МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ДВОЙСТВЕННЫЕ АЛГОРИТМЫ НА ОСНОВЕ АКТИВНОГО МНОЖЕСТВА ДЛЯ ПОСТРОЕНИЯ ОПТИМАЛЬНОЙ РАЗРЕЖЕННОЙ ВЫПУКЛОЙ РЕГРЕССИИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Кротовой Юлии Игоревны

Научный руководитель
к.э.н, доцент _____ А. Р. Файзлиев

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2023

Во введении обосновывается актуальность темы работы, формулируется цель работы и решаемые задачи, отмечается практическая значимость полученных результатов.

Актуальность темы. Построение регрессий и сглаживание данных являются одними из основных задач во многих сферах, особенно в сфере экономики и других прикладных областях, в которых актуальны проблемы исследования динамики процессов. Существует множество различных способов решения проблем в этой предметной области. Одним из подвидов задач построения регрессии является построение регрессии с ограничением на особенность данных, например, монотонность или выпуклость. Когда этот фактор заранее известен, например, из предыдущих свойств динамики значений или из особенностей самих данных, такое условие может значительно улучшить качество построения регрессии.

Целью магистерской работы является разработка двойственного алгоритма на основе активного множества для построения оптимальной разреженной выпуклой регрессии. Основная идея, лежащая в основе метода, состоит в следующем: по имеющимся данным строится регрессия с ограничением на выпуклость. Задача работы - исследовать данную процедуру и доказать оптимальность и сходимость алгоритма.

Объектом исследования являются двойственные алгоритмы и их применение к данным.

Предметом исследования являются методы активного множества для построения выпуклой разреженной регрессии.

Практическая значимость проводимого исследования состоит в том, что на основании построенного алгоритма можно проводить исследования реально существующие данные. Результаты построенной регрессии могут использоваться для анализа динамики ряда или в качестве аналога процедуры сглаживания данных.

Работа прошла апробацию на различных конференциях, в частности, на ежегодной студенческой конференции "Актуальные проблемы математики и механики которую проводил механико-математический факультет СГУ в апреле 2022 года, в секции "Анализ данных в XI Международной молодежной научно-практической конференции "Математическое и компьютерное моде-

лирование в экономике, страховании и управлении рисками ноябрь 2022 года.

Основное содержание работы

В первом разделе рассмотрены основные понятия задачи оптимизации, основные теоремы и леммы.

Экстремальными задачами называют задачи отыскания минимума или максимума функций на заданных множествах. Условимся записывать задачу минимизации функции $f(x)$ на множестве D в виде

$$f(x) \to \min_{x \in D}. \tag{1}$$

Будем рассматривать лишь конечномерные задачи оптимизации, то есть считать, что аргумент функции $x = (x_1, x_2, \ldots, x_n)^\top$ есть вектор конечномерного пространства $\mathbb{R}^n$, а D является некоторым подмножеством из $\mathbb{R}^n$. Экстремальная (оптимизационная) задача (P) представляет собой нахождение $\min f(x)$ при условии, что

$$\varphi_i(x) \leq 0, i = \overline{1,m}, \tag{2}$$

$$x \in S \subseteq R^n \text{ или } Z^n \text{ или } B^n, \tag{3}$$

где $x = (x_1, \ldots, x_n)$ – вектор переменных, f - целевая функция задачи; условия $\varphi_i(x) \leq 0, i = \overline{1,m}, x \in S$ - называются ограничениями задачи.

