Міністерство освіти і науки України

Львівський національний університет імені Івана Франка

Факультет електроніки та комп’ютерних технологій

Кафедра системного проектування

Звіт

Про виконання лабораторної роботи №3

З курсу «Системи машинного навчання»

### Регуляризація

**Виконала:**

Студентка групи ФЕС-32

Філь Дарина

**Перевірив:**

Доцент Колич І.І.

Львів 2024

**Мета:** Засвоїти основи регресійного аналізу з використанням регуляризації.

**Інструменти:** Python, Skilit-learn, Matplotlob, Seaborn.

**Теоретичні відомості**  
**Лінійна регресія**

Лінійна регресія є основним методом машинного навчання для моделювання взаємозв'язків між змінними. Вона дозволяє передбачати значення залежної змінної на основі незалежних змінних.

Формула лінійної регресії

Формула лінійної регресії (з більш ніж однією незалежною змінною) виглядає так:

де:

* залежна змінна;
* вільний член (intercept);
* … , коефіцієнти регресії для незалежної змінної;
* … , незалежні змінні;

**Поліноміальна регресія**

Поліноміальна регресія є узагальненням лінійної регресії, яка дозволяє моделювати нелінійні взаємозв'язки між змінними.

**Формула поліноміальної регресії:**

де:

* залежна змінна;
* вільний член (intercept);
* … , коефіцієнти регресії для незалежної змінної;
* … , незалежні змінні;
* ступінь полінома;

**Регуляризація**

Регуляризація є технікою, яка використовується для запобігання перенавчанню (overfitting) моделей шляхом додавання штрафу до функції втрат. Основні методи регуляризації — це Ridge і Lasso.

**Причини використання регуляризації:**

1. **Запобігання перенавчанню:** регуляризація допомагає моделі загальнювати краще на нових даних, запобігаючи перенавчанню на тренувальних даних.
2. **Стабільність моделі:** регуляризація зменшує варіацію у прогнозах моделі, роблячи її більш стабільною.
3. **Інтерпретованість:** lasso регресія може використовуватися для відсічення неважливих змінних, що покращує інтерпретованість моделі.

**Основні методи регуляризації**

1. **Ridge регресія (L2 регуляризація)**

* **Визначення:** Ridge регресія додає штраф за величину коефіцієнтів регресії до функції втрат. Це допомагає зменшити величину коефіцієнтів, що може знизити ризик перенавчання.
* **Формула:**

де:

* параметр регуляризації, який контролює величину штрафу;
* коефіцієнт регресії;
* **Механізм:** Ridge регресія додає штраф у вигляді суми квадратів коефіцієнтів регресії. Це означає, що модель намагається зменшити величину коефіцієнтів, щоб уникнути перенавчання.
* **Переваги:**
* Допомагає запобігти перенавчанню;
* Підходить для моделей з великим числом ознак;
* **Недоліки:**
* Не може виконувати відсічення ознак (значення коефіцієнтів регресії зменшуються, але не стають нульовими);

1. **Lasso регресія (L1 регуляризація)**

* **Визначення:** Lasso регресія додає штраф за абсолютне значення коефіцієнтів регресії до функції втрат. Це допомагає зменшити кількість ознак, що використовуються моделлю, шляхом відсічення неважливих ознак.
* **Формула:**

де:

* параметр регуляризації, який контролює величину штрафу;
* коефіцієнт регресії;
* **Механізм:** Lasso регресія додає штраф у вигляді суми абсолютних значень коефіцієнтів регресії. Це означає, що деякі коефіцієнти можуть стати нульовими, ефективно виключаючи відповідні ознаки з моделі.
* **Переваги:**
* Виконує відсічення ознак, що спрощує модель і покращує інтерпретованість.
* Допомагає запобігати перенавчанню.
* **Недоліки:**
* Може призводити до високої варіації у випадках, коли існує висока кореляція між ознаками;

**Оцінка моделі**

Після навчання моделі лінійної регресії важливо оцінити її продуктивність. Ось деякі ключові метрики для оцінки моделі:

1. **Середньоквадратична помилка (Mean Squared Error, MSE):**

* **Визначення:** : Середньоквадратична помилка (MSE) є середнім значенням квадратів різниць між фактичними значеннями та передбаченими значеннями. Це міра середньої величини помилки для передбачень моделі.
* **Формула:**

де:

* кількість спостережень;
* фактичне значення;
* передбачене значення;
* Інтерпретація:
* Чим менше значення MSE, тим краща модель
* MSE враховує великі помилки більше, ніж маленькі, оскільки помилки зводяться до квадрату.

