Validierungsdatensatz

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

Ein Validierungsdatensatz besteht aus Datenpunkte, die nicht für das Training eines Modells verwendet wurden. Für eine gelernte Hypothese gibt der mittlere Verlust auf dem Validierungsdatensatz, der Validierungsfehler, an, wie gut die Hypothese die Labels von Datenpunkte außerhalb des Trainingsdatensatzes vorhersagt. Der Validierungsfehler wird zur Modellauswahl verwendet: Die von verschiedenen maschinelles Lernen (ML)-Verfahren gelernten Hypothesen werden verglichen, und diejenige mit dem kleinsten Validierungsfehler wird ausgewählt. Die ausgewählte Hypothese hängt vom Validierungsdatensatz ab, sodass ihre Bewertung einen Testdatensatz erfordert, dessen Datenpunkte weder in das Training noch in die Auswahl eingeflossen sind. Der Umfang des Validierungsdatensatzes bestimmt, wie zuverlässig der Validierungsfehler ist. Wenn Datenpunkte knapp sind, verwendet die k-fache Kreuzvalidierung nacheinander jede Teilmenge des Datensatzes als Validierungsdatensatz und mittelt die k Validierungsfehler.

Definition

Betrachtet seien Wetteraufzeichnungen über drei Tage: Für jeden Tag ist die Lufttemperatur um 06:00 (UTC) das Merkmal und die Lufttemperatur um 15:00 (UTC) desselben Tages das Label. Abb. 1 zeigt einen Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ aus drei solchen Tagen zusammen mit zwei daraus gelernten Hypothesen: der linearen Hypothese $\widehat{h}^{(1)}$, die durch empirische Risikominimierung (ERM) auf dem lineares Modell geliefert wird, und einem Polynom zweiten Grades $\widehat{h}^{(2)}$, das durch alle drei Datenpunkte verläuft und daher einen Trainingsfehler von null erreicht.

Drei weitere Tage, markiert durch offene Dreiecke, wurden vom Training des Modells zurückgehalten: Sie bilden einen Validierungsdatensatz $\mathcal{D}^{(\text{val})}$. Der Validierungsfehler einer auf dem Trainingsdatensatz gelernten Hypothese $\widehat{h}$ ist ihr mittlerer Verlust auf dem Validierungsdatensatz (Shalev-Shwartz and Ben-David, 2014, Abschn. 11.2.1),

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

Hierbei ist $L\left(\mathbf{z}, \widehat{h}\right)$ der Verlust von $\widehat{h}$ auf dem Datenpunkt $\mathbf{z}$. In Abb. 1 erreicht das Polynom $\widehat{h}^{(2)}$ einen Trainingsfehler von null, aber einen Validierungsfehler von 17.3, während die lineare Hypothese $\widehat{h}^{(1)}$ einen Trainingsfehler von 2.9 und einen Validierungsfehler von 3.0 aufweist.

