УДК 519.6
DOI: 10.26102/2310-6018/2024.44.1.033

Детектирование машинно-сгенерированных текстов при помощи адаптивной квантильной регрессии

А.С. Тюрин, П.В. Сараев✉

Липецкий государственный технический университет, Липецк, Российская Федерация

Резюме. В работе рассматривается задача детектирования машинно-сгенерированных текстов при помощи различных инструментов построения регрессионных моделей – классической линейной регрессии, логистической регрессии и квантильной регрессии. Прогресс в области машинного обучения позволяет создавать все более реалистичные тексты, что открывает возможности для их недобросовестного использования. По мере того, как алгоритмы генерации текстов становятся более сложными, возрастает и сложность задачи детектирования таких текстов, что также требует применения более сложных методов математического моделирования и более эффективных численных методов. Рассматриваемый алгоритм адаптивной квантильной регрессии представляет собой инструмент, который позволяет строить модели с акцентом на различные квантили, что делает его особенно полезным для детектирования нетипичных значений, что может указывать на искусственную природу текстов. Также в работе представлено подробное описание исходного открытого набора данных для обучения моделей, представляющего собой сгенерированные тексты при помощи модели GhatGPT и случайные рукописные тексты с различных форумов, приведен анализ проведенных вычислительных экспериментов. Результаты исследования показывают высокую эффективность предложенного метода в данной прикладной области.

Ключевые слова: классификация текстов, квантильная регрессия, адаптивный алгоритм, градиентный спуск, математическое моделирование, численные методы.

Для цитирования: Тюрин А.С., Сараев П.В. Детектирование машинно-сгенерированных текстов при помощи адаптивной квантильной регрессии. *Моделирование, оптимизация и информационные технологии.* 2024;12(1). URL: https://moitvivt.ru/ru/journal/pdf?id=1536 DOI: 10.26102/2310-6018/2024.44.1.033

Detecting machine-generated texts with adaptive quantile regression

A.S. Tyurin, P.V.Saraev✉

Lipetsk State Technical University, Lipetsk, the Russian Federation

Abstract. This paper considers the problem of detecting machine-generated texts using various regression model building tools - classical linear regression, logistic regression and quantile regression. Advances in machine learning are creating increasingly realistic texts, which opens the door to misuse. As text generation algorithms become more sophisticated, the complexity of the task of detecting such texts increases, which also requires more sophisticated mathematical modeling methods and more efficient numerical methods. The proposed adaptive quantile regression algorithm is a tool that allows building models with emphasis on different quantiles, which makes it particularly useful for detecting atypical values that may indicate the artificial nature of the texts. The paper also presents a detailed description of the initial open dataset for model training, which is a set of generated texts using the GhatGPT 3 model and random texts from various forums, and analyzes the computational experiments performed. The results show the high efficiency of the proposed method in this application domain.

Keywords: text classification, quantile regression, adaptive algorithm, gradient descent, mathematical modeling, numerical methods.

For citation: Tyurin A.S., Saraev P.V. Detecting machine-generated texts with adaptive quantile regression. *Modeling, Optimization and Information Technology*. 2024;12(1). URL: https://moitvivt.ru/ru/journal/pdf?id=1536 DOI: 10.26102/2310-6018/2024.44.1.033 (In Russ.).

Введение

Целью данной работы является построение модели детекции машинно-сгенерированных текстов при помощи различных методов построения математических моделей регрессии. В настоящее время большие языковые модели, такие как ChatGPT, способны генерировать достаточно качественные тексты на различные темы – в том числе, тексты на научную тематику, генерируют код на популярных языках программирования. Задача определения текстов на предмет искусственного происхождения актуальна как для медиасферы, так и в сфере науки, так как необходимо удостовериться, что предложенные исследования принадлежат именно автору. В настоящее время уже проведет ряд исследований, которые основаны на различных методах и решающие разные задачи. Так, в работе [1] предлагается решение задачи защиты больших языковых моделей (large language models, LLM) от недобросовестного использования LLM на основе статистических данных. В частности, предлагает фильтрация запросов по фильтру определенных слов, которые могут включать в себя вредоносные запросы. Схожая методика для классификации текстов применяется и в [2]. В [3] рассмотрены основные статистические отличия рукописных и сгенерированных текстов в контексте диалогов с моделью ChatGPT 3.5 по ряду лингвистических метрик. Одной из таких метрик может быть вероятность встретить следующий токен в рассматриваемой цепочке, как показано в работе [4]. Данный метод предлагает использование перекрестной проверки ожидаемости следующего токена в последовательности, т. е. представление о том, насколько неожиданны предсказания следующей лексемы одной модели для другой модели.