1. **Середня абсолютна помилка (МАЕ):**

* **Визначення:** Середня абсолютна помилка (MAE) є середнім значенням абсолютних різниць між фактичними значеннями та передбаченими значеннями. Це міра середньої величини абсолютної помилки для передбачень моделі.
* **Формула:**
* **Інтерпретація:**
* Чим менше значення MAE, тим краща модель.
* MAE є більш інтерпретованою, оскільки виражена в тих же одиницях, що і залежна змінна.

1. **Коефіцієнт детермінації (:**

* **Визначення:** Коефіцієнт детермінації R²показує, яка частка варіації залежної змінної пояснюється незалежними змінними моделі. Це міра того, наскільки добре модель пояснює варіацію в даних.

де:

* середнє значення залежної змінної;
* **Інтерпретація:**
* : Модель ідеально пояснює дисперсію залежної змінної;
* : Модель не пояснює дисперсію залежної змінної;
* Чим ближче значення до 1, тим краще модель пояснює дані;

**Хід роботи**

**Завдання**

1. **Завантаження готових наборів даних з Scikit-learn:** 
   * Завантажити набір sklearn.datasets.fetch\_california\_housing
2. **Поділ даних на тренувальну та тестову вибірки:** 
   * Поділити дані на тренувальний 70% та тестовий 30% набори.
3. **Створити наступні регресійні моделі** 
   * Поліноміальна регресія з ступенем поліному 2 та з регуляризацією: Ridge
   * Поліноміальна регресія з ступенем поліному 20 та з регуляризацією: Ridge
   * Поліноміальна регресія з ступенем поліному 2 та з регуляризацією: Lasso
   * Поліноміальна регресія з ступенем поліному 20 та з регуляризацією: Lasso
4. **Навчання та оцінка моделей:**
   * Для кожної створеної моделі виконати навчання виконати наступні операції

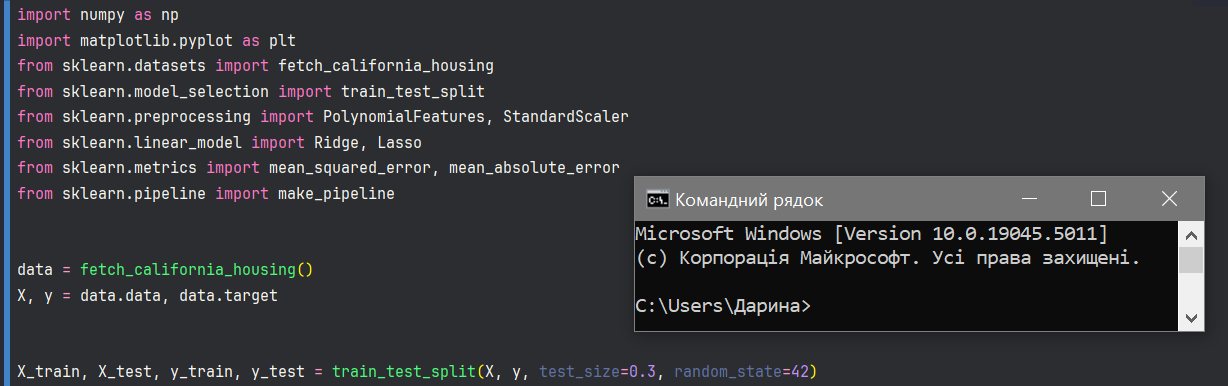
a. Знайти оптимальний параметр регуляризації, який забезпечує найменшу середньоквадратичну помилка (похибку)

b. Вивести коефіцієнти моделі c. Знайти оптимальний параметр регуляризації, який забезпечує найменшу середню абсолютну помилка (похибку)

