Data Leakage

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

Data Leakage tritt auf, wenn eine Methode des maschinellen Lernens Merkmale verwendet, die während des Trainings, jedoch nicht zum Zeitpunkt der Vorhersage verfügbar sind. Beispielsweise deckt bei der Wettervorhersage jeder Datenpunkt ein Zeitintervall ab, und ein Merkmal ist die Temperatur eine Stunde nach dem Vorhersagezeitpunkt. In historischen Daten ist diese spätere Messung verfügbar, bei der tatsächlichen Berechnung einer Vorhersage jedoch nicht, sodass das Merkmal unzulässige Informationen preisgibt. Ein solches durchgesickertes Merkmal lässt den Verlust auf der Trainingsdatensatz und der Testdatensatz nahezu wie null aussehen; dennoch versagt die gelernte Hypothese im produktiven Einsatz, da das Merkmal zum Zeitpunkt der Vorhersage nicht reproduziert werden kann. Der Begriff wird im Sicherheitskontext auch für die unbeabsichtigte Offenlegung sensibler Informationen verwendet.

Definition

Betrachtet sei eine Methode des maschinellen Lernens, die auf einer stündlichen Temperatur-Folge trainiert wird, wobei jeder Datenpunkt ein gleitendes Fenster von drei aufeinanderfolgenden Stunden ist (Abb. 1). Das Label ist die Temperatur der mittleren Stunde, und eines der Merkmale ist die Messung eine Stunde später. In historischen Daten ist diese spätere Messung verfügbar, sodass der Verlust auf der Trainingsdatensatz und der Testdatensatz nahezu wie null aussieht; in Echtzeit muss die mittlere Stunde jedoch vorhergesagt werden, bevor die nächste Stunde gemessen wird, weshalb die gelernte Hypothese nutzlos ist. Der Fehler liegt in Data Leakage: Ein während des Trainings oder der Evaluierung verwendetes Merkmal steht nicht zur Verfügung, wenn eine Vorhersage geliefert werden muss.

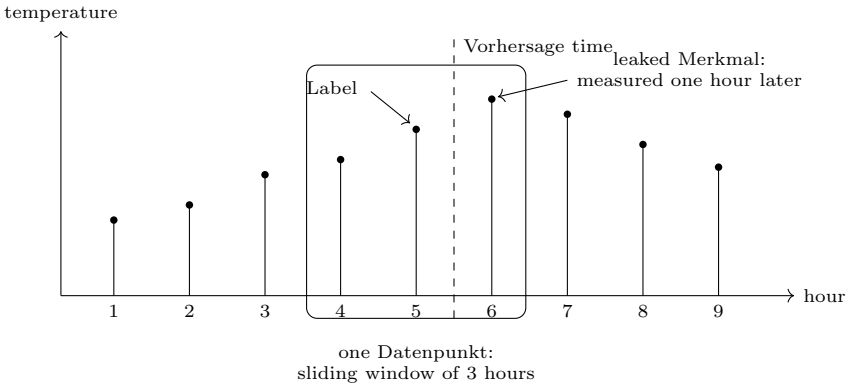


Abbildung 1: Stündliche Temperaturmessungen. Jeder Datenpunkt ist ein gleitendes Fenster von drei Stunden (Kasten); das Label ist die Temperatur der mittleren Stunde. Die Verwendung der Messung eine Stunde später als Merkmal führt zu Data Leakage: Zum Zeitpunkt der Vorhersage (gestrichelte Linie) wurde diese Stunde noch nicht gemessen, sodass das Merkmal nur für historische Daten im Training verfügbar ist

Allgemeiner ausgedrückt ist ein von Data Leakage betroffenes Merkmal eines, dessen Wert zum Zeitpunkt der Vorhersage nicht reproduziert werden kann, sei es, weil er erst später gemessen (wie oben) oder aus dem Label berechnet wird. Data Leakage wird von Kaufman et al. (2012) formalisiert, die auch Methoden zur Erkennung und Vermeidung angeben. Kapoor and Narayanan (2023) geben einen Überblick darüber als Ursache der Reproduzierbarkeitskrise in der auf maschinellem Lernen basierenden Wissenschaft. Testdatensatzkontamination, das Vorhandensein von Datenpunkte der Testdatensatz in der Trainingsdatensatz, ist eine verwandte Form von Data Leakage: Dort ist die Information, die zum Zeitpunkt des Trainings nicht verfügbar sein sollte, die Testdatensatz selbst.

Der Begriff *Data Leakage* wird in einem davon unabhängigen Sicherheitskontext auch für die unbeabsichtigte Offenlegung sensibler Informationen (Datenexfiltration aus einem trainierten System) verwendet – eine Bedeutung, die sich von der oben definierten methodischen unterscheidet.

Siehe auch: Testdatensatzkontamination, Trainingsdatensatz, Validierungsdatensatz, Testdatensatz, Generalisierung, Risiko.

Querverweise

Literatur

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

2. Kaufman et al. (2012). Leakage in Data Mining: Formulation, Detection,

and Avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), pp. 15:1–15:21.