# Testdatensatz

## 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 Testdatensatz enthält Datenpunkte, die zur Bewertung einer von einer maschinelles Lernen (ML)-Methode gelieferten Hypothese verwendet werden. Diese Punkte werden weder für das Training noch für die Modellauswahl genutzt. Ein Testdatensatz wird daher auch als Hold-out-Datensatz bezeichnet. Unter der Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme) ist sein durchschnittlicher Verlust ein unverzerrter Schätzer des Risikos der gelernten Hypothese. Eine Konzentrationsungleichung setzt die Genauigkeit dieser Schätzung in Beziehung zur Größe des Testdatensatzes. Die Schätzung bleibt nur dann gültig, wenn keine Entwurfsentscheidung vom Testdatensatz abhängt. Eine Auswahl auf Basis wiederholter Auswertung verwandelt ihn in einen Validierungsdatensatz; eine Überschneidung mit dem Trainingsdatensatz stellt eine Testdatensatzkontamination dar.

## Definition

Betrachtet sei eine maschinelles Lernen (ML)-Methode, die Wetteraufzeichnungen aus Krems an der Donau (Österreich) nutzt. Sie lernt eine Hypothese $\widehat{h}$, welche die morgige Lufttemperatur um 15:00 Uhr (UTC) aus dem heutigen Messwert zur selben Stunde vorhersagt. Ein Trainingsdatensatz wird verwendet, um eine Hypothese mittels empirische Risikominimierung (ERM) zu lernen. Ein Validierungsdatensatz dient anschließend dazu, Hypothesen aus verschiedenen Modelle, wie etwa einem lineares Modell und der polynomiale Regression, zu vergleichen. Beide Datensätze beeinflussen die ausgewählte Hypothese, sodass ihr Verlust auf ihnen üblicherweise kleiner ausfällt als auf Datenpunkte, welche die Methode nicht genutzt hat (Hastie et al., 2009, Abschn. 7.4; Shalev-Shwartz and Ben-David, 2014, Abschn. 11.2.2). Die Beurteilung von $\widehat{h}$ erfordert daher Datenpunkte, die aus keinem der beiden Datensätze stammen.

Ein Testdatensatz $\mathcal{D}^{(\text{test})}$ ist ein Datensatz aus Datenpunkte, der zur Bewertung einer von einer ML-Methode gelieferten Hypothese verwendet wird. Seine Datenpunkte fließen weder in das Training eines Modells auf dem Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ noch in die Modellauswahl auf Basis eines Validierungsdatensatzes $\mathcal{D}^{(\text{val})}$ ein (Abb. 1). Ein Testdatensatz wird daher auch als Hold-out-Datensatz bezeichnet. Manche Texte verwenden diese Bezeichnung stattdessen für den Validierungsdatensatz (Shalev-Shwartz and Ben-David, 2014,

Abschn. 11.2). Für die Temperaturvorhersage besteht der Testdatensatz aus Tagen, die in keinem der beiden anderen Datensätze enthalten sind.

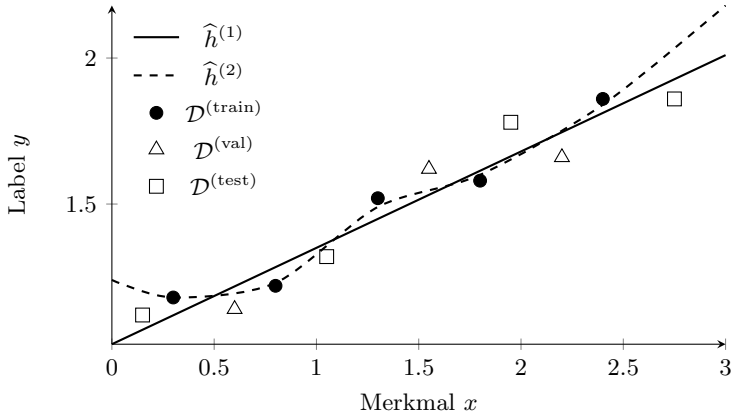


Abbildung 1: Drei Rollen für Datenpunkte. Auf dem Trainingsdatensatz (ausgefüllte Kreise) werden mittels ERM zwei Hypothesen aus unterschiedlichen Modelle gelernt: $\widehat{h}^{(1)}$ aus einem lineares Modell (durchgezogene Kurve) und $\widehat{h}^{(2)}$ aus einer polynomiale Regression dritten Grades (gestrichelte Kurve). Der Validierungsdatensatz (offene Dreiecke) dient zur Auswahl zwischen ihnen. Der Testdatensatz (offene Quadrate) fließt weder in das Training noch in die Modellauswahl ein

Angenommen, $\widehat{h}$ wurde ohne Verwendung von $\mathcal{D}^{(\text{test})}$ gelernt und ausgewählt. Ihr Testfehler ist der durchschnittliche Verlust

