УДК 004.89

ПРИМЕНЕНИЕ КОНТРФАКТИЧЕСКОГО АНАЛИЗА ДЛЯ ДИАГНОСТИКИ ПРЕДВЗЯТОСТИ МОДЕЛЕЙ

Уахид С.К., студент гр. ONT-2, II курс
Билькентский университет, г. Анкара

Применение контрфактического анализа для диагностики предвзятости моделей является важным подходом в области машинного обучения и искусственного интеллекта, который помогает выявлять скрытые формы предвзятости в алгоритмических системах. Контрфактический анализ представляет собой метод, при котором исследуется, как бы изменились предсказания модели, если бы исходные данные были изменены. Этот подход позволяет не только выявить возможную дискриминацию или неравенство в предсказаниях, но и понять, какие факторы или признаки оказывают влияние на решение модели, и какие из них могут привести к предвзятым результатам [1].

Одной из основных проблем машинного обучения, особенно в контексте справедливости, является наличие скрытой предвзятости в данных, которая может быть передана модели на этапе обучения. Предвзятость может проявляться в самых разных формах: от явной дискриминации по чувствительным признакам (таким как пол, возраст, раса и социальный статус) до более тонких, скрытых форм предвзятости, которые сложно заметить в процессе стандартной проверки модели. Контрфактический анализ позволяет «проверить» модель на наличие таких предвзятостей, модифицируя исходные данные таким образом, чтобы протестировать влияние этих изменений на предсказания.

Контрфактический анализ состоит в том, чтобы для одного и того же набора данных изменить один или несколько признаков и посмотреть, как изменится результат модели. Например, если модель прогнозирует вероятность того, что человек получит кредит, можно изменить признак, связанный с его полом или этнической принадлежностью, и посмотреть, как это повлияет на решение модели. Если при изменении только одного признака результат модели меняется существенно, это может свидетельствовать о том, что модель слишком сильно зависит от этого признака для принятия решения, что является проявлением предвзятости.

Примером контрфактического анализа может быть ситуация с системой, которая принимает решения о назначении социальных выплат. Если модель демонстрирует существенные различия в предсказаниях в зависимости от пола или возраста, можно изменить эти признаки и проверить, изменяется ли

решение. Контрфактический тест позволяет оценить, насколько эти признаки оказывают влияние на модель, и если модель активно использует такие признаки для принятия решения, то это может служить сигналом о возможной предвзятости. В случае, если изменения в этих признаках приводят к существенным изменениям в решении, можно сделать вывод о наличии скрытой предвзятости, что требует вмешательства для устранения таких зависимостей [2].

Контрфактический анализ помогает не только выявить явные предвзятости, но и разоблачить более скрытые зависимости, которые могут быть неочевидны на первых этапах анализа модели. Например, если модель ошибочно считает, что более высокий уровень образования или более высокое место работы автоматически означает большее доверие или меньшее количество рисков для определённой группы людей, контрфактический анализ может выявить такие закономерности, проверяя изменения в этих данных и наблюдая, как это отражается на предсказаниях модели.

Использование контрфактического анализа также способствует улучшению интерпретируемости моделей. Когда мы понимаем, как изменение отдельных признаков влияет на поведение модели, это даёт нам более чёткое представление о том, как она функционирует и какие факторы являются определяющими для её решений. Важно отметить, что контрфактический анализ может быть использован в различных типах моделей, включая как традиционные линейные модели, так и более сложные алгоритмы, такие как нейросети или ансамблевые методы.

В случае сложных моделей, таких как глубокие нейронные сети, контрфактический подход может помочь понять, какие признаки, возможно, незаслуженно оказывают влияние на итоговое предсказание, и позволяют уменьшить чрезмерное внимание модели к этим признакам [3].

Кроме того, контрфактический анализ может быть использован для оптимизации и корректировки моделей. Когда предвзятость в модели выявляется с помощью изменений признаков, можно использовать эти данные для пересмотра обучения модели. Например, можно применить методы регуляризации, чтобы ограничить влияние чувствительных признаков, или использовать более сбалансированные данные, которые более точно отражают разнообразие групп в реальном мире.

Контрфактический анализ является также мощным инструментом для мониторинга моделей в реальном времени, что особенно важно в условиях динамически меняющихся данных.

Например, когда модель продолжает работать в реальных условиях и сталкивается с новыми, незамеченными ранее группами данных, контрфактический анализ позволяет следить за тем, как модель продолжает принимать

решения и есть ли риски для её корректности и справедливости. Важно, что такие тесты помогают не только проверять поведение модели в момент её внедрения, но и в процессе её эксплуатации, что крайне важно для поддержания её этической составляющей на протяжении всей её жизненной цикла [4].

Помимо этого, контрфактический анализ может быть интегрирован в процесс проверки и тестирования моделей, что позволяет проводить регулярные аудиты на предмет скрытой предвзятости. Такой подход даёт возможность не только своевременно обнаруживать потенциальные проблемы с дискриминацией, но и систематически улучшать модели с учётом выявленных недостатков. Этот процесс является важной частью обеспечения доверия к искусственному интеллекту, поскольку он помогает гарантировать, что модели не будут воспроизводить или усиливать существующие социальные, экономические или культурные различия.

В заключение, контрфактический анализ представляет собой мощный инструмент для диагностики и устранения предвзятости в моделях машинного обучения. Этот подход помогает выявить скрытые и явные формы предвзятости, улучшить интерпретируемость моделей и адаптировать их к требованиям справедливости.

Контрфактические методы становятся неотъемлемой частью процесса обеспечения этических стандартов в искусственном интеллекте и машинном обучении, что способствует созданию более справедливых и прозрачных систем, соответствующих современным требованиям социальной справедливости [5].

Список литературы:

1. Pearl, J. Causality: Models, Reasoning and Inference / J. Pearl. – 2nd ed. – Cambridge : Cambridge University Press, 2009. – 464 p.

2. Pearl, J. The Book of Why: The New Science of Cause and Effect / J. Pearl, D. Mackenzie. – New York : Basic Books, 2018. – 432 p.

3. Chiappa, S. Path-Specific Counterfactual Fairness / S. Chiappa // Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Vol. 33, № 1. – P. 7801–7808. – DOI 10.1609/aaai.v33i01.33017801.

4. Nabi, R. Fair Inference on Outcomes / R. Nabi, I. Shpitser // Proceedings of the AAAI Conference on Artificial Intelligence. – 2018. – Vol. 32, № 1. – P. 1931–1940.

5. Galhotra, S. Fairness Testing: Testing Software for Discrimination / S. Galhotra, Y. Brun, A. Meliou // Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. – New York : ACM, 2017. – P. 498–510. – DOI 10.1145/3106237.3106277.