МИНОБРНАУКИ РОССИИ

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

Кафедра теории функций и стохастического анализа

АЛГОРИТМЫ РЕШЕНИЯ НЕКОТОРЫХ ЗАДАЧ НЕПАРАМЕТРИЧЕСКОЙ СТАТИСТИКИ И АНАЛИЗ ИХ СХОДИМОСТИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Гудкова Александра Александровича

Научный руководитель
д. ф.-м. н., доцент ____________ С. П. Сидоров

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

Саратов 2019

ВВЕДЕНИЕ

Актуальность данной работы обусловлена тем, что в последнее время повышенный интерес вызывают задачи статистики с ограничениями на форму данных. Интерес вызван в первую очередь их прикладной значимостью. Одной из таких задач является задача построения монотонной регрессии, наилучшим образом приближающей заданный набор значений. Её продолжением является задача построения k-монотонных регрессий, где k - некоторое натуральное число.

Новизна данной работы заключается в том, что алгоритм, рассматриваемый в работе позволяет построить регрессию любого порядка монотонности, но в данной работе алгоритм будет рассматриваться в контексте построения регрессий 3-го порядка монотонности. Данный алгоритм, названный двойственным алгоритмом на основе активного множества, является итерационным и на каждой его итерации использует активное множество для построения вектора, значения которого лежат на кусочно-линейной кривой. Также в работе рассматриваются несколько способов построения регрессии 2-го порядка монотонности, в том числе с помощью двойственного алгоритма на основе активного множества.

Цель бакалаварской работы состоит в разработке алгоритмов построения k-монотонных регрессий, а также в сравнении результатов работы этих алгоритмов. Ещё одной задачей является описание двойственного алгоритма на основе активного множества и описание способа доказательства оптимальности решения, получаемого с помощью данного алгоритма, а также оценка скорости сходимости и сравнение данного алгоритма с другими, уже известными и реализованными методами.

Объектом исследования являются алгоритмы построения k-монотонных регрессий.

Предмет исследования - возможность применения двойственного алгоритма на основе активного множества к реальным статистическим задачам и оценка полученных результатов.

При написании данной работы я придерживался следующего плана действий:

— Привести описание двойственного алгоритма активного множества и идеи, которые использовались при его создании.
— Рассмотреть способ построения 2-монотонной регрессии с помощью данного алгоритма и сравнить результат работы с некоторыми другими алгоритмами.
— Доказать сходимость данного алгоритма к оптимальному решению и оценить скорость сходимости.
— Последний шаг - использовать данный алгоритм для решения некоторых реальных задач построения 3-монотонной регрессии, оценить результаты и сравнить с другими алгоритмами.

Работа прошла апробацию на различных конференциях, в частности, в XIX Международной Саратовской зимней школе «Современные проблемы теории функций и их приложения», посвященной 90-летию со дня рождения академика П. Л. Ульянова, январь 2018 года. На ежегодной студенческой конференции "Актуальные проблемы математики и механики которую проводил механико-математический факультет СГУ в апреле 2019 года, в секции "Анализ данных". В VII Международной молодежной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками», ноябрь 2018 года.

Результаты работы опубликованы в совместных статьях [1–6]

Структура и содержание бакалаврской работы. Выпускная квалификационная работа состоит из введения и трех разделов, в которых рассматриваются алгоритмы решения задачи построения регрессии, а также из заключения, списка использованных источников и приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение содержит основные положения: обоснование актуальности темы работы, формулировку цели, объекта и предмета исследования.

В **первом** разделе основное внимание уделяется основным понятиям, связанным с рассматриваемыми в данной работе алгоритмами и с решением задачи построения k-монотонной регрессии. Такие как:
— k-монотонный вектор и k-монотонная регрессия;
— Δ^k - оператор конечных разностей k-го порядка;

— Δ_n^k - множество всех k-монотонных векторов, размерности n;

А также ставится основная задача построения k-монотонной регрессии:

$$(z-y)^T(z-y) = \sum_{i=1}^{n}(z_i - y_i)^2 \to \min_{z \in \Delta_n^k}, \tag{1}$$

где $y \in R^n$ - заданный вектор, для которого строится k-монотонная регрессия, а $z \in R^n$ - вектор значений k-монотонной регрессии.

