УДК 004.89

ИСПОЛЬЗОВАНИЕ АТАК MEMBERSHIP INFERENCE ДЛЯ ОЦЕНКИ ПРИВАТНОСТИ МОДЕЛЕЙ

Морэтти Л., студент направления Industrial Production Engineering Bachelor's degree program, II курс
Миланский технический университет, г. Милан, Италия

Использование атак типа Membership Inference для оценки приватности моделей машинного обучения представляет собой одно из наиболее исследованных и практически значимых направлений в области анализа уязвимостей интеллектуальных систем. В контексте растущего внимания к вопросам конфиденциальности и нормативной совместимости (например, в рамках регламентов GDPR или HIPAA), способность определить, насколько сильно модель подвержена утечке информации о своей обучающей выборке, становится критически важной. Атаки Membership Inference (MIA) позволяют формализовать и количественно измерить степень приватности модели, проверяя, можно ли по поведению модели определить, использовался ли конкретный образец данных в её обучении. Такие атаки служат как угрозой безопасности, так и диагностическим инструментом, позволяющим разработчикам и исследователям выявлять избыточную зависимость модели от отдельных обучающих примеров [1].

Принцип, лежащий в основе атак MIA, заключается в анализе различий в откликах модели на примеры из обучающей выборки (члены – members) и на ранее не виденные ею данные (не-члены – non-members). Интуитивно, если модель была переобучена на конкретных данных, она будет демонстрировать более уверенные, «острые» предсказания на них, в отличие от примеров, с которыми она не сталкивалась во время обучения. Злоумышленник или исследователь может использовать это различие, чтобы построить классификатор, предсказывающий принадлежность конкретного входа к тренировочному множеству. В простейшем варианте такой классификатор может опираться лишь на уровень уверенности модели (например, максимальное значение softmax-выхода), но на практике часто применяются более сложные схемы с использованием полного распределения вероятностей, градиентов, потерь или скрытых представлений модели [2].

Для оценки приватности через MIA строится целевая модель, которую атакующий может использовать как «чёрный ящик» (black-box), либо как «белый ящик» (white-box), в зависимости от степени доступа. В black-box-сценарии атакующий не знает внутренней архитектуры модели и может лишь

отправлять запросы и наблюдать за выходами. Несмотря на кажущиеся ограничения, исследования показывают, что даже в таком режиме многие модели подвержены успешным атакам, особенно если они сильно переобучены или обучены без использования методов регуляризации. В white-box-сценариях, где доступна информация о весах и внутреннем состоянии модели, атакующий может анализировать более тонкие аспекты, включая значения градиентов, внутренние активации и скорость сходимости на конкретных примерах. Это делает атаку значительно более мощной и позволяет выявлять глубокие утечки информации.

Методология MIA часто включает построение вспомогательных моделей – атакующих (attack models), обучающихся на выходах целевой модели по специально сгенерированному датасету, включающему как тренировочные, так и нетренировочные примеры. Эта процедура может быть полностью автоматизирована и использовать данные, аналогичные тем, на которых обучалась основная модель, либо даже сгенерированные искусственно. Полученная attack model затем тестируется на ранее невиданных примерах, что позволяет определить, насколько точно можно по поведению модели угадать, был ли конкретный пример частью её обучения. Высокая точность такой атаки указывает на низкий уровень приватности модели и её потенциальную уязвимость к утечке информации [3].

Ключевые метрики, применяемые для оценки результатов атак MIA, включают точность атакующего классификатора, AUC (площадь под ROC-кривой), а также membership advantage – разницу между вероятностью правильной идентификации члена и не-члена. Эти метрики позволяют проводить сравнительный анализ различных архитектур, стратегий обучения и методов защиты с точки зрения их устойчивости к подобным атакам. В частности, модели, обученные с использованием дифференциальной приватности, dropout, регуляризации по норме весов или техники ранней остановки, как правило, демонстрируют значительно меньшую уязвимость [4].

Важно понимать, что атаки Membership Inference не только выявляют риски раскрытия данных, но и позволяют понять, насколько обобщающей или переобученной является модель. Высокая чувствительность к индивидуальным примерам часто коррелирует с ухудшением способности к генерализации и снижением доверия к модели в реальных условиях. Таким образом, MIA служит не только средством оценки приватности, но и инструментом валидации качества модели с точки зрения её поведенческой устойчивости и эпистемической надёжности.

Разработка защит от MIA является активной областью исследований. Помимо дифференциальной приватности, применяются техники сглаживания предсказаний, ограничения энтропии выходного распределения, уменьшение

размера модели, использование аугментаций данных и переобучения на синтетических примерах. Некоторые подходы фокусируются на модификации самого API, предоставляющего доступ к модели: например, округление вероятностей, добавление шума в предсказания или предоставление только класса, а не всего распределения. При этом сохраняется баланс между качеством модели и её приватностными свойствами – чрезмерные меры защиты могут ухудшить точность и полезность модели в прикладных задачах [5].

В завершение, использование атак Membership Inference для оценки приватности моделей является неотъемлемой частью современной практики ответственного и доверенного машинного обучения. Эти атаки позволяют выявить слабые места в архитектуре и процессе обучения моделей, определить степень риска утечки персональных данных и принять обоснованные решения относительно необходимости внедрения механизмов защиты. Они также помогают выработать стратегию взаимодействия с предобученными моделями и открытыми ИИ-сервисами, минимизируя возможность раскрытия конфиденциальной информации. В условиях растущего нормативного давления и общественного внимания к вопросам ИИ-этики, устойчивость к MIA становится обязательным атрибутом зрелых ИИ-систем [6].

Список литературы:

1. Papernot, N. Practical Black-Box Attacks against Machine Learning / N. Papernot, P. McDaniel, I. Goodfellow [et al.] // ACM Asia Conference on Computer and Communications Security. – New York : ACM, 2017. – P. 506–519. – DOI 10.1145/3052973.3053009.

2. Song, C. Auditing Data Provenance in Text-Generation Models / C. Song, T. Ristenpart, V. Shmatikov // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2019. – P. 196–206. – DOI 10.1145/3292500.3330885.

3. Zanella-Béguelin, S. Analyzing Information Leakage of Updates to Natural Language Models / S. Zanella-Béguelin, L. Wutschitz, S. Tople [et al.] // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2020. – P. 363–375. – DOI 10.1145/3372297.3417880.

4. Mireshghallah, F. Privacy Regularization: Joint Privacy-Utility Optimization in Language Models / F. Mireshghallah, H. Inan, M. Hasegawa-Johnson [et al.] // Proceedings of NAACL. – Stroudsburg : ACL, 2021. – P. 3799–3807. – DOI 10.18653/v1/2021.naacl-main.301.

5. Kandpal, N. Deduplicating Training Data Mitigates Privacy Risks in Language Models / N. Kandpal, E. Wallace, C. Raffel // Proceedings of the 39th International Conference on Machine Learning. – 2022. – Vol. 162. – P. 10697–10707.

6. Lee, K. Deduplicating Training Data Makes Language Models Better / K. Lee, D. Ippolito, A. Nystrom [et al.] // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. – Stroudsburg : ACL, 2022. – P. 8424–8445. – DOI 10.18653/v1/2022.acl-long.577.