УДК 519.2

Шульгін О. Я., здобувач вищої освіти;
Ніколюк П. К., д-р фіз.-мат. наук, професор,
професор кафедри інформаційних технологій,
Донецький національний університет імені Василя Стуса

ВАЖЛИВІСТЬ АНАЛІЗУ КОЕФІЦІЄНТА РЕГУЛЯРИЗАЦІЇ ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ ПОЛІНОМІАЛЬНОЇ РЕГРЕСІЇ

Ключові слова: регресія, нормалізація, аналіз.

Вступ. Поліноміальна регресія – це метод статистичного аналізу та моделювання, який використовує поліноми для побудови функцій, що описують взаємозв'язок між незалежними та залежними змінними в дослідженні або аналізі даних. У поліноміальній регресії залежна змінна (вихідна змінна) моделюється як поліном від незалежної змінної (вхідної змінної). Вона дає змогу апроксимувати складні залежності між змінними, де зв'язок може бути нелінійним. Регуляризація регресії – це техніка у статистиці та машинному навчанні, яка використовується для зменшення перенавчання (overfitting) та підвищення стійкості моделі, особливо у випадках, коли регресійна модель має велику кількість незалежних змінних або коли дані мають високу розмірність.

Актуальність. Проблема виявлення справжньої закономірності на основі результатів експериментів є дуже важливою і актуальною. Одним із завдань регресійного аналізу є створення моделі, яка дає змогу отримувати оцінки значень залежної змінної на основі значень незалежних показників.

Зазвичай поліноміальна регресія виглядає так:

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon \quad (1),$$

де y – залежна змінна, яку намагаються передбачити або пояснити;

x – незалежна змінна, яка впливає на залежну змінну;

$\beta_0, \beta_1, \beta_2, ..., \beta_k$ – коефіцієнти поліноміальної регресії, які слід визначити під час аналізу даних;

ε – помилка, яка враховує випадкову варіативність, яка не може бути пояснена моделлю [1].

L_1- і L_2-регуляризація – це два тісно пов'язані методи, які застосовують для зменшення ступеня перенавчання моделі, що забезпечує більш якісне прогнозування [2]. Обидва методи додають додатковий член до функції втрат, який враховує регуляризаційний штраф. Він залежить від коефіцієнта λ, який визначає ступінь регуляризації моделі [3].

L_1-регуляризація (Lasso) використовує штраф на основі суми модулів коефіцієнтів регресії (2). Цей метод допомагає досягти розрідженості, тобто встановлює деякі коефіцієнти регресії точно на нуль, зменшуючи кількість незалежних змінних у моделі:

$$L_1 = \sum_i \left(y_i - y(t_i)\right)^2 - \lambda \sum_i |a_i|. \quad (2)$$

L_2-регуляризація (Ridge) використовує штраф на основі суми квадратів коефіцієнтів регресії (3). Цей метод зменшує величину коефіцієнтів регресії, роблячи їх менш чутливими до шуму в даних:

$$L_2 = \sum_i \big(y_i - y(t_i)\big)^2 - \lambda \sum_i a_i^2. \quad (3)$$

Аналіз коефіцієнтів проведемо на прикладі пошуку коефіцієнтів наступного полінома (4):

$$f(x) = \ -6 + 3{,}5x + 0{,}5x^2 + 0{,}4x^3. \quad (4)$$

На відрізку [–7, 7] визначимо 20 випадкових значень, які відповідають значенню полінома з додаванням шуму і використовуються для тренування моделі. Також згенеруємо 20 інших випадкових значень нашого полінома без шуму (червоні точки) для перевірки моделі. В якості параметру λ виберемо значення, рівномірно розподілені в діапазоні від 10^{-4} до 10^3 включно. Також для порівняння додамо випадок, коли $\lambda = 0$ (регуляризація відсутня). Внаслідок моделювання отримаємо дев'ять кривих (рис. 1, 2).

