УДК 004.89

ОЦЕНКА СТЕПЕНИ ЭКСПОНИРОВАНИЯ ЧУВСТВИТЕЛЬНЫХ ПРИЗНАКОВ В ПРОЦЕССЕ ИНФЕРЕНСА

Богданов Д.Г., студент гр. ИТб-221, III курс
Научный руководитель: Николаева Е.А., к. ф.-м. н., доцент
Кузбасский государственный технический университет
имени Т.Ф. Горбачева, г. Кемерово

Оценка степени экспонирования чувствительных признаков в процессе инференса моделей машинного обучения представляет собой критически важное направление в обеспечении доверенности и конфиденциальности современных ИИ-систем. В условиях, когда интеллектуальные модели всё чаще используются для обработки персональных, биометрических, медицинских, финансовых и иных чувствительных данных, становится необходимым не только защищать исходную выборку на этапе обучения, но и глубоко анализировать, в какой степени эти чувствительные характеристики могут быть раскрыты в процессе инференса – то есть при работе модели в продуктивной среде, когда она выдаёт предсказания для новых входов. Если модель даже частично сохраняет в себе следы чувствительных признаков и «выплёскивает» их наружу через выходы или промежуточные представления, это может привести к утечке персональной информации, нарушению законодательства и подрыву доверия пользователей [1].

Под чувствительными признаками в данном контексте понимаются такие переменные, как раса, пол, этническая принадлежность, возраст, состояние здоровья, религиозные или политические взгляды, сексуальная ориентация и другие категории, защищённые как с этической, так и с юридической точки зрения. Даже если такие признаки явно не присутствуют в выходе модели, они могут быть латентно закодированы в предсказаниях, особенно если были тесно скоррелированы с целевыми переменными в процессе обучения. В ряде случаев чувствительные признаки могут быть не только случайно унаследованы, но и сознательно использоваться в модели – например, при наличии прокси-признаков (proxy features), скрыто передающих социально значимую информацию, такую как индекс места проживания, уровень дохода или тип занятости.

Оценка экспонирования этих признаков требует комплексного подхода, сочетающего статистический анализ, тестирование, построение вспомогательных моделей и формальные методы. Одним из наиболее популярных подходов является построение атакующих моделей-декодеров (feature inference models),

задача которых – по выходам модели или её внутренним активациям предсказать значение чувствительного признака [2].

Например, если модель классифицирует кредитные заявки, и на её выходах или softmax-распределениях можно с высокой точностью предсказать этническую принадлежность клиента, это означает, что чувствительный признак экспонирован в процессе инференса. Подобный анализ может проводиться как в white-box-сценарии, с полным доступом к архитектуре модели, так и в black-box-условиях, если сохраняется возможность многократного запроса модели с различными входами.

Математически оценка экспонирования может быть представлена через такие метрики, как точность декодера (inference accuracy), взаимная информация между чувствительным признаком и выходом модели, коэффициенты корреляции или значение AUC в задачах бинарной классификации чувствительного атрибута. Чем выше эти показатели, тем выше вероятность того, что чувствительная информация сохраняется и становится доступной на этапе инференса. Иногда применяется и регрессионный анализ чувствительности, при котором моделируется зависимость выходов от чувствительных признаков при контроле остальных переменных.

Другим подходом является использование контрфактического анализа, в рамках которого для одного и того же входного вектора генерируются его версии, отличающиеся только значением чувствительного признака (например, пол или возраст), и затем сравниваются выходы модели [3].

Существенное расхождение в предсказаниях указывает на то, что чувствительный признак повлиял на результат, и, следовательно, может быть декодирован обратно – прямо или косвенно. Такие эксперименты особенно ценны при анализе fairness-свойств модели, поскольку они демонстрируют наличие несправедливых зависимостей, даже если модель формально не использует чувствительный атрибут как вход.

Важной частью анализа экспонирования является проверка скрытых слоёв модели. В нейросетевых архитектурах внутренние представления (например, эмбеддинги, attention-векторы, скрытые состояния RNN или активации сверточных слоёв) могут неявно кодировать чувствительные признаки. Это особенно заметно в больших языковых и визуальных моделях, где модель обучается извлекать контекстные зависимости. Исследования показывают, что на определённых слоях можно с высокой точностью декодировать пол, этническую принадлежность или наличие определённых заболеваний, даже если эти параметры не входили в обучающие метки. Анализ скрытых представлений позволяет оценить, на каком этапе модели происходит экспонирование и как глубоко оно встроено в архитектуру.

Для снижения степени экспонирования применяются различные защитные стратегии. Одна из них – инвариантное обучение (invariant representation learning), при котором модель оптимизируется так, чтобы внутренние представления были нечувствительны к изменению чувствительных признаков. Это достигается путём введения дополнительных потерь (adversarial loss), наказывающих модель за возможность реконструкции чувствительных атрибутов из промежуточных слоёв. Другой подход – добавление шумов или применение дифференциальной приватности, что снижает вероятность кодирования индивидуальных особенностей обучающей выборки. Также активно используются аудитные инструменты, которые тестируют модель на экспонирование до её внедрения в продуктивную среду [4].

Практически важно также учитывать специфические сценарии использования моделей. В некоторых приложениях, таких как рекомендательные системы, голосовые помощники, оценка рисков или автоматическое принятие решений, экспонирование чувствительных признаков может быть незаметным на уровне одного ответа, но иметь накопительный эффект. Поэтому оценка экспонирования должна включать в себя многократное моделирование взаимодействий, оценку агрегатов поведения модели, а также интерпретируемые объяснения решений, позволяющие экспертам отследить возможные утечки на уровне логики модели.

Таким образом, оценка степени экспонирования чувствительных признаков в процессе инференса является необходимым элементом построения доверенных систем машинного обучения. Она позволяет не только выявлять потенциальные риски утечки персональной информации, но и улучшать архитектуры моделей, делая их более устойчивыми, этичными и соответствующими современным регуляторным требованиям. В условиях, когда ИИ всё глубже внедряется в повседневную жизнь, вопросы экспонирования и справедливости становятся не просто техническими, а системными и общественно значимыми [5].

Список литературы:

1. Tirumala, K. Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models / K. Tirumala, A. Markosyan, L. Zettlemoyer, A. Aghajanyan // Advances in Neural Information Processing Systems. – 2022. – Vol. 35. – P. 38274–38290.

2. Carlini, N. Quantifying Memorization Across Neural Language Models / N. Carlini, D. Ippolito, M. Jagielski [et al.] // International Conference on Learning Representations. – 2023. – P. 1–27.

3. Lehman, E. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? / E. Lehman, S. Jain, K. Pichotta [et al.] // Proceedings of NAACL. – Stroudsburg : ACL, 2021. – P. 946–959. – DOI 10.18653/v1/2021.naacl-main.73.

4. Mattern, J. Membership Inference Attacks against Language Models via Neighbourhood Comparison / J. Mattern, F. Mireshghallah, Z. Jin [et al.] // Findings of the Association for Computational Linguistics: ACL 2023. – Stroudsburg : ACL, 2023. – P. 11330–11343. – DOI 10.18653/v1/2023.findings-acl.719.

5. Hisamoto, S. Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your Machine Translation System? / S. Hisamoto, M. Post, K. Duh // Transactions of the Association for Computational Linguistics. – 2020. – Vol. 8. – P. 49–63. – DOI 10.1162/tacl_a_00299.