УДК: 004.8:519.2

Бевзюк А. Ю., *здобувачка 2 курсу спеціальності 122 Комп'ютерні науки,*
науковий керівник:
Якубич К. О., *асистент кафедри інформаційних технологій*

АПРОКСИМАЦІЯ ФУНКЦІЙ У МАШИННОМУ НАВЧАННІ: ПОРІВНЯННЯ ЛІНІЙНОЇ ТА НЕЛІНІЙНОЇ РЕГРЕСІЇ ДЛЯ КЛАСИФІКАЦІЇ ДАНИХ

Донецький національний університет імені Василя Стуса, м. Вінниця

У сучасному світі обробка даних та машинне навчання стали ключовими в технологічних застосуваннях. Машинне навчання – галузь штучного інтелекту, дає змогу комп'ютерам навчатися на основі даних та виконувати завдання без явного програмування. Останні десятиліття воно стало ключовим складником сфери інформаційних технологій, впливаючи на різноманітні галузі, як-от фінанси, медицина, наука про дані та ін. [1]. Фундаментальним аспектом машинного навчання є апроксимація функцій, що полягає у створенні моделей, які наближено описують поведінку вхідних даних, зокрема зв'язок між незалежними змінними та цільовою змінною. У цій статті порівнюються два основні методи апроксимації функцій: лінійна та нелінійна регресія.

Лінійна регресія – це метод у машинному навчанні, який використовує лінійну функцію для апроксимації залежностей між вхідними та вихідними даними. Вона широко використовувався протягом багатьох років завдяки своїй простоті, інтерпретації та ефективності. Це цінний інструмент для розуміння зв'язків між змінними та прогнозування в різноманітних програмах. Математично це можна представити так:$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon,$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon,$$

де y – цільова змінна, $x_1,\ x_2, \ldots, x_n$ – вхідні ознаки, $\beta_1, \beta_2, \ldots, \beta_n$ – коефіцієнти регресії, ε – похибка [2].

Методи нелінійної регресії дають змогу більш гнучкий підхід, оскільки вони враховують нелінійні взаємозв'язки між змінними та цільовою змінною [3]. Порівняно з лінійною регресією, нелінійні моделі можуть краще апроксимувати складніші залежності в даних, роблячи їх більш ефективними у деяких випадках. Однак ця перевага може призвести до перенавчання, особливо якщо кількість параметрів у моделі велика. Один із поширених підходів – це використання поліноміальної регресії, де відносини між змінними та цільовою змінною моделюються у вигляді поліноміальної функції. Математично це можна виразити так:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \varepsilon,$$

де y – цільова змінна, $x, x^2, x^3, \ldots, x^n$ – це квадрат, куб та інші степені вхідних ознак, $\beta_1, \beta_2, \ldots, \beta_n$ – коефіцієнти регресії, ε – похибка.

Для порівняння лінійної та нелінійної регресії розглянемо завдання класифікації електронних листів на спам та не спам. Для цього ми використаємо рандом-

ний набір даних, що містить різні характеристики електронних листів та відповідні мітки класу.

Відповідно маємо такі дані:

- Набір даних складається з n прикладів, де кожен приклад характеризується певними властивостями електронного листа.
- Кожному прикладу поставлена у відповідність мітка класу, де 0 відповідає не спаму, а 1 – спаму.

