Article

Evaluation Metrics for Machine Unlearning

Cassandra Lindstrom [*]

Posted Date: 24 September 2024

doi: 10.20944/preprints202409.1925.v1

Keywords: machine unlearning; privacy; finance; graph neural network

Article

Evaluation Metrics for Machine Unlearning

Cassandra Lindstrom

Affiliation 1; e-mail@e-mail.com

Abstract: The evaluation of machine unlearning has become increasingly significant as machine learning systems face growing demands for privacy, security, and regulatory compliance. This paper focuses on categorizing and analyzing evaluation metrics for machine unlearning, essential for assessing the success of unlearning processes. We divide the metrics into three key dimensions: unlearning effectiveness, unlearning efficiency, and model utility. Unlearning effectiveness examines the degree to which data is removed from the model, utilizing methods such as data removal completeness, privacy leakage detection, and perturbation analysis to ensure thorough data erasure. Unlearning efficiency considers metrics like time to unlearn, computational cost, and scalability, which are crucial for maintaining system performance in real-time environments. Model utility metrics, including accuracy retention, robustness, and fairness, ensure that unlearning does not compromise the model's predictive capabilities. Through this categorization, we present a comprehensive framework for evaluating machine unlearning, providing a foundation for developing unlearning techniques that balance privacy, performance, and regulatory needs across diverse industries, particularly finance.

Keywords: machine unlearning; privacy; finance; graph neural network

1. Introduction

Machine unlearning has emerged as a critical area of research, addressing the need to remove specific data points or entire datasets from trained models without requiring retraining from scratch. With growing concerns over data privacy, such as those driven by regulations like the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA), unlearning techniques provide a solution for erasing data in compliance with legal and ethical mandates. However, simply removing data is insufficient without a robust evaluation framework to measure the success of unlearning. This paper categorizes and discusses the evaluation metrics for machine unlearning under three key dimensions: unlearning effectiveness, unlearning efficiency, and model utility. By exploring these aspects, we aim to define a comprehensive methodology for evaluating unlearned methods and demonstrate their importance with real-world applications.

2. Literature Review

[1] and [2] has provided a summary of the most relevant research on federated unlearning. [3] has proved that GNN is very successful in representing complex relationships in machine learning. When GNN framework is combined with treasury [4] and crypto trading [5], it becomes very powerful. There are many successful academic and commercial models for machine unlearning. [6] [7] [8] use unique methods called PROJECTOR and GRAPHEDITOR. In PROJECTOR [6], it uses projection techniques to remove specific nodes, ensuring no trace in the model parameters. In GRAPHEDITOR [7], it manages dynamic graphs and enables node/edge deletion and feature updates. The next major categories is the guaranteeing certified unlearning. The most famous is the CEU framework [9] [10], which introduces a single-step update methodology for the removal of specific edges [1].

[2] and [1] summarize approximate unlearning into two classes: data-driven approximation and model-driven approximation. Both approaches aim to remove the influence of specific data points

from machine learning models, but they differ in their methodologies and the specific components of the learning system they target.

While machine unlearning has been widely studied with various models, metric to evaluate machine unlearning has rarely been discussed. [11], [12] and [13] uses relearn time as the main metric to evaluate the unlearning effectiveness. The re-learn time is the number of epochs required for the unlearned model to regain the same accuracy as before. Most other literature indicates that using the re-learn time solely based on reaching or surpassing the original accuracy would be misleading.

3. Unlearning Effectiveness

Unlearning effectiveness measures how well the model has forgotten the target data. The ultimate goal is for the model to behave as if the unlearned data never existed in its training process. Several metrics have been proposed to assess unlearning effectiveness, each focusing on the completeness and integrity of the unlearning process. Below, we explore the main metrics in detail.

3.1. Data Removal Completeness

Data removal completeness evaluates the degree to which the impact of the unlearned data has been eradicated from the model. This can be quantitatively assessed using influence functions, which help in understanding how much a particular data point affects the model's predictions. [14] developed influence functions to estimate the importance of a training example in determining the model's output. This technique can be adapted for unlearning, wherein the goal is to ensure that the influence of unlearned data is diminished or eliminated.

Another common method is to test the model's predictions on the unlearned data points after the unlearning process. If the model behaves similarly to how it would if it had never encountered those data points, the unlearning can be considered successful. For example, if a model trained on a medical dataset is required to forget sensitive patient data, testing it on those records should show no trace of their previous influence on predictions.

