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 РЕШЕНИЯ ЗАДАЧ ПОСТРОЕНИЯ K-МОНОТОННОЙ РЕГРЕССИИ И ИХ ПРИЛОЖЕНИЯ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 248 группы
направления 09.04.03 — Прикладная 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 2021

ВВЕДЕНИЕ

Актуальность темы. В последние годы много внимания уделяется задачам задачам статистики с ограничением на форму данных. Одним из примеров ограничения формы данных является монотонность данных. Построение монотонной регрессии, наилучшим образом приближающей заданный вектор, является одной из наиболее изученных задач статистики с ограничением на форму данных. Продолжением задачи построения монотонной регрессии можно назвать задачу построения k-монотонной регрессии, где k - некоторое натуральное число. Область применения k-монотонных регрессий очень обширна. Например, k-монотонные регрессии используются в непараметрической статистике, при сглаживании эмпирических данных, в формосохраняющем динамическом программировании и при формосохраняющей аппроксимации, а также при решении различных математических задач.

Таким образом, данная тема является актуальной, поскольку k-монотонные регрессии имеют обширную сферу применения, а самой задаче построения k-монотонных регрессий уделяется много внимания в научных работах.

Целью магистерской работы является изучение двойственного алгоритма построения двойственной k-монотонной регрессии, доказательство оптимальности решения, полученного с помощью данного алгоритма, а также оценка его работы в практических задачах.

Объект исследования - двойственный алгоритм построения двойственной k-монотонной регрессии.

Предмет исследования - сходимость двойственного алгоритма построения двойственной k-монотонной регрессии, оптимальность решения, полученного с помощью данного алгоритма.

Для достижения поставленных в работе целей, необходимо решить следующие **задачи:**

— определить основные понятия и задачу построения k-монотонных регрессий;
— рассмотреть двойственный алгоритм построения k-монотонных регрессий для конкретных значений k;
— доказать сходимость двойственного алгоритма к оптимальному реше-

нию задачи построения k-монотонной регрессии для частных случаев;
— рассмотреть двойственный алгоритм построения k-монотонных регрессий в общем случае;
— доказать сходимость двойственного алгоритма к оптимальному решению задачи построения k-монотонной регрессии для общего случая;
— применить двойственный алгоритм построения k-монотонных регрессий при решении практических задач;
— оценить результаты работы алгоритма в практических задачах и описать задачи, в которых алгоритм работает лучше, чем в других.

Практическая значимость исследования в том, что будет изучен алгоритм построения k-монотонных регрессий, а также оптимальность решения, полученного с помощью данного алгоритма. По результатам изучения можно будет сделать вывод об эффективности алгоритма и возможности его применения в различных практических задачах.

Структура и содержание магистерской работы. Выпускная квалификационная работа состоит из введения, трех разделов, заключения, списка использованных источников и приложения. В первом разделе рассматриваются основные понятия и постановка задачи, во втором разделе изучается двойственный алгоритм построения k-монотонной регрессии, а в третьем разделе приведены примеры применения алгоритма в практических задачах.

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 рассматриваемым алгоритмом 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_k^n - множество всех k-монотонных векторов размерности n;

А также ставится основная задача построения k-монотонной регрессии:

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

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).

Во **втором** разделе рассматривается двойственный алгоритм построения k-монотонных ргерессий

Задача (1) может быть переписана в виде задачи выпуклого программирования с линейными ограничениями

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

где минимум берется по всем $z \in \mathbb{R}^n$, таким, что

$$g_i(z) := -\Delta^k z_i \le 0, \quad 1 \le i \le n - k. \qquad (3)$$

В свою очередь, задача (2)–(3) является задачей квадратичного программирования и при этом сильно выпуклой, а значит для неё существует единственное решение.

Данный алгоритм обладает следующими свойствами:

— является алгоритмом полиномиальной сложности, то есть количество операций, требующихся для завершения алгоритма при заданном векторе $y \in R^n$ будет $O(n^p)$, где p - некоторое неотрицательное целое число;
— позволяет получить выпуклое решение;
— позволяет получить оптимальное решение (выполнятся условия Каруша–Куна–Таккера).

Представленный алгоритм использует так называемые активные множества. Активное множество S состоит из блоков вида $[l, r-p] \subset [1, n-k]$, таких, что $[l, r-p] \subset S$, $l-1 \notin S$, $r-p+1, ...r-p+v \notin S$, 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 $l_1 \ge 1$, $r_m \le n-p$, $r_i + p + 1 \le l_{i+1}$, $i \in [1, m-1]$, и m - количество блоков, а p, v - специфичные для каждого порядка монотонности натуральные числа. Если $r_i = l_i$, тогда i-ый блок состоит всего из одной точки.

Значения $z_{r_i}, z_{r_i+1}, \ldots, z_{l_i}, z_{l_i+1}, z_{l_i+k}$, относящиеся к i-му блоку (плюс k точек справа) лежат на прямой линии, и так для любого i.

