

data leakage

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

Data leakage is the use, during the training or the evaluation of a machine learning (ML) method, of information that is unavailable at the time a prediction must be delivered. The most direct route is a feature computed from the label or from quantities that become observable only together with it. Leakage also enters through the machine learning pipeline (ML pipeline), as when a preprocessing step is computed from the entire dataset before the split into training set, validation set, and test set, or when time-ordered data points are assigned to them by random shuffling. The average loss on the validation set or test set then no longer estimates the risk of the learned hypothesis, since the evaluation grants access to information that the deployed method lacks. An error far below a stated baseline is a symptom that warrants a per-feature audit of availability at prediction time.

Definition

Consider a classifier trained to predict, from the weather measurements available on a given morning, whether the coming day will be wet. One of its features is the day’s total rainfall, copied from the same table column that defines the label. The accuracy on the training set and on the test set is then close to one, yet the deployed classifier is useless: on the morning of a new day, that day’s total rainfall has not been measured yet.

Data leakage is the use, during the training or the evaluation of a machine learning (ML) method, of information that is unavailable at the time a prediction must be delivered (Kaufman et al., 2012). Fig. 1 depicts the most direct route: a leaked feature is computed from the label, $x_{\text{leak}} = g(y)$ for some map g , or from quantities that become observable only together with the label — in the opening example, the rainfall total that defines wetness. Leakage also enters through the machine learning pipeline (ML pipeline): a preprocessing step such as data normalization computed from the entire dataset before the split passes summary statistics of the validation set and test set into the training, and splitting time-ordered data points by random shuffling lets the training read the future of the data points it is later evaluated on (Kapoor and Narayanan, 2023). The presence of test set data points themselves in the training set is the sibling defect of test set contamination.

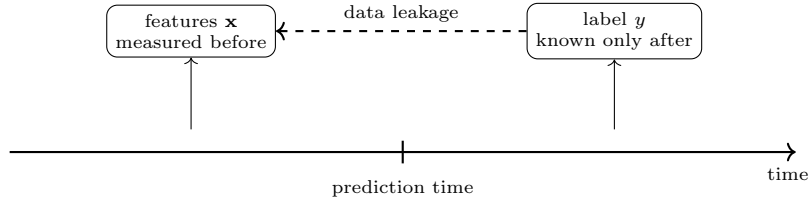


Figure 1: Features must be computable from what is available at prediction time; the label becomes known only later. Data leakage (dashed) routes information from the right of the cutoff into the features used for training or evaluation

The reported error misleads because the average loss on a validation set or test set estimates the risk of the learned hypothesis $\hat{h}$ only if the evaluation reproduces the conditions of prediction. Leakage breaks this premise: the evaluated hypothesis reads a feature that is missing at prediction time, so the reported error refers to a different prediction task than the one the method faces after deployment.

An error far below a stated baseline is a symptom of leakage — near-perfect accuracy on a task known to be hard warrants an audit before the result is trusted (Kaufman et al., 2012). The audit asks, for each feature, whether its value is computable at prediction time. Pipeline leakage is avoided by computing every preprocessing step from the training set only and by splitting time-ordered data points by time (Kapoor and Narayanan, 2023).

Cross References

test set contamination, training set, validation set, test set, risk, generalization, ML pipeline, preprocessing, baseline

References

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.