Данную задачу можно сформулировать следующим образом:

Необходимо по заданному вектору $y \in R^n$ (не обязательно k-монотонному) построить k -монотонный вектор $z \in R^n$, который минимизирует среднеквадратичную ошибку, вычисляемую по формуле (1).

Так же в первом разделе рассматривается жадный алгоритм типа Франка-Вульфа, подробное описание и анализ которого можно найти в статьях [2] и [3].

Чтобы записать данный алгоритм, введем обозначение для приращений: $x_{k+i} = \Delta^{k-1} z_{i+1} - \Delta^{k-1} z_i$, $i = 1, \ldots, n-k$.

И приведем теорему из статьи [7].

Теорема 1. *Выполнение условия $z \in \Delta_n^k$ эквивалентно тому, что существует вектор $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ такой, что z_i, $1 \le i \le n-k$ может быть представлен в следующем виде*

$$z_i = \sum_{j_1=1}^{i}\sum_{j_2=1}^{j_1}\cdots\sum_{j_{k-1}=1}^{j_{k-2}}\sum_{j_k=1}^{j_{k-1}} x_{j_k}, \tag{2}$$

где $x_j \ge 0$ для всех $k+1 \le j \le n$.

Теперь, используя данную теорему, можно записать задачу (1) следующим образом:

$$E(x) := \sum_{i=1}^{n}\left(\sum_{j_1=1}^{i}\sum_{j_2=1}^{j_1}\cdots\sum_{j_{k-1}=1}^{j_{k-2}}\sum_{j_k=1}^{j_{k-1}} x_{j_k} - y_i\right)^2 \to \min_{x \in S}, \tag{3}$$

где $S := \{x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \;\; x_1, \ldots, x_k \in \mathbb{R},\ (x_{k+1}, \ldots, x_n) \in \mathbb{R}_+^{n-k}$ и

$\sum_{j=k+1}^{n} x_j \leq \max \Delta^{k-1} y_i - \min \Delta^{k-1} y_i\}$.

Сам алгоритм можем записать следующим образом:

· Пусть $y = (y_1, \ldots, y_n)^T$ заданный вектор, а N - это максимальное количество итераций;

begin

· И пусть $z^0 = \mathrm{reg}_{k-1}(y)$ - полиномиальная регрессия $k-1$ порядка, наилучшим образом приближающая значения вектора y;

· Выберем в качестве начального приближения $x^0 = (x_1^0, \ldots, x_n^0)^T$ значения, полученные из z^0 с помощью (2);

· Пусть счётчик $t = 0$;

· **while** $t < N$ **do**

  · Вычисляем градиент $\nabla E(x^t)$ в текущей точке x^t;

  · Пусть $\widetilde{x}^t$ - решение линейной оптимизационной задачи: $\langle \nabla E(x^t)^T, x \rangle \to \min\limits_{x \in S}$, где $\langle \cdot, \cdot \rangle$ - скалярное произведение векторов;

  · В качестве следующего приближения выбираем $x^{t+1} = x^t + \alpha_t(\widetilde{x}^t - x^t)$, $\alpha_t = \frac{2}{t+2}$, $t := t+1$;

· Восстанавливаем k-монотонную последовательность $z = (z_1, \ldots, z_n)$ из вектора x^N;

end

Для алгоритма доказана теоретическая скорость сходимости, записывается которая в виде следующей теоремы:

Теорема 2. *Пусть решение x^t получено с помощью жадного алгоритма типа Франка-Вульфа на итерации t. Тогда существует положительная $c(k, y)$, которая не зависит от n, такая, что для любого $t \geq 2$ выполняется:*

$$E(x^t) - E^* \leq \frac{c(k, y) n^{2k - \frac{1}{2}}}{t + 2}, \tag{4}$$

где E^* точное решение задачи (3).