Рассматриваемый алгоритм адаптивной квантильной регрессии разработан в рамках данной работы и представляет из себя модификацию классического алгоритма квантильной регрессии, который решает проблему выбора оптимального уровня квантиля. Так как заранее невозможно угадать оптимальное значение, зачастую приходится провести целый ряд вычислительных экспериментов — попробовать обучить модель со всеми возможными уровнями квантиля и оценить точности полученных моделей на тестовой выборке. Данный подход представляет собой полный перебор, но существуют и более продвинутые методы, направленные на поиск оптимального значения за меньшее количество итераций обучения и проверки модели. Результаты исследования, представленные в [5], предлагают, например, включение дополнительных компонентов в функцию ошибки L1, что способствует более точному и последовательному определению квантилей в определенном диапазоне значений. Также были предложены альтернативные методы, изложенные в [6-8], которые предлагают использовать симплекс-метод, алгоритмы внутренней точки и различные сглаживающие процедуры для определения оптимального квантиля. Альтернативой является предложенный метод динамического изменения квантиля в процессе обучения модели при помощи градиентного спуска. Как показано далее, в контексте данной предметной области становится очевидным преимущество использования предложенного подхода [9] при построении моделей с большим количеством входных параметров при помощи натурального градиентного спуска, так как вычисление

обратной матрицы Фишера требует значительно больших вычислительных ресурсов [10-11].

Исходные данные

В качестве исходных данных использован открытый набор данных ai-text-detection-pile, который предназначен для задач по обнаружению машинно-сгенерированного текста с фокусом на длинные примеры. Набор содержит образцы как рукописного текста, так и текста, сгенерированного искусственным интеллектом при помощи GPT2, GPT3, ChatGPT, GPTJ. В Таблице 1 представлена информация по количеству примеров для языковых моделей. Общее количество строк для машинно-сгенерированного текста достигает 340 000. Для оригинальных рукописных текстов представлены наборы Reddit WritingPromps, содержащий 570 000 строк и другие. Для балансировки наборов текстов была взята только часть рукописных текстов.

Таблица 1 – Состав исходных данных по языковым моделям
Table 1 – Composition of input data on language models

Модель	Набор данных	Количество строк
GPT2	OpenAI gpt2-output-dataset	260 000
GPT3	pairwise-davinci	44 000
GPT3	synthetic-instruct-davinci-pairwise	30 000
GPTJ	synthetic-instruct-gptj-pairwise	44 000
ChatGPT	Scraped from twitter	5 000
ChatGPT	HC3 (ChatGPT Responses)	27 000
ChatGPT	ChatGPT Prompts/emergentmind	500

На Рисунке 1 представлено распределение длин текстов (эссе) рукописных и сгенерированных. После фильтрации аномально коротких и ограничении длины в 6000 символов. Как видно, распределения практически совпадают, что говорит о достаточной сбалансированности наборов.

Рисунок 1 – Распределение длин рукописных и сгенерированных текстов
Figure 1 – Length distribution of handwritten texts and generated texts

Все образцы текстов дополнительно были перемешаны между собой для исключения группировки по принадлежности к конкретной языковой модели. Для

дальнейшей подготовки была использована библиотека Sklearn, в частности, TfidfVectorizer. Этот класс выполняет TF-IDF преобразование над текстами. TF (частота термина) показывает, как часто определенное слово появляется в выбранном образце. Таким образом, TF измеряет важность слова в контексте отдельного текста. IDF (обратная частота документа) показывает уникальность слова в пределах того же текста. Таким образом, некоторые слова могут иметь низкое значение IDF и не будут нести ценности для модели. Для эксперимента были построены несколько наборов обработанных данных – с максимальным количеством признаков (размер словаря) от 5000 до 30000 с шагом в 5000 признаков. Также отдельно сгенерированы наборы данных с двумя вариантами длины n-грамм – от 1 до 2. Таким образом, получились 12 наборов данных.

Построение моделей

Для проведения вычислительных экспериментов были взяты реализации классической линейной регрессии, логистической регрессии и квантильной регрессии из пакета Sklearn. Собственная реализация адаптивной квантильной регрессии выполенена на языке Python. Также для эксперимента взят метод градиентного бустинга в реализации библиотеки CatBoost.