3.2. Privacy Leakage

Privacy leakage refers to how much residual information about the unlearned data can still be extracted from the model. Even after unlearning, there is a possibility that sensitive information remains embedded in the model's weights or parameters, a phenomenon that poses significant privacy risks. Membership inference attacks (Shokri et al., 2017) are a useful tool for evaluating privacy leakage. These attacks try to infer whether specific data points were part of the model's training set by observing the model's outputs on these points. A well-unlearned model should make it impossible for attackers to distinguish whether the data point was ever in the training set.

Membership inference is particularly important in scenarios like social media platforms, where users might request the deletion of personal data. A system that poorly unlearns data could still leak private user information through queries, thus violating user privacy despite apparent compliance with deletion requests.

3.3. Influence Reduction

A more fine-grained measure of unlearning effectiveness is influencing reduction. This metric evaluates the extent to which the gradients associated with the unlearned data points have been neutralized. Researchers often use gradient-based methods to calculate the contribution of each data point to the model's parameter updates. By comparing the gradient profiles before and after unlearning, practitioners can determine whether the data has been fully neutralized from the model’s learning trajectory.

3.4. Perturbation Analysis

Perturbation analysis offers another perspective on unlearning effectiveness. In this approach, small perturbations are introduced to the unlearned data, and the model’s response is examined. If

doi:10.20944/preprints202409.1925.v1

the model’s predictions shift significantly in response to minor changes, it indicates that the data still exerts influence on the model. This can be especially relevant in machine learning models used for high-stakes decision-making, such as credit scoring models, where it is vital that the removed data has no lingering effect on future predictions.

The key idea is to monitor how sensitive the model's predictions are to these minor changes. If the model still reacts significantly to the perturbed data, it indicates that the original data has not been fully removed from the model’s memory. For example, if a minor change in the customer’s income leads to a noticeable shift in the predicted credit score, the model may still retain residual knowledge of the forgotten data. In a successful unlearning process, the model should show minimal or no changes in predictions when confronted with such perturbations, implying that it has genuinely forgotten the data. This technique is useful for high-stakes applications, like healthcare or finance, where lingering effects of sensitive data could lead to privacy violations or biased predictions, undermining both regulatory compliance and ethical standards.

While perturbation analysis has been broadly discussed in machine learning contexts, specific documented examples of its use in the financial industry to evaluate machine unlearning are still emerging. However, the concept can be readily applied to financial models that rely on sensitive personal or transactional data. A relevant hypothetical example could involve a machine learning model used in credit scoring or fraud detection.

Consider a financial institution that uses a machine learning model to assess creditworthiness by analyzing customer data, such as income, debt levels, and transaction history. If a customer requests that their data be removed, perhaps due to GDPR compliance, the model needs to undergo unlearning. In such a case, perturbation analysis could be used to verify whether the customer’s data has been fully unlearned.

Let’s say a credit scoring model uses features like income, loan history, and payment behavior to predict a credit score. After a customer requests data removal, perturbation analysis would involve making small changes to the customer’s financial data, such as adjusting their income by a few percentage points or altering transaction patterns slightly. The model's credit score predictions are then analyzed to see if these small perturbations result in significant shifts in the score.

If, after unlearning, the model’s predictions are still sensitive to these minor adjustments in the customer’s data, it indicates that the unlearning process was incomplete. For instance, if increasing the income of the removed customer by 5% still changes the predicted credit score significantly, it means the model has retained some knowledge of that individual’s profile. Conversely, if the model shows no significant response to these changes, it suggests that the data has been properly forgotten.

Similarly, in fraud detection, financial models analyze transaction data to identify unusual patterns that may indicate fraudulent activities. After unlearning the transaction history of a particular customer, perturbation analysis can be applied by making slight changes to the removed transaction records (e.g., changing the transaction amount or time) and checking if the model still flags them as fraud or non-fraud. If the model’s predictions remain unchanged despite the perturbations, it suggests the customer’s data has been successfully unlearned.

While documented uses of perturbation analysis for machine unlearning in the financial industry are still developing, the technique is gaining relevance as data privacy laws, like GDPR, necessitate secure and verifiable data removal. Financial institutions could increasingly employ perturbation analysis in unlearning scenarios to ensure compliance with data protection regulations and to maintain customer trust by guaranteeing that sensitive financial data is genuinely forgotten from their models.

4. Unlearning Efficiency

While ensuring that the data is forgotten is paramount, the efficiency of the unlearning process is equally important. In many real-world applications, models are frequently updated, and retraining from scratch is computationally prohibitive. Therefore, metrics that measure the resource efficiency of unlearning techniques are crucial. [15], [16], [17] [18] and [19] use unlearn speed to access the

doi:10.20944/preprints202409.1925.v1

unlearning efficiency. It measures the time difference between unlearning and naïve restraining. The larger the difference, the fast the system can restore its privacy, security and utility.