Доказательство данной теоремы приводится в статье [2].

В этом же разделе рассматривается ещё один алгоритм построения k-монотонных регрессий, а именно k-monotone Pool-Adjacent-Violators Algorithm (k-PAVA). Основан данный алгоритм на Pool-Adjacent-Violators Algorithm (PAVA), который позволяет построить 1-монотонную регрессию.

Сам алгоритм записывается следующим образом:

begin

· Пусть $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ заданный вектор. И пусть $z^{[0]} := y$;

· Задаём значение счётчика $t := 1$. k - это порядок монотонности;

repeat

· Задаем значения $j := 0$, $l := 1$;

· **while** $l \leq n - k$ **do**

if $\Delta^k z_l^{[t-1]} < 0$ **then**

$Z_{l,l+k}^{[t]} := \mathrm{reg}_{k-1}\left(Z_{l,l+k}^{[t-1]}\right)$; $l := l + 1$ и вычисляем $\Delta^k z_l^{[t-1]}$;

if $\Delta^k z_l^{[t-1]} \geq 0$ **then**

$z_l^{[t]} := z_l^{[t-1]}$, $j := 0$ и $l := l + 1$;

else

$j := j + 1$ и пусть $Z_{l,l+k+j}^{[t]} := \mathrm{reg}_{k-1}\left(Z_{l,l+k+j-1}^{[t]} \cup z_{l+k+j}^{[t-1]}\right)$;

else

$z_l^{[t]} := z_l^{[t-1]}, l := l + 1$;

· $t := t + 1$;

until $\Delta^k z_l^{[t]} \geq 0$ *для всех* $l = 1, \ldots, n - k$;

· Получаем значения $z^{[t]} = (z_1^{[t]}, \ldots, z_n^{[t]})$;

end

Более подробное описание данного алгоритма можно найти в работе [2]. К недостаткам данного алгоритма можно отнести отсутствие доказательства теоретической сходимости.

Для каждого из алгоритмов приведены примеры работы алгоритмов на некоторых задачах.

Во **втором** и **третьем** разделе рассматривается самый новый алгоритм из перечисленных - двойственный алгоритм на основе активного множества.

В работе доказывается оптимальность данного алгоритма для случая построения 2-монотонной и 3-монотонной регрессии.

Приведем некоторые фрагменты из третьего раздела, в котором рассматривается случай построения 3-монотонной регрессии.

Сначала перепишем задачу (1) для случая $k = 3$ в виде задачи выпуклого программирования с линейными ограничениями:

$$F(z) = \frac{1}{2}z^T z - y^T z \to min, \tag{5}$$

где минимум берётся по всем $z \in R^n$, таким, что

$$g_i(z) = -(z_{i+3} - 3z_{i+2} + 3zi + 1 - z_i) \leqslant 0, 1 \leqslant i \leqslant n - 3. \tag{6}$$

Задача (5)-(6) является задачей квадратичного программирования и к тому же сильно выпуклой, поэтому существует единственное решение данной задачи.

Далее записываются условия Каруша-Куна-Таккера для задачи (5)-(6):

$$\nabla F(z) + \sum_{i=1}^{n-3} \mu_i \nabla g_i(z) = 0, \tag{7}$$

$$g_i(z) \leq 0, \quad 1 \leq i \leq n - 3, \tag{8}$$

$$\mu_i \geq 0, \quad 1 \leq i \leq n - 3, \tag{9}$$

$$\mu_i g_i(z) = 0, \quad 1 \leq i \leq n - 3, \tag{10}$$

где ∇g_i определяется, как градиент g_i, а $\mu = (\mu_1, \ldots, \mu'_{n-3})^T \in \mathbb{R}^{n-3}$ - множитель Лагранжа.

По (7)–(10) можно понять, что полученное решение $\tilde{\mathrm{z}}$ будет разреженным, то есть $\Delta^3\ \tilde{\mathrm{z}}$ будет содержать большое количество нулевых значений.

