Trainingsdatensatz

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 Trainingsdatensatz ist ein Datensatz, der zum Trainieren eines Modells verwendet wird: Er besteht aus Datenpunkte, für die der durch jede infrage kommende Hypothese verursachte Verlust ausgewertet werden kann, sodass Verfahren des Trainings diejenige Hypothese auswählen können, die auf ihm am besten abschneidet. Die empirische Risikominimierung (ERM) beispielsweise lernt die Hypothese, die den mittleren Verlust auf dem Trainingsdatensatz – das empirisches Risiko – minimiert, dessen Minimum der Trainingsfehler ist. Da die gelernte Hypothese so ausgewählt wird, dass sie auf dem Trainingsdatensatz gut abschneidet, kann ein kleiner Trainingsfehler irreführend sein und zu Überanpassung führen. Der Validierungsfehler auf einem zurückgehaltenen Validierungsdatensatz gibt Aufschluss darüber, ob die gelernte Hypothese auch außerhalb des Trainingsdatensatzes gute Vorhersagen liefert.

Definition

Betrachtet seien Wetteraufzeichnungen über drei Tage: Für jeden Tag ist die Lufttemperatur um 06:00 Uhr (UTC) das Merkmal und die Lufttemperatur um 15:00 Uhr (UTC) desselben Tages das Label. Das übergeordnete Ziel eines Verfahrens des maschinelles Lernen (ML)s besteht darin, eine Hypothese zu lernen (oder ein Modell zu trainieren), die nicht nur auf den für das Training verwendeten Datenpunkte einen kleinen Fehler bei der Vorhersage verursacht, sondern auf jedem für die Anwendung typischen Datenpunkt. Der Fehler der Vorhersage wird durch eine Verlustfunktion gemessen: Der Verlust ist eine – typischerweise nichtnegative – Zahl, die den Fehler einer Hypothese auf einem einzelnen Datenpunkt quantifiziert, wie etwa die quadrierte Differenz zwischen dem vorhergesagten und dem wahren Label. Wahrscheinlichkeitstheoretisch wird dieses Ziel als kleines Risiko formalisiert: der erwartete Verlust unter der Wahrscheinlichkeitsverteilung, aus der die Datenpunkte gezogen werden (Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme)). Das Verfahren muss daher die infrage kommenden Hypothesen aus seinem Hypothesenraum anhand von Datenpunkte vergleichen, für die der Verlust ausgewertet werden kann. Bei der Regression oder Klassifikation beispielsweise lässt sich der Verlust nur für Datenpunkte mit bekanntem Label berechnen. Diese Datenpunkte bilden einen Datensatz, der als Trainingsdatensatz bezeichnet wird, $\mathcal{D}^{(\text{train})} = \{\mathbf{z}^{(r)}\}_{r=1}^{m}$.

Ein Verfahren des Trainings ist eine Abbildung $\mathcal{A}$, deren Eingabe der Trainingsdatensatz und deren Ausgabe eine gelernte Hypothese ist, $\widehat{h} = \mathcal{A}(\mathcal{D}^{(\text{train})})$. Das Training wertet diese Abbildung auf dem gegebenen Trainingsdatensatz aus. Beispielsweise minimiert die empirische Risikominimierung (ERM) den mittleren Verlust auf dem Trainingsdatensatz, das empirisches Risiko, um eine Hypothese zu lernen:

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{r=1}^{m} L\left(\mathbf{z}^{(r)}, h\right).$$

Das empirisches Risiko der gelernten Hypothese ist der Trainingsfehler.

Abb. 1 zeigt einen Trainingsdatensatz der drei Tage zusammen mit zwei daraus gelernten Hypothesen: der linearen Hypothese $\widehat{h}^{(1)}$, die von der 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.

