УДК: 621.396.96 DOI: 10.18372/2073-4751.71.17001

Писарчук О.О., д.т.н.,
orcid.org/0000-0001-5271-0248,
Корочкін О.В., к.т.н.,
orcid.org/0000-0002-6569-5849,
Баран Д.Р.,
orcid.org/0000-0002-3251-8897

ВИЗНАЧЕННЯ ПОРЯДКУ ПОЛІНОМІАЛЬНОЇ МОДЕЛІ ДЛЯ ПОБУДОВИ ЛІНІЇ ТРЕНДУ В ЗАДАЧАХ DATA SCIENCE

Київський політехнічний інститут ім. Ігоря Сікорського

avcora@gmail.com

Вступ

На теперішній час значного поширення та практичного застосування набули технології Data Science [1]. В практичному відображенні це, найчастіше, програмна реалізація Back-End компоненти розподілених CRM та ERP програмних систем з властивостями інтелектуальності. Вхідною інформацією для таких систем може бути Big Data масиви із властивостями числових статистичних рядів – статистичні вибірки. Прикладами галузей практичного застосування описаних технології є: оцінювання та прогнозування розвитку економічних показників трейдингових компаній; автоматизовані асистенти водія в автомобільній галузі – сфера Automotive; автоматичні системи керування безекіпажними (безпілотними) транспортними засобами – дронами (UAV); реалізація процесів векторізації цифрових зображень в задачах ідентифікації для технологій Computer Vision [1].

Якість реалізації зазначених аспектів значною мірою визначається точністю розрахунку параметрів трендових залежностей, що, в свою чергу потребує адекватного визначення порядку поліноміальної моделі.

Тому актуальною є задача розвитку методів визначення порядку поліноміальної моделі для побудови лінії тренду в задачах data science.

Практика показує, що апріорно побудувати адекватну математичну модель досліджуваного процесу у вигляді кінцевого аналітичного виразу вдається далеко не завжди. Значною мірою апріорне формування математичних моделей ускладнюється нелінійним характером зміни досліджуваних процесів. У випадках, коли з високим ступенем адекватності не вдається заздалегідь сформувати модель досліджуваного процесу або його адекватний опис призводить до значного ускладнення алгоритмів згладжування, використовують поліноміальну модель [2].

Під час опису процесів і систем поліноміальною моделлю важливим є вибір ступеня полінома, що описує динаміку досліджуваного процесу. Точність поліноміального згладжування визначається двома складовими похибок:

- методичною похибкою, яка зумовлена неадекватним описом досліджуваного процесу поліномом обраного порядку;
- випадковою похибкою, яка зумовлена похибками виміру параметрів досліджуваного процесу.

Зазначені класи похибок знаходяться у взаємному протиріччі. Так, для фіксованого інтервалу спостереження зниження, наприклад, випадкової складової шляхом зменшення порядку апроксимуючого полінома призводить до збільшення методичної похибки. Тому остаточному призначенню порядку апроксимуючого полінома повинен передувати ретельний аналіз вихідних даних для того, щоб забезпечити мінімум суми методичної і випадкової похибок.

Також розглядається один з можливих підходів до оптимального вибору порядку апроксимуючого полінома, який оснований на одночасному аналізі трьох різних критеріїв, що полягають у прийнятті, тим або іншим способом, рішення про включення кожного з ортогональних поліномів до складу апроксимуючого багаточлена.

До недоліків підходу варто віднести таке:

1. Значні обчислювальні витрати під час реалізації підходу, зумовлені необхідністю отримання набору поліномів різного ступеня для перевірки гіпотез про порядок апроксимуючої кривої.

2. Поява зміщеності в кінцевих результатах згладжування, зумовленої перевагою критерію мінімуму дисперсії згладжування при виборі оптимального порядку апроксимуючого полінома.

Зазначені недоліки не дозволяють використовувати традиційні підходи в для побудови лінії тренду в задачах для задач data science.

Мета

Таким чином метою роботи є розроблення підходу до визначення порядку поліноміальної моделі, який знижує недоліки відомих критеріїв.

Вважатимемо, що оптимальне за мінімумом методичної і випадкової похибки згладжування значення порядку апроксимуючого полінома збігається з реальним ступенем кривизни експериментальної кривої. Тоді розв'язання задачі вибору оптимального значення порядку поліноміальної моделі зводиться до визначення ступеня кривизни виміряної вибірки.

