УДК 004.89

ИСПОЛЬЗОВАНИЕ УПРОЩЕННЫХ МОДЕЛЕЙ ДЛЯ ЛОКАЛЬНОЙ ИНТЕРПРЕТАЦИИ СЛОЖНЫХ АРХИТЕКТУР

Соболевский И.П., студент гр. 09-211, II курс
Казанский федеральный университет, г. Казань

Современные архитектуры машинного обучения – от глубоких сверточных и трансформерных сетей до градиентного бустинга и ансамблей – обеспечивают выдающуюся точность и обобщающую способность на множестве задач, от компьютерного зрения до анализа текстов и временных рядов. Однако эти модели часто обладают высокой степенью вычислительной и логической сложности, что делает их поведение трудноинтерпретируемым как для конечного пользователя, так и для разработчика. В условиях, когда ИИ-системы применяются в регулируемых отраслях – медицине, финансовых услугах, юридических решениях, критической инфраструктуре – интерпретируемость становится не просто желательной, а обязательной [1].

Одним из ключевых подходов к решению этой задачи является использование упрощённых моделей для локальной интерпретации – стратегия, при которой поведение сложной модели объясняется через обучение простой, интерпретируемой модели на малой окрестности входа.

Суть подхода локальной интерпретации (local surrogate modeling) состоит в том, что вместо анализа всей архитектуры модели в целом, исследуется её поведение на ограниченном множестве примеров, близких к интересующему входу. Предполагается, что даже если модель в целом ведёт себя сложно и нелинейно, в локальной области её поведение можно аппроксимировать более простой функцией, которая сохраняет смысловую и причинную связь между входом и выходом. Типичным примером является построение линейной модели, решающего дерева или набора правил, которые точно приближают поведение «чёрного ящика» в окрестности интересующего объекта.

Наиболее известной реализацией этого подхода является метод LIME (Local Interpretable Model-agnostic Explanations), в котором вокруг целевой точки создаётся искусственный набор данных, полученный с помощью случайных модификаций исходного входа (например, замены, маскировки или зашумления признаков). Затем сложная модель используется как оракул, чтобы получить предсказания на этом наборе, после чего обучается простая модель (обычно линейная регрессия или дерево решений), аппроксимирующая поведение оригинальной модели в данной области. Эта простая модель и

становится источником интерпретации: она позволяет указать, какие признаки являются наиболее важными в данной ситуации, как они взаимодействуют и к какому выводу они приводят [2].

Подход LIME обладает рядом достоинств: он универсален (работает с любыми моделями), компактен (создаёт лёгкое объяснение) и гибок (может адаптироваться под разные типы данных – текст, изображение, табличные данные). Однако у него есть и ограничения, связанные, прежде всего, с устойчивостью и достоверностью интерпретации. Локальная линейная модель может плохо аппроксимировать оригинальную, если поведение модели в локальной области нелинейно или несгладимо. Кроме того, процесс генерации синтетических точек может приводить к появлению «нереалистичных» входов, на которых предсказания модели не интерпретируемы с точки зрения задачи [3].

Для преодоления этих недостатков разрабатываются более устойчивые методы локального моделирования, такие как Anchors, K-LIME, L2X, ProtoDash, а также методы с регуляризацией на сэмплах, близких к естественным данным.

Отдельный интерес представляют методы построения локальных деревьев решений, которые позволяют объяснять вывод модели в виде логически интерпретируемого пути.

Например, метод MAPLE (Model Agnostic Supervised Local Explanations) обучает дерево, специфичное для окрестности целевой точки, с учётом весов соседей и чувствительности признаков. Такое объяснение легко визуализируется, воспринимается пользователем и может быть документировано.

В медицине, например, такой подход позволяет объяснить диагноз, ссылаясь на комбинацию симптомов, историю болезни и ключевые триггеры, обнаруженные локально, а не на глобальные закономерности всей выборки.

Локальные интерпретаторы также находят применение в задачах отладки и аудита моделей. Например, при обнаружении ошибочного или неожиданного предсказания можно построить локальную интерпретацию и выявить, какие признаки или их комбинации привели к выводу. Если такие признаки не являются логически обоснованными (например, модель принимает решение о кредитоспособности на основе региона, а не дохода), это может указывать на переобучение, предвзятость или утечку данных. В отличие от глобальных метрик важности, локальная интерпретация позволяет сфокусироваться на единичном случае, что особенно ценно при работе с чувствительными или критическими решениями.

Существует также растущий интерес к смешанным архитектурам, в которых локальная интерпретируемая модель становится встроенным компонентом общей структуры. Например, используются гибридные модели, в которых сложная сеть предсказывает результат, а параллельная лёгкая модель

(например, decision tree или логистическая регрессия) обучается на её скрытых признаках и используется как объяснитель. Такие подходы повышают доверие и позволяют контролировать поведение модели не только постфактум, но и во время эксплуатации.

Важно подчеркнуть, что качество локальной интерпретации зависит от метрики близости, стратегии генерации соседей, выбора интерпретируемой модели и точности аппроксимации. Поэтому разработка инструментов и методологий, позволяющих настраивать эти компоненты, становится важной областью инженерии объяснимого ИИ. В современных фреймворках (например, alibi, skater, SHAP, InterpretML) всё чаще реализуются модули локального моделирования, в которых пользователь может задать критерии доверия, метрики соответствия оригиналу и порог качества интерпретации [4].

Таким образом, использование упрощённых моделей для локальной интерпретации сложных архитектур представляет собой эффективную стратегию балансировки точности и прозрачности. Она позволяет предоставлять человеку понятное объяснение решений даже в контексте крайне сложных вычислительных моделей, сохраняя при этом полноту и гибкость анализа [5].

Локальная интерпретируемость становится не просто вспомогательным средством, а обязательным элементом ответственного проектирования, эксплуатации и мониторинга ИИ-систем в условиях реального мира.

Список литературы:

1. Hinton, G. Distilling the Knowledge in a Neural Network / G. Hinton, O. Vinyals, J. Dean // arXiv. – 2015. – arXiv:1503.02531.

2. Craven, M. Extracting Tree-Structured Representations of Trained Networks / M. Craven, J. W. Shavlik // Advances in Neural Information Processing Systems. – 1996. – Vol. 8.

3. Lou, Y. Accurate Intelligible Models with Pairwise Interactions / Y. Lou, R. Caruana, J. Gehrke, G. Hooker // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2013. – P. 623–631. – DOI 10.1145/2487575.2487579.

4. Caruana, R. Intelligible Models for HealthCare : Predicting Pneumonia Risk and Hospital 30-Day Readmission / R. Caruana, Y. Lou, J. Gehrke [et al.] // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2015. – P. 1721–1730. – DOI 10.1145/2783258.2788613.

5. Alvarez-Melis, D. Towards Robust Interpretability with Self-Explaining Neural Networks / D. Alvarez-Melis, T. S. Jaakkola // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.