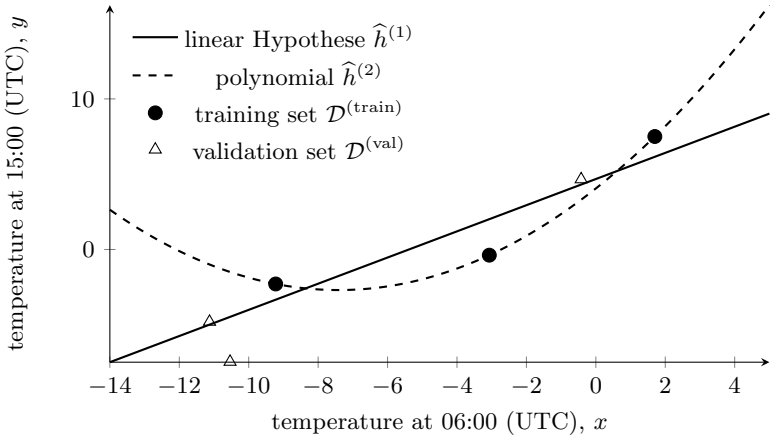


Abbildung 1: Datenpunkte, die tägliche Wetterbedingungen darstellen. Jeder Datenpunkt steht für einen Tag, mit der Lufttemperatur um 06:00 (UTC) als Merkmal x und der Lufttemperatur um 15:00 (UTC) desselben Tages als Label y. Die drei ausgefüllten Kreise bilden den Trainingsdatensatz $\mathcal{D}^{(\text{train})}$; die durchgezogene Kurve ist die lineare Hypothese $\widehat{h}^{(1)}$ und das gestrichelte Polynom die Hypothese $\widehat{h}^{(2)}$, beide gelernt aus $\mathcal{D}^{(\text{train})}$. Die drei offenen Dreiecke wurden vom Training des Modells zurückgehalten und bilden einen Validierungsdatensatz $\mathcal{D}^{(\text{val})}$: Der mittlere Verlust auf ihnen, 3.0 für $\widehat{h}^{(1)}$ und 17.3 für $\widehat{h}^{(2)}$, ist der Validierungsfehler der jeweiligen Hypothese. Daten erzeugt mit `pythondemos/valset.py`

Der Vergleich der beiden Validierungsfehler ist ein Beispiel für Modellauswahl: Die von verschiedenen maschinelles Lernen (ML)-Verfahren gelernten Hypothesen werden verglichen, und diejenige mit dem kleinsten Validierungsfehler wird ausgewählt (Bishop, 2006, Abschn. 1.3). In Abb. 1 wird die Hypothese $\widehat{h}^{(1)}$ ausgewählt. Die ausgewählte Hypothese hängt von $\mathcal{D}^{(\text{val})}$ ab: Ihr Validierungsfehler ist der kleinste der verglichenen Validierungsfehler, sodass der Validierungsfehler der ausgewählten Hypothese nach unten verzerrt ist.

Die Bewertung der Hypothese, die aus dem anhand der Validierungsfehler gewählten Modell gelernt wurde, erfordert einen dritten Datensatz, dessen Datenpunkte weder in das Training noch in die Auswahl eingeflossen sind: den Testdatensatz $\mathcal{D}^{(\text{test})}$ (Hastie et al., 2009, Abschn. 7.2). In Abb. 2 liefern die drei Tage des Testdatensatzes, markiert durch offene Quadrate, einen Testfehler $E^{(\text{test})} = 1.3$ für die ausgewählte Hypothese $\widehat{h}^{(1)}$. Der Testfehler ist die einzige der drei Zahlen, die nicht für eine Auswahl verwendet wurde, sodass er nicht

durch die Auswahl verzerrt ist. Mit nur drei Testtagen ist er jedoch selbst eine verrauschte Schätzung des Risikos von $\widehat{h}^{(1)}$.