Также в третьем разделе доказывается, что сложность у данного алгоритма полиномиальная и приводится доказательство того, что решение, полученное с помощью двойственного алгоритма на основе активного множества, является 3-монотонным и оптимальным (выполняются условия Каруша-Куна-Таккера).

Для дальнейшего описания алгоритма требуется информация о том, что называется активным множеством. Активное множество S состоит из блоков, вида $[l, r-2] \subset [1, n-3]$, таких, что $[l, r-3] \subset S$, $l-1 \notin S$, $r-2 \notin S$, и

$$S = [l_1, r_1] \cup [l_2, r_2] \cup \ldots \cup [l_{m-1}, r_{m-1}] \cup [l_m, r_m],$$

где $l_1 \geq 1$, $r_m \leq n-3$, $r_i + 4 \leq l_{i+1}$, $i \in [1, m-1]$, и m - количество блоков. Если $r_i = l_i$, то и i-ый блок состоит всего из одной точки. Точки $z_{r_i}, z_{r_i+1}, \ldots, z_{l_i}, z_{l_i+1}, z_{l_i+2}, z_{l_i+3}$ находящиеся в i-ом блоке (плюс три точки справа) лежат на прямой линии, и так для каждого i. На каждой итерации алгоритма строится активное множество $S \subset [1, n-3]$ и решается следующая оптимизационная задача:

$$\frac{1}{2}\sum_{i=1}^{n}(z_i - y_i)^2 \to \min, \tag{11}$$

где минимум берётся по всем $z \in \mathbb{R}^n$, удовлетворяющим

$$z_{i+3} - 3z_{i+2} + 3z_{i+1} - z_i = 0 \ \ \forall\, i \in S. \tag{12}$$

Обозначим решение задачи (11)–(12) через $z(S)$.

Теперь можем записать сам алгоритм:

begin

- Входные данные $y \in \mathbb{R}^n$;
- Активное множество $S = \varnothing$;
- Начальное значение для решения $z(S) = y$;
- **while** $z(S) \notin \Delta_n^3$ **do**
  - задаем $S \leftarrow S \cup \{i : \ z_{i+3}(S) - 3z_{i+2}(S) + 3z_{i+1}(S) - z_i(S) < 0\}$;
  - решаем (11)–(12), используя значения из S;
  - Переписываем $z(S)$;
- Возвращаем решение $z(S)$;

end

Доказательство оптимальности полученного решения.

Чтобы доказать оптимальность алгоритма, сначала доказываем несколько вспомогательных лемм.

Лемма 1. *Пусть* z *- оптимальное решение задачи* (5)-(6), *а* y *- вектор, для которого строится* k*-монотонная регрессия. Тогда множители Лагранжа, введенные в* (7)-(10) *могут быть записаны в следующем виде:*

$$\mu_i = -\sum_{j=1}^{i}\left(\sum_{k=j}^{i}(i-k+1)\right)(z_j - y_j), \tag{13}$$

где $1 \le i \le n-3$.

Лемма 2. *Пусть* S*-активное множество и* $1 \in S$, *то есть* $\Delta^3 y_1 < 0$, *и пусть при этом* $2,3,4 \notin S$. *Пусть* z_1, z_2, z_3, z_4 *- значения линейной регрессии, построенной по заданным парам значений* $(1,y_1),(2,y_2),(3,y_3),(4,y_4)$. *Тогда значения соответствующих множителей Лагранжа, определенные в* (13) *будут неотрицательными.*

Лемма 3. *Пусть заданные значения* $y_1,\ldots,y_7$ *такие, что* $1,4 \in S$, *то есть* $\Delta^3 y_1 < 0, \Delta^3 y_4 < 0$, *и при этом* $2,3,5,6,7 \notin S$. *И пусть* z_i *- решения оптимизационной задачи*

$$\frac{1}{2}\sum_{i=1}^{7}(z_i - y_i)^2 \to min, \tag{14}$$

где минимум берется по всем $z = (z_1,\ldots,z_7)$, *таким, что* $\Delta^3 z_1 = \Delta^3 z_4 = 0$. *Тогда значения соответствующих множителей Лагранжа* $\mu_1,\ldots,\mu_7$, *определенные по формуле* (13), *будут неотрицательными.*

