УДК 004.89

К ВОПРОСУ РЕАЛИЗАЦИИ ИТЕРАТИВНЫХ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЭКОЛОГИЧЕСКИХ ЗАДАЧ

К.С. Костенко, ст. группы 3272,
Военная академия связи им. Буденного
г. Санкт-Петербург
Научный руководитель: А.В. Матисов, ст. преподаватель, кафедра информационных и автоматизированных производственных систем.
ФГБОУ ВО «Кузбасский государственный технический университет имени Т.Ф. Горбачева»
г. Кемерово

Для решения экологических задач зачастую требуется математический аппарат для прогнозирования данных. Одним из наиболее точных и простых методов прогнозирования является метод наименьших квадратов с итеративным пересчетом весов (от английского «Iterative Reweighted Least Squares» или сокращенно «IRLS») – это представитель класса предсказательных моделей [1]. Предсказательная модель – это параметрическое семейство функций (1).

$$A = \{g(x, \theta) | \theta \in \ominus\} \quad (1)$$

где $g: X \times \ominus \to Y -$ фиксированная функция, $\theta -$ множество допустимых значений параметра θ.

Функция A (1) должна хорошо аппроксимировать все точки дискетных значений (данных), позволяя тем самым получить значение целевой функции для прогнозируемого значения экземпляра данных x [2]. Для того, чтобы построить предсказательную модель вводится параметрическая функция g (1), которая зависит от объекта x и неизвестного вектора параметров θ. При этом ставится задача детерминирования в параметрическом семействе функций A такое отображение $A: X$, которое позволит решить прикладную задачу с наибольшим аппроксимирующим эффектом. Таким образом, значительно сужается пространство поиска решения: необходимо найти не какую-то неизвестную функцию, зависящую от x, а детерминируем функцию в рамках параметрического семейства функций (1). Параметрическое семейство – это способ задания функции, который учитывает ввод некоторой дополнительной информации о том, каким образом следует задать искомую функциональную зависимость. Кроме аппроксимирующей функции существует ещё одна важная концепция в машинном обучении, которая называется *минимизацией эмпирического риска* (2).

$$\mu(X^l) = argmin_{a \in A} Q(a, X^l) \quad (2)$$

где $X^l -$ это выборка, длина вариационного ряда которой составляет l.

Формула (2) означает, что вводится функционал, зависящий от выборки и от элементов параметрического семейства функции A. Далее необходимо найти такой a из этого семейства, чтобы функционал Q достигал своего минимального значения, то есть имел минимальный риск. Конструкция (2) лежит в основе очень многих методов обучения μ. Одним из частных случаев конструкции является метод наименьших квадратов (3).

$$\mu(X^l) = argmin_\theta \sum_{i=1}^{l} (g(x_i, \theta) - y_i)^2 \quad (3)$$

В формуле (3) функционал минимизации был заменен квадратичной функцией потерь, которая является типичной для задачи регрессии (прогнозирования). Метод наименьших квадратов (3) предполагает поиск такого экземпляра (вектор коэффициентов θ), чтобы была минимизирована сумма квадратов отклонений модельных ответов от истинных (правильных) ответов y_i. Конкатенация метода минимизации эмпирического риска [3] и аппроксимации целевой функции (на которой основана модель машинного обучения) полиномом [4] называется методом наименьших квадратов с итеративным пересчетом весов. Частным прикладным случаем решения этой задачи является полиномиальная регрессия [5], для которой в качестве оптимизационного критерия был выбран метод наименьших квадратов [6]. Модель полиномиальной регрессии может быть обобщенно представлена в виде (4).

$$a(x, \theta) = \theta_0 + \theta_1 x + \cdots + \theta_n x^n \quad (4)$$

Формула (4) описывает полином степени n. Количество полных общих слагаемых полинома будет напрямую зависеть от степени полинома и составляет значение $n + 1$, так как в формуле присутствует свободный член [7]. Обучение полиномиальной регрессии по методу оптимизации наименьшими квадратами в обобщенном виде может быть [8] представлено в виде формулы (5).

$$Q(\theta, X^l) = \sum_{i=1}^{l} (\theta_0 + \theta_1 x_i + \cdots + \theta_n x_i^n - y_i)^2 \rightarrow min_{\theta_0,\ldots,\theta_n} \quad (5)$$

Рассмотрим (рисунки 1 – 4), что происходит с обучающей и контрольной выборкой при увеличении степени полинома n (при обучении и тестировании модели по формуле (5)).

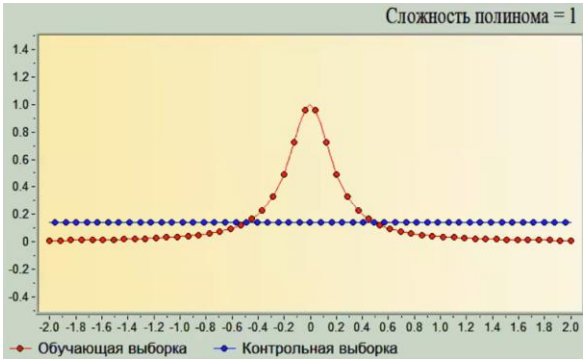

Рисунок 1 – Поведение целевых (прогнозных) значений на обучающей и контрольной выборке (при количественном значении $\theta = 1$)

Из рисунка 1 следует, что приближение к функции данных (обозначена красным цветом на графике) сейчас достигается не очень прецизионным образом. Увеличим степень сложности до значения 3 (то есть в аппроксимирующей функции будет уже три слагаемых θ соответствующих индексов от 0 до 2). Результат приближения представлен на рисунке 2.