Основна частина

Постановка задачі

Нехай результатом є виміряна дискретна за часом рівноточна і рівнодискретна вибірка:

$$y_1, y_2, y_3, \ldots, y_m \qquad (1)$$

Потрібно за вибіркою (1) визначити порядок експериментальної кривої k, що дозволить оптимальним чином сформувати поліноміальну модель у вигляді (2):

$$F(t) = C_0 + C_1 t + C_2 t^2 + \ldots + C_n t^n = \sum_{i=0}^{n_{opt}} C_i t^i$$

де C_i – коефіцієнти апроксимуючого полінома, $n_{opt} = k$ – оптимальний (дорівнює порядку експериментальної кривої) ступінь полінома.

Розв'язання

Ускладнення характеру досліджуваного процесу приводить до появи додаткових ненульових прискорень (похідних вищих порядків) . Ускладнення в моделі вигляду (2) враховується шляхом додавання додаткових коефіцієнтів C_i, суть яких – похідні i–го порядку (C_0 – початкове значення координати, C_1 – швидкість зміни координати, C_2 – прискорення і т. ін.) [1].

Можна стверджувати, що порядок поліноміальної моделі (2) знаходиться в прямій залежності від порядку останньої ненульової (в умовах відсутності похибок вимірів) похідної експериментальної вибірки (1). Таким чином, базовою ознакою для визначення порядку експериментальної моделі в умовах відсутності похибок виміру координат є порядок останньої її ненульової похідної.

У загальному вигляді обчислення похідної за дискретною послідовністю здійснюється згідно з виразом (3):

$$y_j^{(p)} = \frac{y_{j+1}^{(p-1)} - y_j^{(p-1)}}{\Delta t}, \quad j = \overline{1 \ldots m}, \quad p = \overline{1 \ldots n}$$

де $y_j^{(p)}, y_{j+1}^{(p-1)}, y_j^{(p-1)}$ – p–а і $(p-1)$–а похідні в моменти часу j та $j+1$ відповідно; Δt – інтервал отримання вимірів вибірки (1).

Відповідно до базового виразу (3) кінцеві рівняння для пошуку похідних за дискретними даними до п'ятого порядку включно набудуть вигляду:

$$\begin{cases} y_j^{(1)} = \dfrac{y_{j+1} - y_j}{\Delta t}, \\ y_j^{(2)} = \dfrac{y_{j+2} - 2y_{j+1} + y_j}{\Delta t^2}, \\ y_j^{(3)} = \dfrac{y_{j+3} - 3y_{j+2} + 3y_{j+1} - y_j}{\Delta t^3}, \\ y_j^{(4)} = \dfrac{y_{j+4} - 4y_{j+3} + 6y_{j+2} - 4y_{j+1} + y_j}{\Delta t^4}, \\ y_j^{(5)} = \dfrac{y_{j+5} - 5y_{j+4} + 10y_{j+3} - 10y_{j+2} + 5y_{j+1} - y_j}{\Delta t^5}. \end{cases} \qquad (4)$$

Раніше було зазначено, що послідовність (1) отримана експериментальним шляхом, що зумовлює наявність у координатах y_j, $(j = \overline{1 \dots m})$ випадкової похибки вимірів. Тоді модель виміру координати можна подати у вигляді:

$$y_j = y_{j\partial} + \xi, \qquad (5)$$

де $y_{j\partial}$ – дійсне значення координати, яке, як правило, невідоме; ξ – випадкова похибка виміру, що підпорядкована нормальному закону розподілу з відомими числовими характеристиками: математичним сподіванням $m_\xi = 0$ і дисперсією σ_ξ^2 (квадрат средньоквадратичного відхилення (СКВ) похибки), яка характеризує потенційні точності вимірювача. Виходячи із зазначеного, модель p–ї похідної за координатою може бути записана в такий спосіб:

$$y_j^{(p)} = y_{j\partial}^{(p)} + \xi^{(p)}, \qquad (6)$$

де $y_{j\partial}^{(p)}$ – дійсне значення p–ї похідної за координатою; $\xi^{(p)}$ – випадкова похибка розрахунку p–ї похідної з нормальним законом розподілу, нульовим середнім $(m_{\xi^{(p)}} = 0)$ і дисперсією $\sigma_{\xi^{(p)}}^2$, яка характеризує похибку перетворень (3).