За допомогою інтерактивного середовища для числових обчислень та програмування MATLAB, ми спробували реалізувати код для розв'язання цієї задачі (рис. 1).

```
% Генерація випадкових даних для електронних листів
n = 300; % кількість прикладів
x = linspace(0,10,n)'; % характеристики електронних листів
y = 0.5*x + 1 + randn(n,1); % мітки класу (0 - не спам, 1 - спам)

% Лінійна регресія
X_linear = [ones(n,1), x];
beta_linear = X_linear\y;
y_linear = X_linear*beta_linear;

% Нелінійна регресія (метод опорних векторів з ядром RBF)
svm_model = fitrsvm(x,y,'KernelFunction','rbf'); % навчання моделі
y_svm = predict(svm_model, x); % прогнозовані значення

% Розрахунок середньоквадратичної помилки для лінійної регресії
mse_linear = mean((y - y_linear).^2);
% Розрахунок середньоквадратичної помилки для нелінійної регресії
mse_svm = mean((y - y_svm).^2);

% Вивід результатів
fprintf('Середньоквадратична помилка для лінійної регресії: %.4f\n', mse_linear);
fprintf('Середньоквадратична помилка для нелінійної регресії: %.4f\n', mse_svm);

% Графіки
figure;
scatter(x,y, 'filled');
hold on;
plot(x,y_linear,'r','LineWidth',2); % лінійна регресія
plot(x,y_svm,'g','LineWidth',2); % нелінійна регресія
xlabel('Характеристика');
ylabel('Мітка класу');
legend('Дані','Лінійна регресія','Нелінійна регресія');
title('Класифікація електронних листів: Лінійна vs Нелінійна регресія');
```

Рис. 1. Реалізація задачі

Даний код створює різні моделі для класифікації електронних листів на спам та не спам, а також оцінює їх ефективність. Спочатку генерується набір випадкових даних, які відображають характеристики електронних листів. Під час цього враховується, що деякі електронні листи можуть бути класифіковані як спам, тоді як інші – як не спам. Далі використовується два різні підходи до класифікації, спочатку лінійна регресія, яка апроксимує залежність між характеристиками електронного листа та його класом. Далі втілений нелінійний підхід, використовуючи

метод опорних векторів з ядром RBF, щоб змоделювати більш складні залежності між характеристиками та класами [2].

Після навчання обох моделей ми оцінюємо їх ефективність, обчислюючи середньоквадратичну помилку для кожної. Це дає змогу нам порівняти, наскільки добре кожна модель передбачає класи електронних листів.

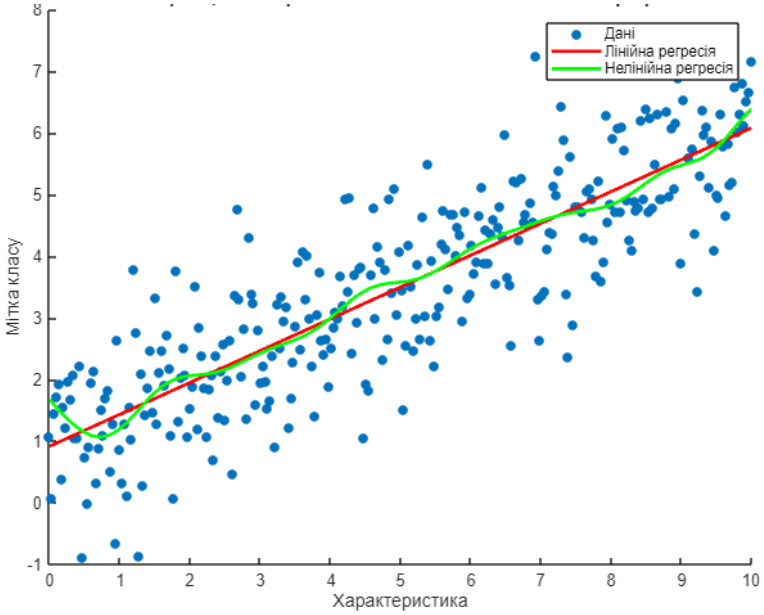

Рис. 2. Графік, який відображає лінійну та нелінійну регресію

На рис. 1 ми можемо бачити результат роботи програми, а саме побудований графік, який візуалізує як вихідні дані, так і прогнозовані значення для кожної моделі. Це дає нам змогу краще зрозуміти, як кожен підхід впорався з класифікацією електронних листів.

```
Середньоквадратична помилка для лінійної регресії: 0.8487
Середньоквадратична помилка для нелінійної регресії: 0.8411
```

Рис. 3. Результати, виведені в консоль

Також у консоль виводяться результати, які описують середньоквадратичну помилку для лінійної та нелінійної регресії. Це допомагає нам порівняти їх точність та ефективність у вирішенні задачі класифікації електронних листів на спам та не спам. Загалом отримані нами результати свідчать про те, що обидві моделі можуть бути використані для класифікації електронних листів на спам та не спам, проте нелінійна регресія може бути трохи ефективнішою в цьому випадку.