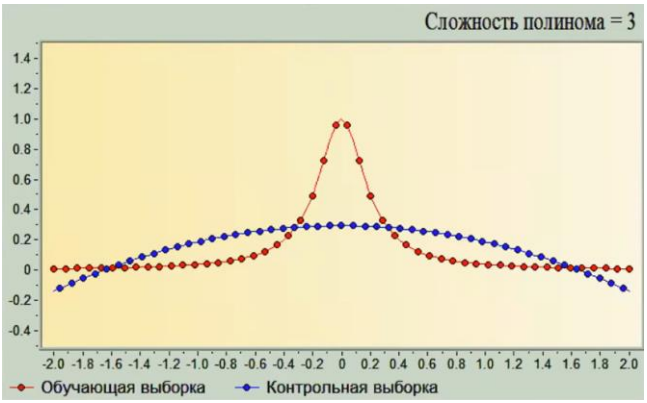

Рисунок 2 – Поведение целевых (прогнозных) значений на обучающей и контрольной выборке (при количественном значении $\theta = 3$)

Приближение аппроксимации становится всё ближе к истинным данным, но пока что это ещё слабо заметно на выборке [9]. Увеличив степень полинома до значения в 23 [7], начинаем наблюдать практически точное описывание данных (рисунок 3).

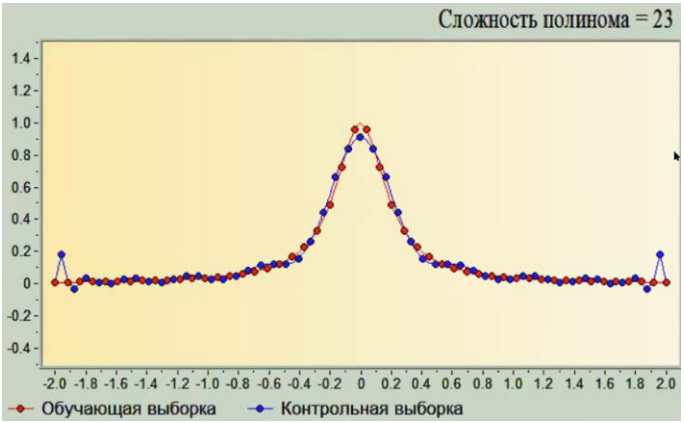

Рисунок 3 – Поведение целевых (прогнозных) значений на обучающей и контрольной выборке (при количественном значении $\theta = 23$)

Отметим, что увеличение степени полинома до бесконечности не приведет к лучшим результатам [1, 8]. Наиболее верным доказательством этих слов будет рисунок 4, на котором степень приобретает показатель 37.

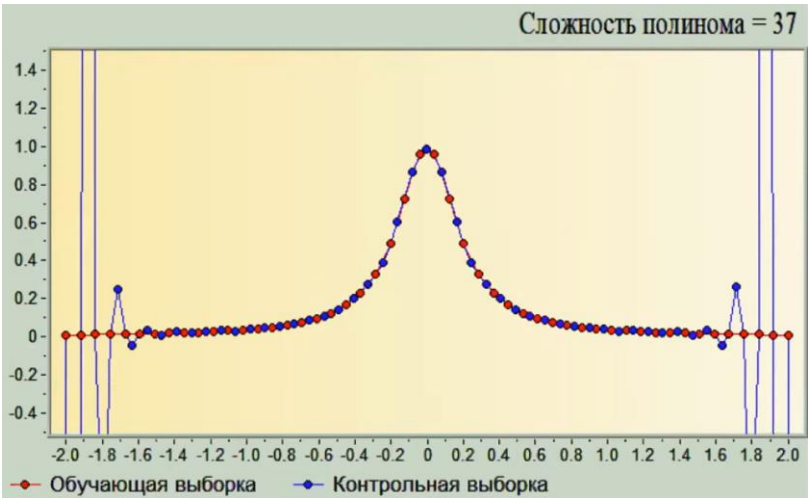

Рисунок 4 – Переобучение алгоритма при количественном значении $\theta = 37$

На рисунке 4 представлен эффект переобучения [9] – случай, при котором модель начинает искать несуществующие зависимости в данных и становится всё больше и больше отклоняться от истинных значений. Таким образом, для того чтобы наиболее точно описывать исследуемые данные, необходимо постоянно следить за точностью разрабатываемой модели машинного обучения, отслеживая любые негативные отклонения от высокого значения.