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

Hierbei ist $L\left(\mathbf{z}, \widehat{h}\right)$ der Verlust, eine Zahl, die den Fehler auf einem einzelnen Datenpunkt quantifiziert. Unter der Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme) sind die Punkte in $\mathcal{D}^{(\text{train})}$, $\mathcal{D}^{(\text{val})}$ und $\mathcal{D}^{(\text{test})}$ unabhängige Ziehungen aus derselben Wahrscheinlichkeitsverteilung $p(\mathbf{z})$. Der Testfehler ist dann ein unverzerrter Schätzer des Risikos von $\widehat{h}$ (Hastie et al., 2009). Das Risiko ist der erwartete Verlust auf einem Datenpunkt, der unabhängig von allen drei Datensätzen aus $p(\mathbf{z})$ gezogen wird. Da der Validierungsfehler zur Auswahl von $\widehat{h}$ herangezogen wird, unterschätzt er tendenziell das Risiko der ausgewählten Hypothese.

Die Größen der drei Datensätze sind Entwurfsentscheidungen. Unter der u.i.v.-Annahme mittelt der Testfehler über $|\mathcal{D}^{(\text{test})}|$ unabhängige Verlustwerte. Liegt der Verlust in $[0, 1]$, beschränkt eine Konzentrationsungleichung die Abweichung zwischen Testfehler und Risiko. Für ein Konfidenzniveau

$\delta \in (0, 1)$ beträgt diese Abweichung mit einer Wahrscheinlichkeit von mindestens $1 - \delta$ höchstens $\sqrt{\ln(2/\delta)/(2\,|\mathcal{D}^{(\text{test})}|)}$ (Shalev-Shwartz and Ben-David, 2014, Thm. 11.1). Löst man diese Schranke nach $|\mathcal{D}^{(\text{test})}|$ auf, erhält man die Größe des Testdatensatzes, die für eine gewünschte Abweichung und ein gegebenes Konfidenzniveau erforderlich ist.

Der Validierungsdatensatz erfordert eine Konzentrationsungleichung, die jede während der Modellauswahl verglichene Hypothese gleichzeitig abdeckt, sodass seine erforderliche Größe mit deren Anzahl wächst (Shalev-Shwartz and Ben-David, 2014, Thm. 11.2). Der Trainingsdatensatz erhält die verbleibenden Datenpunkte. Es gibt keine feste Regel, welche die Anteile vorgibt. Abb. 2 zeigt ein gängiges Beispiel: 50 Prozent für das Training sowie jeweils 25 Prozent für die Validierung und das Testen (Hastie et al., 2009, Abschn. 7.2).

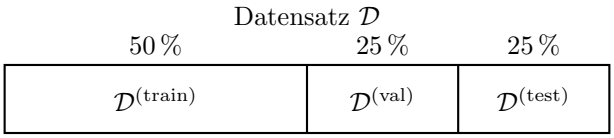


Abbildung 2: Ein Datensatz $\mathcal{D}$, aufgeteilt in den Trainingsdatensatz $\mathcal{D}^{(\text{train})}$, den Validierungsdatensatz $\mathcal{D}^{(\text{val})}$ und den Testdatensatz $\mathcal{D}^{(\text{test})}$, mit den typischen Anteilen nach Hastie et al. (2009), Abschn. 7.2

Ein Testfehler bleibt nur dann eine gültige Risikoschätzung, wenn der Testdatensatz einmalig verwendet wird, nachdem alle Entscheidungen festgelegt wurden. Wird $\mathcal{D}^{(\text{test})}$ zum Vergleich von Modelle herangezogen, verwandelt ihn dies in einen Validierungsdatensatz und seinen Fehler in einen Validierungsfehler (Goodfellow et al., 2016, Abschn. 5.3). Werden einzelne seiner Datenpunkte in den Trainingsdatensatz übernommen, liegt eine Testdatensatzkontamination vor (Kapoor and Narayanan, 2023, S. 4, L1).

Bei einem ML-Wettbewerb vergleichen die Veranstalter die von den Teilnehmenden eingereichten Modelle auf einem gemeinsamen Testdatensatz (Hastie et al., 2009, Abschn. 11.9). Die Veranstalter halten die Test-Labels unter Verschluss und berechnen den jeweiligen Fehler. Die Teilnehmenden können diese Labels nicht nutzen, um ihre Modelle zu überarbeiten. Der gemeldete Testfehler kann daher als Schätzung für das Risiko dienen, sofern die Veranstalter wiederholte Rückmeldungen auf demselben Testdatensatz begrenzen.

Synonyme: Hold-out-Datensatz, Testdaten.

Siehe auch: Trainingsdatensatz, Validierungsdatensatz, Modellauswahl, Risiko, Generalisierung, Testdatensatzkontamination, Data Leakage, ERM.

## Querverweise

## Literatur


1. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

2. Kapoor and Narayanan (2023). Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. Patterns, 4(9), pp. 100804.

3. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, 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.