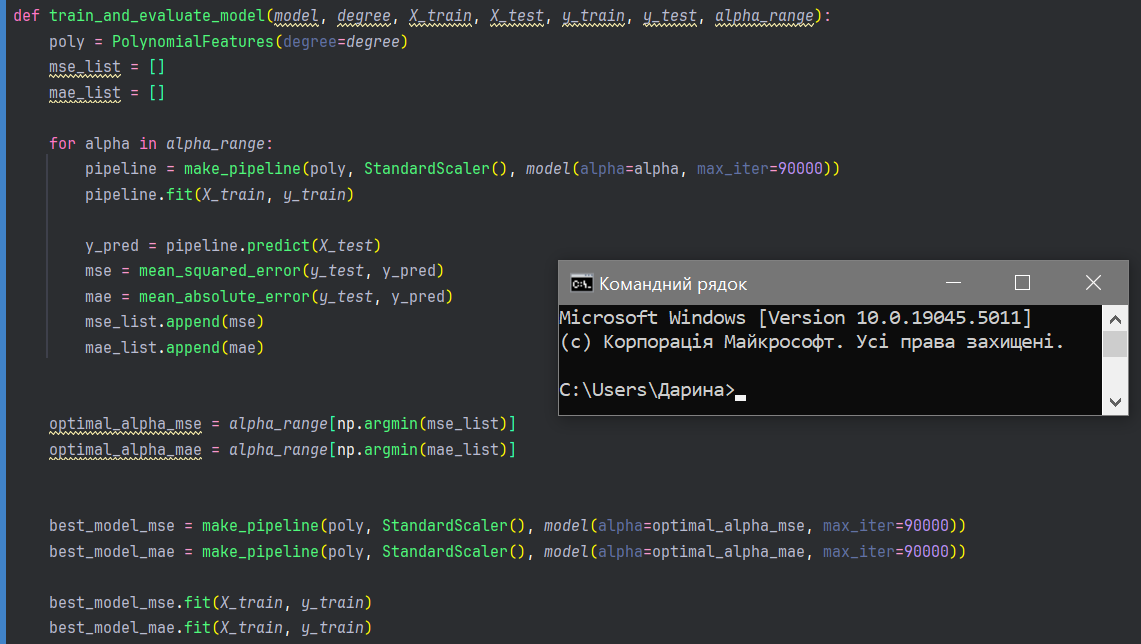
d. Вивести коефіцієнти моделі та порівняти з попередніми

* Оцінити продуктивності моделі на тестових даних.
* Результати оцінки похибок передбачення від коефіцієнта параметра регуляризації зобразити на графіках (середньоквадратична помилка та середня абсолютна помилка)

