УДК ***378.02:37.016***

САБИТОВ Б.Р., АЛМАСБЕКОВА З., КАРАГУЛОВ Ш.К.
КНУ им. Ж. Баласагына, Бишкек
SABITOV B.R., ALMASBEKOVA Z., KARAGULOV SH.K.
J. Balasagyn KNU, Bishkek

ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ В ПРИКЛАДНЫХ ЗАДАЧАХ

Прикладдык маселелерде машиналык окутууну колдонуу

The use of machine learning in applied tasks

Аннотация: *В настоящей статье исследуется построение линейных и нелинейной моделей с применением технологий машинного обучения для прикладных задач. Изучаются различные модели и их аппроксимирующее свойства для реальных процессов с применением полиномов различных степеней.*

Аннотация: *Берилген макалада прикладдык маселелер үчүн машиналык окутуунун технологиясын колдонуп сызыктуу жана сызыктуу эмес моделдерди тургузуу изилденет. Ар кандай тартиптеги көп мүчөлөрдү колдонуп моделдер жана алардын жакындаштыруу касиеттери изилденет.*

Annotation:*This article explores the construction of linear and non-linear models using machine learning technologies for applied problems. We study various models and their approximating properties for real processes using polynomials of various degrees.*

Ключевые слова: *Прикладные задачи, регрессия, метод наименьших квадратов, python, модели, регуляризация, машинное обучение*

Урунттуу сөздөр: *Прикладдык маселелер, регрессия, эң кичине чарчы усулу, python, моделдер, регуляризация, машинналык окутуу.*

Keywords: *Applied tasks, regression, least-squares method, python, models, regularization, machine learning.*

Во многих работах задачи построения линейных моделей изучаются с применением линейной регрессии [1]. Обычно изучаемая модель представляется только линейными функциями, зависящихся от признаков и представляет собой линейную модель. Класс таких задач решается методом наименьших квадратов [2]. В данной работе мы используем в качестве изучаемой модели некоторую нелинейную функцию, входящие признаки в которых, имеют вид различных функций степеней p. Полином данной модели в нашем случае будет выглядеть следующим образом:

$$\forall h \in H, h(x) = \omega_0 + \omega_1 x + \omega_2 x^2 + \ldots + \omega_p x^p = \sum_{i=0}^{p} \omega_i x^i \quad (1)$$

Введя соответствующие обозначения

$$x_1 = x, x_2 = x^2, \text{и.т.д.} \quad (2)$$

наша нелинейная модель, с помощью введенных обозначений (2) будет выглядеть следующим образом

$$\forall h \in H, h(x) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 \ldots + \omega_p x_p = \sum_{i=0}^{p} \omega_i x^i \quad (3)$$

В данном случае мы снова получаем линейную модель для исходной нелинейной. К данной задаче можно применить снова алгоритм — метода наименьших квадратов. Таким образом полиномиальная *регрессия* относится к тому же классу линейных задач. Для прикладных задач особенно важно изучить аппроксимирующие свойства полиномов различных степеней для базы данных различных моделей. Технологии машинного обучения позволяют данные моделей разбивать на обучающие и тестовые, что позволяет построить новую модель на основе уже обученных моделей. Неизвестные коэффициенты ω, в данном случае вычисляются с применением следующей формулы

$$\vec{\omega} = (X^T X)^{-1} X^T \vec{y}, \quad (4)$$

которая получится из условий минимизации некоторого функционала $L(\omega_0, \omega_1, \ldots \omega_n)$ и называется нормальным уравнением. Структура функционала представляет с собой

следующую среднеквадратичную ошибку

$$L(X,\vec{y},\vec{\omega}) = \frac{1}{2n}||\vec{y} - X\vec{\omega}||_2^2$$

Ниже приведены графики анализа с применением регрессионного анализа для нелинейных данных сгенерированного в виде, параболы с гауссовым шумом. Как видно из рисунка, наиболее подходящим является прогноз в виде параболической кривой.

Рис.1. Применение линейной регрессии с полиномами различных 2 степени

При увеличении степени полинома, в данном случае точность прогноза растет. Например, для полинома степени 100 наша построенная модель имеет склонность к переобучению.

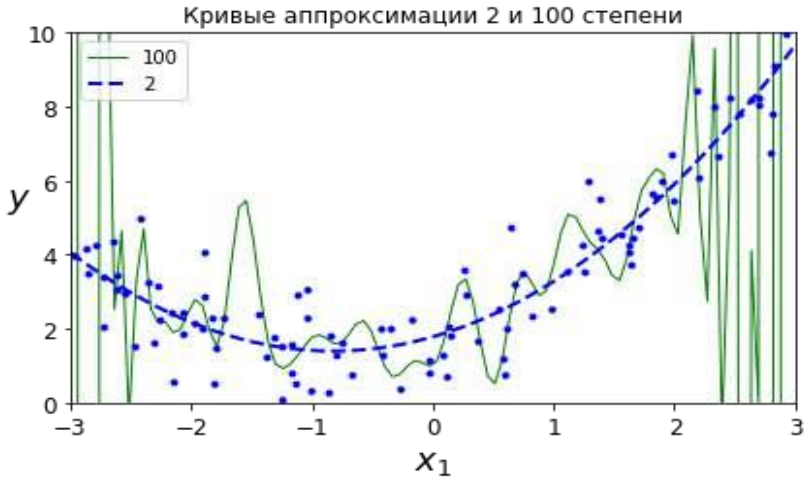

Рис.2. Применение линейной регрессии с полиномами 2 и 100 степеней.