Теорема 1. ***(Теорема Куна-Таккера в локальной форме).*** *Точка $x^* \in Q$– оптимальное решение задачи выпуклого программирования в том и только в том случае, когда существуют такие числа $\lambda_i > 0, i = \overline{1,m}$, что*

$$-f'(x^*) = \sum_{i=1}^{m} \lambda_i \varphi_i'(x^*),$$
$$\lambda_i \varphi_i(x^*) = 0, i = \overline{1,m}.$$

Теорема 2. ***(Теорема Куна-Таккера в локальной форме, линейный случай).*** *Точка $x^* \in Q$ - оптимальное решение задачи выпуклого программирования с линейными ограничениями в том и только в том случае, когда*

существуют такие числа $\lambda_i \geq 0, i = \overline{1,m}$, *что*

$$-f'(x^*) = \sum_{i=1}^{m} \lambda_i a_i,$$
$$\lambda_i ((a_i, x) - b_i) = 0, i = \overline{1,m}.$$

Функцию

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i \varphi_i(x),$$

определенную при всех x и λ, назовем функцией Лагранжа.

Пара (x^*, λ^*) называется седловой точкой функции

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*) \, \forall x \in R^n, \forall \lambda \geq 0.$$

Теорема 3. ***(Теорема Куна-Таккера в нелокальной форме).*** *Вектор* $x^* \in Q$ *является оптимальным решением задачи выпуклого программирования тогда и только тогда, когда существует такой вектор* λ^*, *что пара* (x^*, λ^*) *является седловой точкой функции Лагранжа.*

Во втором разделе представлены различные методы активного множества. Методы активного множества являются мощным инструментом для решения задач оптимизации, таких как задача квадратичного программирования. Они позволяют решать задачи оптимизации с большим количеством переменных и ограничений быстрее, чем другие методы. В основе методов активного множества лежит идея постепенного уменьшения размера активного множества переменных и ограничений до тех пор, пока не будет достигнуто оптимальное решение.

Активное множество представляет собой метод поиска решения для задачи квадратичного программирования QP - *Quadratic Programming*, который использует процесс итерации для построения активного множества и поиска оптимального решения.

В разделе рассмотрены несколько известных методов активного множества для решения выпуклых QP. Методы прямого и двойного активного множества позволяют решать общие выпуклые задачи QP, тогда как метод

проекции градиента и метод основного двойственного активного множества в основном использовались для решения выпуклых BQP. Алгоритм *Primal active-set method* для выпуклой задачи квадратичного программирования может быть представлен следующей схемой ниже.

Algorithm 1 Primal active-set method for convex QP

1: Инициализация: выбор начального приближения x_0 и пустого множества активных ограничений A_0

while *не выполнен критерий остановки* **do**

end

2: Решить задачу QP с учетом текущего множества активных ограничений:

$$\min_x \frac{1}{2} x^T Q x + c^T x \quad \text{при условиях} \quad Ax \leq b, \quad x_j = 0 \quad \text{для} \quad j \notin A_k$$

3: Вычислить множество активных ограничений A_{k+1}:

$$A_{k+1} = \{i \in \{1, \ldots, m\} \mid a_i^T x_{k+1} = b_i\}$$

4: Если $A_{k+1} = A_k$, то выход из цикла

5: Иначе, обновить текущее приближение и множество активных ограничений:

$$x_{k+1} = x_k + \alpha p_k, \quad A_{k+1} = A_k \cup \{i \in \{1, \ldots, m\} \mid a_i^T p_k < 0\}$$

где p_k - направление спуска, выбранное из активных ограничений, а α - длина шага, определяемая с помощью условия минимума по направлению p_k:

$$\alpha = \frac{b_i - a_i^T x_k}{a_i^T p_k}$$

для $i \in A_k$ таких, что $a_i^T p_k < 0$

6: Возвращение значения x_{k+1}

Здесь Q - квадратная матрица размера $n \times n$, c - вектор размера n, A - матрица размера $m \times n$, b - вектор размера m. Ограничения задаются в виде линейных неравенств $Ax \leq b$ и условий на переменные $x_j = 0$ для $j \notin A_k$.