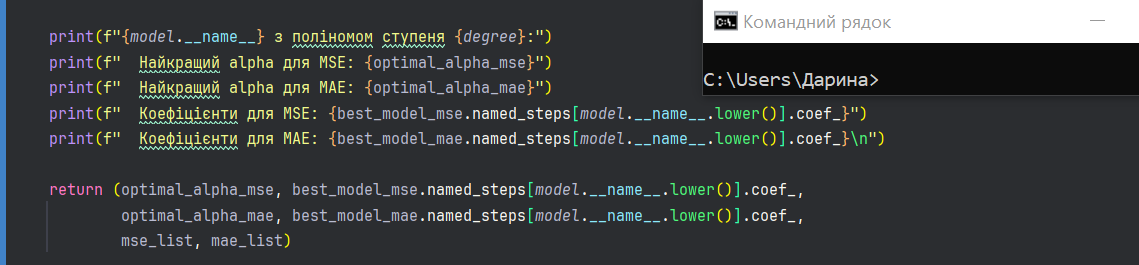
1. **Оформити звіт**

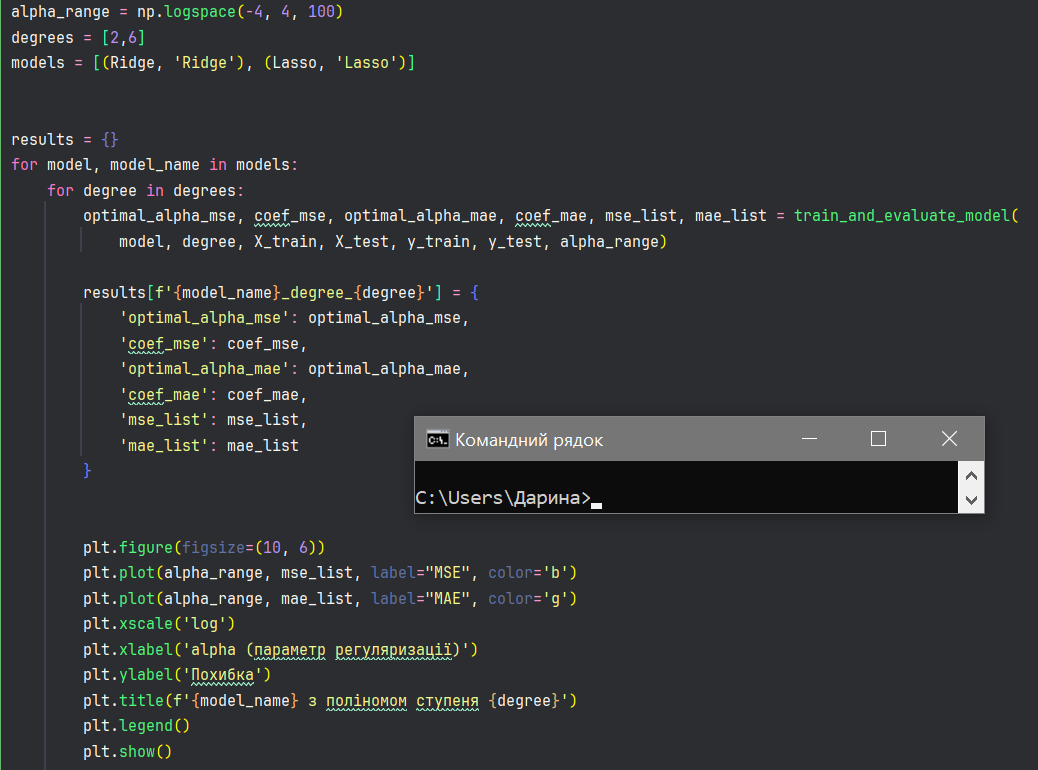
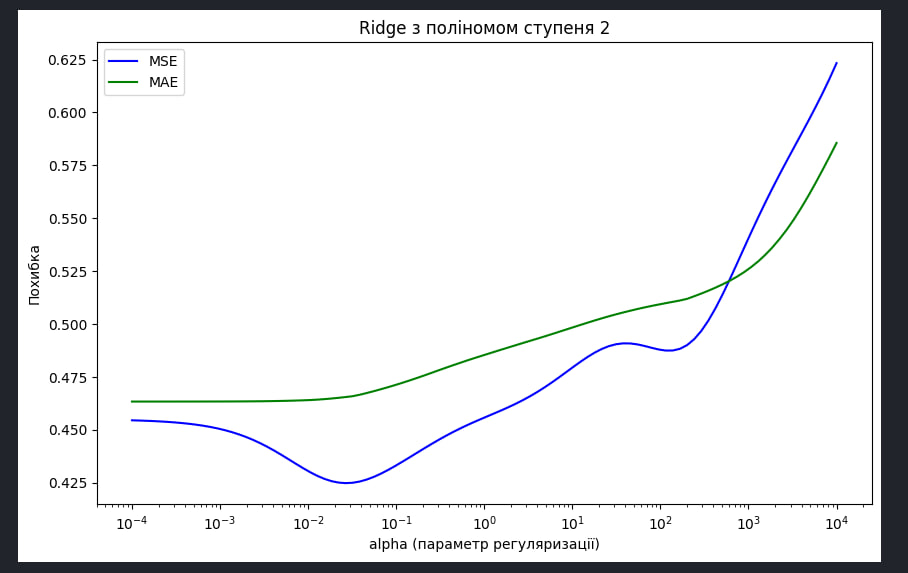
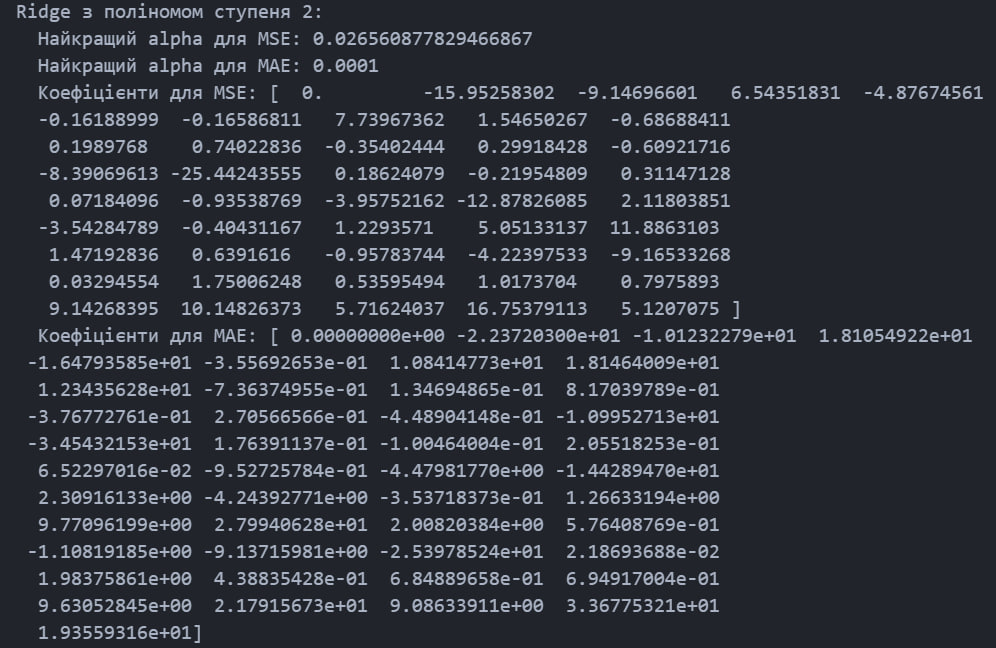
****