На каждой итерации алгоритма выбирается некоторое активное множество $S \subset [1, n-k]$ и решается следующая задача оптимизации:

$$\frac{1}{2}\sum_{i=1}^{n}(z_i - y_i)^2 \to \min, \tag{4}$$

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

$$\Delta^k z_i = 0 \;\; \forall\, i \in S. \tag{5}$$

Отметим, что решение задачи (4)–(5) существует и оно единственное. Будем обозначать его как $z(S)$.

Запишем алгоритм для построения k-монотонной регрессии для произвольного k.

The dual active-set algorithm for convex regression

begin

- Входные данные $y \in \mathbb{R}^n$;
- Активное множество $S = \varnothing$;
- Начальное приближение $z(S) = y$;
- **while** $z(S) \notin \Delta_k^n$ **do**
  - Меняем активное множество $S \leftarrow S \cup \{i :\ \Delta^k z_i(S) < 0\}$;
  - Решаем вспомогательную задачу (4)–(5), используя значения из активного множества S;
  - Переписываем вектор $z(S)$;
- Возвращаем решение $z(S)$;

end

Помимо самого алгоритма в данном разделе доказывается теорема об оптимальности данного алгоритма.

Рассмотрим пример доказательства оптимальности решения для алгоритма построения 3-монотонной регрессии.

Задача (1) записывается в виде задачи выпуклого программирования с линейными ограничениями

$$F(z) = \frac{1}{2} z^T z - y^T z \to \min, \tag{6}$$

где минимум берется по всем $z \in \mathbb{R}^n$, таким, что

$$g_i(z) := -\left(\Delta_{i+1} z_{i+3} - z_{i+2}(\sum_{j=i}^{i+2} \Delta_j) + z_{i+1}(\sum_{j=i}^{i+2} \Delta_i) - \Delta_{i+1} z_i\right) \leq 0, \tag{7}$$

при $1 \leq i \leq n-3$. В свою очередь, задача (6)–(7) является задачей квадратичного программирования и при этом сильно выпуклой, а значит для неё существует единственное решение.

Пусть $\hat{z}$ - единственное глобальное решение задачи (6)–(7), тогда существуют множители Лагранжа $\mu = (\mu_1, \ldots, \mu_{n-3})^T \in \mathbb{R}^{n-3}$, такие, что

$$\nabla F(z) + \sum_{i=1}^{n-3} \mu_i \nabla g_i(z) = 0, \tag{8}$$

$$g_i(z) \leq 0, \quad 1 \leq i \leq n-3, \tag{9}$$

$$\mu_i \geq 0, \quad 1 \leq i \leq n-3, \tag{10}$$

$$\mu_i g_i(z) = 0, \quad 1 \leq i \leq n-3, \tag{11}$$

где ∇g_i - это градиент функции g_i. Уравнения (8)–(11) - это условия Каруша–Куна–Таккера. Из (8) следует, что

$$\frac{\partial}{\partial z_j}\left[\frac{1}{2}\sum_{i=1}^{n}(z_i - y_i)^2 + \sum_{i=1}^{n-3} \mu_i \left(-\Delta_{i+1} z_{i+3} + z_{i+2}(\Delta_i + \Delta_{i+1} + \Delta_{i+2}) - \right.\right.$$

$$\left.\left. - z_{i+1}(\Delta_i + \Delta_{i+1} + \Delta_{i+2}) + \Delta_{i+1} z_i\right)\right] = 0,$$

Если просуммировать данные равенства, то получим одно из условий оптимальности решения

$$\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} y_i. \tag{12}$$

При доказательстве оптимальности решения используются вспомогательные леммы, приведем некоторые из них.

Лемма 1. *Пусть* z *- оптимальное решение задачи* (6)-(7)*, а* y *- вектор, для которого строится* k*-монотонная регрессия. Тогда множители Лагранжа, введенные в* (8)-(11) *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 $1 \leq i \leq n-3$*.*

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 соответствующие множители Лагранжа, определенные в* (13) *будут неотрицательными.*

Теорема 1. *Для любого начального активного множества* $S \subset S^*$*, алгоритм сходится к оптимальному решению задачи* (1) *не более, чем за* $n - |S|$ *итераций.*

Доказательство. Идея доказательства оптимальности решения следующая: рассмотрим работу алгоритма по итерациям и докажем, что для решения, полученного на каждой итерации выполняются условия Каруша–Куна–Таккера.

Общее количество итераций можно получить с помощью следующих рассуждений: на каждой итерации алгоритма активное множество S пополняется по крайней мере одной точкой из $[1, n-3]$, которая до этого не принадлежала множеству S. Алгоритм завершает работу, когда $z(S)$ становится 3-монотонным. Если $S = [1, n-3]$, то блок будет всего 1, а следовательно 3-монотонность не будет нарушаться. Если $|S| < n-3$, то количество итераций будет меньше, чем $n - |S|$, где $|S|$ - количество элементов в активном множестве S.