Множество активных ограничений A_k определяется как множество индексов ограничений, которые выполняются как равенства на текущем приближении x_k. Направление спуска p_k выбирается из множества активных ограничений, и шаг α определяется так, чтобы удовлетворять ограничениям и минимизировать функцию в направлении p_k. Остановка алгоритма происходит, когда множество активных ограничений не меняется на последней итерации.

В третьем разделе описаны основные алгоритмы для решения задачи работы.

Существуют различные методы для построения выпуклой регрессии, включая *PAVA (Pool - Adjacent - Violators Algorithm)* и *PDAS (Piecewise - Deterministic - Approximate - Aggregate Slope)* методы.

$PAVA$ является итерационным методом, который работает следующим образом:

1. Инициализировать функцию регрессии фиксированным начальным значением.
2. Найти первую пару соседних точек, которые нарушают монотонность регрессии и поменять их значения между собой, сохраняя при этом остальные точки на своих местах.
3. Повторять шаг 2 до тех пор, пока монотонность регрессии не будет восстановлена.
4. Вернуть регрессионную функцию, определенную на основе значений регрессии для каждой точки данных.

$PDAS$ является стохастическим методом, который работает следующим образом:

1. Инициализировать функцию регрессии случайным образом.
2. Выбрать случайную пару точек данных и вычислить значение регрессии для этой пары точек.
3. Обновить регрессию, чтобы уменьшить расхождение между вычисленным значением регрессии и фактическим значением.
4. Повторять шаги 2 и 3 до тех пор, пока монотонность регрессии не будет достигнута или пока не будет достигнуто максимальное количество итераций.

5. Вернуть регрессионную функцию, определенную на основе значений регрессии для каждой точки данных.

$PAVA$ и $PDAS$ методы оба имеют свои преимущества и недостатки. $PAVA$ является более эффективным, но не всегда сходится к оптимальному решению, тогда как PDAS имеет большую вероятность нахождения оптимального решения, но может быть более вычислительно сложным. Выбор метода зависит от конкретной задачи и требуемой точности результата.

В четвертом разделе описан алгоритм для построения выпуклой разреженной регрессии с помощью двойственного алгоритма активного множества. Применение выпуклой регрессии часто связано с подбором выпуклых данных, когда предполагается, что существует неизвестная выпуклая функция отклика $\kappa(t)$ независимой переменной t. Сосредоточимся на одномерном случае и предполагаем, что $\kappa(t)$ выпукла, т. е. разделенная разность второго порядка не отрицательна

$$\kappa(t') - 2\kappa(t'') + \kappa(t''') \geq 0, \quad \forall t' < t'' < t'''.$$

Для линейно-упорядоченной последовательности наблюдаемых значений объясняющей переменной $t_1 < \cdots < t_n$ соответствующая последовательность наблюдаемых значений функции отклика может быть представлена как

$$y_i = \kappa(t_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{4}$$

где ϵ_i - ошибка наблюдения. Из-за ошибок ожидаемая выпуклость (положительная разделенная разность второго порядка) $y_{i+2} - 2y_{i+1} + y_i \geq 0$ может нарушаться для некоторых индексов i. Задача направлена на восстановление нарушенной выпуклости путем нахождения поправки наименьшего изменения к наблюдаемым значениям. Формально это можно сформулировать как задачу квадратичной оптимизации следующим образом:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{n} w_i (z_i - y_i)^2, \Delta^2 z_i \geq 0, \tag{5}$$

где $w \in \mathbb{R}^n_{++}$ — вектор весов. Пусть z^* — решение задачи. Активные ограничения предполагают, что компоненты z^* разбиты на блоки последовательных

