verlust von $\widehat{h}^{(2)}$ über alle 1824 Datenpunkte, 26.93: Drei Datenpunkte decken das Überanpassung auf, ohne das Risiko genau festzulegen. Eine Hypothese mit einem größeren Trainingsfehler kann die bessere sein: Die lineare Hypothese $\widehat{h}^{(1)}$ hat einen Trainingsfehler von 2.22, aber einen Validierungsfehler von 2.44.

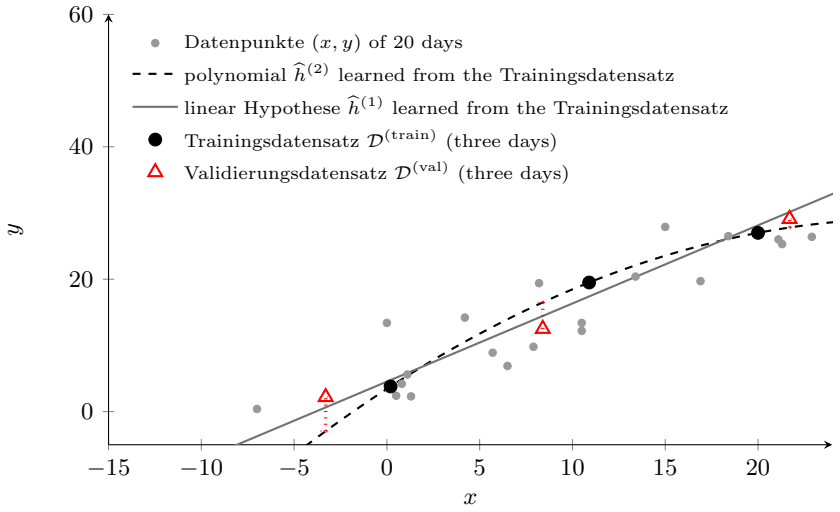

Abbildung 1: Zwanzig Datenpunkte (graue Punkte) aus einem Datensatz von 1824 täglichen Wetterbeobachtungen in Krems an der Donau (Österreich), 2020 bis 2024. Jeder Datenpunkt steht für einen Tag, mit der Lufttemperatur um 06:00 (UTC) als seinem Merkmal x und der Lufttemperatur um 15:00 (UTC) desselben Tages als seinem Label y. Die drei schwarzen Punkte bilden das Trainingsdatensatz. Die gestrichelte Kurve ist das aus ihnen gelernte Polynom zweiten Grades $\widehat{h}^{(2)}$; es verläuft durch alle drei Punkte, sodass sein Trainingsfehler gleich null ist. Die durchgezogene Gerade ist die lineare Hypothese $\widehat{h}^{(1)}$, die aus demselben Trainingsdatensatz gelernt wurde. Die drei roten Dreiecke bilden das Validierungsdatensatz, und die gepunkteten Balken sind die Fehler der Vorhersage des Polynoms auf ihnen. Der Mittelwert der drei quadrierten Balken ist der Validierungsfehler des Polynoms. Daten erzeugt mit `pythondemos/valerr.py`

Was der Validierungsfehler (1) schätzt, wird durch eine Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme) für die Erzeugung der Daten festgelegt. Die Datenpunkte sind Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) mit einer gemeinsamen Wahrscheinlichkeitsverteilung. Diese Wahrscheinlichkeitsverteilung definiert das Risiko $\bar{L}\big(\widehat{h}\big)$, den Erwartungswert von $L\left(\mathbf{z}, \widehat{h}\right)$ über einen daraus gezogenen Datenpunkt. Jeder Summand des Durchschnitts (1) hat diesen Erwartungswert, sodass E_v die Realisierung einer ZV mit dem Erwartungswert $\bar{L}\big(\widehat{h}\big)$ ist: Für eine feste Hypothese $\widehat{h}$ ist er ein unverzerrter Schätzer von $\bar{L}\big(\widehat{h}\big)$. Seine Varianz ist die Varianz eines einzelnen Verlustes geteilt durch $|\mathcal{D}^{(\text{val})}|$, d. h., sie nimmt mit wachsendem $|\mathcal{D}^{(\text{val})}|$ ab.