В качестве метрик сравнения моделей использованы метрики F1 и Accuracy. F1 является стандартной метрикой для оценки бинарных классификаторов. Метрика Accuracy является легко интерпретируемой и вычисляет точность подмножества: набор значений, предсказанный для выборки, должен точно отвечать соответствующему набору меток в тестовых данных:

$$accuracy(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1} 1\,(y_i = \hat{y}_i),$$

где n – размер выборки, i – номер конкретного наблюдения, 1(x) – индикаторная функция. Также для визуального контроля качества модели используется график precision-recall, как показано на Рисунке 1, где прямой угол изгиба графика соответствовал бы идеальной модели. Данные по метрикам precision-recall получены при помощи базовой модели логистической регрессии на наборе данных в 15000 признаков и значением n-граммы, равным 1. Все вычислительные эксперименты производились на компьютере с 16 Гб оперативной памяти и 16 потоками центрального процессора. При этом наборы данных, содержащие все строки и 30000 признаков, не помещаются в оперативную память. Для проведения вычислительных экспериментов все наборы данных разделены на обучающую и тестовую выборки в соотношении 80 % и 20 %.

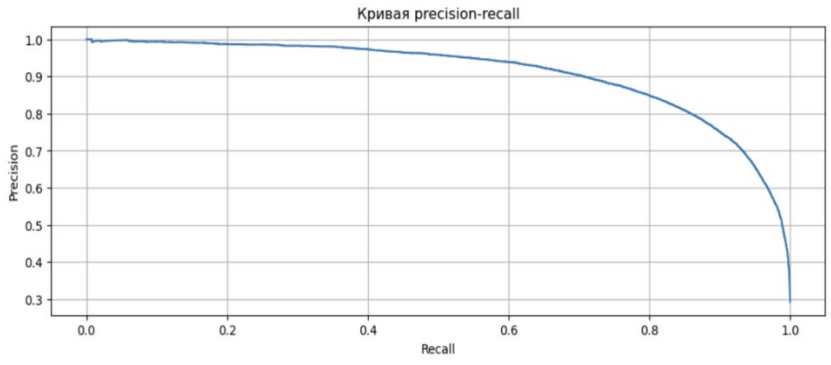

Рисунок 1 – График метрики precision-recall
Figure 1 – Precision-recall illustration

Для выбора оптимального размера набора данных построен ряд моделей при помощи логистической регрессии. Графики метрик приведены на Рисунке 2.

Рисунок 2 – Зависимость метрик модели от длины словаря
Figure 2 – Dependence of model metrics on dictionary length

Как видно на графике, размер словаря свыше 15000 признаков не вносит существенного улучшения в точность модели, поэтому в дальшейшем ограничим размер признаков до 15000. Также значительное увеличение размера набора данных приносит увеличение значения n-грамм. На Рисунке 3 представлены показатели метрик в зависимости от длины n-грамм.

Рисунок 3 – Зависимость метрик модели от длин n-грамм
Figure 3 – Dependence of model metrics on n-gram lengths

Таким образом, максимальная длина n-грамм достигает значения 2, так как дальнейшее увеличение длины не несет существенного улучшения точности для модели. Примеры n-грамм представлены на Рисунке 4.

```
[40]: ['government officials',
       'government said',
       'government spending',
       'grand jury',
       'grand theft',
       'great britain',
       'great deal',
       'great depression',
       'great number',
       'great time',
       'great way',
       'green bay',
       'greenhouse gas',
       'greenhouse gas emissions',
       'grim reaper',
       'grocery store',
       'group members',
       'group people',
       'group said',
       'groups people',
       'growing number',
       'growth rate',
       'growth tracking',
       'gtx 1080',
       'gun control']
```

Рисунок 4 – Примеры n-грамм
Figure 4 – N-gram examples

Далее были построены модели градиентного бустинга, квантильной регрессии, линейной регрессии и адаптивной квантильной регрессии на полном наборе данных с 15000 признаками.

Анализ результатов

Результаты проведения вычислительных экспериментов приведены в Таблице 2. Здесь представлены метрики точности F1 и Accuracy для различных методов, а также скорость обучения моделей в секундах на наборе данных в 200000 строк и 15000 параметров.