Time to unlearn is the most straightforward efficiency metric and measures the duration it takes to complete the unlearning process. In practice, this metric is especially relevant in large-scale systems where unlearning requests may be frequent. In a financial services application, for instance, regulators might require firms to remove sensitive data from models that drive algorithmic trading. If the unlearning process takes too long, it could lead to delays in compliance and significant operational risks. You can find examples of using time to unlearn at [11], [12] and [13].

Methods like approximate unlearning [14] and federated unlearning aim to reduce this time by only modifying parts of the model directly related to the unlearned data rather than retraining the entire model from scratch. The more efficient the unlearning, the better suited the approach is for practical, large-scale applications.

Beyond time, computational cost refers to the hardware and energy resources consumed during unlearning. High-dimensional models, especially deep neural networks, require considerable computational power, and reducing this cost is critical in environments where models are continually updated. Federated learning scenarios, for example, demand low-latency unlearning processes that can run efficiently on decentralized devices with limited computational resources. In such cases, unlearning should incur minimal computational overhead, making lightweight methods like "local unlearning" in federated systems highly desirable. To quantify computational cost, researchers track metrics such as memory usage, energy consumption, and processing time on GPUs or CPUs. These metrics are especially relevant for large-scale, cloud-based models, where cost-efficient operations are essential for both economic and environmental sustainability.

Scalability measures how well the unlearning process adapts to growing amounts of data or increasingly complex models. Efficient unlearning should maintain performance even as the model scales up in size. For instance, an image recognition model used in self-driving cars might need to unlearn specific objects or features. If the unlearning technique cannot handle large-scale model updates without a significant drop in performance, it becomes impractical for real-world use. Techniques like machine learning pruning and efficient gradient updates are often employed to maintain scalability while minimizing computational costs.

5. Model Utility

Once data has been unlearned, it's critical to ensure that the remaining model continues to perform effectively. A key challenge in machine unlearning is maintaining the model's utility, i.e. the ability to generate accurate predictions on unseen data without the unlearned data points. Several metrics help in assessing model utility post-unlearning. Research in [20], [21] and [15] emphasize that performance of the unlearning model should be consistent before and after the process. By removing data from the trained model may deteriorate its performance, which should be avoided. Therefore, it is motivated to evaluate the utility of the unlearned model to ensure it is functionable after applying model.

Accuracy retention refers to how much of the model's original predictive accuracy is preserved after the unlearning process. Ideally, the unlearning procedure should affect only the predictions related to the unlearned data while leaving the model's overall accuracy intact. For example, if a financial forecasting model unlearns data from a certain time period, the model should still accurately predict market trends from other periods. Various approaches, such as selective retraining and incremental learning, have been proposed to ensure minimal loss in model accuracy. Selective retraining focuses only on the parts of the model influenced by the unlearned data, thus preserving the model's knowledge of the remaining dataset.

Robustness refers to the model's stability and reliability after the unlearning process. If the model becomes too sensitive or brittle following data removal, it indicates that the unlearning process has compromised its generalization capabilities. One way to assess robustness is through adversarial testing, where the model is exposed to slightly perturbed inputs to check whether its predictions remain consistent. A robust model should be able to maintain performance across different input variations, even after unlearning.

Fairness is another critical metric in evaluating model utility post-unlearning. The removal of data can introduce unintended biases or exacerbate existing ones. This is particularly important in applications where fairness is paramount, such as hiring algorithms or lending decisions. For instance, if a machine learning model used for job recruitment unlearns data from a specific demographic group, the remaining model should not display biased outcomes against that group. Techniques like fairness-aware unlearning focus on ensuring that unlearning does not compromise the model's fairness, making it crucial for applications where equitable outcomes are a legal and ethical necessity.

Machine learning models are frequently updated with new data. Consistency across updates measures how smoothly the model integrates new data without significant shifts in behavior after unlearning. For example, in recommender systems, unlearning user preferences for specific products should not cause the system to lose its ability to make relevant recommendations for other users. A consistent model maintains its behavior across various updates and unlearning events, thus preserving its overall reliability.

6. Conclusions

This paper presents a structured framework for evaluating machine unlearning through three distinct categories of metrics: unlearning effectiveness, efficiency, and model utility. Evaluation metrics play a critical role in determining the success of machine unlearning techniques, and our framework emphasizes their importance in ensuring that the unlearning process is both thorough and efficient, without compromising the performance of the model. Unlearning effectiveness metrics, such as data removal completeness, privacy leakage detection, and perturbation analysis, help verify that the data has been genuinely erased from the model. Metrics for unlearning efficiency, including time to unlearn and computational cost, ensure that the process is computationally feasible, even in large-scale, dynamic environments like financial markets. Finally, model utility metrics assess how well the model maintains its accuracy, robustness, and fairness after unlearning, ensuring that the removal of data does not degrade the model's overall performance.