In der Praxis ist die gelernte Hypothese $\widehat{h} = \mathcal{A}(\mathcal{D}^{(\text{train})})$ der Ausgabe eines Verfahrens zum Training $\mathcal{A}$, das ein Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ einliest. Unter der u.i.v.-Annahme ist das Trainingsdatensatz selbst die Realisierung einer ZV,

und damit auch die gelernte Hypothese $\widehat{h}$: Eine andere Realisierung von $\mathcal{D}^{(\text{train})}$ hätte eine andere Hypothese geliefert. Der Validierungsfehler ist dann erwartungstreu bedingt auf die gelernte Hypothese. Der Trainingsfehler besitzt keine solche Eigenschaft: Bedingt auf $\widehat{h}$ sind die Datenpunkte von $\mathcal{D}^{(\text{train})}$ keine u.i.v. Realisierungen der Wahrscheinlichkeitsverteilung eines Datenpunkts mehr, da $\widehat{h}$ so gewählt wurde, dass ihr durchschnittlicher Verlust auf ihnen klein wird.

Sind die Datenpunkte von $\mathcal{D}^{(\text{val})}$ und $\mathcal{D}^{(\text{train})}$ Realisierungen von u.i.v. Zufallsvariablen, so ist $\mathcal{D}^{(\text{val})}$ unabhängig von $\widehat{h}$ (Shalev-Shwartz and Ben-David, 2014, Abschn. 11.2.1). Der bedingter Erwartungswert von E_v gegeben $\widehat{h}$ ist dann ihr Risiko,

$$\mathbb{E}\big\{E_v \mid \widehat{h}\big\} = \bar{L}\big(\widehat{h}\big).$$

Die Verlustfunktion $L(\cdot, \cdot)$ wird im Gegensatz dazu auf einem einzelnen Datenpunkt ausgewertet.

Wie groß muss ein Validierungsdatensatz sein, damit E_v nahe am Risiko liegt? Die oben berechnete Varianz gibt die typische Abweichung an; eine Konzentrationsungleichung liefert eine Garantie, die mit einer vorgegebenen Wahrscheinlichkeit gilt. Für eine feste Hypothese $\widehat{h}$, einen Verlust mit Werten in $[0, 1]$ und einen Konfidenz-Parameter $\delta \in (0, 1)$ liefert die Hoeffding-Ungleichung mit einer Wahrscheinlichkeit von mindestens $1 - \delta$ über die Ziehung von $\mathcal{D}^{(\text{val})}$,

$$\big|E_v - \bar{L}\big(\widehat{h}\big)\big| \leq \sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}^{(\text{val})}|}}, \tag{2}$$

siehe Shalev-Shwartz and Ben-David (2014), Satz 11.1. Die Schranke (2) hängt allein von der Größe des Validierungsdatensatz und von δ ab, nicht aber vom Hypothesenraum $\mathcal{H}$, aus dem $\widehat{h}$ gelernt wurde. Dies unterscheidet sich von der Kontrolle der Abweichung des Trainingsfehlers vom Risiko. Die Lösung der ERM wurde durch den Vergleich aller Hypothesen aus $\mathcal{H}$ auf dem Trainingsdatensatz ausgewählt, sodass eine Schranke wie (2) für den Trainingsfehler von $\widehat{h}$ für jede Hypothese aus $\mathcal{H}$ gleichzeitig gelten muss, und eine solche Schranke wächst mit der Größe von $\mathcal{H}$ (Shalev-Shwartz and Ben-David, 2014, Abschn. 4.1 und 11.2.1).

