УДК 004.89

ИНТЕРПРЕТАЦИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ С ПОМОЩЬЮ ДЕРЕВЬЕВ РЕШЕНИЙ И ПРАВИЛ

Агафонов В.А., студент гр. 053, I курс
Томский государственный университет, г. Томск

Одной из важнейших задач, стоящих перед современным машинным обучением, является повышение прозрачности и доверия к моделям, особенно когда речь идёт о критически значимых областях – здравоохранении, правосудии, финансах, кибербезопасности и прочих.

Несмотря на широкое распространение мощных предсказательных архитектур, таких как глубокие нейросети и ансамбли, их внутреннее поведение остаётся во многом непонятным не только конечным пользователям, но и самим разработчикам.

В этих условиях возникает растущий интерес к методам интерпретации, которые позволяли бы объяснять предсказания сложных моделей понятным, логически структурированным образом. Одним из наиболее продуктивных и интуитивно доступных подходов здесь является интерпретация через деревья решений и логические правила, представляющая собой важный путь к приближению моделей машинного обучения к человеческому мышлению.

Основная идея этого направления заключается в том, чтобы аппроксимировать поведение сложной модели, обученной на исходных данных, интерпретируемой моделью-суррогатом – как правило, это дерево решений или набор логических правил. Такая интерпретируемая структура может быть построена либо на основе исходных признаков (feature space), либо в пространстве скрытых представлений (latent space), формируя карту решений, которую можно прочитать, визуализировать и проанализировать. При этом достигается важная цель: сохранение предсказательной способности исходной модели при обеспечении когнитивной доступности и логической объяснимости её поведения.

Простейший и наиболее очевидный способ реализации данной идеи – это построение решающего дерева-суррогата (surrogate decision tree). В этом подходе мы фиксируем поведение обученной модели «чёрного ящика» (например, нейросети или градиентного бустинга) на большом наборе входных данных и затем обучаем дерево решений на тех же данных, но с использованием предсказаний исходной модели в качестве целевой переменной.

В результате получаем дерево, каждая вершина которого соответствует конкретному признаку, порогу или логическому условию, а каждый путь от

корня к листу – правилу типа «если–то». Такое дерево можно визуализировать и использовать для анализа: какие признаки наиболее значимы, как устроена логика ветвлений, где происходят ошибки, каков объём интерпретируемого подпространства. Сравнение предсказаний дерева и исходной модели позволяет оценить степень соответствия, а локальный анализ даёт возможность объяснить отдельные предсказания.

Важным преимуществом деревьев является интерпретируемость на естественном языке: они позволяют трансформировать модель в набор понятных логических условий. Это особенно ценно в задачах, где вывод должен быть понятен не только специалисту по данным, но и пользователю без технической подготовки – например, врачу, судье, сотруднику службы поддержки. Правило «если возраст > 50 и уровень глюкозы > 140, то риск диабета высокий» гораздо более интуитивно, чем плотность скрытого слоя в трансформере.

Другим важным подходом является генерация логических правил через декомпозицию модели – в частности, на основе алгоритмов rule extraction (RE). Здесь целью является не просто обучение суррогата, а извлечение компактных, логически непротиворечивых и когнитивно осмысленных правил, описывающих поведение модели или её отдельных компонентов.

Примерами таких методов являются:

- TREPAN – построение дерева решений по внутренним активациям нейросети;
- Anchors – построение якорных правил, которые локально описывают поведение модели с высокой уверенностью;
- DeepRED – извлечение правил из многослойных персептронов с сохранением структуры логики;
- CORELS, BRL, SLIM – методы построения компактных, однозначных и интерпретируемых логических правил на уровне всей выборки.

Эти алгоритмы позволяют получить объяснения в виде наборов «если–то» с приоритетами, весами и возможностью контроля сложности. Более того, такие правила могут использоваться в юридических и нормативных системах как часть документации по поведению модели, обеспечивая прозрачность, верифицируемость и возможность аудита.

Важным направлением является также локальная аппроксимация с помощью деревьев. В этом случае на небольшом окрестностном подмножестве данных (вблизи конкретного объекта) строится дерево решений, описывающее поведение модели в данной точке пространства. Такой подход аналогичен локальным линейным методам (LIME), но с логикой деревьев, и особенно эффективен, когда линейная аппроксимация недостаточна [1]. Локальные деревья можно визуализировать пользователю вместе с объектом, показать путь

прохождения по дереву, с указанием значений признаков и граничных условий – это позволяет объяснять решение модели в терминах конкретного примера [2].

В последние годы также развиваются гибридные подходы, в которых используются объяснимые компоненты внутри сложных моделей – например, attention-механизмы с логическим ограничением, композиционные модели с логикой выбора, слои интерпретируемых подрешений. Такие архитектуры стремятся соединить высокую выразительность нейросетей с прозрачностью деревьев и правил. Более того, в задачах, где предобученные модели остаются «чёрным ящиком», деревья и правила могут использоваться для верификации поведения: сравнение логики модели с экспертной логикой, поиск противоречий, проверка на соответствие политике, формализация нормативных ограничений [3].

Инструментальная поддержка данного направления также активно развивается. Библиотеки такие как Skater, Alibi, InterpretML, RuleFit, SkopeRules, PyCaret, а также модули в TensorFlow и PyTorch, позволяют встраивать интерпретируемость в процессы обучения, логгировать правила, генерировать отчёты и отчуждаемые объяснения.

Они находят применение в MLOps-сценариях, юридической отчётности, медицинских интерфейсах и даже в образовании – как средство обучения пониманию логики ИИ [4].

Таким образом, использование деревьев решений и логических правил для интерпретации моделей машинного обучения остаётся одним из наиболее эффективных и универсальных подходов. Он сочетает вычислительную точность с понятийной доступностью, формальную выразительность с когнитивной наглядностью [5].

В условиях нормативного и этического давления на индустрию ИИ, этот подход не только помогает «заглянуть внутрь модели», но и мостом соединяет машинное рассуждение с человеческим пониманием, делая искусственный интеллект по-настоящему подотчётным и доверенным.

Список литературы:

1. Breiman, L. Random Forests / L. Breiman // Machine Learning. – 2001. – Vol. 45, № 1. – P. 5–32. – DOI 10.1023/A:1010933404324.

2. Ribeiro, M. T. Anchors : High-Precision Model-Agnostic Explanations / M. T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the AAAI Conference on Artificial Intelligence. – 2018. – Vol. 32, № 1.

3. Letham, B. Interpretable Classifiers Using Rules and Bayesian Analysis / B. Letham, C. Rudin, T. H. McCormick, D. Madigan // The Annals of Applied Statistics. – 2015. – Vol. 9, № 3. – P. 1350–1371. – DOI 10.1214/15-AOAS848.

4. Friedman, J. H. Predictive Learning via Rule Ensembles / J. H. Friedman, B. E. Popescu // The Annals of Applied Statistics. – 2008. – Vol. 2, № 3. – P. 916–954. – DOI 10.1214/07-AOAS148.

5. Quinlan, J. R. Induction of Decision Trees / J. R. Quinlan // Machine Learning. – 1986. – Vol. 1. – P. 81–106. – DOI 10.1007/BF00116251.