УДК 004.89

СТАТИСТИЧЕСКИЕ ПОДХОДЫ К ВЫЯВЛЕНИЮ И ИЗМЕРЕНИЮ ГРУППОВОЙ ПРЕДВЗЯТОСТИ

Носов А.Т., студент гр. 034, I курс
Томский государственный университет, г. Томск

Статистические подходы к выявлению и измерению групповой предвзятости в моделях машинного обучения играют важнейшую роль в обеспечении справедливости и этичности алгоритмических решений. Групповая предвзятость может проявляться, когда модель систематически ошибается для определённой группы пользователей, что приводит к неравному обращению с ними [1].

Например, это может касаться различий в ошибках для разных демографических групп, таких как различия по возрасту, полу, этнической принадлежности или социальному статусу.

Важно понимать, что выявление и измерение предвзятости не ограничивается только прямыми признаками дискриминации, такими как расовая или половая предвзятость, но и охватывает более тонкие формы, которые могут не быть очевидными без применения сложных статистических методов.

Одним из основных подходов к выявлению групповой предвзятости является использование метрик справедливости, которые позволяют оценить, насколько равномерно распределяются ошибки или результаты предсказаний модели между различными группами. Статистический анализ таких метрик может включать вычисление показателей различий между группами, таких как разница в положительных исходах, ошибках первого рода (ложноположительных) и ошибках второго рода (ложных отрицательных). Если одна группа получает больше положительных исходов или меньше ошибок по сравнению с другой, это может свидетельствовать о наличии предвзятости в модели. Например, если модель, предназначенная для классификации заявок на кредит, чаще одобряет заявки от мужчин, чем от женщин, это может указывать на наличие гендерной предвзятости [2].

Для более комплексного анализа и количественного измерения групповой предвзятости широко применяются методы гипотезы и параметрические статистические тесты. Например, можно использовать t-тесты или ANOVA (анализ дисперсии) для проверки того, есть ли статистически значимые различия между группами по результатам предсказаний.

Такие методы позволяют выявить, существует ли статистически значимая разница в распределении ошибок или предсказаний модели для различных

групп, а также оценить, насколько эти различия влияют на качество модели. В случае обнаружения значительных различий, статистические тесты позволяют понять, насколько велик вклад определённых признаков (например, пола или расы) в эти различия и могут ли они служить причиной предвзятости.

Дополнительно к простым статистическим тестам, методы регрессии также играют ключевую роль в анализе групповой предвзятости. Логистическая регрессия или линейная регрессия могут быть использованы для анализа того, как определённые признаки (например, возраст, этническая принадлежность или уровень образования) влияют на вероятность того, что модель сделает ошибку для различных групп. Если в регрессионной модели присутствуют сильные зависимости между чувствительными признаками и ошибками, это указывает на наличие предвзятости, которую необходимо устранить. Одним из примеров является использование многофакторной регрессии, где можно учесть не только индивидуальные признаки пользователей, но и их взаимодействия, что позволяет более точно идентифицировать скрытые формы предвзятости [3].

Важной составляющей статистического анализа групповой предвзятости является также оценка справедливости на основе многоклассовых или многогрупповых моделей. В таких моделях важно не только анализировать ошибки в каждой отдельной группе, но и учитывать, как различные группы взаимодействуют между собой. Например, при обучении модели для классификации на нескольких группах (например, для разных возрастных групп или этнических категорий) необходимо исследовать не только ошибки внутри каждой группы, но и взаимодействие между группами, чтобы понять, каким образом модель может оказаться предвзятой относительно одной группы в контексте других. В таких случаях может применяться многомерный анализ, включающий методы кластеризации и основных компонент (PCA), чтобы выделить скрытые взаимосвязи между признаками и ошибками.

Кроме того, в рамках статистических подходов часто используется метод постобработки для корректировки результатов модели с целью устранения предвзятости. Например, если обнаружено, что модель значительно хуже работает для одной из групп, можно применить весовое корректирование, которое позволяет уменьшить влияние ошибок на предсказания модели для этой группы. Также можно использовать калибровку вероятностейили методы переподборки данных, чтобы привести результаты модели в соответствие с требованиями справедливости. Статистические методы, такие как переподборка с учётом пропорций или исправление вероятности (например, с использованием подходов как Исходные пропорции ошибок или Calibrated Equalized Odds), могут быть использованы для перераспределения или изменения веса ошибок в тех группах, которые были выявлены как более подверженные предвзятости.

Статистические методы также позволяют моделировать и прогнозировать влияние изменения данных на результат. Например, с помощью методов сэмплирования можно провести тестирование модели на изменённом наборе данных, чтобы проверить, как она будет вести себя при изменении соотношения пользователей из разных групп. Это помогает предсказать, как новые данные (например, введение новых пользователей с иными демографическими характеристиками) могут повлиять на справедливость модели в будущем. Такой подход становится особенно важным в динамично изменяющихся областях, таких как здравоохранение или финансовые услуги, где модели должны адаптироваться к новым данным, сохраняя при этом равенство и справедливость [4].

Кроме того, статистические методы могут использоваться для оценки долгосрочных последствий моделей, особенно в контексте социального воздействия. Например, если модель демонстрирует перекос в ошибках между группами, это может привести к усилению социальной несправедливости с течением времени. Статистическое моделирование позволяет предсказать, как такие перекосы могут изменяться с течением времени и какие последствия могут возникнуть в долгосрочной перспективе для различных групп. Например, в банковской сфере это может повлиять на доступность кредитов для определённых социальных групп, что в дальнейшем может повлиять на их финансовое положение и социальное благосостояние.

В заключение, статистические подходы к выявлению и измерению групповой предвзятости играют важную роль в создании этичных и справедливых моделей машинного обучения. С помощью методов анализа ошибок, статистических тестов, регрессионных моделей и методов постобработки можно не только выявить предвзятость в моделях, но и разработать стратегии для её устранения, что способствует созданию более справедливых и прозрачных систем искусственного интеллекта.

В конечном итоге, такие подходы помогают улучшить качество моделей, повысить доверие пользователей и гарантировать, что алгоритмические решения не усиливают социальные различия и не приводят к дискриминации [5].

Список литературы:

1. Hutchinson, B. 50 Years of Test (Un)fairness: Lessons for Machine Learning / B. Hutchinson, M. Mitchell // Proceedings of the Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2019. – P. 49–58. – DOI 10.1145/3287560.3287600.

2. Barocas, S. Big Data’s Disparate Impact / S. Barocas, A. D. Selbst // California Law Review. – 2016. – Vol. 104, № 3. – P. 671–732. – DOI 10.15779/Z38BG31.

3. Romei, A. A Multidisciplinary Survey on Discrimination Analysis / A. Romei, S. Ruggieri // The Knowledge Engineering Review. – 2014. – Vol. 29, № 5. – P. 582–638. – DOI 10.1017/S0269888913000039.

4. Zliobaite, I. A Survey on Measuring Indirect Discrimination in Machine Learning / I. Zliobaite // arXiv preprint. – 2015. – arXiv:1511.00148.

5. O'Neil, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy / C. O'Neil. – New York : Crown, 2016. – 272 p.