In Abb. 1 war das Validierungsdatensatz fest vorgegeben. Ein numerisches Experiment zeigt, wie stark der Validierungsfehler schwankt, wenn das Validierungsdatensatz stattdessen zufällig gewählt wird. Betrachtet sei die Hypothese

$$\widehat{h}(x) = 5.97 + 1.04\, x,$$

die mittels lineare Regression aus vierzig Datenpunkte derselben Stationsaufzeichnung gewonnen wurde, nämlich jedem neunten Tag des Jahres 2024. Die Datenpunkte von 305 weiteren Tagen desselben Jahres, von denen keiner für das Training verwendet wurde, werden in einem Datensatz $\mathcal{D}^{(\text{pool})}$ zusammengefasst. In diesem Experiment dient dieser Pool lediglich als Sammlung, aus der jedes Validierungsdatensatz zufällig gezogen wird. Er ist kein Testdatensatz $\mathcal{D}^{(\text{test})}$, das separat zurückgehalten wird, um eine einzige ausgewählte Hypothese zu beurteilen, und das nur einmalig eingelesen wird.

Abb. 2 zeigt den Mittelwert und die Standardabweichung des Validierungsfehlers von $\widehat{h}$ über wiederholte Ziehungen eines Validierungsdatensatz vorgegebener Größe aus $\mathcal{D}^{(\text{pool})}$. Die Standardabweichung nimmt mit zunehmender Größe des Validierungsdatensatz ab, wie es die obige Berechnung der Varianz vorhersagt: Sie fällt mit $1/\sqrt{|\mathcal{D}^{(\text{val})}|}$. Dies entspricht auch der Form der Schranke (2), obwohl diese Schranke selbst hier nicht anwendbar ist, da der quadratischer Verlust Werte außerhalb von $[0, 1]$ annimmt.

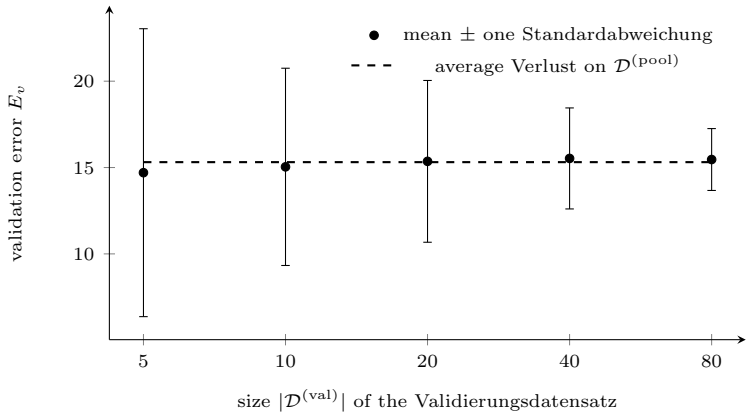

Abbildung 2: Streuung des für die Hypothese $\widehat{h}(x) = 5.97 + 1.04\,x$ berechneten Validierungsfehlers. Die Validierungsfehler werden für verschiedene Validierungsdatensätze berechnet, die für eine gegebene Größe zufällig aus dem Vorrat $\mathcal{D}^{(\text{pool})}$ von Datenpunkte ausgewählt werden, die nicht zum Lernen von $\widehat{h}$ verwendet wurden. Jeder Balken erstreckt sich über eine Standardabweichung zu beiden Seiten des Mittelwerts. Die gestrichelte Kurve ist der mittlere Verlust, berechnet über ganz $\mathcal{D}^{(\text{pool})}$. Die Standardabweichung nimmt ab, wenn $|\mathcal{D}^{(\text{val})}|$ wächst. Daten erzeugt mit `pythondemos/valerr.py`