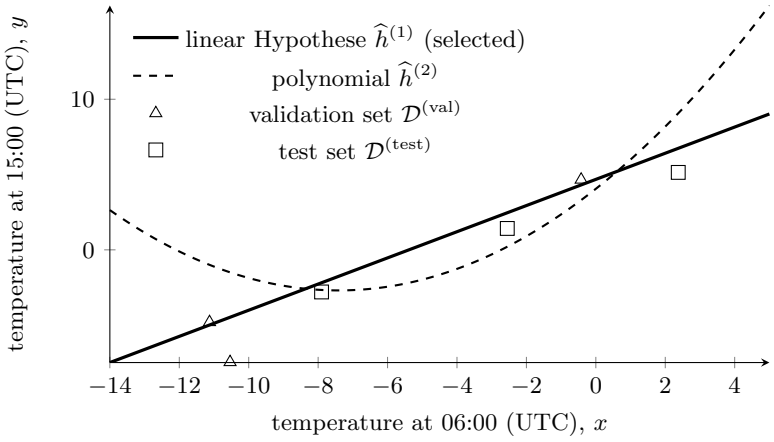


Abbildung 2: Modellauswahl mit einem Validierungsdatensatz. Die dicke durchgezogene Kurve ist die ausgewählte Hypothese $\widehat{h}^{(1)}$ und das gestrichelte Polynom die verworfene Hypothese $\widehat{h}^{(2)}$, beide aus Abb. 1; der Validierungsfehler über dem Validierungsdatensatz $\mathcal{D}^{(\text{val})}$ (offene Dreiecke) wählt die Hypothese $\widehat{h}^{(1)}$ aus (3.0 gegenüber 17.3 für das Polynom $\widehat{h}^{(2)}$). Der Testdatensatz $\mathcal{D}^{(\text{test})}$ (offene Quadrate) bewertet die ausgewählte Hypothese, die darauf einen Testfehler $E^{(\text{test})} = 1.3$ erzielt. Daten erzeugt mit `pythondemos/valset.py`

Der Validierungsfehler ist ein Mittelwert aus $|\mathcal{D}^{(\text{val})}|$ Verluste, einem pro Validierungs-Datenpunkt. Wie genau er das Risiko einer gelernten Hypothese abbildet, hängt daher vom Umfang $|\mathcal{D}^{(\text{val})}|$ ab, und ein zu kleiner Validierungsdatensatz liefert einen unzuverlässigen Validierungsfehler (Bishop, 2006, Abschn. 1.3).

Gegeben seien ein Verlust mit Werten in $[0, 1]$, eine feste Hypothese $\widehat{h}$ und ein vom Anwender gewählter Konfidenz-Parameter δ in $(0, 1)$. Die Hoeffding-Ungleichung (Shalev-Shwartz and Ben-David, 2014, Thm. 11.1) garantiert dann mit einer Wahrscheinlichkeit von mindestens $1 - \delta$, dass das Risiko von $\widehat{h}$ innerhalb eines Unsicherheitsbandes der halben Breite ϵ um den gemessenen Validierungsfehler liegt. Die Garantie setzt voraus, dass

$$|\mathcal{D}^{(\text{val})}| \geq \frac{\ln(2/\delta)}{2\epsilon^2}.$$

Abb. 3 zeigt diese untere Schranke als Funktion der halben Breite ϵ für drei Werte von δ; sie gilt für jeden Validierungsfehler, nicht nur für die des Wetterbeispiels. Eine Verengung des Bandes ist aufwendig: Eine Halbierung von ϵ vervierfacht den erforderlichen Umfang. Damit der Validierungsfehler mit $\delta = 0.1$

höchstens um $\epsilon = 0.1$ vom Risiko abweicht, genügt es, $|\mathcal{D}^{(\text{val})}| \geq 150$ Datenpunkte zurückzuhalten – der Punkt auf der gestrichelten Kurve bei $\epsilon = 0.1$ in Abb. 3 –, unabhängig vom Hypothesenraum und vom Umfang von $\mathcal{D}^{(\text{train})}$. Die Schranke gilt für eine einzelne feste Hypothese. Eine Schranke für die ausgewählte Hypothese muss für alle verglichenen Hypothesen gleichzeitig gelten, was einen größeren Validierungsdatensatz erfordert (Shalev-Shwartz and Ben-David, 2014, Thm. 11.2).

