УДК 004.89

ОЦЕНКА УСТОЙЧИВОСТИ МОДЕЛИ К ПЕРЕКРЕСТНОЙ КОРРЕЛЯЦИИ МЕЖДУ ПРИВАТНЫМИ И ПУБЛИЧНЫМИ ДАННЫМИ

Жэнг Э., студент направления Computational Modelling and Simulation
MSc programme, II курс
Технический университет Дании, Копенгаген, Дания

В современных системах машинного обучения, особенно в сценариях с участием пользовательских или чувствительных данных, одной из всё более актуальных проблем становится оценка устойчивости модели к перекрёстной корреляции между приватными и публичными данными. Под этой формулировкой понимается способность модели сохранять приватность входных данных даже в тех случаях, когда в открытом доступе имеются фрагменты информации, статистически или семантически связанные с этими приватными компонентами [1].

Это явление представляет собой фундаментальный риск, особенно в эпоху масштабной цифровизации, когда множество разнородных источников данных – от социальных сетей до медицинских записей и профилей покупок – могут быть объединены для анализа.

Перекрёстная корреляция между публичными и приватными данными – это ситуация, при которой на основе доступной внешней информации возможно с высокой вероятностью сделать выводы о конфиденциальных характеристиках объекта [2].

Например, знание о том, что пользователь часто ищет рецепты безглютеновой пищи, может коррелировать с предполагаемым диагнозом (целиакия), даже если сама модель о диагнозе явно не сообщает. Или знание публичных данных о возрасте, профессии и географии может позволить сужать пространство возможных значений чувствительных признаков, таких как уровень дохода или политические взгляды. Это делает задачу оценки устойчивости к таким перекрёстным связям важной не только с точки зрения формальной приватности, но и в контексте реальной безопасности пользователя.

Сложность оценки устойчивости к перекрёстной корреляции заключается в латентном характере утечки. Модель может формально не выдавать приватную информацию напрямую и даже соответствовать базовым стандартам приватности (например, дифференциальной приватности), но всё равно оставаться уязвимой к инференции – извлечению информации на основе

статистических связей. Особенно это характерно для моделей, обученных на больших и разнообразных датасетах, где приватные и публичные признаки нередко переплетены в процессе коррелированного обучения [3].

Одним из ключевых направлений оценки такой устойчивости является анализ взаимной информации между выходами модели и приватными признаками при наличии частичной внешней информации. Это может быть реализовано через введение шумных предсказателей, которые моделируют потенциального атакующего: они получают публичные признаки, сгенерированные моделью, и пытаются извлечь скрытые, приватные характеристики. Чем выше точность таких атакующих моделей, тем ниже устойчивость базовой модели к перекрёстной корреляции.

Также важным инструментом является контрфактическое тестирование. В этом методе проводится серия экспериментов, в которых публичные данные варьируются при фиксированных приватных признаках (и наоборот), а затем анализируется стабильность выходов модели [4].

Если модель чувствительна к косвенным связям между публичным и приватным, она будет проявлять нестабильность, которая указывает на потенциальную уязвимость к извлечению скрытых зависимостей.

Подобный анализ особенно эффективен при работе с генеративными языковыми моделями и системами рекомендаций, где выходы модели напрямую зависят от пользовательского контекста.

В условиях ограниченного доступа к обучающей выборке применяется эмпирическая оценка с помощью атак Attribute Inference, адаптированных под сценарии с публичной информацией. Такие атаки строят гипотезы о приватных признаках, опираясь не на параметры модели, а на видимую реакцию модели на варьируемые публичные входы. В некоторых случаях используется синтетическая генерация профилей с разной степенью пересечения публичной и приватной информации, чтобы зафиксировать порог, при котором вероятность успешной инференции становится недопустимо высокой [5].

Дополнительно, устойчивость может оцениваться через метрики чувствительности – например, влияние небольших изменений в публичных признаках на вероятность раскрытия приватной информации. Высокая чувствительность говорит о слабой изоляции данных и требует пересмотра архитектуры или методов обучения модели.

Существуют и архитектурные подходы к снижению перекрёстной корреляции, в том числе декорреляционное обучение (decorrelation learning), при котором модель обучается так, чтобы выходы были статистически независимы от чувствительных признаков при фиксированных публичных. Также используется adversarial training с дискриминатором, стремящимся предсказать приватные атрибуты, а основная модель обучается так, чтобы сделать эти

предсказания максимально неинформативными. Это позволяет минимизировать латентную утечку и повысить устойчивость модели к перекрёстной инференции [6].

Особое значение оценка устойчивости приобретает в правовом контексте. В соответствии с нормативами, такими как GDPR или HIPAA, утечка информации может быть признана не только при прямом раскрытии, но и при возможности логического восстановления персональных данных. Это означает, что даже косвенные зависимости, ускользающие от стандартных тестов приватности, могут быть признаны нарушением конфиденциальности. В такой ситуации формальная устойчивость к перекрёстной корреляции становится не просто техническим показателем, а необходимым критерием легитимности модели.

Таким образом, оценка устойчивости модели к перекрёстной корреляции между приватными и публичными данными – это сложная, многомерная задача, сочетающая в себе элементы теории информации, статистического тестирования, машинного обучения и правового анализа. Без надлежащей оценки и валидации в этом аспекте невозможно говорить о реальной безопасности ИИ-систем, особенно в контексте персонализированных сервисов, медицинских решений и цифровых платформ. Разработка комплексных инструментов для такой оценки – приоритетное направление для построения доверенного и этически ответственного искусственного интеллекта [7].

Список литературы:

1. Wood, A. Differential Privacy: A Primer for a Non-Technical Audience / A. Wood, M. Altman, A. Bembenek [et al.] // Vanderbilt Journal of Entertainment & Technology Law. – 2018. – Vol. 21, № 1. – P. 209–276.

2. Nissim, K. Smooth Sensitivity and Sampling in Private Data Analysis / K. Nissim, S. Raskhodnikova, A. Smith // ACM Symposium on Theory of Computing. – New York : ACM, 2007. – P. 75–84. – DOI 10.1145/1250790.1250803.

3. Kasiviswanathan, S. P. What Can We Learn Privately? / S. P. Kasiviswanathan, H. K. Lee, K. Nissim [et al.] // SIAM Journal on Computing. – 2011. – Vol. 40, № 3. – P. 793–826. – DOI 10.1137/090756090.

4. Blum, A. A Learning Theory Approach to Noninteractive Database Privacy / A. Blum, K. Ligett, A. Roth // Journal of the ACM. – 2013. – Vol. 60, № 2. – Article 12. – DOI 10.1145/2450142.2450148.

5. Hsu, J. Differential Privacy: An Economic Method for Choosing Epsilon / J. Hsu, M. Gaboardi, A. Haeberlen [et al.] // IEEE Computer Security Foundations Symposium. – Piscataway : IEEE, 2014. – P. 398–410. – DOI 10.1109/CSF.2014.35.

6. Near, J. P. Programming Differential Privacy / J. P. Near, C. Abuah. – San Francisco : OpenMined, 2021. – 292 p.

7. European Parliament and Council. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. – Brussels : Official Journal of the European Union, 2016. – L 119. – P. 1–88.