Über die Schätzung des Risikos einer einzelnen Hypothese hinaus wird der Validierungsfehler mit dem Trainingsfehler und über verschiedene Modelle hinweg verglichen. Ein Trainingsfehler weit unterhalb des Validierungsfehlers deutet auf Überanpassung hin (Hastie et al., 2009, Abschn. 7.2). Ein Validierungsfehler nahe am Trainingsfehler zeigt kein Überanpassung; liegen beide zudem weit über dem Fehler, den ein größeres Modell auf demselben Validierungsdatensatz erreicht, lautet die Diagnose Underfitting, und das Modell ist zu klein für die Aufgabe. Die beiden Fehler zusammen erlauben daher die Diagnose einer Methode des ML und leiten deren Überarbeitung an: durch ein größeres Modell im zweiten Fall und durch Regularisierung, mehr Datenpunkte für das Training oder ein kleineres Modell im ersten Fall.
Der Vergleich der Validierungsfehler mehrerer Kandidaten-Modelle ist die Modellauswahl (Bishop, 2006, Sect. 1.3). Ein Modell hat keinen eigenen Validierungsfehler. Die Formulierung *Validierungsfehler eines Modells* steht verkürzend

für den Validierungsfehler der Hypothese, die $\mathcal{A}$ aus diesem Modell liefert, da E_v der mittlere Verlust einer einzelnen Hypothese ist. Ausgewählt wird das Modell, dessen Hypothese den kleinsten der verglichenen Validierungsfehler erzielt. Dieser kleinste Validierungsfehler ist als Schätzwert für das Risiko nach unten verzerrt, da das Minimum mehrerer verrauschter Schätzwerte im Mittel unterhalb der geschätzten Größe liegt, weshalb ein Testdatensatz $\mathcal{D}^{(\text{test})}$ für die abschließende Bewertung zurückgehalten wird (Bishop, 2006, Sect. 1.3; Hastie et al., 2009, Sect. 7.2).

Was der Validierungsfehler kostet, sind Datenpunkte: Ein für $\mathcal{D}^{(\text{val})}$ zurückgehaltener Datenpunkt kann nicht für das Training verwendet werden. Was er einbringt, ist ein Schätzwert, den der Trainingsfehler nicht liefern kann, berechnet auf Datenpunkte, die das Training nie eingelesen hat. Der Preis dieses Schätzwerts wird doppelt gezahlt, da er selbst verrauscht ist. Seine Standardabweichung fällt nur mit $1/\sqrt{|\mathcal{D}^{(\text{val})}|}$, sodass eine Halbierung des Rauschens viermal so viele Datenpunkte kostet: Auf den Kremser Aufzeichnungen streut ein Validierungsdatensatz von zwanzig Datenpunkte um 4.68 Quadratgrad um das Risiko und eines von achtzig um 1.79. Eine Abweichung, die kleiner als diese Streuung ist, lässt sich an einem einzelnen Validierungsdatensatz nicht ablesen. Für die Hypothese des Experiments, die über dem lineares Modell gelernt wurde, liegt der Trainingsfehler um 1.70 Quadratgrad unter dem Risiko, während ein Mittelwert über zwanzig Datenpunkte um etwa 4.22 schwankt. Ein einzelnes Validierungsdatensatz dieser Größe kann daher das Überanpassung eines lineares Modells überhaupt nicht auflösen. Es löst jedoch das Überanpassung der polynomiale Regression neunten Grades auf, deren Trainingsfehler weit unter ihrem Risiko liegt. Wo zu wenige Datenpunkte vorhanden sind, um sie für ein Validierungsdatensatz dieser Größe aufzuwenden, ist die k-fache Kreuzvalidierung die übliche Antwort: Jeder Datenpunkt dient einmal zur Validierung und ansonsten zum Training, zum Preis des Lernens einer Hypothese pro Teilmenge anstelle einer einzigen insgesamt.

Siehe auch: Validierungsdatensatz, Validierung, Trainingsdatensatz, Testdatensatz, Trainingsfehler, Verlust, Risiko, Hypothese, ERM, Überanpassung, Modellauswahl, Generalisierungsabstand, bedingter Erwartungswert, u.i.v.-Annahme, unverzerrter Schätzer.

Querverweise

Literatur

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

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

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