УДК 004.89

ИНСТРУМЕНТЫ ОЦЕНКИ ДОВЕРИЯ К ИНТЕРПРЕТАЦИЯМ ПРЕДСКАЗАНИЙ МОДЕЛЕЙ

Шевченко П.И., студент гр. 23.М02-мкн, I курс
Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург

С ростом числа приложений искусственного интеллекта в критически важных областях – от медицины и финансов до правосудия и оборонной сферы – стремительно увеличивается значимость не только объяснений решений моделей машинного обучения, но и оценки доверия к самим этим объяснениям. Объяснение, предоставляемое моделью, становится смысловым интерфейсом между человеком и машиной, но если это объяснение нестабильно, искажено, не воспроизводимо или слишком чувствительно к внешним условиям, оно может не только дезинформировать пользователя, но и породить ложное чувство уверенности в системе. Таким образом, необходимы специализированные инструменты, позволяющие оценивать надёжность, устойчивость и правдоподобность интерпретаций, получаемых от моделей – как в офлайн-анализе, так и в условиях реального применения [1].

Классическая задача интерпретируемости долгое время фокусировалась на разработке методов объяснения: SHAP, LIME, Integrated Gradients, Attention Visualization, Surrogate Models и других. Однако по мере интеграции этих методов в системы принятия решений, стало ясно, что не всякое объяснение является качественным или заслуживающим доверия. Например, два объяснения одного и того же решения, полученные от разных интерпретаторов или с разными параметрами, могут радикально различаться. Или объяснение может быть стабильным, но вводить в заблуждение, выделяя признаки, не имеющие причинной связи с решением. Поэтому возникает новый уровень задачи – оценка доверия к объяснению, которая включает в себя как количественные метрики, так и качественные эвристики, а также инженерные средства их визуализации и анализа.

Одной из важнейших характеристик интерпретации является стабильность. Хорошее объяснение должно сохранять свою структуру при незначительных изменениях входных данных. Например, если к изображению добавлен слабый шум или в текст вставлено слово, не влияющее на смысл, интерпретация не должна меняться кардинально. Метрики стабильности могут строиться на основе расстояния между векторами важности признаков для близких

входов (например, с помощью косинусной меры, корреляции или KL-дивергенции). В условиях высокой размерности используются также структурные меры различия, такие как расстояние по деревьям решений или изменение порядка значимости признаков. Стабильные объяснения повышают доверие, поскольку демонстрируют, что модель опирается на устойчивую логику [2].

Второй ключевой аспект – согласованность (consistency). Разные методы интерпретации одной и той же модели на одних и тех же данных должны давать сходные объяснения. Если SHAP и LIME показывают радикально разные приоритеты признаков, это указывает либо на сложность самой модели, либо на неустойчивость используемых интерпретаторов. Метрики согласованности включают корреляционные показатели между оценками значимости признаков, визуальную симметрию карт активации и меры структурного подобия между логическими объяснениями. Кроме того, важно оценивать согласованность интерпретаций с экспертными знаниями – то есть насколько объяснение модели соответствует разумным или ожидаемым паттернам, известным в предметной области.

Следующий важный показатель – локальная точность объяснения (local fidelity). Он измеряет, насколько хорошо объясняющая модель (например, локальная линейная аппроксимация в LIME) воспроизводит поведение исходной модели вблизи анализируемого примера. Если интерпретация говорит, что признак X играет главную роль, но удаление X из входа не меняет предсказания модели, то доверие к такому объяснению снижается. Метрики точности интерпретации могут включать ошибки аппроксимации, изменение предсказаний при варьировании важнейших признаков, а также тестирование на контрфактических примерах [3].

Отдельно выделяется вопрос воспроизводимости интерпретации. При повторном анализе одного и того же примера с теми же параметрами объяснение должно быть идентичным или минимально отличаться. Это особенно важно в условиях MLOps – при внедрении интерпретируемости в CI/CD пайплайны и при документировании решений для нормативных целей. Невоспроизводимая интерпретация может дискредитировать всю систему, даже если её точность остаётся высокой. Инструментальная поддержка воспроизводимости требует логгирования не только входов и предсказаний, но и используемых версий интерпретаторов, параметров генерации объяснений и состояния модели [4].

Отдельной проблемой становится оценка доверия со стороны пользователя. В интерфейсных решениях доверие к объяснению может оцениваться с помощью пользовательских метрик: времени, затраченного на анализ, субъективной оценки понятности, кликов по объяснению, реакций в диалоговых системах. Эти метрики, хотя и не математически строгие, предоставляют важные

сведения о том, как интерпретации влияют на принятие решений и уровень уверенности человека в ИИ [5].

Для поддержки оценки доверия к интерпретациям создаются специализированные библиотеки и платформы. Среди них можно выделить IBM AI Fairness 360, Alibi Explain, InterpretML, Captum, Fiddler AI, Arize AI и TruEra, каждая из которых предлагает средства не только генерации объяснений, но и анализа их качества. Они позволяют отслеживать дрейф объяснений, строить метрики согласованности, логгировать объяснения в реальном времени и визуализировать их с учётом контекста применения модели. Некоторые из этих инструментов поддерживают Model Lineage Tracking – отслеживание истории изменений модели и сопоставление этого с изменением структуры объяснений.

Особое значение приобретает сценарное тестирование интерпретаций, когда создаются контролируемые ситуации, в которых ожидаемое объяснение известно заранее. Например, можно сконструировать искусственные примеры, в которых только один признак влияет на решение, и проверить, будет ли интерпретатор выделять именно его. Это позволяет валидировать интерпретаторы как программные компоненты, аналогично тестированию обычного ПО.

Таким образом, доверие к интерпретации – это не столько вопрос субъективного восприятия, сколько предмет системного анализа, включающего оценку стабильности, согласованности, локальной точности, воспроизводимости и соответствия контексту.

Инструменты, обеспечивающие такую оценку, становятся неотъемлемой частью инженерии доверенного ИИ, позволяя не только объяснять поведение модели, но и гарантировать, что это объяснение само по себе заслуживает доверия. В условиях стремительно усложняющихся моделей и растущей нормативной нагрузки такие инструменты играют всё более важную роль в обеспечении прозрачности, подотчётности и безопасности интеллектуальных решений.

Список литературы:

1. Alvarez-Melis, D. On the Robustness of Interpretability Methods / D. Alvarez-Melis, T. S. Jaakkola // arXiv. – 2018. – arXiv:1806.08049.

2. Ghorbani, A. Interpretation of Neural Networks is Fragile / A. Ghorbani, A. Abid, J. Zou // Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Vol. 33, № 1. – P. 3681–3688. – DOI 10.1609/aaai.v33i01.33013681.

3. Yeh, C.-K. On the (In)fidelity and Sensitivity of Explanations / C.-K. Yeh, C.-Y. Hsieh, A. Suggala [et al.] // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.

4. Dombrowski, A.-K. Explanations Can Be Manipulated and Geometry is to Blame / A.-K. Dombrowski, M. Alber, C. Anders [et al.] // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.

5. Slack, D. Fooling LIME and SHAP : Adversarial Attacks on Post Hoc Explanation Methods / D. Slack, S. Hilgard, E. Jia [et al.] // Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. – New York : ACM, 2020. – P. 180–186. – DOI 10.1145/3375627.3375830.