УДК 004.89

МЕТОДЫ ОЦЕНКИ ИНДИВИДУАЛЬНОЙ СПРАВЕДЛИВОСТИ В ПРЕДСКАЗАНИЯХ МОДЕЛЕЙ

Гаврилов И.А., студент гр. 23226, II курс
Новосибирский государственный университет, г. Новосибирск

Методы оценки индивидуальной справедливости в предсказаниях моделей машинного обучения имеют решающее значение для обеспечения того, чтобы алгоритмические решения, принимаемые в различных сферах – от кредитования до здравоохранения и правосудия – соответствовали этическим стандартам и справедливости. В отличие от групповой справедливости, которая ориентирована на равенство результатов для разных групп пользователей, индивидуальная справедливость фокусируется на том, чтобы каждому человеку или отдельной единице данных обеспечивалась справедливая и последовательная обработка. Это означает, что два индивидуальных наблюдения, которые схожи по своей сущности, должны получать схожие предсказания, независимо от того, к какой группе они принадлежат. Такая форма справедливости важна, потому что она помогает избежать ситуаций, когда небольшие различия в данных, не имеющие отношения к сути задачи, могут привести к значительным различиям в предсказаниях модели, что может быть несправедливым [1].

Одним из ключевых понятий в контексте индивидуальной справедливости является принцип, согласно которому аналогичные индивиды должны получать одинаковые решения или предсказания, если их характеристики (особенно те, которые не связаны с риском или результатом задачи) одинаковы. Например, в системе, которая решает, кому следует одобрить кредит, два человека, имеющие аналогичные финансовые данные, но отличающиеся, скажем, возрастом или полом, должны получать одинаковые решения о кредите. В противном случае, если предсказания модели будут сильно различаться для таких схожих пользователей, это может свидетельствовать о нарушении принципа индивидуальной справедливости и, как следствие, дискриминации. Стремление к индивидуальной справедливости помогает гарантировать, что несущественные различия между индивидами не станут причиной необоснованных различий в результатах.

Для оценки индивидуальной справедливости в предсказаниях моделей используется несколько различных подходов и метрик. Одна из наиболее распространённых методик – это параметрическое определение справедливости, которое предполагает использование расстояний между объектами в

пространстве признаков. В этой модели считается, что если два объекта (или индивида) близки в многомерном пространстве признаков, их предсказания должны быть схожими. Для этого часто используются такие метрики, как евклидово расстояниеили расстояние Манхэттена, которые позволяют измерить, насколько похожи два объекта в контексте их признаков [2].

Если модель принимает одинаковые решения для объектов, которые находятся рядом друг с другом в этом пространстве, это будет свидетельствовать о её справедливости. Напротив, если расстояние между объектами мало, но модель выдаёт разные результаты, это может свидетельствовать о наличии предвзятости.

Другим подходом является анализ с использованием принципа контрфактической справедливости. Этот метод основывается на гипотетических сценариях, в которых рассматривается, как изменится предсказание модели, если изменить один или несколько признаков объекта, оставаясь при этом в рамках того же контекста. Идея заключается в том, что если два индивидуальных наблюдения различаются по одному признаку (например, возраст или пол), но все остальные признаки остаются неизменными, то модель должна выдать одинаковое решение для этих двух наблюдений, если различие в признаках не имеет существенного влияния на результат. Контрфактическая справедливость помогает выявить, насколько модель зависит от признаков, которые не должны оказывать влияния на её предсказания.

Одним из примеров инструмента, используемого для проверки индивидуальной справедливости, является метод измерения отклонений предсказаний модели для схожих индивидов. В этом случае для каждой пары похожих наблюдений вычисляется разница в их предсказаниях. Если разница слишком велика, это может свидетельствовать о том, что модель несправедливо обращается с этими схожими наблюдениями. Такая проверка помогает находить и устранять недочёты в модели, которые могут проявляться на индивидуальном уровне [3].

Применение методов объяснимого ИИ в сочетании с индивидуальной справедливостью играет важную роль в обеспечении транспарентности и объяснимости предсказаний моделей. Например, если модель принимает решение о выдаче кредита, важно не только, чтобы решение было справедливым, но, и чтобы оно было объяснимо для каждого индивидуального пользователя. В таких случаях использование методов визуализации или объяснения модели, таких как SHAP (Shapley Additive Explanations) или LIME (Local Interpretable Model-Agnostic Explanations), позволяет понять, какие признаки были наиболее важными для принятия решения и почему. Если объяснение модели для двух схожих пользователей различается, это может быть признаком нарушения принципа индивидуальной справедливости, и требуется вмешательство для корректировки модели.

Особое внимание стоит уделить оценке отклонений при изменении чувствительных признаков, таких как пол, возраст или этническая принадлежность. Если модель демонстрирует значительные различия в предсказаниях при изменении этих признаков, это может указывать на предвзятость, даже если остальные признаки остаются одинаковыми. Таким образом, важно не только анализировать индивидуальную справедливость для нейтральных признаков, но и контролировать, как чувствительные признаки влияют на результаты модели.

На практике для реализации индивидуальной справедливости часто применяют регуляризацию в процессе обучения модели. Регуляризация помогает уменьшить зависимость модели от нерелевантных признаков, которые могут вызвать несправедливые различия в предсказаниях для схожих объектов. Это позволяет модели лучше фокусироваться на действительно значимых для задачи признаках и минимизировать влияние признаков, которые могут создавать предвзятость [4].

Важным аспектом является также постоянный мониторинг и аудит модели в процессе её эксплуатации, чтобы убедиться в том, что она сохраняет индивидуальную справедливость в реальных условиях. Это особенно важно для систем, которые продолжают обучаться на новых данных, поскольку изменения в данных могут непредсказуемо повлиять на предсказания модели, нарушая принцип индивидуальной справедливости. Регулярный анализ и тестирование модели на соответствие принципам справедливости позволяют оперативно выявлять и устранять любые отклонения.

Таким образом, оценка индивидуальной справедливости в предсказаниях моделей является необходимым шагом для обеспечения этичности и прозрачности алгоритмических решений. Применение различных статистических и машинных методов для выявления и устранения несправедливости помогает гарантировать, что модели будут действовать честно и последовательно для каждого пользователя, независимо от его демографических характеристик. Внедрение таких практик способствует повышению доверия к алгоритмическим системам и делает их более ответственными и справедливыми [5].

Список литературы:

1. Lahoti, P. ifair: Learning Individually Fair Data Representations for Algorithmic Decision Making / P. Lahoti, K. P. Gummadi, G. Weikum // IEEE International Conference on Data Engineering. – 2019. – P. 1334–1345. – DOI 10.1109/ICDE.2019.00121.

2. Gillen, S. Online Learning with an Unknown Fairness Metric / S. Gillen, C. Jung, M. Kearns, A. Roth // Advances in Neural Information Processing Systems. – 2018. – Vol. 31. – P. 2605–2614.

3. Yona, G. Probably Approximately Metric-Fair Learning / G. Yona, G. N. Rothblum // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 5680–5688.

4. Joseph, M. Fairness in Learning: Classic and Contextual Bandits / M. Joseph, M. Kearns, J. H. Morgenstern, A. Roth // Advances in Neural Information Processing Systems. – 2016. – Vol. 29. – P. 325–333.

5. Chierichetti, F. Fair Clustering through Fairlets / F. Chierichetti, R. Kumar, S. Lattanzi, S. Vassilvitskii // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5029–5037.