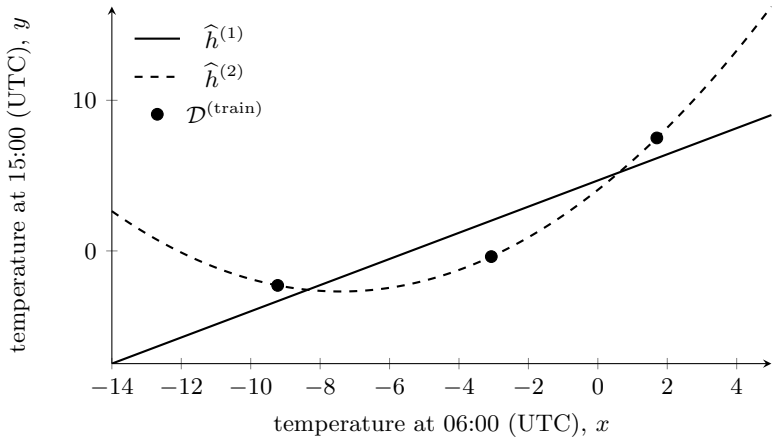

Abbildung 1: Datenpunkte, die tägliche Wetterbedingungen darstellen: Jeder Datenpunkt entspricht einem Tag, mit der Lufttemperatur um 06:00 Uhr (UTC) als Merkmal und der Lufttemperatur um 15:00 Uhr (UTC) desselben Tages als Label. Die drei ausgefüllten Kreise bilden den Trainingsdatensatz $\mathcal{D}^{(\text{train})}$. Beide Kurven sind aus $\mathcal{D}^{(\text{train})}$ gelernte Hypothesen: Die lineare Hypothese $\widehat{h}^{(1)}$ (durchgezogene Kurve) minimiert den mittleren quadratischen Fehler auf $\mathcal{D}^{(\text{train})}$ über dem lineares Modell, während das gestrichelte Polynom zweiten Grades $\widehat{h}^{(2)}$ durch alle drei Datenpunkte verläuft und einen Trainingsfehler von null erreicht. Beide verursachen einen kleinen mittleren Verlust auf $\mathcal{D}^{(\text{train})}$, doch zwischen und jenseits der drei Kreise liegen die beiden Kurven weit auseinander: Bei der höchsten dargestellten 06:00-Uhr-Temperatur unterscheiden sich die beiden vorhergesagten 15:00-Uhr-Temperaturen um rund sieben Grad. Daten erzeugt mit `pythondemos/trainset.py`

Da $\widehat{h}$ so gewählt wird, dass das empirisches Risiko klein wird, ist der Trai-

ningsfehler typischerweise kleiner als der Verlust, den dieselbe Hypothese auf Datenpunkte außerhalb von $\mathcal{D}^{(\text{train})}$ verursacht (Hastie et al., 2009, Abschn. 7.4). Ein kleiner Trainingsfehler garantiert daher keinen kleinen Verlust außerhalb des Trainingsdatensatzes. Abb. 1 veranschaulicht dies: Das Polynom $\widehat{h}^{(2)}$ erreicht einen Trainingsfehler von null, die lineare Hypothese $\widehat{h}^{(1)}$ hingegen nicht. Die Abbildung zeigt zwar keine Datenpunkte außerhalb von $\mathcal{D}^{(\text{train})}$, wohl aber, dass die beiden Kurven für Werte des Merkmals zwischen und jenseits der drei Datenpunkte des Trainings unterschiedliche Labels vorhersagen. Höchstens eine der beiden Kurven kann daher auf einem neuen Datenpunkt mit einem solchen Wert des Merkmals einen kleinen Verlust verursachen, und der Trainingsfehler gibt keinen Aufschluss darüber, welche dies ist (siehe Überanpassung).

Eine zweite Teilmenge von Datenpunkte, der Validierungsdatensatz $\mathcal{D}^{(\text{val})}$, wird daher beim Training des Modells zurückgehalten (Abb. 2): Der mittlere Verlust auf $\mathcal{D}^{(\text{val})}$, der Validierungsfehler, bewertet die gelernte Hypothese auf Datenpunkte, auf denen sie nicht trainiert wurde. Ein Vergleich der beiden Fehler ermöglicht eine Diagnose des Verfahrens des MLs: Sind beide klein, wird die gelernte Hypothese voraussichtlich auch außerhalb von $\mathcal{D}^{(\text{train})}$ gute Vorhersagen treffen. Ein kleiner Trainingsfehler in Verbindung mit einem weit darüber liegenden Validierungsfehler deutet auf Überanpassung hin; eine typische Ursache ist ein für die Anzahl der Datenpunkte in $\mathcal{D}^{(\text{train})}$ zu großer Hypothesenraum (Bishop, 2006, Abschn. 1.1; Hastie et al., 2009, Abschn. 7.2).

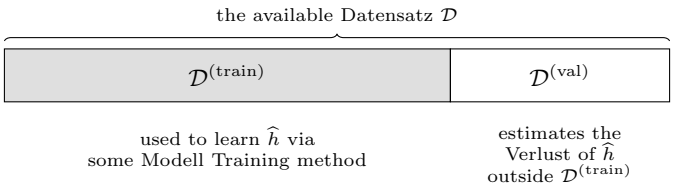

Abbildung 2: Der verfügbare Datensatz, aufgeteilt in zwei disjunkte Teilmengen. Nur die schattierte Teilmenge, der Trainingsdatensatz $\mathcal{D}^{(\text{train})}$, geht in das Verfahren des Trainings $\mathcal{A}$ ein, das eine gelernte Hypothese $\widehat{h}$ liefert; der verbleibende Validierungsdatensatz dient dazu abzuschätzen, wie sich diese Hypothese auf Datenpunkte verhält, auf denen sie nicht trainiert wurde

Welche Datenpunkte in $\mathcal{D}^{(\text{train})}$ enthalten sind, entscheidet darüber, was gelernt wird. Die Wahl des Trainingsdatensatzes ist daher für die vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI) von Bedeutung: Ein Trainingsdatensatz, in dem eine Gruppe von Datenpunkte unterrepräsentiert ist, kann zu einer gelernten Hypothese führen, die auf dieser Gruppe einen größeren Verlust verursacht – eine Quelle von Unfairness (siehe Fairness (maschinelles Lernen)) (Barocas et al., 2023; Buolamwini and Gebru, 2018; Mehrabi et al., 2021).

Siehe auch: Training, Datensatz, Validierungsdatensatz, Testdatensatz, Datenpunkt, ERM, Hypothese, Verlust, Trainingsfehler, Validierungsfehler, Hypo-

thesenraum, Überanpassung.

Querverweise

Literatur

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

2. Mehrabi et al. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), pp. Art. no. 115.

3. Barocas et al. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press.

4. Buolamwini and Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pp. 77–91.

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