

ВИКОРИСТАННЯ BOOTSTRAP-ТЕХНОЛОГІЇ ДЛЯ КЛАСИФІКАЦІЇ ДАНИХ МЕТОДАМИ ЛІНІЙНОЇ ТА ЛОГІСТИЧНОЇ РЕГРЕСІЇ

Дарина Довголюк, Тетяна Телелеха, Марія Бринчук

USING BOOTSTRAP TECHNIQUES FOR DATA CLASSIFICATION WITH LINEAR
AND LOGISTIC REGRESSION

Daryna Dovholiuk, Tetiana Telelekha, Mariia Brynchuk

Волинський національний університет імені Лесі Українки, просп. Волі, 13, Луцьк,
43025, Україна

Abstract. *The paper focuses on exploring bootstrap techniques in data analysis through the application of linear and logistic regression methods, with implementation in R.*

Лінійна та логістична регресії є традиційними методами прогнозування та класифікації в емпіричних дослідженнях, включаючи сучасні методи штучного інтелекту. На основі емпіричних даних обчислюють параметри та будують моделі, які застосовують для прогнозування майбутніх значень, або для класифікації нових випадків. В умовах, коли модель ще не достатньо випробувана на нових даних, а потреба застосування є нагальною, застосовують bootstrap-технології для дослідження якості моделі.

Лінійна регресія. У статистичному аналізі часто виникає потреба в описі зв'язку між двома змінними, позначеними як X та Y , які спостерігаються для n об'єктів. Для кожного об'єкта маємо відповідні значення X_j та Y_j . Щоб змодельовати взаємозв'язок між ними, використовують наближену лінійну залежність виду:

$$Y_j \approx b_0 + b_1 X_j$$

де, b_0 і b_1 — невідомі коефіцієнти, які потрібно оцінити за даними X_j та Y_j , для $j = 1, \dots, n$.

Такий підхід формує основу моделі простої лінійної регресії, яку широко застосовують для прогнозування значень, виявлення трендів у даних, а також перевірки статистичних гіпотез щодо залежності між змінними.

Сьогодні лінійна регресія широко застосовується в поведінкових біологічних, і соціальних науках, в сфері економіки та фінансів для виявлення можливих зв'язків між змінними. Крім того, вона стала фундаментальним інструментом у сфері штучного інтелекту та машинного навчання.

У програмному середовищі RGui лінійна регресія реалізовується за допомогою функції `lm()`. Для аналізу результатів та здійснення прогнозів використовують функції `summary()` і `predict()` відповідно.

Логістична регресія. Логістична регресія – це один з найефективніших методів аналізу, який переважно використовується для бінарної класифікації, тобто передбачення до якого з двох можливих класів належить певне спостереження (наприклад: «так/ні», «хворий/здоровий», «успіх/невдача»). Такий підхід дозволяє відповісти на запитання: «Чи має пацієнт захворювання?», «Чи є випадок тяжким?», «Чи отримає абітурієнт стипендію?» тощо.

На відміну від лінійної регресії, яка прогнозує безперервні числові значення, логістична регресія моделює ймовірність того, що подія настане. Для цього вона використовує спеціальну математичну функцію — логістичну (або сигмоподібну) функцію, яка приймає будь-яке дійсне число і перетворює його на значення, близьке до 1 при великих значеннях вхідної змінної, або до 0 — при малих. Це дає змогу пояснити результат як ймовірність належності спостереження до позитивного класу.

Математично сигмоподібна функція має вигляд:

$$y = \frac{1}{1 + e^{-x}}$$

де, y — ймовірність належності до позитивного класу, а $x = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$ — рівняння лінійної регресії з коефіцієнтами β .

Щоб прийняти остаточне рішення про належність до певного класу, у логістичній регресії встановлюють поріг класифікації — критичне значення ймовірності, після якого спостереження відносять до позитивного або негативного класу. Типовим є поріг 0,5 (тобто 50%), проте в прикладних задачах це значення може змінюватись відповідно до контексту.

