УДК 004.89

ПОСТРОЕНИЕ ТЕСТОВЫХ СЦЕНАРИЕВ ОЦЕНКИ КОНФИДЕНЦИАЛЬНОСТИ ДЛЯ ДОВЕРЕННЫХ МОДЕЛЕЙ

Мюллер Э., студент направления MSc Machine Learning, II курс
Королевский технологический институт, Стокгольм, Швеция

Обеспечение конфиденциальности данных при использовании моделей машинного обучения является одной из приоритетных задач в современных исследованиях и практических приложениях искусственного интеллекта. В условиях все более широкого внедрения предобученных и кастомизированных моделей в разнообразные критически важные области, включая медицину, финансы, государственное управление и другие сферы с повышенными требованиями к безопасности, становится необходимым не только разрабатывать надежные методы защиты, но и создавать адекватные инструменты оценки этих методов. В этой связи построение тестовых сценариев, направленных на всестороннюю и объективную оценку конфиденциальности доверенных моделей, приобретает ключевое значение [1].

Тестовые сценарии оценки конфиденциальности представляют собой формализованные и систематизированные подходы к моделированию различных условий эксплуатации и атакующих воздействий, способных привести к раскрытию чувствительной информации, содержащейся в моделях. Их задача – воспроизвести реалистичные ситуации, в которых могут проявляться уязвимости, а также проверить эффективность встроенных в модели механизмов защиты и контроля. Такой подход обеспечивает возможность не только выявления слабых мест в системе, но и сравнения различных архитектур и методов с точки зрения уровня обеспечиваемой приватности.

Разработка тестовых сценариев требует глубокого понимания потенциальных угроз и видов атак, применяемых в области машинного обучения, таких как membership inference, model inversion, reconstruction attacks и другие методы извлечения данных. Каждый из этих видов атак имеет свои особенности, которые должны учитываться при построении соответствующих тестов. Важно, чтобы сценарии охватывали как случайные, так и направленные атаки, а также отражали условия как централизованного, так и распределенного обучения, учитывая особенности федеративных и приватных вычислительных сред [2].

Ключевым аспектом является определение критериев и метрик оценки, которые позволяют количественно измерить уровень конфиденциальности и степень устойчивости моделей к атакам. Эти метрики могут включать

вероятность успешного извлечения данных, степень информации, подверженной утечке, время и ресурсы, необходимые для реализации атаки, а также влияние защитных механизмов на качество модели. Совмещение количественных и качественных показателей обеспечивает более полное представление о безопасности и доверии к модели.

Важной составляющей построения тестовых сценариев является использование репрезентативных наборов данных и реалистичных условий эксплуатации. Это позволяет не только получить объективные результаты, но и учесть влияние специфики прикладных задач на уровень конфиденциальности.

Например, в медицинских приложениях требуется особая тщательность в моделировании атак, связанных с восстановлением персональных медицинских данных, тогда как в финансовой сфере актуальны сценарии, связанные с раскрытием транзакционной информации [3].

Автоматизация процессов тестирования играет существенную роль в практическом применении разработанных сценариев.

Интеграция инструментов тестирования в конвейеры разработки и развертывания моделей позволяет регулярно проводить оценки безопасности, своевременно выявлять новые угрозы и обеспечивать соблюдение требований нормативных актов. Это особенно важно в условиях быстрого обновления моделей и появления новых методов атак.

Кроме технических аспектов, построение тестовых сценариев тесно связано с этическими и правовыми вопросами, так как они помогают обеспечить прозрачность и ответственность при разработке и использовании моделей машинного обучения. Результаты тестирования служат важным элементом аудита и сертификации, что повышает доверие к ИИ-системам со стороны пользователей и регуляторов [4].

Таким образом, создание комплексных и адаптивных тестовых сценариев для оценки конфиденциальности доверенных моделей является фундаментальным направлением в обеспечении безопасности современных систем искусственного интеллекта.

Этот подход способствует выявлению и минимизации рисков утечки данных, поддерживает высокие стандарты защиты информации и способствует развитию надежных и этически ответственных технологий машинного обучения [5].

Список литературы:

1. Lyu, L. Privacy and Robustness in Federated Learning: Attacks and Defenses / L. Lyu, H. Yu, Q. Yang // IEEE Transactions on Neural Networks and Learning Systems. – 2022. – Vol. 33, № 9. – P. 4719–4735. – DOI 10.1109/TNNLS.2020.3020825.

2. Mothukuri, V. A Survey on Security and Privacy of Federated Learning / V. Mothukuri, R. M. Parizi, S. Pouriyeh [et al.] // Future Generation Computer Systems. – 2021. – Vol. 115. – P. 619–640. – DOI 10.1016/j.future.2020.10.007.

3. Liu, B. When Machine Learning Meets Privacy: A Survey and Outlook / B. Liu, M. Ding, H. Xue [et al.] // ACM Computing Surveys. – 2021. – Vol. 54, № 2. – Article 31. – DOI 10.1145/3436755.

4. Rigaki, M. A Survey of Privacy Attacks in Machine Learning / M. Rigaki, S. Garcia // ACM Computing Surveys. – 2023. – Vol. 56, № 4. – Article 101. – DOI 10.1145/3624010.

5. Hu, H. Membership Inference Attacks on Machine Learning: A Survey / H. Hu, Z. Salcic, L. Sun [et al.] // ACM Computing Surveys. – 2022. – Vol. 54, № 11s. – Article 235. – DOI 10.1145/3523273.