Published on *The National Law Review* https://natlawreview.com

Technological and Legal Defenses Against Privacy Attacks on Machine Learning Models

Article By:

Erin Jane Illman

Sinan Pismisoglu

Machine learning (ML) models are a cornerstone of modern technology, allowing models to learn from and make predictions based on vast amounts of data. These models have become integral to various industries in an era of rapid technological innovation, driving unprecedented advancements in automation, decision-making, and predictive analysis. The reliance on large amounts of data, however, raises significant concerns about privacy and data security. While the benefits of ML are manifold, they are not without accompanying challenges, particularly in relation to privacy risks. The intersection of ML with privacy laws and ethical considerations forms a complex legal landscape ripe for exploration and scrutiny. This article will explore privacy risks associated with ML, privacy in the context of California's privacy legislation, and countermeasures to these risks.

Privacy Attacks on ML Models

There are several distinct types of attacks on ML models, four of which target the privacy of protected information**.**

1. **Model Inversion Attacks** constitute a sophisticated privacy intrusion where an attacker endeavors to reconstruct original input data by reverse-engineering a model's output. A practical illustration might include an online service recommending films based on previous viewing habits. Through this method, an attacker could deduce an individual's past movie choices, uncovering private information such as race, religion, nationality, and gender. This type of information can be used to perpetuate social engineering schemes (or the use of known information to build (sham) trust and ultimately extract sensitive data from an individual). In other contexts, such an attack on more sensitive targets can lead to substantial privacy breaches, exposing information such as medical records, financial details, or personal preferences. This exposure underscores the importance of robust safeguards and understanding the underlying ML mechanisms.
2. **Membership Inference Attacks** involve attackers discerning whether an individual's personal information was utilized in training a specific algorithm, such as a recommendation system or health diagnostic tool. An analogy might be drawn to an online shopping platform, where an attacker infers that a person was part of a customer group based on recommended

products, thereby learning about shopping habits or even more intimate details. These types of attacks harbor significant privacy risks, extending across various domains like healthcare, finance, and social networks. The accessibility of Membership Inference Attacks, often not requiring intricate knowledge of the target model's architecture or original training data, amplifies their threat. This reach reinforces the necessity for interdisciplinary collaboration and strategic legal planning to mitigate these risks.

3. **Reconstruction Attacks** aim to retrieve the original training data by exploiting the model's parameters. Imagine a machine learning model as a complex, adjustable machine that takes in data (like measurements, images, or text) and produces predictions or decisions. The parameters are the adjustable parts of this machine that are fine-tuned to make it work accurately. During training, the machine learning model adjusts these parameters so that it gets better at making predictions based on the data it is trained on. These parameters hold specific information about the data and the relationships within the data. A Reconstruction Attack exploits these parameters by analyzing them to work backward and figure out the original training data. Essentially, the attacker studies the settings of the machine (parameters) and uses them to reverse-engineer the data that was used to set those parameters in the first place.
   For instance, in healthcare, ML models are trained on sensitive patient data, including medical histories and diagnoses. These models fine-tune internal settings or parameters, creating a condensed data representation. A Reconstruction Attack occurs when an attacker gains unauthorized access to these parameters and reverse-engineers them to deduce the original training data. If successful, this could expose highly sensitive information, such as confidential medical conditions.
4. **Attribute Inference Attacks** constitute attempts to guess or deduce specific private attributes, such as age, income, or health conditions, by analyzing related information. Consider, for example, a fitness application that monitors exercise and diet. An attacker employing this method might infer private health information by analyzing this data. Such attacks have the potential to unearth personal details that many would prefer to remain confidential. The ramifications extend beyond privacy, with potential consequences including discrimination or bias. The potential impact on individual rights and the associated legal complexities emphasizes the need for comprehensive legal frameworks and technological safeguards.