Правильний вибір порогу має вирішальне значення особливо в сферах, де вартість помилки є високою, як-от медицина. Зміна цього значення безпосередньо впливає на частоту хибнопозитивних (false positives) і хибнонегативних результатів (false negatives).

Логістична регресія широко застосовується в машинному навчанні: для виявлення спаму, прогнозування відтоку клієнтів, передбачення ймовірності натискання на рекламу тощо. Її головною перевагою є відносна простота навчання порівняно з іншими моделями машинного навчання та штучного інтелекту.

У середовищі RGui логістичну регресію можна реалізувати за допомогою функції `glm()`, а для перегляду результатів використовуються `summary()` та `predict()`.

Bootstrap. Bootstrap — це статистичний метод, який дозволяє оцінювати варіативність і надійність статистичних оцінок шляхом випадкового вибіркового повторення з наявного набору даних скінченну кількість разів. Цей підхід особливо корисний, коли неможливо або складно зробити припущення про розподіл генеральної сукупності.

Назва «bootstrap» походить від англійського виразу “pulling oneself up by one’s bootstraps” («витягуватися себе за допомогою власних тяг»), що базується на історії XVIII століття про барона Мюнхгаузена, який, за легендою, врятував себе від утоплення, витягнувшись за допомогою власного взуття (тяг) (Efron & Tibshirani, 1994). Дійсно, створення численних вибірок на основі однієї може видаватися малоімовірним або навіть штучним підходом. Втім, метод bootstrap є корисним інструментом для оцінки статистичної невизначеності, зокрема в умовах складної або невідомої структури даних.

Переваги bootstrap проявляються також у тому, що ми можемо оцінити якість отриманої моделі (або статистичної оцінки) без припущень про базовий розподіл. Маючи цю інформацію, ми можемо оцінювати параметри вибірки, будувати довірчі інтервали та перевіряти гіпотези щодо будь-якого статистичного функціоналу, який нам потрібен.

Процедура bootstrap передбачає багаторазове (100 - 500 разів) формування нових вибірок з поверненням із початкових даних, побудову моделі на кожній із них та наступний аналіз розподілу оцінених параметрів з наступною перевіркою на наявних даних. Це дозволяє визначити довірчі інтервали для коефіцієнтів, оцінити дисперсії та похибки.

Проте, варто зазначити, що пряма оцінка помилки прогнозу за bootstrap може бути упередженою, оскільки bootstrap-вибірка містить ті самі спостереження, що й початкова вибірка, а тому результат може бути оптимістичним. Для усунення цього ефекту використовують bootstrap із залишенням декількох спостережень оцінка проводиться лише на тих прикладах, які не потрапили в конкретну bootstrap - вибірку.

Виконання практичного завдання з побудовою лінійної та логістичної регресії продемонструвало можливості програми R для реалізації як регресійних алгоритмів, так і bootstrap технології.

Бібліографія

1. Майборода Р.Є. Методичні рекомендації по курсу: Регресійний аналіз та асимптотична статистика. Київ: КНУ, 2020. – 40 с.
2. Павлів Н. Р., Левицький О. В. Методи машинного навчання для задач класифікації у медицині: Вісник Національного університету «Львівська політехніка»: Комп'ютерні науки та інформаційні технології. Львів: ЛНУ, 2023. – 220–236 с.
3. R Core Team. An Introduction to R.– Vienna: R Foundation for Statistical Computing, 2023. – URL: <https://cran.r-project.org/doc/manuals/R-intro.pdf> (Дата звернення: 18 трав. 2025.)
4. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: with Applications in R. – 2nd ed. – New York: Springer, 2021. – 431 p.
5. Bootstrapping in Statistics with R. URL: <https://www.datacamp.com/tutorial/bootstrapping>. (Дата звернення: 18 трав. 2025.)
6. Solomon D. Bootstrapping: The Basics. URL: <https://drew-solomon.medium.com/bootstrapping-the-basics-4dbd7ca965f1> (Дата звернення: 18 трав. 2025.)
7. Saito T., Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. – 2015. – Vol. 10, № 3. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8137228> (Дата звернення: 18 трав. 2025.)