УДК 004.89

МЕТОДЫ ОБЪЯСНЕНИЯ ПРЕДСКАЗАНИЙ МОДЕЛЕЙ В ЗАДАЧАХ С ВЫСОКОЙ РАЗМЕРНОСТЬЮ

Мещеряков О.Г., студент гр. М05-212, II курс
Московский физико-технический институт, г. Москва

Современные модели машинного обучения широко применяются в задачах с высокой размерностью данных – таких, где число признаков может исчисляться сотнями тысяч или даже миллионами. Это характерно для геномики, нейровизуализации, анализа естественного языка, многоканальных сенсорных потоков и высокоразрешающих изображений. Однако именно в этих случаях проблема интерпретируемости становится наиболее острой. Человек, взаимодействующий с моделью, не в состоянии проанализировать значимость и взаимодействие столь большого числа признаков [1].

Более того, сами методы объяснения, разработанные для задач со сравнительно небольшим числом входов, сталкиваются в условиях высокой размерности с фундаментальными ограничениями: вычислительной неустойчивостью, переусложнённостью визуализаций, когнитивной перегрузкой и отсутствием интуитивной структуры в объяснениях. Поэтому в последние годы активно развиваются специализированные методы интерпретации, адаптированные под специфику высокоразмерных задач [2].

Одной из основных трудностей является так называемое «проклятие размерности»: по мере увеличения числа признаков пространство данных разреживается, взаимные зависимости между признаками становятся всё более сложными, а статистические методы теряют устойчивость. Даже при использовании устойчивых моделей, таких как градиентный бустинг или сверточные нейросети, объяснение, основанное на всех признаках одновременно, оказывается либо недостоверным, либо невозможным в силу своей избыточной сложности. Поэтому первым шагом в интерпретации таких моделей обычно становится агрегация, отбор или переопределение признаков.

Это может быть реализовано с помощью пороговой фильтрации важности признаков (например, с использованием модифицированных SHAP или permutation importance), группировки признаков по семантике или анатомии (в медицине, биологии), а также выделения значимых компонент в проекционном или латентном пространстве модели. Часто модель воспринимается не как преобразователь отдельных признаков, а как порождающая механизм,

выявляющий скрытые, более компактные факторы, которые затем становятся объектом интерпретации.

В задачах, где данные высокой размерности связаны с изображениями или текстом, важную роль играют методы интерпретации в скрытом пространстве модели. Например, в трансформерах интерпретация может осуществляться через механизмы внимания, которые позволяют визуализировать, какие участки текста или изображения влияют на конкретное предсказание. Однако в высокоразмерных входах даже attention-матрицы могут быть чересчур плотными, что требует введения иерархических или свернутых представлений.

В нейровизуализации или обработке медицинских изображений применяется анализ латентных слоёв и понижение размерности (например, с помощью PCA или UMAP) для последующего построения семантически интерпретируемых карт активации. В таких случаях результат интерпретации не выражается напрямую через отдельные пиксели или слова, а строится на уровне семантических групп или концептов, выделенных моделью.

В дополнение к этому, важную роль играют методы локальной интерпретации с регуляризацией. Классические подходы типа LIME и SHAP не масштабируются напрямую на случаи с десятками тысяч признаков, поскольку создают слишком громоздкие объяснения. Поэтому разрабатываются их модификации, позволяющие получить редуцированные, но стабильные объяснения. Например, использование L1-регуляризации позволяет ограничить число признаков, входящих в локальное объяснение, а байесовские модификации SHAP учитывают априорные распределения значимости и подавляют неинформативные вкладчики. Эти методы позволяют выявить компактный набор признаков, вносящих наибольший вклад в решение, при этом избегая переобъяснения или включения шумовых факторов.

Некоторые подходы идут дальше, предлагая интерпретацию не в терминах исходных признаков, а в терминах более высокоуровневых понятий, таких как концепты, логические правила или семантические кластеры. Например, в биоинформатике это может быть переход от отдельных генов к сигнальным путям, в текстах – от слов к темам, в изображениях – от пикселей к объектам. Такая интерпретация возможна благодаря концептуальному разложению или онтологической агрегации, где каждый фрагмент объяснения соотносится с понятием, понятным человеку, и может быть проверен в доменной практике. Это особенно важно в тех областях, где вывод модели должен быть не только точным, но и объяснимым – например, в медицине, где объяснение может быть сопоставлено с клиническим протоколом.

В задачах высокой размерности важно также учитывать вопрос устойчивости интерпретации. Даже при использовании мощных моделей, нестабильность объяснения (то есть изменение значимости признаков при малых

флуктуациях входа) может дискредитировать результат в глазах пользователя. Поэтому интерпретируемость должна быть сопряжена с метриками доверия: стабильностью, воспроизводимостью, консистентностью между похожими примерами. Это требует внедрения процедур многократного тестирования объяснений, анализа вариативности значений важности, а также применения специальных инструментов аудита интерпретируемости, которые автоматически отслеживают расхождения в объяснениях модели при различных сценариях.

Не менее важной проблемой является визуализация результатов интерпретации в условиях высокой размерности. Прямое отображение вкладов тысяч признаков невозможно, поэтому используются методы масштабируемой визуализации: динамические тепловые карты, интерактивные графы зависимостей, иерархические дашборды, где пользователь может «погружаться» в уровни детализации, начиная с агрегированных мета-признаков и доходя до конкретных входов. Кроме того, растёт интерес к генерации текстовых объяснений, где поведение модели описывается на естественном языке, с опорой на структурные или семантические паттерны, выявленные на тренировочных данных.

Таким образом, интерпретация предсказаний в задачах с высокой размерностью требует комплексного подхода, сочетающего методы отбора признаков, работы в скрытом пространстве, регуляризированных объяснений, семантической агрегации и масштабируемой визуализации. Она не может быть сведена к одномерной визуализации важности, но должна быть частью многоуровневой инфраструктуры доверия, встроенной в жизненный цикл модели [3].

В условиях растущей регуляторной нагрузки, социальной ответственности и этических ожиданий от ИИ, такие методы становятся не только средством объяснения, но и механизмом обеспечения подотчётности и надёжности сложных машинных решений [4–6].

Список литературы:

1. Fong, R. C. Interpretable Explanations of Black Boxes by Meaningful Perturbation / R. C. Fong, A. Vedaldi // 2017 IEEE International Conference on Computer Vision. – Venice : IEEE, 2017. – P. 3449–3457. – DOI 10.1109/ICCV.2017.371.

2. Petsiuk, V. RISE : Randomized Input Sampling for Explanation of Black-box Models / V. Petsiuk, A. Das, K. Saenko // Proceedings of the British Machine Vision Conference. – 2018.

3. Zintgraf, L. M. Visualizing Deep Neural Network Decisions : Prediction Difference Analysis / L. M. Zintgraf, T. S. Cohen, T. Adel, M. Welling // International Conference on Learning Representations. – 2017.

4. Adebayo, J. Sanity Checks for Saliency Maps / J. Adebayo, J. Gilmer, M. Muelly [et al.] // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.

5. Hooker, S. A Benchmark for Interpretability Methods in Deep Neural Networks / S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.

6. Fong, R. Understanding Deep Networks via Extremal Perturbations and Smooth Masks / R. Fong, M. Patrick, A. Vedaldi // 2019 IEEE/CVF International Conference on Computer Vision. – Seoul : IEEE, 2019. – P. 2950–2958.