**Государственное бюджетное общеобразовательное учреждение**

**города Москвы «Школа №1532»**

**Непараметрическая оценка регрессии методом Розенблатта - Парзена**

ГБОУ Школа №1532, 10 класс

Фукс Даниил

Руководитель: учитель информатики

ГБОУ Школа 1532,

Сергиенко Антон Борисович

Москва, 2024

Оглавление

[Введение 3](#_Toc158724739)

[Цель и задачи работы 4](#_Toc158724740)

[Гипотеза 5](#_Toc158724741)

[Методика выполнения работы 5](#_Toc158724742)

# Введение

**Актуальность работы**

В современном мире стоимость проведения экспериментов значительно возросла, поэтому для экономии ресурсов и достижения необходимых прогнозов важно уметь предсказывать результаты экспериментов. Существуют два основных подхода к построению графиков зависимости на основе выборочных данных: параметрическая и непараметрическая оценка. Параметрическая модель имеет недостатки, такие как:

* Предположение о функциональной форме: параметрическая оценка требует предположения о функциональной форме зависимости между независимыми и зависимыми переменными. Если это предположение неверно, то модель может давать неточные или неправильные результаты.
* Ограниченность модели: параметрическая модель имеет ограниченное количество параметров, которые могут быть оценены. Это ограничивает способность модели улавливать сложные зависимости в данных.
* Чувствительность к выбросам: параметрическая оценка регрессии может быть чувствительна к выбросам в данных. Одиночные аномальные значения могут сильно влиять на оценку параметров и приводить к неточным результатам.

Непараметрическая оценка регрессии в свою очередь имеет недостатки, например, требуемый объем данных, однако она обладает лучшей гибкостью, т.е. не требуют предварительного задания функциональной формы зависимости между переменными, также непараметрические методы обычно менее чувствительны к выбросам, поскольку они не полагаются на предположения о распределении данных или о функциональной форме модели. Главные минусы непараметрической оценки регрессии методом Розенблатта-Парезна заключаются в требуемом объеме данных, а также в сложности вычислении оптимальных параметров.

# Цель и задачи работы

Разработка и анализ эффективности моделирования многомерной многозначной функции на основе экспериментальных данных через использования непараметрической оценки регрессии.

Основными задачами является:

1. Разработать непараметрическую модель с наилучшим значением параметра для многозначной функции (в которой одному значению соответствует несколько значений ).
2. Реализовать предложенный метод в программном продукте.
3. Построить графики.

Методы и методики решения основных задач: «стандартная» непараметрика, среда программирования: PyCharm.

# Гипотеза

Автором предполагается, что непараметрическую оценку регрессии можно применять не только для случаев однозначной функции (когда одному значению соответствует несколько значений ), но и применить ее для случаев многозначных функций с использованием модификаций оценки регрессии.

# Методика выполнения работы

Главная формула, которая будет использоваться для вычисления предположительной формулы:

– элемент выборки,

– точка, для которой ищется точка ,

– элемент выборки,

– настраиваемый параметр

Параметр «выключает» точки, который расположены слишком далеко от искомого значения. Это позволяет учитывать только нужные нам данные. Т.к. – настраиваемый параметр, то при непараметрическая оценка регрессии будет представлять из себя среднее арифметическое, т.к. будут учитываться все точки и делиться на их количество.

Такой вариант не подходит, т.к. данная модель может выдавать слишком неточные значения, из-за чего параметр требуется подбирать под каждую выборку. Чтобы определять, насколько параметр подходит под выборку, требуется определять «ошибку», которая определяется по формуле:

– элемент выборки,

– смоделированное значение,

– количество элементов выборки

Автор работы написал код, который вычисляет зависимость «ошибки» от параметра . Был получен следующий результат:

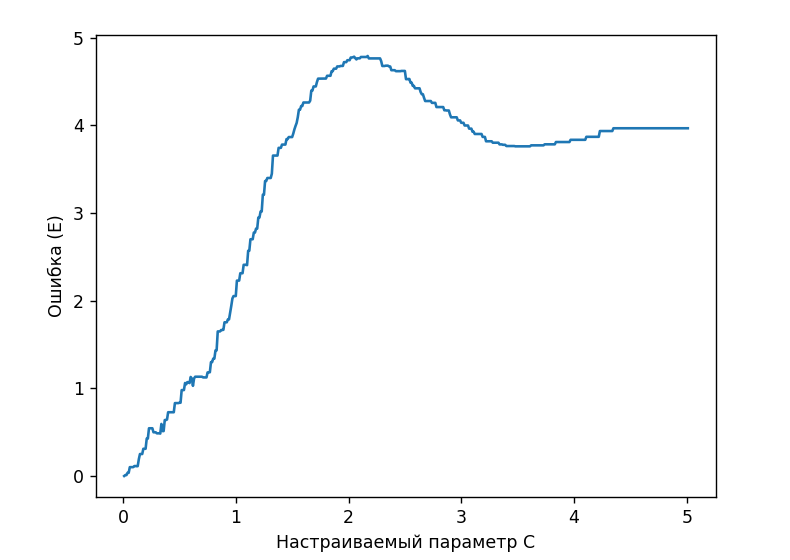


Рисунок 1 – график зависимости ошибки от настраиваемого параметра без «выключения» точек

На графике видно, что при такой модели при → 0 ошибка будет минимальна.

Так происходит из-за точек, которые у нас уже имеются. При слишком малом функция моделирования точки создает только те же значения, что уже есть в выборке. Иными словами, функция возвращает имеющееся значение, а если значения нет, то возвращает ноль.

Изображение выглядит как диаграмма, линия, График, снимок экрана

Автоматически созданное описание

Рисунок 2 – Аппроксимация без «выключения» имеющихся значений

Синим цветом обозначены точки, которые представленны в выборке. Красная линия построена на основе предположенных функцией точек. По графику видно, что он принимет значение не равное нулю только при том значении x, при котором эксперемент производился, т.е. данные есть в выборке. В ином случае функция возвращает ноль.

Для того, чтобы такого не происходило, требуется не включать в выборку точки, которые уже существуют. Если их исключить, то получится следующий график:

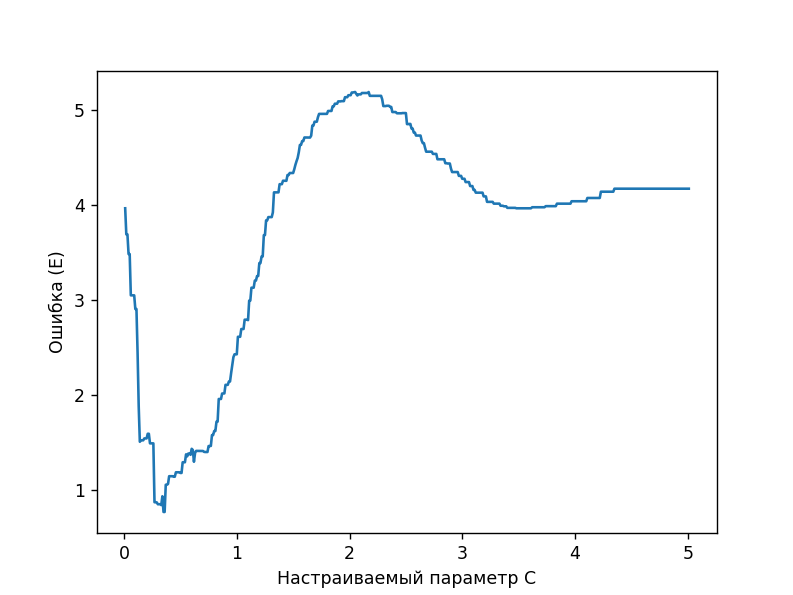


Рисунок 3 – график зависимости ошибки от настраиваемого параметра () с «выключением» точек

На нем видно, что теперь при → 0 ошибка не минимальна. Далее автор подобрал наилучшее значение параметра , при котором ошибка будет минимальна. Затем, опираясь на этот параметр, были смоделированы новые точки и получен следующий график:

Изображение выглядит как карта, линия, диаграмма, текст

Автоматически созданное описание

Рисунок 4 – Аппроксимация с «выключением» имеющихся значений

На этом моменте формула готова, но только для однозначной функции. Иногда приходится работать с функцией, которая при одинаковом значении X имеет два и более различных значений Y. Такие функции называются многозначными. Для этого автором предлагается следующая модификация непараметрической оценки регрессии: