УДК 004.89

ИНСТРУМЕНТЫ ОЦЕНКИ СТАБИЛЬНОСТИ ИНТЕРПРЕТАЦИЙ ПРИ ИЗМЕНЕНИИ ВХОДНЫХ ДАННЫХ

Омонов Ю.А., студент гр. БКНАД211, II курс
Национальный исследовательский университет «Высшая школа экономики»,
г. Москва

Стабильность интерпретации модели машинного обучения – один из центральных критериев доверия к её объяснениям. Даже если предсказание модели является корректным и соответствует фактам, оно не будет принято пользователем, если сопровождающее его объяснение оказывается непредсказуемым, изменчивым или нерепрезентативным при незначительных модификациях входных данных.

Особенно это критично в сценариях, где объяснение используется для принятия решений – будь то медицинская диагностика, судебная экспертиза, кредитный скоринг или автономное управление. В таких условиях ключевую роль начинают играть инструменты оценки стабильности интерпретаций, позволяющие количественно и визуально проверить, насколько устойчивы объяснительные механизмы модели к небольшим флуктуациям, шумам или модификациям входа [1].

Под стабильностью интерпретации понимается согласованность выходного объяснения при варьировании входных данных в пределах, не изменяющих суть задачи. Если, например, модель классифицирует изображение кошки с высокой уверенностью, то объяснение (например, карта важности признаков или Grad-CAM-визуализация) также должно продолжать указывать на морду, уши или другие семантически значимые части животного, даже если изображение повёрнуто на несколько градусов, немного зашумлено или подверглось мягкой фильтрации. Нарушение этой согласованности может указывать как на переобученность модели, так и на нестабильность самого интерпретатора.

Современные инструменты оценки стабильности интерпретации развиваются в трёх основных направлениях: метрическое сравнение интерпретаций, визуальная диагностика расхождений, и автоматизированное стресс-тестирование интерпретируемости [2].

Первый подход предполагает формализацию сравнения двух (или более) интерпретаций, полученных на основе близких входов.

Используются различные метрики расстояния между объяснениями: [3]

- косинусная близость или корреляция между векторами важности признаков (в SHAP, LIME и других методах);
- Jaccard-индекс для пересечения наиболее важных признаков в двух интерпретациях; Wasserstein-дистанция (earth mover's distance) между картами важности в визуальных методах (например, между Grad-CAM картами);
- структурное сравнение деревьев, в случае интерпретации моделей в виде логических правил.

Чем меньше расстояние между интерпретациями при минимальных изменениях входа, тем выше стабильность объяснителя. Эти метрики могут быть агрегированы по выборке и отображены в виде распределений, выявляя аномалии или чувствительные области пространства данных.

Второй подход – визуальная диагностика расхождений – особенно актуален в компьютерном зрении и обработке текста. Например, платформа Captum для PyTorch позволяет сравнивать Grad-CAM карты или карты значимости признаков side-by-side при оригинальном и модифицированном входе. Аналогично, инструменты типа SHAP позволяют интерактивно сравнивать бар-чарты или waterfall-графики значимости признаков для разных случаев. Если при незначительном изменении входа в объяснении радикально меняются важные признаки, это может быть визуально зафиксировано и использовано для последующего анализа причин нестабильности.

Третий и наиболее системный метод – это автоматизированное тестирование стабильности. Существуют фреймворки, позволяющие сгенерировать серию изменений входа по заранее заданным правилам – например, добавление гауссовского шума, удаление случайных слов, сдвиг изображения, изменение яркости или контраста. На каждый из таких входов модель выдаёт предсказание, а интерпретатор – объяснение. Далее анализируется, насколько различаются объяснения: строятся графы изменений, тепловые карты расхождений или гистограммы устойчивости. Такие методы позволяют оценить robustness интерпретатора на множестве сценариев, включая реальные атаки, ошибки датчиков, неполные данные и прочие искажения.

В задачах с текстом применяется подход на основе удаления или перефразирования токенов, при котором сравниваются интерпретации исходного текста и модифицированного (например, с синонимами или заменённым порядком слов). Если важность ключевых слов сохраняется, объяснение считается устойчивым. При этом, даже если модель продолжает правильно классифицировать модифицированный текст, но объяснение сдвигается на иные слова – это может быть индикатором структурной нестабильности.

Дополнительным направлением является оценка «интерпретационной чувствительности», то есть измерение того, насколько сильно объяснение

зависит от малых изменений в данных по сравнению с самим предсказанием. Метрика интерпретационной чувствительности может быть определена как отношение изменения интерпретации к изменению выхода модели. Высокое значение указывает на то, что объяснение реагирует сильнее, чем сама модель – ситуация крайне нежелательная для надёжной интерпретируемости.

Некоторые системы MLOps и платформы доверенного ИИ (например, Fiddler, Arize, Truera) уже включают в себя модули автоматического мониторинга стабильности объяснений, как часть наблюдаемости модели (model observability). Это позволяет в продакшн-средах отслеживать, как изменяется структура объяснений при появлении новых данных, дрейфе распределений или миграции моделей. Например, можно настроить сигнал тревоги, если SHAP-объяснение на новых данных уходит за пределы доверительного интервала, построенного на тренировочной выборке.

В условиях нормативного контроля и требований подотчётности (например, в ЕС, США и некоторых азиатских юрисдикциях) стабильность интерпретации становится не просто желательной, а необходимой. В частности, регулирующие органы требуют, чтобы объяснения ИИ-систем были не только корректными, но и воспроизводимыми при повторном предъявлении аналогичных ситуаций. Инструменты оценки стабильности играют здесь роль технического механизма валидации соблюдения таких требований [4].

Таким образом, стабильность интерпретаций при изменении входных данных – это критически важное измерение доверия к объяснимому искусственному интеллекту [5].

Инструменты для её оценки позволяют выявлять слабые места в моделях и интерпретаторах, корректировать подходы к обучению, а также обеспечивать соответствие нормативным и этическим стандартам.

По мере усложнения архитектур и роста значимости ИИ в реальной жизни, такие инструменты становятся не факультативным, а обязательным компонентом ответственной инженерии интеллектуальных систем.

Список литературы:

1. Kindermans, P.-J. The (Un)reliability of Saliency Methods / P.-J. Kindermans, S. Hooker, J. Adebayo [et al.] // Explainable AI : Interpreting, Explaining and Visualizing Deep Learning. – Cham : Springer, 2019. – P. 267–280. – DOI 10.1007/978-3-030-28954-6_14.

2. Rebuffi, S.-A. There and Back Again : Revisiting Backpropagation Saliency Methods / S.-A. Rebuffi, R. Fong, X. Ji, A. Vedaldi // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. – Seattle : IEEE, 2020. – P. 8836–8845.

3. Tomsett, R. Sanity Checks for Saliency Metrics / R. Tomsett, D. Harborne, S. Chakraborty [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Vol. 34, № 4. – P. 6021–6029. – DOI 10.1609/aaai.v34i04.6064.

4. Bansal, N. How Well Do Explanation Methods Explain the Behavior of Black-Box Models? / N. Bansal, C. Agarwal, A. Nguyen // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. – Stroudsburg : ACL, 2020. – P. 3072–3082.

5. Ancona, M. Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks / M. Ancona, E. Ceolini, C. Öztireli, M. Gross // International Conference on Learning Representations. – 2018.