Последующие 3 леммы достаточно объемные, поэтому формулировки тут не приведены. В леммах 4 и 5 доказывается, что значения множителей Лагранжа не убывают при добавлении в активное множество новых элементов. А в лемме 6 доказано, что значения μ неотрицательны при решении частного класса задач.

Основным же результатом данного раздела является теорема о сходимости алгоритма к точному решению задачи.

Теорема 3. *Для любого наперед заданного активного множества $S \subset S^*$, алгоритм сходится к точному решению задачи* (1) *не более, чем за $n - |S|$ итераций. Где n - размерность задачи, S^* - активное множество, получившееся на последней итерации алгоритма.*

Каждый раздел содержит **практическую** часть, в которой рассматриваются примеры работы алгоритма. В последнем подразделе второго раздела приведены результаты сравнения всех алгоритмов на тестовой задаче, а именно оценка времени работы, ошибка алгоритма, вычисляемая по формуле $\frac{1}{n}\sum_{i=1}^{n}(z_i - y_i)^2$, кардинальность, количество итераций.

В качестве визуализации работы двойственного алгоритма на основе активного множества рассмотрим пример:

Пусть исходный вектор $y = (y_1, y_2, \ldots, y_n)$ задан следующим образом:

$$y_i = \frac{(x_i)^3}{100} + \varphi_i, i = 1, \ldots, 101,$$

где вектор x заполнен значениями от -50 до 50, а φ_i - случайные величины, распределенные по нормальному закону с параметрами 0 и 10, то есть верно: $\varphi_i \sim N(0, 10), i = 1, \ldots, 101$.

В результате получаем вектор значений $z = (z_1, \ldots, z_{101})$. Если соединить все их прямыми линиями, то получим кривую, которая и является регрессией 3-го порядка монотонности. Её можно видеть на рисунке 1. Так же на рисунке 1 изображены значения y_i в виде полых кругов.

Как видно по рисунку 1, кривая достаточно точно приближает заданный набор значений. Решение было найдено всего за 3 итерации алгоритма.