компонентов, лежащих на одной прямой, т.е. с одинаковыми первыми разностями. Пусть $z(t)$ — выпуклая функция, удовлетворяющая интерполяционному условию:

$$z(t_i) = z_i^*, \quad \forall i \in [1, n].$$

Здесь и далее множество индексов $[i, i+1, \ldots, j-1, j]$ обозначается $[i, j]$ и называется сегментом индексов. Из-за блочной структуры z^* форма $z(t)$ напоминает кусочно-линейная функцию, что позволяет предположить, что она может иметь нарушения выпуклости на определенных интервалах t. Эта особенность задачи часто подвергается критике и мотивирует необходимость сглаживания решения выпуклой регрессии. Рассмотрим следующую регуляризованную задачу выпуклой регрессии:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{n} w_i (z_i - y_i)^2 + \sum_{i=1}^{n-2} \mu_i (z_{i+2} - 2z_{i+1} + z_i)^2, \Delta^2 z_i \geq 0, \tag{6}$$

где $\mu \in \mathbb{R}_+^{n-1}$ - вектор штрафных параметров. Штрафной член в (6) предназначен для сглаживания функций, интерполирующих решение этой задачи. Это объясняет, почему мы называем (6) задачей сглаженной выпуклой регрессии *(SCR - Smoothed Convex Regression)*. Заметим, что поскольку (6) является задачей квадратичной оптимизации со строго выпуклой целевой функцией, ее решение существует и единственно. При $\mu = 0$ задача (6), очевидно, сводится к (5).

Будем называть $\Delta^2 z_i \geq 0$ в SCR ограничением $i \in [1, n-2]$. Каждая итерация нашего алгоритма связана с выбором активного множества $S \subseteq [1, n-1]$ и решением соответствующей подзадачи

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{n} w_i (z_i - y_i)^2 + \sum_{i=1}^{n-2} \mu_i (z_{i+2} - 2z_{i+1} + z_i)^2, \Delta^2 z_i = 0, \quad \forall i \in S. \tag{7}$$

Обозначим его единственное оптимальное решение через $z(S)$. Для представления эффективного способа решения этой подзадачи рассмотрим условия оптимальности. Они также будут использоваться в следующем разделе для изучения свойств сходимости алгоритма.

Активное множество S предполагает, что существуют наборы последовательных индексов вида $[l, r] \subseteq [1, n]$ такие, что $[l, r-2] \subseteq S$, $l-2 \notin S$ и $r \notin S$. Будем называть такой набор блоком. Обратим внимание, что блок может быть одноэлементным, когда $l = r$. Тогда общее количество блоков, обозначенное через m, равно $n - |S|$. Блочное разбиение (сегментация) $[1, n]$, индуцированное S, может быть представлено как

$$[1, n] = [l_1, r_1], [l_2, r_2], \ldots, [l_m, r_m], \tag{8}$$

где $l_1 = 1$, $r_m = n$, $r_{i+1} = l_{i+1}$, $\forall i \in [1, m-2]$.

$$\begin{cases} w_1(z_1 - y_1) + \mu_1(z_3 - 2z_2 + z_1) = 0, \\ w_2(z_2 - y_2) + \mu_2(z_4 - 2z_3 + z_2) - 2\mu_1(z_3 - 2z_2 + z_1) = 0, \\ w_3(z_3 - y_3) + \mu_3(z_5 - 2z_4 + z_3) - 2\mu_2(z_4 - 2z_3 + z_2) + \mu_1(z_3 - 2z_2 + z_1) = 0, \\ \ldots \\ w_i(z_i - y_i) + \mu_i(z_{i+2} - 2z_{i+1} + z_i) - 2\mu_{i-1}(z_{i+1} - 2z_i + z_{i-1}) + \\ \qquad + \mu_{i-2}(z_i - 2z_{i-1} + z_{i-2}) = 0, \\ \ldots \\ w_n(z_n - y_n) + \mu_{n-2}(z_n - 2z_{n-1} + z_{n-2}) = 0. \end{cases} \tag{9}$$