Таблица 2 – Сравнение методов по скорости сходимости и точности
Table 2 – Comparison of methods in terms of convergence speed and accuracy

Метод	F1	Accuracy	Время обучения, сек
Адаптивная квантильная регрессия	**0,83**	**0,91**	548
Quantile regression + градиентный спуск (полный перебор квантилей)	0,82	**0,91**	10 960
CatBoost	0,82	0,89	670
Линейная регрессия	0,79	0,81	300
Логистическая регрессия	0,80	0,87	426

Как видно из таблицы, предложенный метод адаптивной квантильной регрессии гораздо быстрее классической реализации по временным затратам и превосходит его по точности, так как позволяет изменять уровень квантиля на достаточно малое значение. В случае полного перебора всех возможных значений квантиля с шагом в 0,05 потребовалось 3 часа, в то время как адаптивной реализации потребовалось всего 9 минут. Преимущество адаптивного подхода тем сильнее выражено, чем меньше шаг подбора квантиля и чем больше размер набора данных по параметрам. При использовании большего объема оперативной памяти возможно обучение моделей с более высоким значением n-грамм, что также может повысить качество прогнозов.

Заключение

В ходе данного исследования были рассмотрены различные методы построения математических моделей для решения задачи детектирования машинно-сгенерированных текстов на открытом наборе данных, содержащем как тексты ChatGPT, так и рукописные образцы. Вычислительные эксперименты показывают, что применение квантильной регрессии дает прирост в точности модели, но при этом значительно увеличивает время, так как необходимо провести поиск оптимального уровня квантиля среди всех возможных значений. Время поиска такого оптимального значения многократно увеличивается с уменьшением шага сетки поиска до 0,01 и менее, что в контексте рассматриваемой задачи может достигать нескольких суток. Метод адаптивной квантильной регрессии значительно ускоряет процесс поиска оптимального значения квантиля и позволяет решить задачу за приемлемое время на оборудовании среднего уровня.

Полученные модели позволяют достаточно точно различать рукописные и машинно-сгенерированные тексты, но остается открытым вопрос применения более продвинутых методов векторизации текстов. Так, на конфигурации с 16 ГБ ОЗУ возможна работа с набором данных в 200000 строк и 15000 параметров, в то время как другие реализации векторизации могут привнести в параметры больше информации, что позволит строить более точные модели или использовать большее количество наблюдений и параметров.

СПИСОК ИСТОЧНИКОВ

1. He Y., Qiu J., Zhang W., Yuan Z. Fortifying Ethical Boundaries in AI: Advanced Strategies for Enhancing Security in Large Language Models. URL: http://arxiv.org/abs/2402.01725 (дата обращения: 03.02.2024).
2. Seo Ji-Hoon, Lee Ho-Sun, Choi Jin-Tak. Classification Technique for Filtering Sentiment Vocabularies for the Enhancement of Accuracy of Opinion Mining. *International journal of u- and e-service, science and technology.* 2015;8(10):11–20. DOI: 10.14257/ijunesst.2015.8.10.02.
3. Sandler M., Choung H., Ross A., David P. A Linguistic Comparison between Human and ChatGPT-Generated Conversations. URL: https://arxiv.org/pdf/2401.16587.pdf (дата обращения: 05.02.2024).
4. Hans A., et al. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. URL: https://arxiv.org/pdf/2401.12070.pdf (дата обращения: 04.02.2024).
5. Zheng Qi, Peng Limin, He Xuming. Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics.* 2015;43(5):2225–2258. DOI: 10.1214/15-AOS1340.
6. Barrodale I., Roberts F.D.K. An Improved Algorithm for Discrete l_1 Linear Approximation. *SIAM Journal on Numerical Analysis.* 1973;10(5):839–848. DOI: 10.1137/0710069.
7. Chen C. An Adaptive Algorithm for Quantile Regression. В сборнике: *Theory and Applications of Recent Robust Methods by ICORS2003: International Conference on Robust Statistics – 2003, 13–18 июля 2003 года, Антверпен, Бельгия.* Базель: Springer Basel AG; 2004. С. 39–48.
8. Chen C. A Finite Smoothing Algorithm for Quantile Regression. *Journal of Computational and Graphical Statistics.* 2007;16(1):136–164. DOI: 10.1198/106186007X180336.