Список литературы:

1. Томас Кормен, Чарльз Лейзерсон. Алгоритмы: построение и анализ, 3-е издание – М.: ООО И.Д. Вильямс. 2013. – 1328 с.
2. Свидетельство о государственной регистрации программы для ЭВМ № 2022682567 Российская Федерация. Интеллектуальная система второго медицинского мнения для превентивного предсказания заболеваний сердечно-сосудистой системы: № 2022682189: заявл. 18.11.2022: опубл. 24.11.2022 / П. А. Пылов, А. В. Балуева, А. В. Протодьяконов.
3. *Майтак Р. В., Пылов П. А.* Параметризация гиперпараметров в прикладных задачах машинного обучения на основе ядерных функций // Россия молодая: СБОРНИК МАТЕРИАЛОВ XIV ВСЕРОССИЙСКОЙ, НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ МОЛОДЫХ УЧЕНЫХ С МЕЖДУНАРОДНЫМ УЧАСТИЕМ, Кемерово, 2023.
4. *Пылов П. А., Майтак Р. В., Протодьяконов А. В.* Прогнозирование вектора ответов наборов данных на основе изотонических особенностей в

задаче регрессии // Инновации в информационных технологиях, машиностроении и автотранспорте: Сборник материалов VI Международной научно-практической конференции, Кемерово, 2022.

5. *Пылов П. А., Майтак Р. В., Протодьяконов А. В.* Анализ потенциала органических материалов для эффективного производства высококачественного твердого топлива // Актуальные проблемы общества, экономики и права в контексте глобальных вызовов: сборник материалов XX Международной научно-практической конференции., Москва, 17 мая 2023 года. Том Часть 2. – Санкт-Петербург: Печатный цех, 2023. – С. 129-132.

6. *Пылов П. А., Майтак Р. В., Протодьяконов А. В.* Параметризация гиперпараметров в прикладных моделях машинного обучения на основе ядерных функций // Актуальные проблемы общества, экономики и права в контексте глобальных вызовов: сборник материалов XX Международной научно-практической конференции., Москва, 17 мая 2023 года. Том Часть 2. – Санкт-Петербург: Печатный цех, 2023. – С. 43-49.

7. *Пылов, П. А.* Интерпретируемые модели машинного обучения для анализа сейсмоакустических данных // Обработка информации и математическое моделирование: материалы Всероссийской научно-технической конференции с международным участием, Новосибирск, 19–20 апреля 2023 года. – Новосибирск: Сибирский государственный университет телекоммуникаций и информатики, 2023. – С. 196-198. – DOI 10.55648/978-5-91434-085-5-2023-130-132.

8. *Дягилева А. В., Пылов П. А., Майтак Р. В.* Разработка метода автоматизированного сейсмоакустического мониторинга на базе компьютерного анализа ядерных функций // Вестник научного центра по безопасности работ в угольной промышленности. 2023. № 2.

9. *Пылов П. А., Майтак Р. В., Протодьяконов А. В.* Исследовательская модель сильного искусственного интеллекта для решения задачи оптического распознования символов // Инновации в информационных технологиях, машиностроении и автотранспорте: Сборник материалов VI Международной научно-практической конференции, Кемерово, 2022.

10. *Пылов П. А., Майтак Р. В., Протодьяконов А. В.* Оценка уровня надежности вероятностных метрик в прикладных задачах искусственного интеллекта // Инновации в информационных технологиях, машиностроении и автотранспорте: Сборник материалов VI Международной научно-практической конференции, Кемерово, 2022.

11. *Пылов П. А*. Аналитика возможностей визуализации данных в разнообразных темах оформления на основе библиотек matplotlib и seaborn // Россия молодая: Сборник материалов XII Всероссийской научно-практической конференции молодых ученых с международным участием. Кемерово, 2020.

12. *Пылов П. А., Протодьяконов А. В.* Экстракция признаков в моделях последовательного глубокого обучения // Россия молодая: Сборник материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 19–21 апреля 2022 года / Редколлегия: К.С. Костиков (отв.

ред.) [и др.]. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2022. – С. 31525.1-31525.3.

13. *Пылов П. А., Протодьяконов А. В.* Модификация нейронной сети XGBOOST в задачи детекции мошеннических банковских транзакций // Россия молодая: Сборник материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 2022.

14. *Пылов П. А., Садовников В. Е., Протодьяконов А. В., Бобровских А. И.* Значимость правильного выбора типа лидера на результат работы команды на примере разработки инновационного проекта автомобилестроительной компании // Россия молодая: Сборник материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 2022.

15. *Пылов П. А., Протодьяконов А. В., Бобровских А. И.* Teamlead как разработчик и юридический лидер команды в одном лице // Россия молодая: Сборник материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 2022.

16. *Пылов П. А., Протодьяконов А. В.* Демонстрация алгоритма спектральной кластеризации в моделях искусственного интеллекта на основе совместимости спектров // Инновации в информационных технологиях, машиностроении и автотранспорте (ИИТМА-2020): сборник материалов IV Международной научно-практической конференции с онлайн-участием, Кемерово, 2020.