

membership inference attack

Author and editors

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

Editors: Konstantina Oliumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

A membership inference attack decides whether a given data point belonged to the training set of a published hypothesis. The signal it uses is memorization: empirical risk minimization (ERM) drives the loss down on the training set and nowhere else, so a small loss at a candidate data point is evidence of membership. An adversary queries the hypothesis and compares the loss of its prediction against a threshold. What the attack can learn is therefore bounded by the gap between training error and validation error, which makes it a privacy cost of overfitting rather than a separate failure. Differential privacy (DP) bounds the leak by construction, at a cost in accuracy.

Definition

A clinic trains a hypothesis on the records of a few hundred patients, which are personal data, and publishes it. A patient asks whether their own record was among them. A membership inference attack is the privacy attack that answers such a question: an adversary with access to $\hat{h}$ decides whether a given data point belonged to the training set (Shokri et al., 2017).

The signal is memorization. Empirical risk minimization (ERM) drives the loss down on the training set and nowhere else, so a small loss at a candidate data point is evidence of membership. The adversary queries $\hat{h}$ at the feature vector and compares the loss of its prediction against a threshold. Fig. 1 shows why the attack works: a hypothesis flexible enough to pass through five data points reproduces each of their labels exactly, while a data point it never saw is missed by a visible margin.

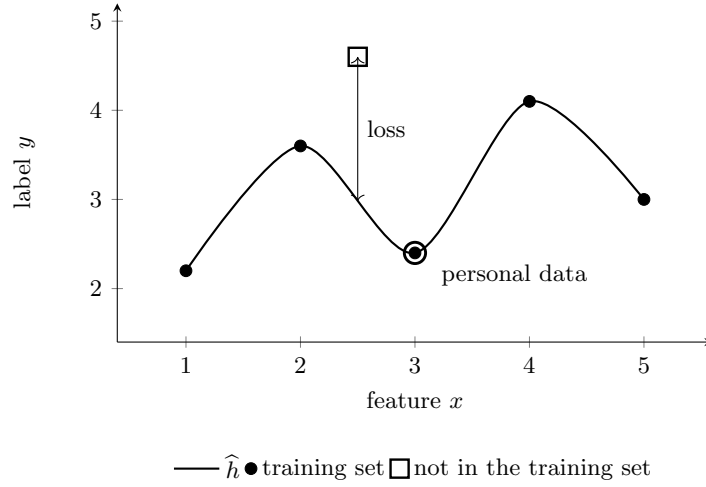


Figure 1: Why membership is visible. The hypothesis $\hat{h}$ passes through every data point of the training set, so the circled one, a record of personal data, incurs no loss. The square data point was not in the training set and is missed by the marked margin. An adversary who can evaluate the loss reads membership off that difference

The attack is therefore bounded by what separates the training error from the validation error: a hypothesis that generalizes leaks little, and one that memorizes leaks much (Shokri et al., 2017). Differential privacy (DP) bounds the leak by construction, at a cost in accuracy.

Cross References

attack, privacy attack, personal data, DP, overfitting, training error, validation error, training set, ERM

References

1. Shokri et al. (2017). Membership Inference Attacks Against Machine Learning Models. In: 2017 IEEE Symp. Secur. Privacy, pp. 3–18.