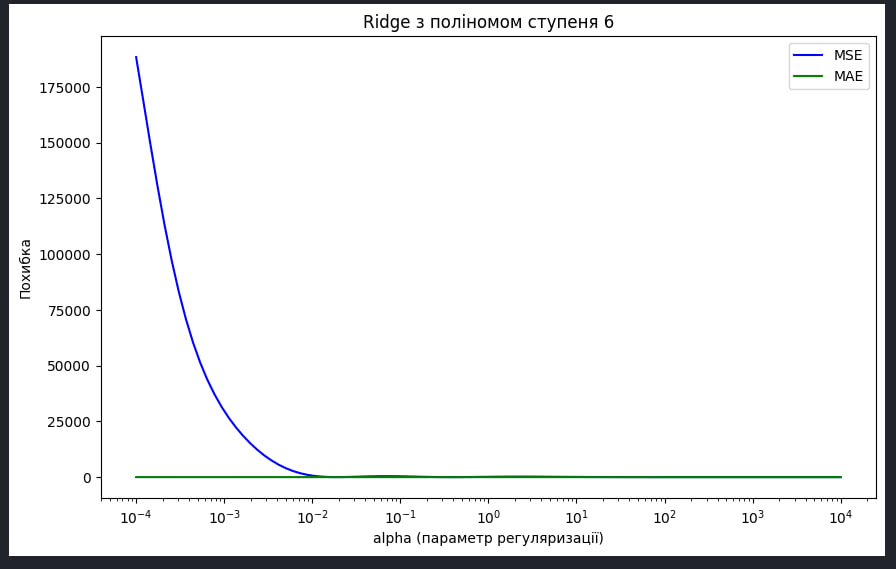
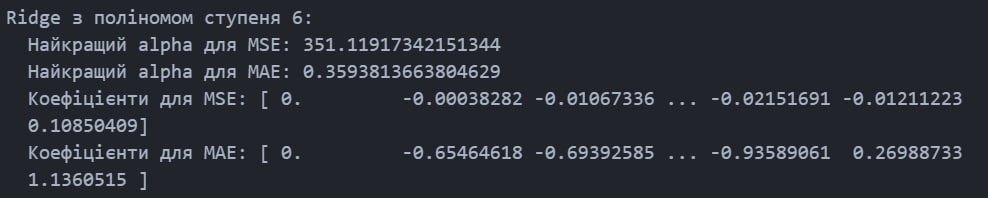
**Рис. 1** Завантажування бібліотек т поділ тестових даних