ML Privacy under California Privacy Laws

Organizations hit by attacks targeting ML models, like the ones described, could find themselves directly violating California laws concerning consumer data privacy. The California Consumer Privacy Act (CCPA) enshrines the right of consumers to request and obtain detailed information regarding the personal data collected and processed by a business entity. This fundamental right, however, is not without potential vulnerabilities. Particularly, Model Inversion Attacks, which reverse-engineer personal data, pose a tangible risk. By enabling unauthorized access to such information, these attacks may impede or compromise the exercise of this essential right. The CCPA further affords consumers the right to request the deletion of personal information, mandating businesses to comply with such requests. Membership Inference Attacks can reveal the inclusion of specific data within training sets, potentially undermining this right. The exposure of previously deleted data could conflict with the statutory obligations under the CCPA. To safeguard consumers' personal information, the CCPA also obligates businesses to implement reasonable security measures. Successful attacks on ML models, such as those previously described, might be construed as a failure to fulfill this obligation. Such breaches could precipitate non-compliance, attracting potential legal liabilities.

The California Privacy Rights Act (CPRA) amends the CCPA and introduces rigorous protections for Sensitive Personal Information (SPI). This category encompasses specific personal attributes, including, but not limited to, financial data, health information, and precise geolocation. Attribute Inference Attacks, through the unauthorized disclosure of sensitive attributes, may constitute a direct contravention of these provisions, signifying a significant legal breach. Focusing on transparency, the CPRA sheds light on automated decision-making processes, insisting on clarity and openness. Unauthorized inferences stemming from various attacks could undermine this transparency, thereby impacting consumers' legal rights to comprehend the underlying logic and implications of decisions that bear upon them. Emphasizing responsible data stewardship, the CPRA enforces data minimization and purpose limitation principles. Attacks that reveal or infer personal information can transgress these principles, manifesting potential excesses in data collection and utilization beyond the clearly stated purposes by exposing data that is not relevant for the intended purposes of the models. For example, an attacker could use a model inversion attack to reconstruct the face image of a user from their name, which is not needed for the facial recognition model to function. Moreover, an attacker could use an attribute inference attack to disclose the political orientation or sexual preference of a user from their movie ratings, which is not stated or agreed by the user when using the movie recommendation model.

Mitigating ML Privacy Risk

Considering California privacy laws, as well as other state privacy laws, legal departments within organizations must develop comprehensive and adaptable strategies. These must encompass clear and enforceable agreements with third-party vendors, establish internal policies reflecting state law mandates, and conduct data protection impact assessments and actionable incident response plans to mitigate potential breaches. Continuous monitoring of evolving legal landscapes at the state and federal level ensures alignment with existing obligations and prepares organizations for future legal developments.

The criticality of technological defenses cannot be overstated. Implementing safeguards such as advanced encryption, stringent access controls, and other measures forms a robust shield against privacy attacks and legal liabilities. More broadly, the intricacies of complying with the CCPA and CPRA require an in-depth understanding of technological functionalities and legal stipulations. A cohesive collaboration among legal and technical experts and other stakeholders, such as business leaders, data scientists, privacy officers, and consumers, is essential to marry legal wisdom to technological and practical acumen. Interdisciplinary dialogue ensures that legal professionals comprehend the technological foundations and practical use case of ML while technologists grasp the legal parameters and ethical considerations embedded in the CCPA and CPRA**.**

Staying ahead of technological advancements and legal amendments requires constant vigilance. The CPRA's emphasis on transparency and consumer rights underscores the importance of effective collaboration, adherence to industry best practices, regular risk assessments, and transparent engagement with regulators and stakeholders, and other principles, i.e., accountability, fairness, accuracy, and security that govern artificial intelligence. Organizations should adopt privacy-by-design and privacy-by-default approaches that embed privacy protections into the design and operation of ML models.

The Future of ML Privacy Risks

The intersection of technology and law, as encapsulated by privacy attacks on ML models, presents a vibrant and complex challenge. Navigating this terrain in the era of the CCPA and CPRA demands

an integrated, meticulous approach, weaving together legal strategies, technological safeguards, and cross-disciplinary collaboration.

Organizations stand at the forefront of this evolving landscape, bearing the responsibility to safeguard individual privacy and uphold legal integrity. The path forward becomes navigable and principled by fostering a culture that embraces compliance, vigilance, and innovation and by aligning with the specific requirements of the CCPA and CPRA. The challenges are numerous and the stakes significant, yet with prudent judgment, persistent effort, and a steadfast dedication to moral values, triumph is not merely attainable, it becomes a collective duty and a communal achievement.

Listen to this post

National Law Review, Volume XIII, Number 229