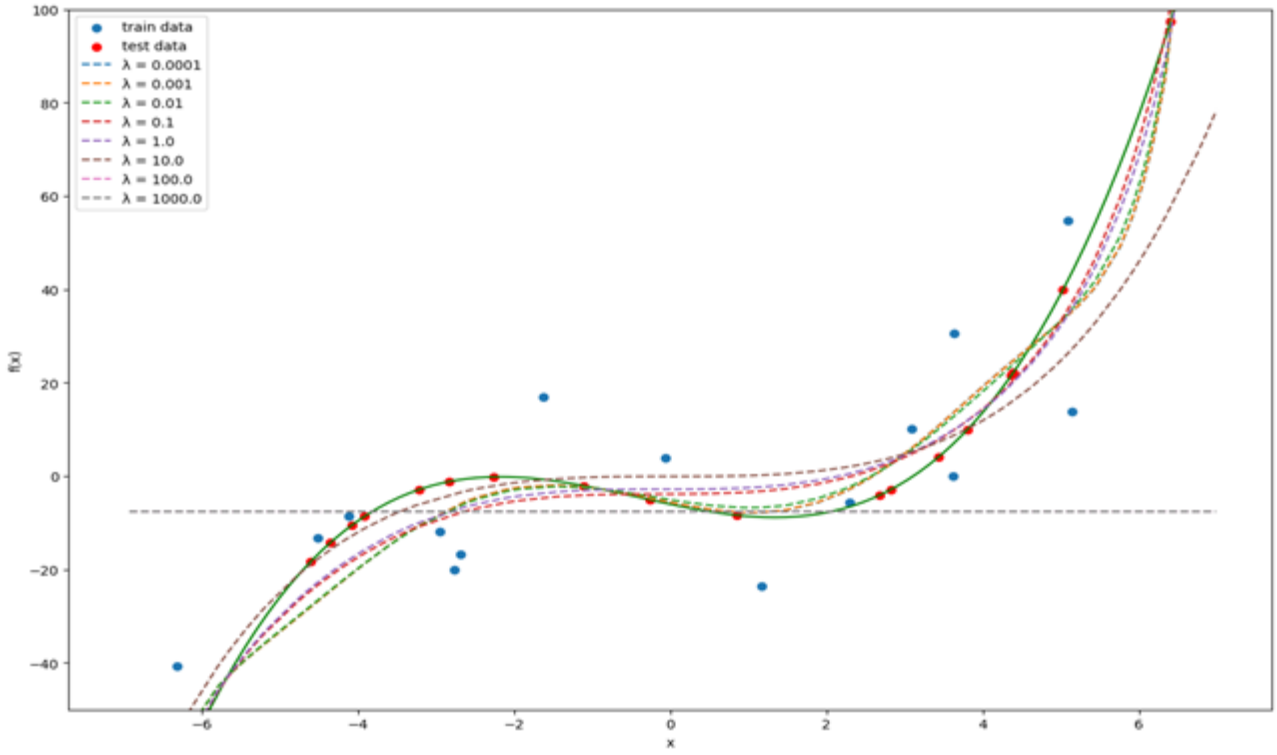

Рисунок 1 – Результати моделювання L_1-регуляризації для різних значень λ

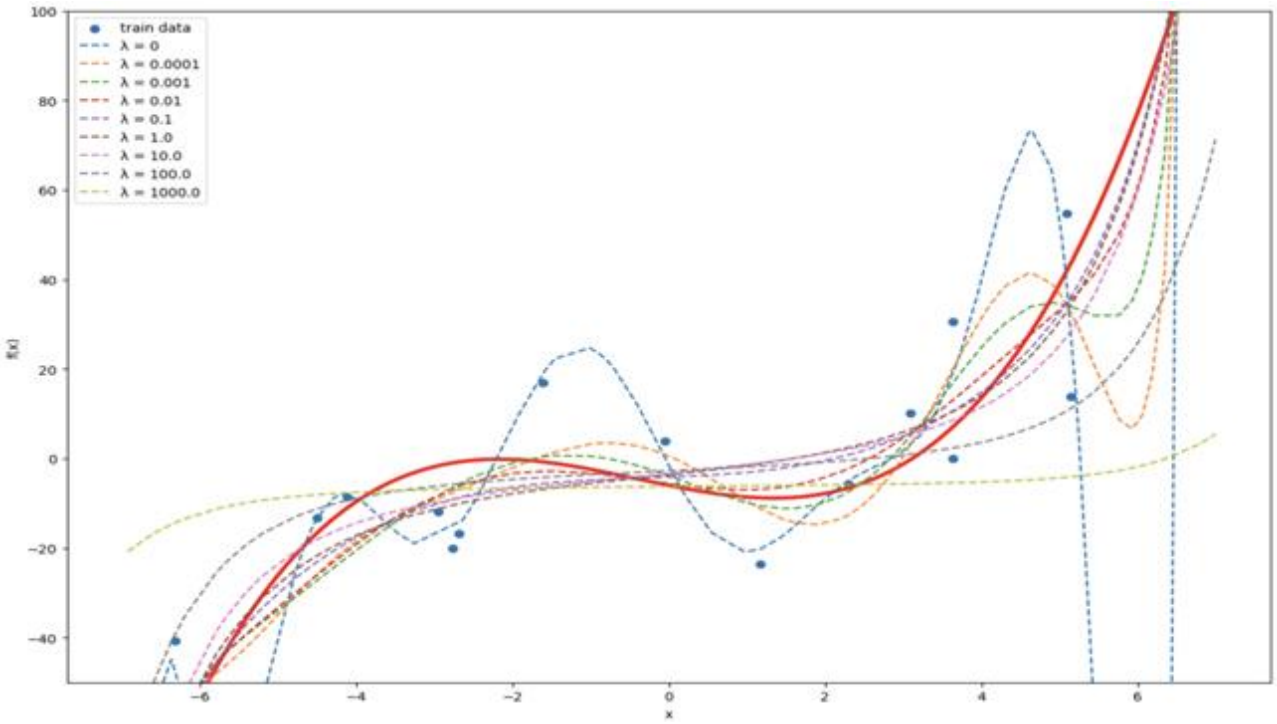

Рисунок 2 – Результати моделювання L_2-регуляризації для різних значень λ

Висновки

Аналізуючи результати, можна побачити, що надто малий коефіцієнт призводить до перенавчання моделі (це особливо добре видно на прикладі L_2-регуляризації). Натомість надто великий коефіцієнт призводить до недонавчання моделі. Тому оптимальним буде починати пошук значення коефіцієнту в межах від 0,1 до 1. Але остаточний вибір, звичайно, залежить від конкретної задачі. Також варто додати, що існує третій вид регуляризації – Elastic Net. Він поєднує в собі підходи, використані в L_1- і L_2-регуляризації. Під час вирішення задачі варто спробувати застосувати Elastic Net і також провести аналіз вибору коефіцієнта регуляризації.

Список використаних джерел

1. Поліноміальна регресія. URL: https://uk.wikipedia.org/wiki/Поліноміальна_регресія (дата звернення: 06.11.2023).

2. Пулеко І. В., Обіход С. В. Особливості застосування алгоритмів лінійної регресії у службі машинного навчання Microsoft Azure *Комп'ютерні технології: інновації, проблеми, рішення:* Тези доповідей III Всеукраїнської науково-технічної конференції, 26–27 листопада 2020 р. Житомир: Житомирська політехніка, 2020. С. 79–80. URL: https://conf.ztu.edu.ua/wp-content/uploads/2021/01/tezy-dopovidej-kt2020_os-2.pdf (дата звернення: 06.11.2023).

3. Test Run – L_1 and L_2 Regularization for Machine Learning. URL: https://learn.microsoft.com/en-us/archive/msdn-magazine/2015/february/test-run-l1-and-l2-regularization-for-machine-learning (дата звернення: 06.11.2023).