Data reconstruction from machine learning models via inverse estimation and Bayesian inference

Agus Hartoyo[1,3][0000−0002−8285−8689], Dominika Ciupek[1][0000−0002−1411−3060], Maciej Malawski[1,2][0000−0001−6005−0243], and Alessandro Crimi[2][0000−0001−5397−6363]

[1] Sano - Centre for Computational Personalized Medicine, 30-054 Kraków, Poland
{a.hartoyo,d.ciupek,m.malawski}@sanoscience.org
https://sano.science

[2] AGH University of Krakow, 30-054 Kraków, Poland
alecrimi@agh.edu.pl

[3] Telkom University, 40257 Bandung, Indonesia

Abstract. The ability to recover original datasets from machine learning models offers profound benefits for understanding model behavior, validating models, and ensuring robustness. This paper addresses the novel challenge of reconstructing training data solely from a trained model via inverse estimation and Bayesian inference. The problem is framed as an inverse problem, where the goal is to infer the features of the original dataset that would result in the model producing observed outputs. To address this challenge, we propose a novel theoretical framework that rigorously establishes correlations between key variables in a multivariate Bayesian setting. Using a derivative-based approach, we derive expressions quantifying how variations in prior assumptions and model accuracy impact the divergence between true and estimated posteriors. Specifically, we examine the concurrent behavior of the partial derivatives of these relationships with respect to independent variables, providing deeper insights into how errors in prior assumptions and model predictions amplify deviations in data reconstruction.

The theoretical framework establishes that the fidelity of the recovered data is primarily governed by two key factors: (1) the accuracy of the assumed prior, and (2) the accuracy of the machine learning model's predictions. These expressions suggest that minimizing errors in both prior assumptions and model predictions directly reduces the divergence between true and estimated posteriors, thereby improving the quality of data reconstruction.

In the empirical part of the study, we employ Markov Chain Monte Carlo (MCMC) sampling using the Metropolis-Hastings algorithm to estimate the posterior distributions. The MCMC process generates samples by proposing candidate values based on the model's `predict_proba` function, with each candidate being accepted or rejected based on its likelihood and prior probability. This allows us to efficiently explore the posterior distribution, ensuring that the empirical posterior approximates the true posterior.

Empirical evaluation using multiple benchmark datasets (including Retinal Ganglion Cell Stimulus Types, Sirtuin6 Small Molecules, and Heart Disease datasets) and a range of machine learning algorithms (e.g., deep neural networks, gradient boosting machines, and decision tree classifiers) reveals strong correlations between the accuracy of assumed priors, model prediction accuracy, and the fidelity of data reconstruction. These empirical results corroborate our theoretical predictions, showing that improved priors and more accurate models significantly reduce the KLD between true and estimated posteriors, thereby enhancing data reconstruction quality. We further analyze the correlation plots across various datasets, showing consistent patterns of improved fidelity with better priors and models.

Furthermore, we demonstrate the practical utility of this approach by creating synthetic models trained on the reconstructed data. These synthetic models closely replicate the performance of the original models on the same test sets. The strong correlation between the accuracies of the synthetic and base models, as well as their comparable absolute performance, indicates that our reconstruction method preserves key data characteristics. These findings underscore the robustness of our inverse estimation method and its potential to generate high-fidelity synthetic datasets that retain essential properties of the original data.

This work makes several contributions to advancing theoretical and practical techniques for data reconstruction and model introspection within machine learning. First, it provides a rigorous theoretical framework that bridges the gap between model performance and data fidelity, offering valuable insights into how priors and model accuracy influence data recovery. Second, the use of MCMC sampling in the empirical evaluation ensures robust posterior estimation, effectively capturing the posterior distributions across high-dimensional datasets. Finally, the empirical validation across diverse datasets and algorithms highlights the generality of the approach, suggesting that the method can be extended to various machine learning models and datasets. Our findings set the groundwork for future studies exploring synthetic data generation, model analysis, and introspection in broader machine learning contexts.

Keywords: Data reconstruction · Machine learning · Inverse estimation · Bayesian inference · Synthetic data generation · Synthetic models · MCMC sampling.