УДК 004.89

РАЗРАБОТКА УНИВЕРСАЛЬНОГО ИНТЕРФЕЙСА ДЛЯ ИНТЕРПРЕТАЦИИ МОДЕЛЕЙ НА ПРАКТИКЕ

Кузьмина В.Р., студент гр. 09-361, IV курс
Казанский федеральный университет, г. Казань

Разработка универсального интерфейса для интерпретации моделей машинного обучения является одной из ключевых задач современной науки и инженерии в области искусственного интеллекта [1].

С ростом сложности и распространённости моделей глубокого обучения, ансамблевых методов и трансформеров возникает острая необходимость в создании инструментов, которые бы позволяли эффективно, понятно и доступно объяснять решения этих моделей в самых различных прикладных областях – от медицины и финансов до промышленности и юриспруденции.

Универсальный интерфейс выступает своеобразным мостом между сложной внутренней логикой моделей и конечными пользователями, будь то специалисты по данным, разработчики или конечные заказчики, нуждающиеся в прозрачности и доверии к работе ИИ-систем.

В первую очередь, разработка такого интерфейса требует интеграции различных методов интерпретации, включая локальные и глобальные объяснения, визуализации важности признаков, тепловые карты, attention-механизмы, а также методы, основанные на теории вкладов, например, значения Шепли. Важной особенностью является необходимость универсальности – интерфейс должен поддерживать широкий спектр моделей, включая нейронные сети, деревья решений, градиентный бустинг, а также гибридные и кастомные архитектуры. Это предполагает построение модульной архитектуры, где различные интерпретаторы и визуализаторы можно подключать и настраивать в зависимости от задачи и специфики данных [2].

Не менее важен аспект интерактивности, позволяющий пользователям самостоятельно исследовать и анализировать данные и предсказания модели. Такой интерфейс должен обеспечивать динамическое взаимодействие – фильтрацию и сегментацию данных, настройку параметров интерпретации, проведение what-if анализа, сравнение поведения модели на различных подвыборках. Визуальные компоненты, включая диаграммы, графы и интерактивные панели, должны быть интуитивно понятными и адаптированными под разные типы пользователей: от технических специалистов до бизнес-аналитиков и руководителей.

Технически универсальный интерфейс строится на основе современных веб-технологий, позволяющих обеспечить кроссплатформенность, масштабируемость и удобство развертывания. Использование библиотек для визуализации данных (D3.js, Plotly, Vega), фреймворков для разработки пользовательских интерфейсов (React, Angular, Vue.js) и серверных технологий для обработки больших объёмов данных обеспечивает плавную работу даже с высокоразмерными наборами и сложными вычислительными процессами. Кроме того, необходима интеграция с ML-фреймворками и платформами (TensorFlow, PyTorch, scikit-learn, MLflow), что позволяет в реальном времени получать предсказания и объяснения, а также отслеживать версии моделей и изменения данных.

Особое внимание уделяется стандартизации форматов данных и протоколов обмена между компонентами системы. Это необходимо для обеспечения совместимости и повторного использования интерпретаторов и визуализаторов в различных приложениях и организациях. Важным элементом является поддержка расширяемости – разработчики должны иметь возможность создавать новые модули и подключать их без нарушения общей архитектуры.

Кроме технических аспектов, при разработке универсального интерфейса необходимо учитывать и организационно-этические вопросы. В ряде отраслей, таких как здравоохранение, банковское дело и государственное управление, требования к прозрачности и объяснимости моделей регулируются законодательством. Интерфейс должен помогать соблюдению этих норм, предоставляя отчёты, протоколы аудита и возможности для независимой экспертизы. Это повышает доверие пользователей и снижает риски неправомерного использования технологий [3].

Наконец, универсальный интерфейс должен обеспечивать поддержку обучения и адаптации пользователей. Интерактивные подсказки, встроенная документация, кейсы и примеры помогают специалистам разных уровней быстро освоить инструменты интерпретации и применять их в своей работе. Это способствует более широкому распространению практик объяснимого ИИ и повышению качества принимаемых решений.

Таким образом, создание универсального интерфейса для интерпретации моделей является сложной, многогранной задачей, требующей междисциплинарного подхода и сочетания передовых технологий визуализации, интерактивности и стандартизации [4].

Успешная реализация таких решений способна кардинально повысить прозрачность, контролируемость и доверие к современным интеллектуальным системам, что становится необходимым условием их эффективного и безопасного внедрения в практику [5, 6].

Список литературы:

1. Abdul, A. Trends and Trajectories for Explainable, Accountable and Intelligible Systems / A. Abdul, J. Vermeulen, D. Wang [et al.] // Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. – New York : ACM, 2018. – P. 1–18. – DOI 10.1145/3173574.3174156.

2. Wang, D. Designing Theory-Driven User-Centric Explainable AI / D. Wang, Q. Yang, A. Abdul, B. Y. Lim // Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. – New York : ACM, 2019. – P. 1–15. – DOI 10.1145/3290605.3300831.

3. Mohseni, S. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems / S. Mohseni, N. Zarei, E. D. Ragan // ACM Transactions on Interactive Intelligent Systems. – 2021. – Vol. 11, № 3–4. – Article 24. – DOI 10.1145/3387166.

4. Hoffman, R. R. Metrics for Explainable AI : Challenges and Prospects / R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman // arXiv. – 2018. – arXiv:1812.04608.

5. Hase, P. Evaluating Explainable AI : Which Algorithmic Explanations Help Users Predict Model Behavior? / P. Hase, M. Bansal // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – Stroudsburg : ACL, 2020. – P. 5540–5552. – DOI 10.18653/v1/2020.acl-main.491.

6. Lakkaraju, H. Faithful and Customizable Explanations of Black Box Models / H. Lakkaraju, O. Bastani // Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. – New York : ACM, 2019. – P. 131–138. – DOI 10.1145/3306618.3314229.