UMEÅ UNIVERSITET

NAVIGATING MODEL ANONYMITY AND ADAPTABILITY

Ayush Kumar Varshney

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Salens namn eller beteckning, byggnad BIO.A.206 - Aula Anatomica on Måndagen, 26 Maj, 2025 kl. 10:00.

Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor, Eyke Hüllermeier

Chair of Artificial Intelligence and Machine Learning, Institute of Informatics, Ludwig-Maximilians-Universität München, Germany.

Institutionen för Datavetenskap

Organization
Umeå University
Department of Computing Science

Document type
Doctoral thesis

Date of publication
05 May 2025

Author
Ayush Kumar Varshney

Title
Navigating Model Anonymity and Adaptability

Abstract

In the rapidly evolving era of Artificial Intelligence (AI), privacy-preserving techniques have become increasingly important, particularly in areas such as deep learning (DL). Deep learning models are inherently data-intensive and often rely on large volumes of sensitive personal information to achieve high performance. The increasing usage of data has raised critical concerns around data privacy, user control, and regulatory compliance. As data-driven systems grow in complexity and scale, safeguarding individual privacy while maintaining model utility has become a central challenge in the field of machine learning and deep learning. Traditional privacy-preserving frameworks focus either on protecting privacy at the database level (e.g., k-anonymity) or during the inference/output from the model (e.g., differential privacy), each introducing trade-offs between privacy and utility. While these approaches have contributed significantly to mitigating privacy risks, they also face practical limitations, such as vulnerability to inference attacks or degradation in model performance due to the added noise.

In this thesis, we take a different approach by focusing on anonymous models (i.e., models that can be generated by a set of different datasets) in model space with integral privacy. Anonymous models create ambiguity to the intruder by ensuring that a trained model could plausibly have originated from various datasets, at the same time it gives the flexibility of choosing models which do not cause much utility loss. Since exhaustively exploring the model space to find recurring models is computationally intensive, we introduce a relaxed variant called Δ-Integral Privacy, where two models are considered recurring if they are within a bounded Δ distance. Using this notion, we present practical frameworks for generating integrally private models for both machine learning and deep learning settings. We also provide probabilistic guarantees, demonstrating that under similar training environments, models tend to recur with high probability when optimized using mean samplers (e.g., SGD, Adam). These recurring models can be further utilized as an ensemble of private model that can estimate prediction uncertainty, which can be used for privacy-preserving concept drift detection in streaming data. We further extend our investigation to distributed settings, particularly Federated Learning (FL), where a central server only aggregates client-side model updates without accessing raw data. With strong empirical evidences supported by theoretical guarantees, we can claim that our frameworks with integral privacy are robust alternatives to conventional privacy-preserving methods.

In addition to generating anonymous models, this thesis also focus on developing approaches which enable users to remove their data and their influence from a machine learning model (machine unlearning) under their right-to-be-forgotten. As the demand for data removal and compliance with right-to-be-forgotten regulations intensifies, the need for efficient, auditable, and realistic unlearning mechanisms becomes increasingly urgent. Beyond regulatory compliance, machine unlearning plays a vital role in removing copyrighted content as well as mitigating security risks. In federated learning, when the clients share same feature space and participate in a huge number, we argue that if the server employs integrally private aggregation mechanism then it can plausibly deny client participation in training the global model up to certain extent. Hence, reducing the computational requirements for frequent unlearning requests. However, when there are limited number of clients with complimentary features, the server must employ unlearning mechanisms to deal with the model compression and high communication cost. This thesis shows that machine unlearning can be made efficient, effective and auditable, even in complex, distributed and generative environments. Our work spans across multiple dimensions, including privacy, auditability, computational efficiency, and adaptability across diverse data modalities and model architectures.

Keywords: Data privacy, Anonymous models, Machine Unlearning, Federated Learning, Federated Unlearning, and Generative Machine Unlearning.

Language
English

ISBN
print: 978-91-8070-671-1
PDF: 978-91-8070-672-8

ISSN
0348-0542

Number of pages
180 papers