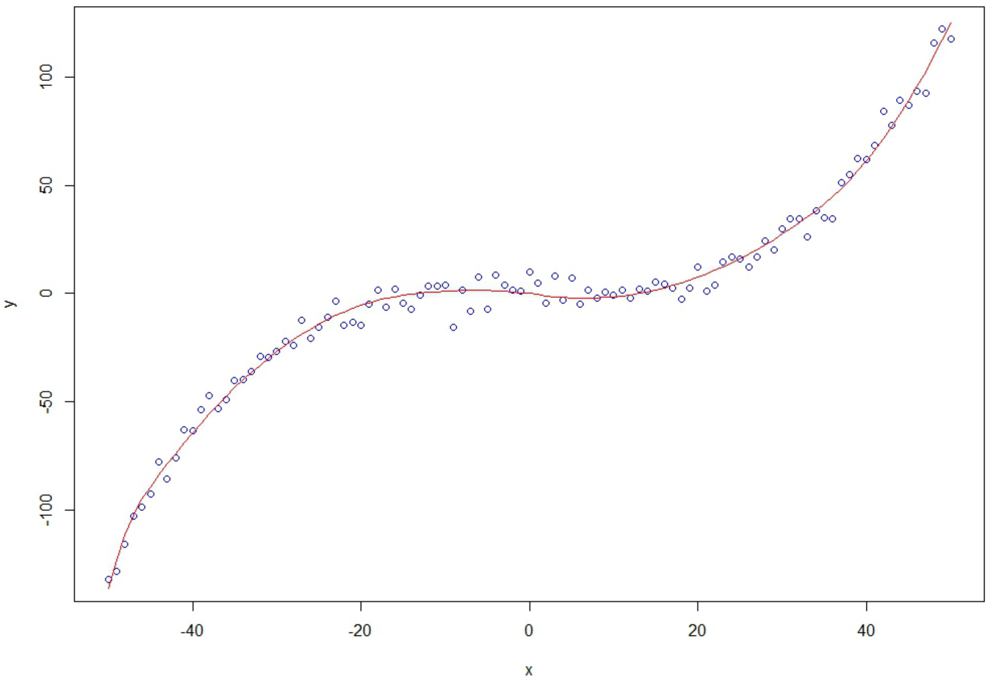

Рисунок 1 – Визуализация работы алгоритма

ЗАКЛЮЧЕНИЕ

В данной работе были рассмотрены некоторые алгоритмы построения k-монотонных регрессий. Для каждого алгоритма приведены результаты работы на тестовых задачах. Также было произведено сравнение работы алгоритмов на тестовой задаче.

Следует отметить, что k-FWA и двойственный алгоритм на основе активного множества тратят намного меньшее количество времени и итераций на поиск решения, но при этом не позволяют контролировать его кардинальность. Алгоритм k-FWA - итерационный алгоритм, на каждой итерации которого решение уже является k-монотонным, однако его точность хуже, чем у предыдущих алгоритмов. Чтобы допиться больше точности, требуется большее количество итераций. Это так же подтверждается теоретическими результатами в виде теоремы 2.

Для двойственного алгоритма на основе активного множества доказана его сходимость к точному решению, а также приведена оценка скорости сходимости.

Все алгоритмы были реализованы на языке R, исходные коды программ приведены в приложении.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Gudkov, A. A. A dual active set algorithm for optimal sparse convex regression / A. A. Gudkov , S. V. Mironov, S. P. Sidorov, S. V. Tyshkevich, // Вестн. Сам. гос. техн. ун-та. Сер. Физ.-мат. науки – 2019. – Т. 23, № 1. – С. 227-246

2 Sidorov, S. P. Algorithms for sparse k-monotone regression / S. P. Sidorov, A. R. Faizliev, A. A. Gudkov, S. V. Mironov // Lecture Notes in Computer Science. – Vol. 10848. – Delft, The Netherlands: Springer International Publishin, 2018. – Pp. 546-556.

3 Гудков, А. А. Алгоритм типа алгоритма Франка - Вульфа для построения монотонной регрессии / А. А. Гудков, А. Р. Файзлиев, С. В. Миронов, С. П. Сидоров // Современные проблемы теории функций и их приложения: материалы 19-й международной Саратовской зимней школы, посвященной 90-летию со дня рождения академика П.Л. Ульянова. – Саратов, Россия: ООО Изд-во «Научная книга», 2018. – С. 111-114.

4 Faizliev, A. R. Greedy Algorithm for Sparse Monotone Regression / A. R. Faizliev, A. A. Gudkov , S. V. Mironov, M. A. Levshunov // Lecture Notes in Computer Science. – 2017. – Vol. 10836. – Pp. 23-31.

5 Гудков, А. А. О сходимости жадного алгоритма для решения задачи построения монотонной регрессии / А. А. Гудков, С. В. Миронов, А. Р. Файзлиев // Изв. Сарат. ун-та. Нов.сер. Сер. Математика. Механика. Информатика. – 2017. – Т. 17, № 4. – С. 431-440.

6 Гудков, А. А. Построение k-монотонных регрессий с использованием жадного алгоритма на основе метода Франка-Вульфа / А. А. Гудков // Научные исследования студентов саратовского государственного университета : материалы итоговой студенческой научной конференции. – Саратов, Россия: Изд-во Сарат. ун-та, 2017. – С. 12-14.

7 Toader G. The representation of n-convex sequences / G. Toader. – 1981. – Vol. 10, no. 1. – Pp. 113-118.