Оптимальность решения следует из доказанных ранее лемм.

На каждой итерации алгоритма в активное множество добавляются точки, в которых нарушается 3-монотонность. Если одна из таких точек i

изолированная (то есть $i-3, i-2, i-1, i+1, i+2, i+3 \notin S$), то алгоритм заменяет $z_i, z_{i+1}, z_{i+2}, z_{i+3}$ значениями линейной регрессии, построенной по $(i, z_i), (i+1, z_{i+1}, (i+2, z_{i+2}), (i+3, z_{i+3})$. По лемма 2 соответствующие значения множителей Лагранжа будут неотрицательными.

При доказательстве неотрицательности других ситуаций, возникающих при работе алгоритма, так же используются вспомогательные леммы. □

В **третьем** разделе алгоритм используется при решении различных практических задач и оценивается целесообразность его применения.

Рассмотрим задачу построения 3-монотонной регрессии для данных с постоянным шагом: $x = (-100, -99, \dots, 99, 100) \in \mathbb{R}^{201}$, и при этом $y \in \mathbb{R}^{201}$, $y_i = \frac{x_i^3}{1000} + \frac{x_i^2}{100} + \frac{x_i}{100} + \varphi_i$, где $\varphi_i \sim N(0,5)$ - случайная величина, распределенная по нормальном закону с параметрами 0 и 5. Результат работы можно увидеть на рисунке 1. Как можно видеть по рисунку 1, значения решения, полученные с помощью алгоритма и соединенные на графике линией, точно повторяют форму исходных данных, изображенных в виде точек, что может свидетельствовать о его применимости к аналогичным задачам.

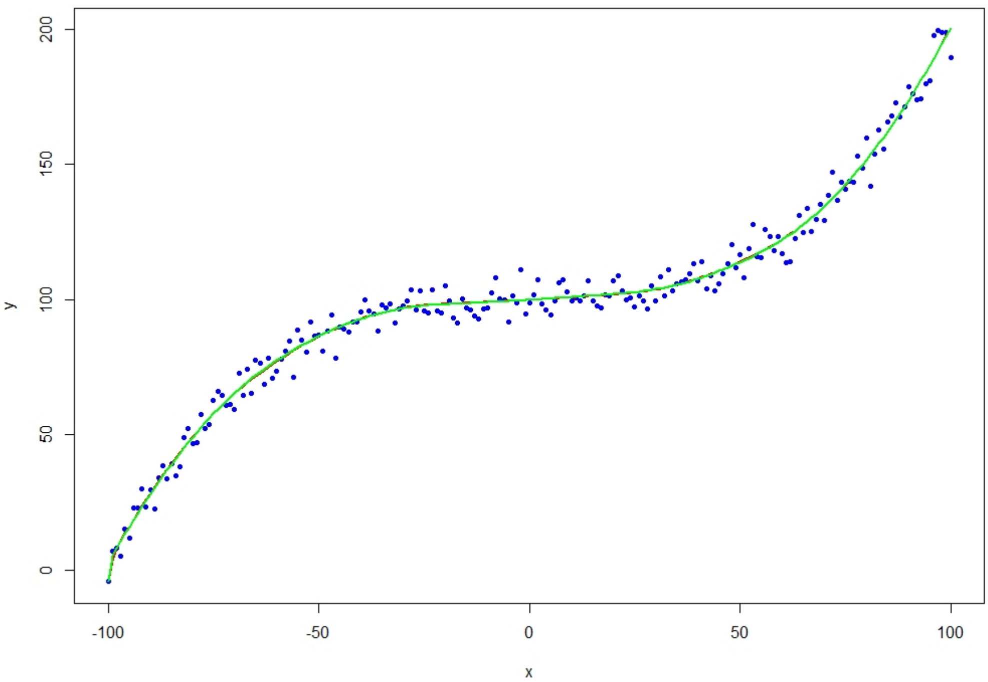

Рисунок 1 – Визуализация работы алгоритма

ЗАКЛЮЧЕНИЕ

В данной работе рассмотрен двойственный алгоритм построения k-монотонных регрессий. Для алгоритма приведены результаты работы в практических задачах. Рассмотрены как сильные, так и слабые стороны алгоритма, и приведены рассуждения о его применимости к различным задачам. Можно сделать вывод, что алгоритм показывает хорошие результаты в задачах построения k-монотонных регрессий.

Для двойственного алгоритма на основе активного множества доказана его сходимость к точному решению, а также приведена оценка скорости сходимости. Данный алгоритм имеет полиномиальную сложность и сходится к оптимальному решению задачи не более, чем за n итераций, где n - размерность задачи.

Алгоритм реализован на языке R, исходные коды программ приведены в приложении.