Отметим, что в обеих случаях коэффициенты регрессии вычислялись с помощью нормального уравнения. Ниже приведены результаты анализа нелинейных данных для различных прикладных задач, с использованием нелинейных данных вида $\sin(x)$, построены аппроксимирующие свойства полиномами различных степеней.

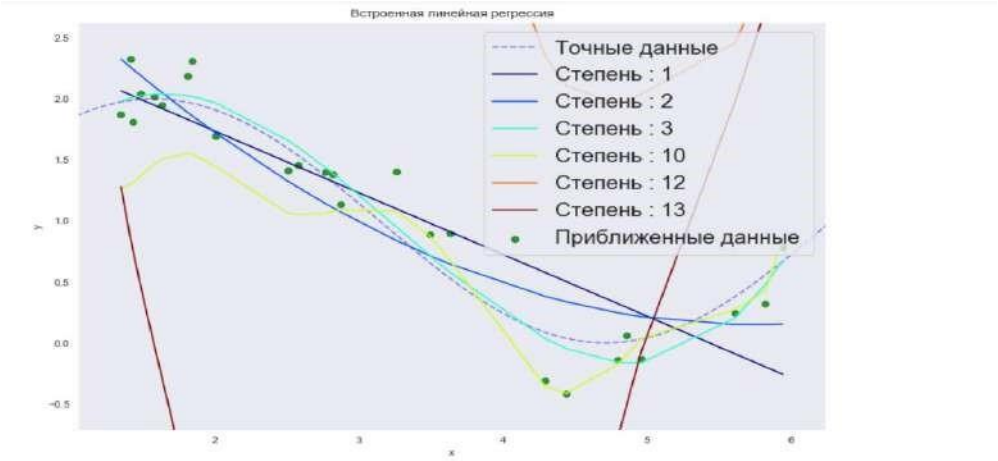

Рис.3. Применение линейной регрессии с полиномами различных степеней для функции $L(x, \omega) =$

$$\omega_0 + \omega_1 * sin(x)$$

Обычно при увеличении степени полинома, среднеквадратичная ошибка уменьшается. В нашем случае полином третьей степени интерполирует вместо экстраполяции данных. Таким образом, сложные модели, у которых степеней свободы достаточно много, могут попросту запомнить весь тренировочный набор, полностью теряя обобщающую способность. Видно наша матрица стала слабо обусловленной, а задача - некорректно поставленной. При вычислении обратной матрицы мы имеем большую дисперсию. При небольшом изменении начальной матрицы, обратная будут сильно отличаться друг от друга. Изучение собственных чисел полученной матрицы с помощью numpy показывает, что присутствуют комплексно значные собственные значения для нашей матрицы.

Известно, что для симметричных и положительно определенных матриц собственные значения должны быть действительными числами, что не совпадает с теорией. Наша матрица стала в данном случае несимметричной.

Метод регуляризации А.Н. Тихонова

Мы выше видели выше, что матрица $(X^T X)$ сингулярная и обратная матрица не существуют. В таких случаях обычно используются различные методы регуляризации. В работе, данная технология изучается уже с применением метода регуляризации А.Н.Тихонова[3].

Метод регуляризации в пространстве L_1

Изучим норму L_1:

$$\mathrm{R}(\vec{\omega}) = ||\vec{\omega}||_1 = \sum_{j=1}^{m} |\omega_j|$$

Тогда функционал $\mathrm{L}(\mathrm{X}\vec{y}, \vec{\omega})$ в задаче

$$\min_{\omega} \mathrm{L}(\mathrm{X}, \vec{y}, \vec{\omega})$$

имеет вид

$$\mathrm{L}(\mathrm{X}, \vec{y}, \vec{\omega}) = \frac{1}{2n} \sum_{i=1}^{n} (\vec{x}_i^T \vec{\omega} - y_i)^2 + \lambda \sum_{j=1}^{m} |\omega_j|$$

Выпишем необходимое условие существования минимума данного функционала:

$$\frac{d}{d\omega_j} \mathrm{L}(\mathrm{X}, \vec{y}, \vec{\omega}) = \frac{1}{n} \sum_{i=1}^{n} (\vec{x_i}^T \vec{\omega} - y_i) \vec{x_i} + \lambda sign(\omega_j) = 0$$

Для поиска приближенного решения можно использовать метод градиентного спуска. Для искомых весов исходного функционала используем метод градиентного спуска [4]:

$$\vec{\omega}_{new} := \vec{\omega} - \alpha \frac{dL}{d\vec{\omega}}$$

В данном случае α параметр регуляризации антиградиента, которая отвечает за скорость спуска функционала к минимуму.