By focusing on these evaluation metrics, this paper highlights their central role in developing and refining unlearning algorithms, especially in sensitive applications where privacy is paramount. In the financial industry, for instance, effective evaluation metrics can help ensure that credit scoring models or fraud detection systems can forget specific user data while continuing to function accurately and efficiently. As machine unlearning techniques evolve, the metrics presented here will guide the development of future algorithms that not only meet regulatory and privacy demands but also preserve the utility and efficiency of machine learning models. Future research should continue to refine these metrics, addressing trade-offs between unlearning effectiveness and efficiency, particularly in more complex, decentralized, or federated learning environments.

References

1. N. Li, C. Zhou, Y. Gao, H. Chen, A. Fu, Z. Zhang and Y. Shui, "Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects," *arXiv preprint arXiv:2403.08254*, 2024.
2. T. Shaik, X. Tao, H. Xie, L. Li, X. Zhu and Q. Li, "Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy," *arXiv preprint arXiv:2305.06360*, 2023.
3. Z. Wang, Y. Zhu, Z. Li, Z. Wang, H. Qin and X. Liu, "Graph neural network recommendation system for football formation," *Applied Science and Biotechnology Journal for Advanced Research*, vol. 3, p. 33–39, 2024. doi: 10.5281/zenodo.12198843
4. Z. Li, B. Wang and Y. Chen, "Incorporating economic indicators and market sentiment effect into US Treasury bond yield prediction with machine learning," *Journal of Infrastructure, Policy and Development*, vol. 8, p. 7671, 2024.

5. Z. Li, B. Wang and Y. Chen, "A Contrastive Deep Learning Approach to Cryptocurrency Portfolio with US Treasuries," *Journal of Computer Technology and Applied Mathematics*, vol. 1, pp. 1-10, 2024.
6. W. Cong and M. Mahdavi, "Efficiently forgetting what you have learned in graph representation learning via projection," in *International Conference on Artificial Intelligence and Statistics*, 2023.
7. W. Cong and M. Mahdavi, "Grapheditor: An efficient graph representation learning and unlearning approach".
8. Y. Wei, X. Gu, Z. Feng, Z. Li and M. Sun, "Feature Extraction and Model Optimization of Deep Learning in Stock Market Prediction," *Journal of Computer Technology and Software*, vol. 3, 2024.
9. E. Chien, C. Pan and O. Milenkovic, "Certified graph unlearning," *arXiv preprint arXiv:2206.09140*, 2022.
10. K. Wu, J. Shen, Y. Ning, T. Wang and W. H. Wang, "Certified edge unlearning for graph neural networks," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.
11. V. S. Chundawat, A. K. Tarun, M. Mandal and M. Kankanhalli, "Zero-shot machine unlearning," *IEEE Transactions on Information Forensics and Security*, vol. 18, p. 2345–2354, 2023.
12. A. Golatkar, A. Achille, A. Ravichandran, M. Polito and S. Soatto, "Mixed-privacy forgetting in deep networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.
13. A. K. Tarun, V. S. Chundawat, M. Mandal and M. Kankanhalli, "Fast yet effective machine unlearning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
14. A. Ginart, M. Guan, G. Valiant and J. Zou, "Making AI Forget You: Data Deletion in Machine Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
15. Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE symposium on security and privacy*, 2015.
16. Z. Izzo, M. A. Smart, K. Chaudhuri and J. Zou, "Approximate data deletion from machine learning models," in *International Conference on Artificial Intelligence and Statistics*, 2021.
17. C. Guo, T. Goldstein, A. Hannun and L. Van Der Maaten, "Certified Data Removal from Machine Learning Models," in *International Conference on Machine Learning*, 2020.
18. Y. Wu, E. Dobriban and S. Davidson, "Deltagrad: Rapid retraining of machine learning models," in *International Conference on Machine Learning*, 2020.
19. J. Brophy and D. Lowd, "Machine unlearning for random forests," in *International Conference on Machine Learning*, 2021.
20. L. Graves, V. Nagisetty and V. Ganesh, "Amnesiac machine learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
21. A. Sekhari, J. Acharya, G. Kamath and A. T. Suresh, "Remember what you want to forget: Algorithms for machine unlearning," *Advances in Neural Information Processing Systems*, vol. 34, p. 18075–18086, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.