УДК 004.89

ИНТЕГРАЦИЯ ИНТЕРПРЕТИРУЕМОСТИ В ПРОЦЕСС ОБУЧЕНИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Гришин Д.С., студент гр. 602, II курс
Московский государственный университет имени М. В. Ломоносова,
г. Москва

Интерпретируемость моделей машинного обучения на сегодняшний день является одним из ключевых требований к современным интеллектуальным системам, особенно в областях, где принятие решений связано с высокими рисками, такими как медицина, финансы, юриспруденция и государственное управление.

Традиционно задачи интерпретации рассматривались как постфактумный этап – после того, как модель была обучена, специалисты применяли различные методы объяснения, чтобы понять, как модель приходит к своим выводам. Однако в последнее время наблюдается заметный сдвиг парадигмы: интерпретируемость всё чаще становится интегральной частью самого процесса обучения модели, что позволяет создавать не только более прозрачные, но и устойчивые и надежные системы.

Интеграция интерпретируемости в обучение моделей – это комплексный подход, в рамках которого методы и принципы объяснимости внедряются на этапах подготовки данных, построения архитектуры модели, оптимизации параметров и валидации. Такой подход обеспечивает, что модель изначально проектируется и обучается с учётом возможности последующего глубокого анализа и понимания её поведения. Это позволяет избежать проблем, связанных с «чёрным ящиком», когда модель демонстрирует высокую точность, но при этом сложно объяснима или содержит скрытые ошибки, вызывающие нежелательные последствия.

Одним из ключевых направлений в этой области является разработка моделей с встроенными интерпретируемыми компонентами, таких как внимание (attention mechanisms), прозрачные слои, селективные маски и модульные архитектуры. Например, в нейронных сетях внимание позволяет выявлять, на каких частях входных данных модель концентрирует свои усилия при принятии решения, что уже само по себе служит формой объяснения. При обучении такие механизмы могут быть дополнительно регуляризованы или направлены на повышение семантической осмысленности выделяемых признаков, что усиливает интерпретируемость без значительных потерь в производительности.

Другой важный подход – обучение с ограничениями и регуляризациями, которые направлены на формирование простых, устойчивых и объяснимых моделей. К ним относятся, например, методы sparsity regularization (регуляризация разреженности), ограничение сложности модели, минимизация взаимной информации между признаками и скрытыми представлениями, а также использование структурированных архитектур с ограниченным числом параметров. Такие методы позволяют создавать модели, которые легче поддаются интерпретации как с точки зрения количественной оценки важности признаков, так и в терминах причинно-следственных связей [1].

Интеграция интерпретируемости непосредственно в процесс обучения также включает в себя применение многоцелевая оптимизации, где помимо стандартной функции потерь добавляются критерии, связанные с прозрачностью и понятностью модели. Например, оптимизация может одновременно учитывать точность предсказаний и согласованность выделяемых интерпретаторами признаков с экспертными знаниями, что способствует выработке моделей, способных объяснять свои решения в терминах, понятных специалистам предметной области [2].

Особое значение приобретает использование объяснительных моделей-суррогатов, обучаемых совместно с основными, более сложными моделями. В таком случае основная модель обеспечивает высокое качество предсказаний, а суррогат – локальные или глобальные объяснения, адаптированные и оптимизированные в ходе обучения. Эта парадигма позволяет сохранять баланс между точностью и интерпретируемостью, при этом поддерживая возможность обратной связи, когда результаты интерпретации влияют на корректировку основной модели [3].

Кроме архитектурных и алгоритмических аспектов, важным элементом интеграции является развитие инструментария для мониторинга и адаптации интерпретируемости в процессе эксплуатации моделей. Это позволяет отслеживать стабильность и адекватность интерпретаций при изменении данных, выявлять случаи деградации объяснимости и своевременно принимать меры по обновлению или перенастройке моделей. Такой динамический контроль способствует поддержанию высокого уровня доверия и предотвращению неожиданных ошибок в работе ИИ-систем [4].

С практической точки зрения интеграция интерпретируемости в обучение требует скоординированных усилий в области инженерии данных, разработки моделей, настройки процессов обучения и построения интерфейсов для визуализации и анализа. Современные ML-фреймворки постепенно внедряют соответствующие модули и API, позволяя исследователям и разработчикам создавать кастомные решения с поддержкой объяснимости на всех этапах жизненного цикла модели [5].

В заключение, интеграция интерпретируемости в процесс обучения моделей машинного обучения – это стратегически важное направление, которое меняет подходы к созданию и использованию ИИ-систем. Такой подход позволяет не только повысить уровень доверия и ответственности, но и создать модели, обладающие высокой адаптивностью, устойчивостью и способностью к самоконтролю [6].

В результате это способствует более широкому и безопасному внедрению искусственного интеллекта в разнообразные сферы человеческой деятельности, где прозрачность и понимание решений имеют критическое значение.

Список литературы:

1. Ross, A. S. Right for the Right Reasons : Training Differentiable Models by Constraining Their Explanations / A. S. Ross, M. C. Hughes, F. Doshi-Velez // Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. – Melbourne : IJCAI, 2017. – P. 2662–2670. – DOI 10.24963/ijcai.2017/371.

2. Chen, C. This Looks Like That : Deep Learning for Interpretable Image Recognition / C. Chen, O. Li, C. Tao [et al.] // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.

3. Li, O. Deep Learning for Case-Based Reasoning through Prototypes : A Neural Network that Explains Its Predictions / O. Li, H. Liu, C. Chen, C. Rudin // Proceedings of the AAAI Conference on Artificial Intelligence. – 2018. – Vol. 32, № 1.

4. Kim, B. Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability / B. Kim, R. Khanna, O. O. Koyejo // Advances in Neural Information Processing Systems. – 2016. – Vol. 29.

5. Koh, P. W. Understanding Black-Box Predictions via Influence Functions / P. W. Koh, P. Liang // Proceedings of the 34th International Conference on Machine Learning. – 2017. – Vol. 70. – P. 1885–1894.

6. Протодьяконов, А. В. Асимптотический анализ поведения прикладных моделей машинного обучения : учебное пособие / А. В. Протодьяконов, А. В. Дягилева, П. А. Пылов. – Вологда : Инфра-Инженерия, 2023. – 144 с. – ISBN 978-5-9729-1455-5. – EDN APHQME.