У випадку рівності нулю дійсного значення p–ї похідної (що раніше було прийнято за базову ознаку порядку виміряної послідовності) модель (6) трансформується до вигляду:

$$y_j^{(p)} = \xi^{(p)}. \qquad (7)$$

Вираз (7), з одного боку, демонструє не показовість ознаки рівності нулю вищих похідних для визначення порядку виміряної кривої, що зумовлено наявністю похибок виміру, а, з іншого боку, визначає необхідність і зміст нової ознаки.

Загалом компоненти вибірки (1) варто розглядати як рівнодискретні та рівноточні значення (реалізації) деякої випадкової функції, що характеризує рух об'єкта спостереження. Під час роботи з випадковими величинами прийнято оперувати їх вірогіднісними характеристиками (математичне сподівання, дисперсія і т. ін.), які є більш показовими для аналізу, ніж окремі реалізації (виміри). У зв'язку з цим слід визначити дисперсію перетворення (3) виміряної випадкової вибірки (1). У цьому разі похибку перетворень можна визначити двома способами: теоретично (використовуючи граничні теореми теорії ймовірності); експериментально (використовуючи вирази для визначення числових характеристик випадкових величин за експериментальними даними).

Використовуючи теорему про дисперсію суми (різниці) для рівнодискретних та рівноточних випадкових величин, припускаючи що кореляційна залежність між сусідніми вимірами дорівнює нулю, дисперсія перетворень (3) визначатиметься за виразом:

$$\sigma^2_{y^{(3)}T} = \frac{2\sigma_y^2}{\Delta t^{2(p)}}, \qquad (8)$$

де σ_y^2 – дисперсія похибок виміру координат вибірки (1), яка, як правило, відома і характеризує потенційні точності вимірника.

Експериментальне визначення числових характеристик випадкових величин основано на обробці серії реалізацій. Для розглянутої задачі вважатимемо, що вирази (3) і відповідно формули (4) послідовно застосовуються до кожного компонента вибірки (1). У результаті цього отримаємо підвибірки:

$$y_j^{(1)}, (j = \overline{1 \dots m-1}),\ y_j^{(2)}, (j = \overline{1 \dots m-2}), \dots, y_j^{(p)}, (j = \overline{1 \dots m-p}), \qquad (9)$$

Підвибірки (10) також являють собою рівнодискретні та рівноточні реалізації випадкових функцій, які характеризують зміну в часі похідних координат. Оскільки вибірка, що характеризується вибіркою (1), являє собою нелінійний процес (нестаціонарну випадкову функцію), отже, і вибірки (9) також характеризуватимуть нестаціонарні випадкові процеси, однак їхня не лінійність буде зменшуватиметься із зростанням порядку похідної. Таким чином, можна стверджувати, що для будь-якої виміряної нелінійної вибірки знайдеться така похідна, яка буде стаціонарним випадковим процесом. Очевидно, що стаціонарність функції, яка описує зміну p–ї похідної, буде матиме місце тільки тоді, коли дійсна складова моделі (6) буде

постійною величиною. Дана обставина відображає ознаку, за якою можна визначити порядок експериментальної кривої. Тому для визначення експериментальної похибки визначення похідної за координатою використовуватимемо вирази для аналізу стаціонарних випадкових дискретних функцій.

Експериментальне визначення дисперсії перетворень (3) реалізується згідно з виразом:

$$\sigma^2_{y^{(p)}E} = \frac{1}{(m-p)-1}\sum_{j=1}^{m-p}(y_j^{(p)} - m_{y^{(p)}}), \quad (10)$$

де $y_j^{(p)}$ – компоненти вибірок p–ї похідної (див. (9)); $m_{y^{(p)}}$ – математичне сподівання для p–ї похідної вибірок (9), визначене відповідно до виразу.

Вираз (10) дасть правильний результат тільки тоді, коли аналізована вибірка з набору (9) характеризуватиме стаціонарний випадковий процес.

$$m_{y^{(p)}} = \frac{1}{m-p}\sum_{j=1}^{m-p} y_j^{(p)} .$$

В інших випадках спостерігатиметься зміщеність оцінок, що розраховуються. Відстежити наявність зміщеності в оцінках експериментально розрахованої дисперсії перетворення (3) можна реалізувати шляхом порівняння її значення з теоретично розрахованою дисперсією. Слід зазначити, що між теоретично (8) і експериментально (10) розрахованими значеннями дисперсій стаціонарної випадкової функції завжди буде деяка відмінність, зумовлена обмеженістю обсягу випадкової вибірки, яка використовується у виразі (10). Проте ця різниця буде набагато меншою, ніж у випадку коли аналізована функція буде нестаціонарною, а оцінка дисперсії, отримана відповідно до (10), – зміщеною.