9. Тюрин А.С. Адаптивная квантильная регрессия. *Моделирование, оптимизация и информационные технологии.* 2024;12(1). URL: https://moitvivt.ru/ru/journal/pdf?id=1514. DOI: 10.26102/2310-6018/2024.44.1.016 (дата обращения: 07.02.2024).
10. Duan T., Avati A., Ding D.Y., Thai K.K., Basu S., Ng A., Schuler A. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. В сборнике: *ICML 2020: 37th International Conference on Machine Learning: Proceedings of the 37th International Conference on Machine Learning, 13-18 июля 2020 года, Вена, Австрия.* 2020. С. 2690–2700.
11. Тюрин А.С., Сараев П.В. Построение квантильной регрессии с использованием натурального градиентного спуска. *Прикладная математика и вопросы управления.* 2023;(2):43–52. DOI: 10.15593/2499-9873/2023.2.04.

REFERENCES

1. He Y., Qiu J., Zhang W., Yuan Z. Fortifying Ethical Boundaries in AI: Advanced Strategies for Enhancing Security in Large Language Models. URL: http://arxiv.org/abs/2402.01725 [Accessed 3rd February 2024].
2. Seo Ji-Hoon, Lee Ho-Sun, Choi Jin-Tak. Classification Technique for Filtering Sentiment Vocabularies for the Enhancement of Accuracy of Opinion Mining. *International journal of u- and e-service, science and technology.* 2015;8(10):11–20. DOI: 10.14257/ijunesst.2015.8.10.02.
3. Sandler M., Choung H., Ross A., David P. A Linguistic Comparison between Human and ChatGPT-Generated Conversations. URL: https://arxiv.org/pdf/2401.16587.pdf [Accessed 5th February 2024].
4. Hans A., et al. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. URL: https://arxiv.org/pdf/2401.12070.pdf [Accessed 4th February 2024].
5. Zheng Qi, Peng Limin, He Xuming. Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics.* 2015;43(5):2225–2258. DOI: 10.1214/15-AOS1340.
6. Barrodale I., Roberts F.D.K. An Improved Algorithm for Discrete l_1 Linear Approximation. *SIAM Journal on Numerical Analysis.* 1973;10(5):839–848. DOI: 10.1137/0710069.
7. Chen C. An Adaptive Algorithm for Quantile Regression. In: *Theory and Applications of Recent Robust Methods by ICORS2003: International Conference on Robust Statistics – 2003, 13–18 July 2003, Antwerp, Belgium.* Basel: Springer Basel AG; 2004. P. 39–48.
8. Chen C. A Finite Smoothing Algorithm for Quantile Regression. *Journal of Computational and Graphical Statistics.* 2007;16(1):136–164. DOI: 10.1198/106186007X180336.
9. Tyurin A.S. Adaptive quantile regression. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii = Modeling, Optimization and Information Technology.* 2024;12(1). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1514. DOI: 10.26102/2310-6018/2024.44.1.016 [Accessed 7th February 2024].
10. Duan T., Avati A., Ding D.Y., Thai K.K., Basu S., Ng A., Schuler A. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. In: *ICML 2020: 37th International Conference on Machine Learning: Proceedings of the 37 th International Conference on Machine Learning, 13-18 July 2020, Vienna, Austria.* 2020. P. 2690–2700.

11. Tyurin A.S., Saraev P.V. Construction of quantile regression using natural gradient descent. *Prikladnaya matematika i voprosy upravleniya = Applied Mathematics and Control Sciences.* 2023;(2):43–52. (In Russ.). DOI: 10.15593/2499-9873/2023.2.04.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Тюрин Алексей Сергеевич, доцент, кафедра автоматизированных систем управления Липецкого государственного технического университета, Липецк, Российская Федерация.
e-mail: leha2148@gmail.com
ORCID: 0009-0000-1313-5826

Alexey S. Tyurin, Associate Professor, Department of Automated Control Systems, Lipetsk State Technical University, Lipetsk, the Russian Federation.

Сараев Павел Викторович, доктор технических наук, доцент, профессор, Липецкий государственный технический университет, Липецк, Российская Федерация.
e-mail: psaraev@yandex.ru
ORCID: 0000-0002-1373-2521

Pavel V. Saraev, Doctor of Technical Sciences, Associate Professor, Professor, Lipetsk State Technical University, Lipetsk, the Russian Federation.

Статья поступила в редакцию 10.03.2024; одобрена после рецензирования 21.03.2024; принята к публикации 29.03.2024.

The article was submitted 10.03.2024; approved after reviewing 21.03.2024; accepted for publication 29.03.2024.