УДК 004.89

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ШЕПЛИ ДЛЯ КОЛИЧЕСТВЕННОЙ ОЦЕНКИ ВКЛАДА ПРИЗНАКОВ

Ермаков С.Ю., студент гр. 617, II курс
Московский государственный университет имени М.В. Ломоносова,
г. Москва

Оценка вклада отдельных признаков в предсказание модели машинного обучения является центральной задачей интерпретируемости, особенно в условиях растущего интереса к доверенному искусственному интеллекту. Среди множества существующих подходов наибольшую строгость и математическую обоснованность демонстрируют методы, основанные на концепции значений Шепли, заимствованной из теории кооперативных игр. Они позволяют не просто указать, какие признаки повлияли на конкретное предсказание, но и количественно оценить вклад каждого признака, учитывая сложные взаимодействия между ними и все возможные подмножества признакового пространства. Это делает методы Шепли особенно ценными для задач, где необходима прозрачность, воспроизводимость и математическая интерпретируемость объяснений [1].

Метод Шепли был предложен Ллойдом Шепли в 1953 году для оценки вклада участников в совместной игре. В терминах машинного обучения каждый «игрок» – это признак, а «игра» – это процесс предсказания модели. Значение Шепли для признака определяется как средний прирост (или убыль) качества предсказания, возникающий при включении этого признака в различные подмножества других признаков. Иными словами, мы представляем себе все возможные порядки, в которых признаки могут быть поданы в модель, и фиксируем, насколько конкретный признак изменяет предсказание в каждом из таких порядков. После усреднения по всем возможным комбинациям получаем величину, отражающую «справедливый» вклад данного признака [2].

Главное достоинство значений Шепли – их аксиоматическая строгость. Они удовлетворяют ряду ключевых свойств: эффективности (все вклады суммируются в изменение предсказания), симметрии (одинаковые по влиянию признаки получают одинаковую оценку), нулевого вклада (если признак не влияет на предсказание ни в одном подмножестве, его значение Шепли равно нулю), и аддитивности (при объединении двух моделей значения суммируются). Эти свойства придают объяснениям на основе Шепли не только

количественную выразительность, но и правовую и этическую защищённость в контекстах, где требуется нормативная отчётность.

Практическая реализация метода Шепли в задачах машинного обучения может быть затруднена из-за комбинаторной сложности: число подмножеств признаков экспоненциально возрастает с ростом размерности. Для преодоления этого барьера разработаны приближённые методы, такие как KernelSHAP, TreeSHAP и DeepSHAP. KernelSHAP использует подгонку локальной линейной модели по выборке случайных масок признаков и соответствующих предсказаний модели. Он универсален и может применяться к любым моделям, включая «чёрные ящики», но требует большого числа прогонов модели для оценки. TreeSHAP оптимизирован для градиентных бустингов и деревьев решений; он использует структуру дерева для точного и быстрого вычисления значений Шепли, избегая необходимости в явном перечислении подмножеств. DeepSHAP комбинирует идеи из Layer-wise Relevance Propagation и SHAP для нейросетевых моделей, позволяя учитывать иерархию слоёв и нелинейные зависимости между признаками [3].

Методы Шепли находят широкое применение в задачах с высокой регуляторной или социальной значимостью. В кредитном скоринге значения Шепли позволяют объяснить, почему одному заёмщику был одобрен кредит, а другому – нет, с точным указанием вклада каждой характеристики (доход, стаж, кредитная история и пр.). В здравоохранении они помогают врачам интерпретировать рекомендации диагностических моделей, показывая, какие биомаркеры или симптомы сыграли ключевую роль в постановке диагноза. В задачах обнаружения мошенничества Шепли-метрики позволяют не только детектировать подозрительные транзакции, но и объяснить, какие признаки (частота операций, геолокация, шаблоны действий) вызвали тревогу модели [4].

Применение методов Шепли позволяет решать и более тонкие задачи: выявление взаимодействий признаков, оценка важности групп признаков, обнаружение предвзятости (bias) в отношении социально значимых характеристик. Например, анализ значений Шепли может показать, что модель систематически учитывает пол или возраст даже при их формальном исключении из входных данных – за счёт коррелирующих прокси-признаков. Кроме того, метод применим в оценке дрейфа модели: сравнение распределений значений Шепли на старой и новой выборках позволяет отслеживать, как изменяется логика модели во времени [5].

Одной из сложностей остаётся интерпретация самих значений Шепли в высокоразмерных пространствах. Для решения этой проблемы используются агрегационные методы: усреднение по классам, построение тепловых карт важности, группировка по смысловым категориям. Также набирает популярность визуализация объяснений в виде водопадных графов (waterfall plots), где

предсказание представлено как сумма вклада признаков, начиная от базового значения. Это позволяет пользователю видеть, как каждый признак сдвигает результат вверх или вниз, формируя интуитивное объяснение.

Инструментальная поддержка методов Шепли активно развивается. Библиотеки SHAP, InterpretML, Alibi, Captum, ExplainX, а также платформы вроде Google What-If Tool, Fiddler AI и Azure Interpretability SDKпозволяют интегрировать интерпретацию с помощью Шепли в MLOps-пайплайны, CI/CD-модели и интерфейсы для конечных пользователей. Все чаще значения Шепли становятся частью модел-карт (model cards) – стандартов документирования моделей, принятых в крупных ИТ-компаниях и регулируемых секторах.

Таким образом, использование методов Шепли для количественной оценки вклада признаков представляет собой не только математически строгий, но и практически мощный подход к интерпретации моделей. Он сочетает аксиоматичность, адаптивность к различным архитектурам, устойчивость к коррелированным данным и широкую применимость в реальных задачах.

В условиях растущего запроса на доверие, нормативную подотчётность и прозрачность ИИ, методы Шепли формируют фундамент интерпретируемой инженерии машинного обучения, позволяя превратить сложные модели в предсказуемые и объяснимые инструменты принятия решений.

Список литературы:

1. Shapley, L. S. A Value for n-Person Games / L. S. Shapley // Contributions to the Theory of Games II / ed. by H. W. Kuhn, A. W. Tucker. – Princeton : Princeton University Press, 1953. – P. 307–317.

2. Štrumbelj, E. Explaining Prediction Models and Individual Predictions with Feature Contributions / E. Štrumbelj, I. Kononenko // Knowledge and Information Systems. – 2014. – Vol. 41, № 3. – P. 647–665. – DOI 10.1007/s10115-013-0679-x.

3. Janzing, D. Feature Relevance Quantification in Explainable AI : A Causal Problem / D. Janzing, L. Minorics, P. Blöbaum // Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. – 2020. – Vol. 108. – P. 2907–2916.

4. Aas, K. Explaining Individual Predictions when Features are Dependent : More Accurate Approximations to Shapley Values / K. Aas, M. Jullum, A. Løland // Artificial Intelligence. – 2021. – Vol. 298. – Article 103502. – DOI 10.1016/j.artint.2021.103502.

5. Sundararajan, M. The Many Shapley Values for Model Explanation / M. Sundararajan, A. Najmi // Proceedings of the 37th International Conference on Machine Learning. – 2020. – Vol. 119. – P. 9269–9278