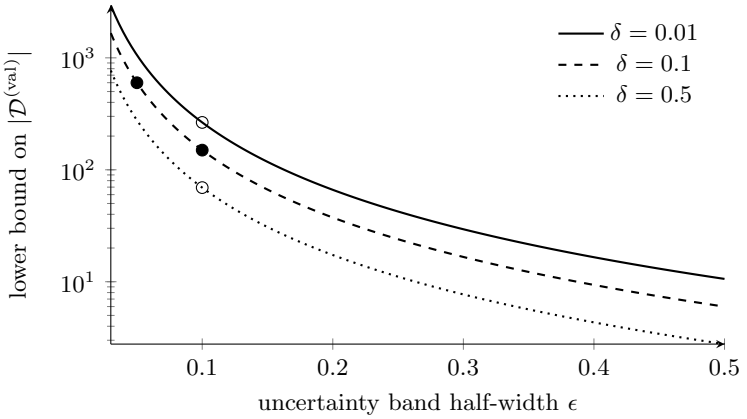


Abbildung 3: Die untere Schranke $\ln(2/\delta)/(2\epsilon^2)$ für den Umfang $|\mathcal{D}^{(\text{val})}|$ als Funktion der halben Breite ϵ des Unsicherheitsbandes um den gemessenen Validierungsfehler, für drei Werte des Konfidenz-Parameters δ: 0.01 (durchgezogene Kurve), 0.1 (gestrichelte Kurve) und 0.5 (gepunktete Kurve). Die vertikale Achse ist logarithmisch. Eine Halbierung von ϵ von 0.1 auf 0.05 vervierfacht den erforderlichen Umfang (ausgefüllte Kreise auf der gestrichelten Kurve, 150 auf 600), während eine Verringerung des Konfidenz-Parameters δ von 0.5 auf 0.01 bei $\epsilon = 0.1$ ihn um etwa das Vierfache erhöht (offene Kreise, 70 auf 265)

Die Schranke lässt sich durch Simulation überprüfen. Abb. 4 gibt an, wie oft der Validierungsfehler einer festen Hypothese um mehr als ϵ vom Risiko abweicht, wenn der Validierungsdatensatz genau die Anzahl an Datenpunkte enthält, welche die Schranke verlangt. Der dort verwendete Verlust nimmt die Werte 0 und 1 mit gleicher Wahrscheinlichkeit an, sodass seine Varianz die größtmögliche für einen Verlust mit Werten in $[0, 1]$ ist. Jeder Anteil wird über 200,000 Validierungsdatensätze gemessen. Jeder gemessene Anteil bleibt unter seinem δ sowie unter $\delta/4$, sodass der von der Schranke geforderte Umfang größer ist als der Umfang, den die gemessenen Abweichungen verlangen. Ein achtmal kleinerer Validierungsdatensatz verletzt die Garantie hingegen. Bei $\epsilon = 0.1$ und $\delta = 0.1$ verlangt die Schranke 150 Datenpunkte. Mit 19 davon liegen 36 von 100 Validierungsfehler weiter als ϵ vom Risiko entfernt.

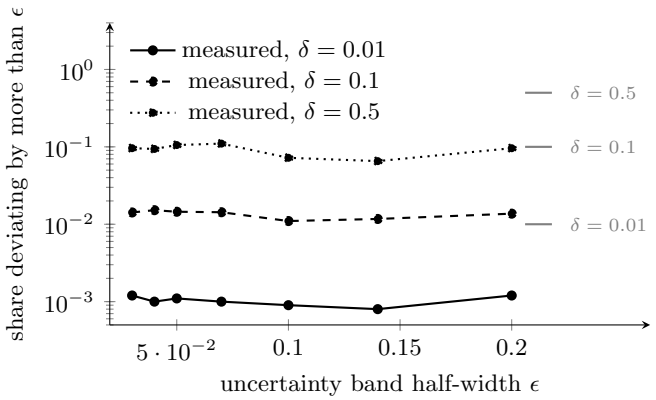


Abbildung 4: Gemessener Anteil der Validierungsdatensätze, deren Validierungsfehler um mehr als die halbe Breite ϵ vom Risiko abweicht, wenn der Validierungsdatensatz genau die $\lceil \ln(2/\delta)/(2\epsilon^2) \rceil$ Datenpunkte enthält, die Abb. 3 erfordert. Jeder Anteil wird über 200,000 Validierungsdatensätze für einen Verlust mit Werten in $\{0, 1\}$ und einem Risiko von $1/2$ gemessen. Die drei Konfidenz-Parameter aus Abb. 3 behalten ihre Darstellungsweise bei: $\delta = 0.01$ (durchgezogene Kurve), 0.1 (gestrichelte Kurve) und 0.5 (gepunktete Kurve). Die drei grauen Segmente am rechten Rand, die mit ihrem jeweiligen δ beschriftet sind, markieren die Nominalwerte. Jeder gemessene Anteil bleibt mindestens um den Faktor vier unter seinem δ. Die vertikale Achse ist logarithmisch skaliert. Daten erzeugt durch `pythondemos/valset.py`