****

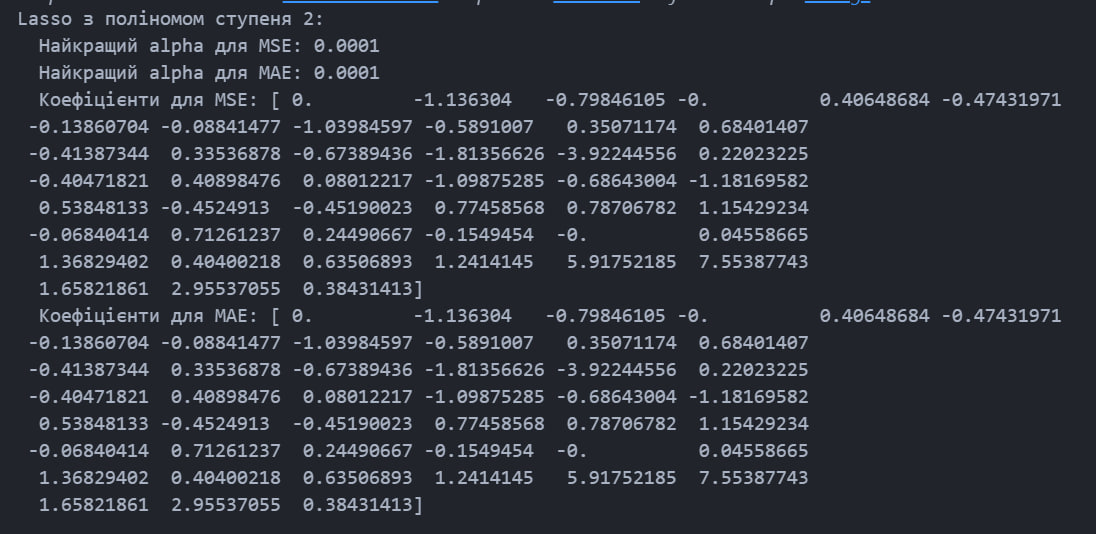
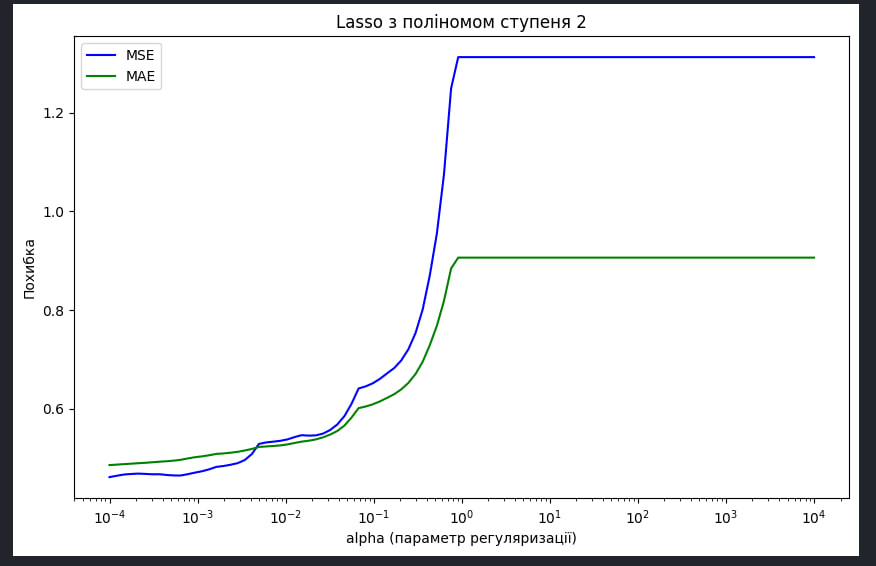
**Рис. 2** Функція для тренування та оцінки моделі підбирає значення альфа для MAE/MSE. Усі дані для оцінювання зберігаються в масивах, а наприкінці результати зберігаються у словнику result.

****

**Рис. 3** Продовження коду функції train\_and\_evaluate\_model  
  
  
**Рис. 4** Додавання даних оцінювання до словника `result` та вивід графіків для моделей, які використовують регуляризацію Ridge (L2) та Lasso (L1) з поліномами степеня 2 та 6.  


**Рис. 5-6** Оцінка моделі з поліномом ступеня 2 за допомогою регуляризації Ridge показує графік, що ілюструє залежність похибки (MSE/MAE) від значення alpha. Високі значення alpha можуть призвести до недостатньої адаптації моделі, тоді як низькі значення можуть викликати перенасичення даними, що, в свою чергу, збільшує похибку.

**Рис. 7-8** Оцінка моделі поліному 6 ступеня з використанням Ridge регуляризації

**Рис. 9-10** Оцінка моделі поліному 2 ступеня з використанням Lasso регуляризації

**Висновок:** У цій лабораторній роботі я навчилася працювати з методами регуляризації L1/L2 (Lasso та Ridge). На основі виконаних завдань та отриманих результатів я помітила, що значення alpha безпосередньо впливає на MSE/MAE при низьких та високих значеннях. Зокрема, низькі значення alpha призводять до перенасичення даними, тоді як високі значення викликають їх нестачу, що суттєво збільшує похибку.