Для нормализации ошибки параметров применим следующее преобразование:

$$\bar{\mu}_{.j}=\frac{1}{n}\sum_{i=1}^{n} x_{ij}$$
$$\bar{\sigma}_{.j}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{\mu}_{.j})^2}$$
$$\vec{x}_{new}=\frac{\vec{x}-\vec{\mu}}{\vec{\sigma}}$$

Такое преобразование называется стандартизацией, распределение каждого признака теперь имеет нулевое математическое ожидание и единичную дисперсию. В данном случае необходимо запустить процесс еще раз. Давайте, теперь реализуем метод регуляризации Тихонова в пространстве $\boldsymbol{L_2}$. Для ее реализации выпишем регуляризирующий функционал Тихонова, которая в общем виде выглядит как добавление нового члена к средне квадратичной ошибке:

$$\mathrm{L}(\mathrm{X},\vec{y},\vec{\omega})=\frac{1}{2n}||\vec{y}-\mathrm{X}\vec{\omega}||_2^2+||W\vec{\omega}||^2$$

В качестве матрицы Тихонова, которая выражается как произведение некоторого числа на единичную матрицу возьмём в виде:

$$\mathrm{W}=\frac{\lambda}{2}\mathrm{E}$$

где, λ называется параметром регуляризации. Если возьмем производную получим новую функцию стоимости по параметрам модели. Приравняв к нулю, которую определяем $\vec{\omega}$, как точное решение нашей задачи минимизации функционала в виде нормального уравнения аналогичной (4).

$$\vec{\omega}=(X^TX+\lambda E)^{-1}X^T\vec{y}$$

Полученная регрессия называется Тихоновской или гребневой регрессией. А гребнем в данном случае подразумевается диагональная матрица, которую мы прибавляем к исходной матрице XTX.В результате у нас получается регулярная матрица, обратная матрица которой существует. В целом решение уменьшает дисперсию, и становится смещенным.

Определив коэффициенты и применяя технологию сглаживающих функционалов Тихонова, построим кривую регрессии с применением библиотек Python.

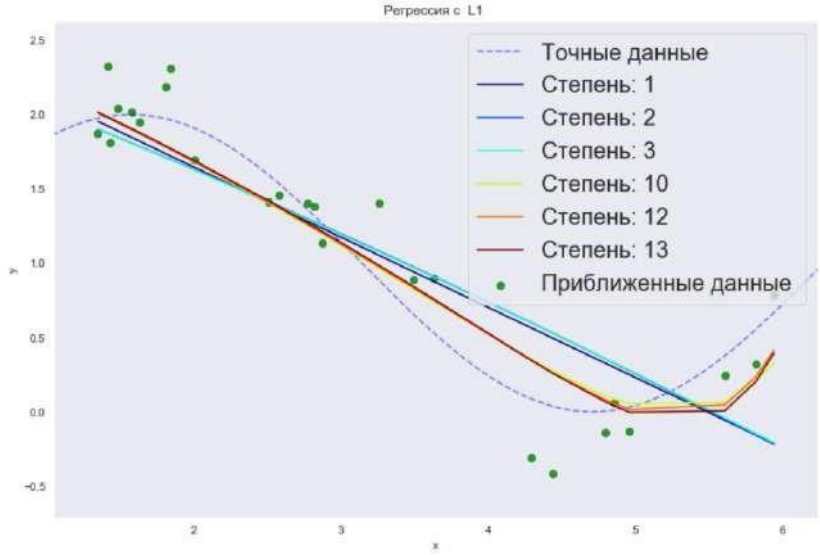

Рис.4. Результат применения метода регуляризации А.Н. Тихонова для полиномиальной регрессии.

В данном случае большая часть коэффициентов близка или равна к нулю. Таким образом применение описанного способа построения регрессии, дает улучшенные результаты для нелинейных данных, которые описывают многие прикладные задачи.

Заключение

В данной работе были изучены некоторые модели для прикладных задач и их построения с применением технологий регрессионного анализа. Использованы параметрические методы линейной регрессии, которые дают толчок к решению непараметрических моделей. Изучено, также вопрос полиномиальной аппроксимации с различными степенями для построения некоторых сложных моделей.

Список цитируемых источников

1. Р. Норман, Г.С. ДрайперПрикладной регрессионный анализ. - М.: Наука,2016
2. Губанов В.С. Обобщенный метод наименьших квадратов. -М.: Наука,1997
3. А.Н. Тихонов, В.И. Арсенин Методы решения некорректных задач. -М.: Наука,1979
4. Ф.П.Васильев Численные методы решения экстремальных задач. - М.: Наука,1988

Рецезент:

Бийбосунов Б.И. *- доктор физико-математических наук, профессор*