

УДК 004.89

## **ИНСТРУМЕНТЫ КОНТРОЛЯ СДВИГА ДАННЫХ ДЛЯ ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ В РЕАЛЬНЫХ УСЛОВИЯХ**

Кустов А.С., Оператор научной роты, III курс  
Военная академия связи имени Маршала Советского Союза С. М. Буденного,  
г. Санкт-Петербург

Внедрение предобученных моделей в реальные производственные условия неизбежно сталкивается с одним из наиболее фундаментальных вызовов современной машинной инженерии – сдвигом данных (data shift). Являясь результатом несовпадения между распределением данных, на которых модель обучалась, и данными, с которыми она взаимодействует в эксплуатационной среде, сдвиг существенно снижает качество предсказаний, увеличивает вероятность ошибок и подрывает доверие к системе.

Особенно уязвимыми оказываются масштабные предобученные модели, которые разрабатываются в изолированных условиях – зачастую на обобщённых, публичных или синтетических датасетах – и впоследствии используются в задачах с иным контекстом, структурой и статистикой входа. Именно поэтому разработка и внедрение инструментов контроля сдвига данных становится краеугольным камнем обеспечения надёжности и воспроизводимости поведения таких моделей в условиях реального мира [1].

Сдвиг данных может проявляться в различных формах. Среди них наиболее часто встречаются covariate shift – изменение распределения признаков при сохранении зависимости между входами и метками; label shift – изменение распределения самих меток (например, когда одни классы становятся более или менее частотными); concept shift – изменение функциональной зависимости между входами и выходом, например, из-за изменения нормативов, пользовательского поведения или формата целевой задачи; а также domain shift, когда наблюдается радикальное отличие между источником и целевой областью (например, модель, обученная на новостях, используется для юридических текстов). В каждой из этих ситуаций поведение модели может деградировать, иногда – незаметно, но с накопительным эффектом, ведущим к системным сбоям [2].

В условиях реального времени и постоянно обновляющегося потока данных критически важным становится наличие автоматизированных инструментов, способных обнаруживать, отслеживать и интерпретировать сдвиги данных, чтобы оперативно принимать решение о необходимости адаптации

модели или вмешательства пользователя. Одним из базовых классов таких инструментов являются детекторы сдвига, построенные на статистических и геометрических подходах. Они сравнивают текущие входные данные с эталонным распределением (например, распределением признаков, извлечённых из валидационного набора), используя метрики расстояния: Kullback-Leibler divergence, Wasserstein distance, Maximum Mean Discrepancy, Energy distance и др. Эти методы эффективны, если данные численно представлены в векторном пространстве, что особенно характерно для предобученных моделей, у которых есть чётко выраженное латентное пространство признаков [3].

Однако в более сложных случаях, например при работе с текстами или изображениями, необходимо использовать глубокие методы сравнения, которые опираются на эмбединги модели или на активации промежуточных слоёв. Так, в языковых моделях может сравниваться распределение sentence embeddings или логитов, а в системах компьютерного зрения – распределения карт внимания или признаков последнего сверточного слоя.

Кроме того, вводится понятие drift score – агрегированного показателя изменения, который обновляется по мере накопления новых данных и сигнализирует о приближении к порогу допустимого сдвига. Такие показатели могут быть встроены в MLOps-инфраструктуру и служить триггерами для запуска повторной валидации модели, дообучения или интервенции оператора [4].

Особое внимание в современных инструментах уделяется локализации сдвига. Недостаточно просто зафиксировать факт того, что данные изменились; необходимо понять, какие именно компоненты данных, признаки или подгруппы вызвали это отклонение. Для этого применяются методы subgroup drift detection, основанные на стратификации данных по категориям, кластерам или скрытым переменным [5].

В частности, такие подходы используются при анализе предвзятости: например, если модель, ранее демонстрировавшая равномерную производительность на данных о мужчинах и женщинах, начинает ошибаться только на одном из подмножеств, это указывает на частичный сдвиг, который может быть незаметен при агрегированной оценке.

Ещё одним важным направлением является оценка чувствительности модели к зафиксированному сдвигу. Не каждый сдвиг данных является критическим – иногда модель сохраняет работоспособность и предсказательную силу, несмотря на статистические отличия. Поэтому инструменты должны включать метрики воздействия (impact metrics): насколько изменились предсказания модели после наступления сдвига, увеличилось ли количество отказов, снизилась ли уверенность, изменилась ли интерпретируемость вывода.

Такие метрики могут быть основаны на сравнении выводов до и после наступления изменения, с привлечением специальных эталонных кейсов или использованием симуляции альтернативных сценариев.

Современные фреймворки контроля сдвига всё чаще реализуются как интерактивные и объяснимые системы, которые не только выявляют проблему, но и помогают человеку принять решение. Например, они визуализируют зоны сдвига в пространстве признаков, маркируют «новые» данные, не представленные в обучении, и предлагают возможные меры: пересэмплирование, дообучение на новой квазилинейной репрезентации, расширение обучающего корпуса.

Некоторые системы используют методы мета-обучения, чтобы предсказать, как модель поведёт себя при продолжении сдвига, и тем самым помогают планировать интервенции на упреждающей стадии.

Интеграция контроля сдвига в жизненный цикл предобученных моделей обеспечивает континуум доверия, в рамках которого модель не просто деплоится один раз, а находится под постоянным наблюдением, диагностикой и, при необходимости, адаптацией.

Такой подход особенно важен в условиях динамически изменяющейся среды: медицинские протоколы обновляются, язык эволюционирует, рыночные условия меняются. Модель, игнорирующая эти изменения, рано или поздно станет источником риска.

Наконец, по мере роста нормативных инициатив в области регулируемого ИИ – от ISO/IEC 42001 до европейского AI Act – контроль сдвига данных начинает играть юридическую роль, становясь элементом доказательства устойчивости, объяснимости и допустимости модели. Его включение в отчётность, сертификацию и аудит моделей позволяет организациям формировать обоснованную позицию относительно корректности и допустимости использования предобученных систем в изменяющихся условиях [6].

Таким образом, инструменты контроля сдвига данных являются не просто инженерной надстройкой над моделью, но важнейшим компонентом доверенного ИИ.

Они обеспечивают непрерывную адаптацию, обоснованность решений и прозрачность поведения моделей в реальном времени. Их дальнейшее развитие – как в техническом, так и в нормативном направлении – станет краеугольным фактором устойчивого внедрения предобученных ИИ-систем в чувствительные, критические и социально значимые сферы.

### Список литературы:

1. Quinonero-Candela, J. Dataset Shift in Machine Learning / J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence. – Cambridge : MIT Press, 2009. – 248 p.

2. Gama, J. A Survey on Concept Drift Adaptation / J. Gama, I. Zliobaite, A. Bifet [et al.] // ACM Computing Surveys. – 2014. – Vol. 46, № 4. – Article 44. – DOI 10.1145/2523813.
3. Lu, J. Learning under Concept Drift : A Review / J. Lu, A. Liu, F. Dong [et al.] // IEEE Transactions on Knowledge and Data Engineering. – 2019. – Vol. 31, № 12. – P. 2346–2363. – DOI 10.1109/TKDE.2018.2876857.
4. Lipton, Z. C. Detecting and Correcting for Label Shift with Black Box Predictors / Z. C. Lipton, Y.-X. Wang, A. Smola // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 3122–3130.
5. Rabanser, S. Failing Loudly : An Empirical Study of Methods for Detecting Dataset Shift / S. Rabanser, S. Günnemann, Z. C. Lipton // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.
6. Page, E. S. Continuous Inspection Schemes / E. S. Page // Biometrika. – 1954. – Vol. 41, № 1/2. – P. 100–115. – DOI 10.1093/biomet/41.1-2.100.