Секция «40.2 Логика»

Поиск формальных объяснений для моделей искусственного интеллекта

Научный руководитель – Григорьев Олег Михайлович

Зданевич Андрей Вячеславович

Аспирант

Московский государственный университет имени М.В.Ломоносова, Философский факультет, Москва, Россия

E-mail: andrechats@gmail.com

Специфика современных моделей ИИ делает проблематичным интуитивное понимание их решений. Это представляет проблему как для пользователей, так и для разработчиков. Для её решения разрабатываются различные алгоритмы генерации объяснений. В разработке как этих решений, так и самого понятия объяснения полезной оказывается логическая формализация.

Для простоты рассмотрим случай классификации. Классификатор $\mathcal{M}$ представляет собой функцию классификации $k : \mathbb{F} \to \mathcal{K}$, где $\mathcal{K}$ есть множество классов $\{c_1, ..., c_n\}$, а $\mathbb{F}$ есть пространство всех возможных входных данных для классификации. Для целей формализации удобно представить, что каждая точка X в пространстве $\mathbb{F}$ есть множество литералов $\{\lambda_1, ..., \lambda_m\}$ где каждый литерал λ_i имеет вид $(f_i = v_i)$. Каждое f_i есть признак из множества признаков $\mathcal{F}$, ему сопоставляется значение v_i из множества его значений $\mathbb{D}_i$.

Объяснительной проблемой для модели $\mathcal{M}$ является упорядоченная двойка (X, c), где $k(X) = c$. Смысл объяснения заключается в том, чтобы показать, какие значения признаков были наиболее важными при классификации. Для этого вводится понятие абдуктивного объяснения.

Слабым абдуктивным объяснением $WAXp$ является подмножество $S \subseteq X$ такое, что значений в нём достаточно для объясняемого результата классификации.

$$WAXp(S, (X, c)) \Leftrightarrow \forall X'_{\in\mathbb{F}}(S \subseteq X' \Rightarrow (k(X) = k(X')))$$

S является абдуктивным объяснением AXp если и только если оно является минимальным таким подмножеством.

Для некоторых классификаторов существуют логические репрезентации: для каждого класса c_i есть формула M_{c_i} такая, что

$$k(X) = c_i \Leftrightarrow X\&\neg M_{c_i} \vDash \bot$$

То есть непротиворечивость этой формулы зависит от множества литералов X. Ясно, что в таком случае верно и $X \vDash M_{c_i}$. По теореме о полноте получаем $X \vdash M_{c_i}$. Это позволяет сформулировать особое определение абдуктивного объяснения для таких случаев:

$$WAXp(S, (X, c)) \Leftrightarrow ((S \subseteq X)\&(S \vdash M_{c_i}))$$

Абдуктивным объяснением является минимальное подмножество такого S.

Таким образом для генерации объяснения необходимо найти минимальное подмножество посылок вывода $\lambda_1, ..., \lambda_n \vdash M_{c_i}$, при котором сохраняется выводимость. Это обращает задачу поиска объяснения в задачу оптимизации вывода.

Однако этот подход связан с технической трудностью. Ясно, что любое полученное с помощью оптимизации подмножество посылок является слабым абдуктивным объяснением. Однако, чтобы показать, что оно является абдуктивным объяснением, то есть

минимальным подмножеством, необходимо показать, что дальнейшая оптимизация невозможна. Возможность подобной демонстрации серьёзно ограничивается особенностями построения вывода в конкретном исчислении.

Одним из вариантов решения может быть использование системы релевантного натурального вывода с характеристикой зависимости. Проблемой является возможность включения в характеристику зависимости выводимой формулы не необходимых посылок, например, с помощью правила введения конъюнкции. Предположительно, ограничение таких правил позволит исключить возможность включения в зависимости выводимой формулы тех посылок, которые не обосновывают полезных для вывода формул. Тогда можно получить абдуктивное объяснение, исключив из множества посылок все такие посыки, которые не участвовали в цепочке зависимостей финальной формулы.

Источники и литература

1) Marques-Silva, J. Logic-Based Explainability: Past, Present & Future. arXiv, June 4, 2024.
2) Ignatiev, A.; Narodytska, N.; Asher, N.; Marques-Silva, J. From Contrastive to Abductive Explanations and Back Again. In AIxIA 2020 – Advances in Artificial Intelligence; Baldoni, M., Bandini, S., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2021; Vol. 12414, pp 335–355.
3) Amgoud, L. Non-Monotonic Explanation Functions. In Symbolic and Quantitative Approaches to Reasoning with Uncertainty; Vejnarová, J., Wilson, N., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2021; Vol. 12897, pp 19–31.
4) Béjar, R.; Morgado, A.; Planes, J.; Marques-Silva, J. On Logic-Based Explainability with Partially Specified Inputs. arXiv.org.
5) Ignatiev, A.; Narodytska, N.; Marques-Silva, J. Abduction-Based Explanations for Machine Learning Models. AAAI 2019, 33 (01), 1511–1519.