Ein typischer ML-Ablauf teilt den verfügbaren Datensatz daher dreifach auf. Auf einem Trainingsdatensatz werden die Hypothesen gelernt, ein Validierungsdatensatz wählt unter ihnen aus, und ein Testdatensatz bewertet die ausgewählte. Der Testdatensatz wird nur einmal verwendet.

Wenn Datenpunkte knapp sind, ist diese Aufteilung kostspielig: Die obige Schranke verlangt 150 Datenpunkte bei $\epsilon = 0.1$ und $\delta = 0.1$, was ein Datensatz von 200 Tagen nicht allein für die Validierung aufwenden kann. Abhilfe schafft die k-fache Kreuzvalidierung, die den Datensatz in k Teilmengen unterteilt. Für jede Teilmenge nacheinander wird diese als Validierungsdatensatz und die übrigen $k - 1$ Teilmengen als Trainingsdatensatz verwendet. Jeder der k Validierungsfehler wird daher auf anderen Datenpunkte und für eine andere gelernte Hypothese berechnet. Jede Teilmenge liefert einen verrauschten Validierungsfehler, einen Mittelwert der Verluste über die wenigen Datenpunkte der Teilmenge, und die Mittelung der k teilmengenweisen Validierungsfehler ergibt eine zuverlässigere Schätzung des Risikos (Hastie et al., 2009, Abschn. 7.10.1; Stone, 1974). Werden die drei Tage des Trainingsdatensatzes und die drei Validierungstage aus Abb. 1 zu einem Datensatz zusammengefasst, ergeben drei Teilmengen zu je zwei Tagen teilmengenweise Validierungsfehler von 11.21, 2.80 und 3.29 mit einem Mittelwert von 5.77. Über 300 wiederholte Datensätze von

sechs Tagen beträgt die Standardabweichung des Dreifach-Mittelwerts 14.4, gegenüber 55.8 bei einer einzelnen Aufteilung, die zwei Tage zurückhält.

Jede Teilmenge erfordert einen eigenen Durchlauf des Trainings, insgesamt k anstelle des einzelnen Durchlaufs bei einer einfachen Aufteilung. Weder eine einfache Aufteilung noch die k-fache Kreuzvalidierung vermeidet das Schätzrauschen, das durch die Verwendung eines kleinen Validierungsdatensatzes entsteht. Drei Validierungs-Datenpunkte sind zu wenig, um zwei Hypothesen zuverlässig in eine Rangfolge zu bringen. Über 20,000 Ziehungen von drei Validierungstagen bevorzugt der Validierungsfehler für das Paar aus Abb. 2, dessen Risiken 4.29 Quadratgrad für die Gerade und 23.46 für das Polynom betragen, in 12 von 100 Ziehungen die Hypothese mit dem größeren Risiko. Bei zwei Geraden, deren Risiken sich um 1.68 Quadratgrad unterscheiden, geschieht dies in 24 von 100 Ziehungen. Der große Abstand im Risiko bei der Auswahl in Abb. 2 macht diese Auswahl daher nicht sicher; was der Validierungsfehler aus drei Datenpunkte liefert, ist eine Präferenz, keine Garantie.

Siehe auch: Validierung, Trainingsdatensatz, Testdatensatz, Datenpunkt, Risiko, Hypothese, Verlust, Validierungsfehler, Trainingsfehler, k-fache Kreuzvalidierung, Leave-One-Out-Kreuzvalidierung (LOO-CV), Modellauswahl.

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.

4. Stone (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), pp. 111–133.