УДК 004.89

ПОДДЕРЖКА ИНТЕРПРЕТИРУЕМОСТИ В ДОВЕРЕННЫХ ФРЕЙМВОРКАХ МАШИННОГО ОБУЧЕНИЯ

Чернов В.С., студент гр. 23222, II курс
Новосибирский государственный университет, г. Новосибирск

В современных условиях стремительного развития искусственного интеллекта и его интеграции в ответственные и высокорискованные сферы деятельности – такие как медицина, финансы, автономные транспортные системы и государственное управление – возрастают требования не только к эффективности и точности моделей машинного обучения, но и к их прозрачности, объяснимости и интерпретируемости. Интерпретируемость моделей становится важнейшим фактором доверия со стороны пользователей, регулирующих органов и общества в целом. В этой связи доверенные фреймворки машинного обучения обязаны обеспечивать комплексную поддержку механизмов интерпретируемости, которые позволяют понять, обосновать и контролировать поведение моделей на всех этапах их жизненного цикла. Настоящая статья посвящена анализу подходов, архитектурных решений и технических средств, реализующих поддержку интерпретируемости в доверенных ML-фреймворках [1].

Поддержка интерпретируемости в доверенных фреймворках подразумевает внедрение как глобальных, так и локальных методов объяснения результатов моделей. Глобальная интерпретируемость позволяет получить общее представление о работе модели, выявить ключевые признаки и понять общие закономерности, используемые при принятии решений. Локальная интерпретируемость направлена на объяснение конкретного предсказания в отдельном случае, что особенно важно при анализе аномальных или спорных решений. Для достижения этих целей фреймворки должны обеспечивать интеграцию различных алгоритмов и техник, таких как методы важности признаков (feature importance), контрфактические объяснения, локальные модели приближения (например, LIME, SHAP), визуализации активаций и слоёв нейросетей, а также attention-механизмы в трансформерах и других архитектурах.

Важнейшим архитектурным требованием является модульность и расширяемость компонентов интерпретируемости. Доверенные фреймворки должны предоставлять стандартизированные API и интерфейсы, позволяющие интегрировать новые методы и инструменты объяснения без необходимости

переработки основной инфраструктуры. Это обеспечивает гибкость и адаптивность в условиях быстрого развития исследовательских направлений и появления новых техник XAI (Explainable AI). Кроме того, модульность способствует воспроизводимости и повторному использованию объяснений в различных приложениях и на разных этапах разработки [2].

Обязательным элементом поддержки интерпретируемости является обеспечение верифицируемости и прозрачности самих механизмов объяснения. В доверенных системах недостаточно лишь предоставлять объяснения – они должны быть проверяемыми и воспроизводимыми, что достигается за счёт формализации алгоритмов генерации объяснений и внедрения систем контроля качества. Важным аспектом является интеграция логгирования и аудита объяснительных сессий, что позволяет проводить ретроспективный анализ и выявлять возможные несоответствия или ошибки в выводах модели.

В рамках доверенных фреймворков необходимо предусмотреть поддержку интерпретируемости на всех этапах жизненного цикла модели – от подготовки данных и обучения до развёртывания и мониторинга в продуктивной среде. Особенно значима интерпретируемость при мониторинге поведения модели в реальном времени, где выявление отклонений, drift-а данных или ошибок в предсказаниях может сопровождаться автоматической генерацией объяснений для быстрого реагирования и корректировки. Это требует тесной интеграции интерпретируемых компонентов с системами управления и оркестрации вычислительных процессов [3].

Безопасность и защита данных также являются важными факторами, влияющими на поддержку интерпретируемости. В доверенных фреймворках должны быть реализованы механизмы ограничения доступа к объяснениям, чтобы предотвратить возможность утечки конфиденциальной информации через интерпретируемые данные. Например, контрфактические объяснения могут раскрывать чувствительные особенности обучающей выборки, что требует применения дифференциальной приватности или иных методов защиты при их генерации и хранении [4].

Кроме технических и архитектурных аспектов, существенным элементом поддержки интерпретируемости является соответствие нормативным и этическим стандартам. Доверенные фреймворки должны обеспечивать возможность генерации отчётов и метрик, демонстрирующих соблюдение требований к объяснимости и прозрачности моделей, что особенно актуально в условиях растущего законодательства в области ИИ (например, AI Act в ЕС). Это способствует формированию доверия как у пользователей, так и у регулирующих органов, облегчая процессы сертификации и аудита.

В результате, поддержка интерпретируемости в доверенных фреймворках машинного обучения представляет собой многоаспектную задачу, объединяющую методы XAI, архитектурные решения, механизмы верификации и соблюдение нормативных требований. Комплексное внедрение таких механизмов способствует созданию более прозрачных, этически ответственных и надёжных систем искусственного интеллекта, способных успешно функционировать в ответственных и регулируемых сферах применения. В дальнейшем развитие этой области будет стимулировать улучшение инструментов интерпретируемости и расширение их функциональности, обеспечивая баланс между сложностью моделей и потребностями пользователей в прозрачности и доверии [5].

Список литературы:

1. Bach, S. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation / S. Bach, A. Binder, G. Montavon [et al.] // PLOS ONE. – 2015. – Vol. 10, № 7. – e0130140. – DOI 10.1371/journal.pone.0130140.

2. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable / C. Molnar. – 2nd ed. – Munich : Leanpub, 2022. – 330 p.

3. Samek, W. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models / W. Samek, G. Montavon, A. Vedaldi [et al.]. – Cham : Springer, 2019. – 439 p. – DOI 10.1007/978-3-030-28954-6.

4. Ribeiro, M. T. Anchors: High-Precision Model-Agnostic Explanations / M. T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the AAAI Conference on Artificial Intelligence. – 2018. – Vol. 32, № 1. – P. 1527–1535.

5. Пылов, П. А. Экстракция признаков в моделях последовательного глубокого обучения / П. А. Пылов, А. В. Протодьяконов // Россия молодая : сб. материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 19–21 апреля 2022 года. – Кемерово : Кузбасский государственный технический университет имени Т. Ф. Горбачева, 2022. – С. 31525.1–31525.3. – EDN MZEHGM.