Отже, у машинному навчанні вибір між лінійною та нелінійною регресією залежить від складності та природи даних, а також від конкретної задачі, що вирішується. Хоча лінійна регресія проста і легко інтерпретується, вона може бути неефективною для моделювання нелінійних зв’язків. Нелінійні методи регресії, навпаки, забезпечують більш гнучкий підхід і можуть точніше відображати складні зв’язки в даних.

Список використаних джерел

1. Smola A., Vishwanathan S. V. N. Introduction to Machine Learning: Cambridge University Press, 2008. 234 p. URL: https://alex.smola.org/drafts/thebook.pdf (дата звернення: 17.04.2024).

2. Кононова К. Ю. Машинне навчання: методи та моделі: підручник для бакалаврів, магістрів та докторів філософії спеціальності 051 «Економіка». Харків: ХНУ імені В. Н. Каразіна, 2020. 301 с. URL: https://comsys.kpi.ua/upload/Rainforcement%20Learning%20.pdf (дата звернення: 17.04.2024).

3. What Is Nonlinear Regression? *Nonlinear Regression.* URL: https://ch.mathworks.com/discovery/nonlinear-regression.html (дата звернення: 17.04.2024).

4. Метод інтерполяції для прогнозування метрик використання хмарних обчислень в статистичному навчанні / Н. А. Потапова, Л. О. Волонтир, І. П. Частоколенко, М. С. Григоренко. *Наука і техніка сьогодні.* 2024. № 4(32). С. 1192–1205.

УДК 519.6:004.42:004.043

Козачок А. О., *здобувачка 2 курсу спеціальності 122 Комп'ютерні науки,*
науковий керівник:
Хмелівський Ю. С., *асистент кафедри інформаційних технологій*

РОЛЬ МЕТОДІВ ОБЧИСЛЕНЬ У ВИРІШЕННІ ЗАДАЧ АНАЛІЗУ ДАНИХ

Донецький національний університет імені Василя Стуса, м. Вінниця

У сучасному світі обробка та аналіз даних є важливим кроком у прийнятті рішень у різних сферах діяльності, зокрема в бізнесі, науці, медицині та соціальних науках. Масштаби зібраної інформації стрімко зростають і потребують відповідних методів обробки та аналізу. Роль обчислювальних методів у вирішенні завдань аналізу даних полягає в їх здатності ефективно обробляти великі обсяги інформації та забезпечувати точність, швидкість і достовірність отриманих результатів [1].

Важливим аспектом обчислювальних методів у вирішенні задач аналізу даних є їх внесок у створення та розвиток алгоритмів обробки даних. Обчислювальні методи, як-от машинне навчання, статистичний аналіз і обробка сигналів, можуть допомогти створити алгоритми, здатні розпізнавати образи і виявлення кореляцій у великих наборах даних. Наприклад, алгоритми класифікації даних, як-от опорні векторні машини та нейронні мережі, можуть відрізнити дані від одного класу до іншого [2].

Другий аспект методів обчислень полягає в їх здатності працювати з великими обсягами даних. Традиційні методи аналізу даних, як-от ручне кодування чи статистичні методи, можуть бути неефективними під час обробки великих обсягів даних через обмежену швидкість та ресурси. Методи обчислень, які базуються на паралельних обчисленнях або використанні великих обчислювальних кластерів, можуть значно прискорити аналіз даних, забезпечуючи швидкий доступ до інформації та високу продуктивність обробки.

На прикладі (рис. 1) показано, як простий метод аналізу даних може значно полегшити життя в сучасному світі, допомагаючи вирішувати навіть складні завдання. Приміром, шляхом аналізу великого обсягу медичних даних можна розробити прості моделі для прогнозу можливого ризику серцево-судинного захворювання у пацієнтів. Це дає змогу вчасно виявляти особливості, що можуть вка-