Алгоритм начинается с любого активного набора, такого что $S \subseteq S^*$. Самый простой из допустимых вариантов - $S = \varnothing$. На каждой итерации он решает трехдиагональную систему линейных уравнений (9), а затем расширяет множество S, дополнительно активизируя ограничения в (6), для которых выполняется строгая выпуклость $\Delta^2 z > 0$ нарушена. Это, как и в алгоритме PAV, предполагает объединение соответствующих смежных блоков, что объясняет, почему мы называем наш алгоритм $SPAV$. Слияние связано с обновлением коэффициентов, определяющих линейную систему (9). Соответствующее количество арифметических операций пропорционально количеству новых активных ограничений. В отличие от обычных алгоритмов активного набора, $SPAV$ может расширить активный набор более чем одним элементом одновременно. Изложенный алгоритм формально можно выразить

следующим образом.

Algorithm 2 SPAV

begin

input $y \in \mathbb{R}^n$, $\mu \in \mathbb{R}^n_+$, $S \subseteq S^*$

find $z(S)$ that solves (9)

while $z(S)$ *is not convex* **do**

set $S \leftarrow S \cup \{r_i : y_i - 2y_{i+1} + y_{i+2} \leq 0\}$, if $\mathrm{r}_i, r_{i+2} \in S : r_{i+1} \in S$

$\underline{f}indz(S)thatsolves(9)$

output $z(S)$

end

Вычислительная сложность алгоритма $SPAV$ составляет $O(n^2)$. Эта оценка основана на следующих двух наблюдениях. Во-первых, активное множество S расширяется в цикле $while$ за счет включения хотя бы одного индекса, а это означает, что число итераций цикла $while$ не превосходит $n-1$. Во-вторых, вычислительная сложность решения четырехдиагональной линейной системы (9) равно $O(n)$.

В разделе доказана основная теорема о сходимости алгоритма.

Теорема 4. *Для любого начального $S \subseteq S^*$ алгоритм $SPAV$ сходится к оптимальному решению задачи SCR не более чем за $n-1-|S|$ итерации. Более того, после последней итерации $S = S^*$.*

Также в данном разделе реализован алгоритм на языке $Python$, который решает задачу построения разреженной выпуклой регрессии по сгенерированным данным. Методика алгоритма и доказательство его сходимости приведены ранее. Рассмотрены примеры работы алгоритма на различных наборах данных.

В **заключении** описаны основные результаты работы.

В первом разделе содержится общая характеристика задач оптимизации, основные определения и обозначения понятий теории оптимизации, вводятся понятия функции Лагранжа и условий Каруша-Куна-Таккера. Данный раздел является вводным и служит для обеспечения систематизированного и замкнутого изложения основного материала.

Во втором разделе описана методика активного множества и различные вариации применения этого метода.

В третьем разделе перечислены основные алгоритмы, выбран оптимальный для задачи алгоритм $SPAV$, показаны его преимущества перед другими методами.

В четвертом разделе разработан алгоритм для решения задачи построения регрессии с ограничением на монотонность. Разработан алгоритм, доказана его сходимость и оптимальность.

В пятом разделе рассмотрены сгенерированные данные, интерпретированные в качестве выпуклой функции с добавлением шума. По этим данным была построена выпуклая регрессия с помощью метода $SPAV$. Разработан быстрый алгоритм с двойным активным множеством для решения поставленной задаче. В работе доказана его конечная сходимость к оптимальному решению. Вычислительные эксперименты подтвердили ряд важных преимуществ алгоритма $SPAV$, в частности его масштабируемость, что позволяет рассматривать его как практический алгоритм для решения крупномасштабных задач. Эффективность алгоритма обусловлена его способностью расширять активное множество, добавляя сразу большую часть ограничений.

Работа прошла апробацию на различных конференциях, в частности, на ежегодной студенческой конференции "Актуальные проблемы математики и механики которую проводил механико-математический факультет СГУ в апреле 2022 года, в секции "Анализ данных в XI Международной молодежной научно-практической конференции "Математическое и компьютерное моделирование в экономике, страховании и управлении рисками ноябрь 2022 года.