Таким чином, порядок експериментальної моделі можна визначити за порядком p -ї похідної, для якої абсолютна різниця між теоретичним $\sigma_{y^{(p)}T}$ і експериментальним $\sigma_{y^{(p)}E}$ значенням СКВ похибки визначення p–ї похідної $\Delta^{(p)}$ буде мінімальною.

$$\Delta^{(p)} = \left|\sigma_{y^{(p)}E} - \sigma_{y^{(p)}T}\right| \quad (11)$$

Відповідно до викладеного оптимальне значення порядку поліноміальної моделі визначається умовою:

$$\begin{cases} n_{opt} = k = p, \ при \\ \Delta^{(p)} = \left|\sigma^2_{y^{(p)}E} - \sigma^2_{y^{(p)}T}\right| = \min \ із \ \Delta^{(p)}, \ p = \overline{1\ldots m} \end{cases} \quad (12)$$

Алгоритм визначення порядку поліноміальної моделі включає такі етапи:

1. За компонентами вибірки (1) визначити підвибірки (9) згідно з виразами (4).

2. Визначити теоретичне та експериментальне значення дисперсій розрахунку p -х похідних, використовуючи вирази (8) і (10) відповідно.

3. Розрахувати різниці (11) для кожної пари СКВ похибок визначення p -х похідних.

4. Прийняти рішення про оптимальний порядок поліноміальної моделі відповідно до умови (12).

Слід зазначити, що сформований у такий спосіб алгоритм досить простий у реалізації і вимагає значно менше обчислювальних витрат у порівнянні з відомими підходами, коли для визначення порядку експериментальної кривої слід одержати декілька згладжуючих поліномів і проаналізувати низку критеріїв для вироблення остаточного рішення.

Оцінювання якості функціонування й ефективності використання розробленого *алгоритму визначення порядку поліноміальної моделі* проводилась методом імітаційного моделювання. Під час дослідження покладалося, що обробці підлягала вибірка в 21 вимір координати y з темпом оновлення інформації $\Delta t = 2сек$ і СКВ похибки вимірювання $\sigma_y = 0.1y.од$. Аналіз розробленого алгоритму здійснювався для кривих до п'ятого порядку включно. Результати досліджень подані в табл. 1, 2.

Дані табл.1 демонструють процес визначення оптимального значення порядку поліноміальної моделі за експериментальними даними при використанні умови (12).

Таблиця 1.

Параметр	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	n_{opt}
$k=1$	0.024	0.027	0.033	0.037	0.039	1
$k=2$	23.62	0.027	0.033	0.037	0.039	2
$k=3$	$16*10^2$	67.47	0.033	0.037	0.039	3
$k=4$	$86*10^3$	$60*10^2$	25*10	0.037	0.039	4
$k=5$	$42*10^5$	$40*10^4$	$28*10^3$	$12*10^2$	0.039	5

У табл.1 подані значення параметра $\Delta^{(p)}$ для похідних до п'ятої включно ($p=\overline{1\ldots5}$) при зміні порядку виміряної кривої ($k=\overline{1\ldots5}$). Останній стовпець табл.1 характеризує прийняте рішення про оптимальне значення порядку згладжуючого полінома n_{opt}.

Аналіз поданих у табл.1 даних дозволяє зробити висновок, що умова (12) є працезданою, а розроблений алгоритм реалізує формування оптимального (рівного порядку експериментальної кривої) значення поліноміальної моделі за наявності похибок вимірювання.

Для оцінювання ефективності використання розробленого алгоритму визначення порядку поліноміальної моделі під час розв'язання кінцевої задачі – побудова високоточної аналітичної моделі – були проведені дослідження, результати яких подані в табл.2.

Таблиця 2.

k	Без оптимального призначення порядку полінома		З оптимальним призначенням порядку полінома	
	Середина	Краї	Середина	Краї
1	0.0011	0.0033	0.0010	0.0032
2	0.0012	0.0034	0.0012	0.0034
3	0.0013	$9*10^3$	0.0020	0.0053
4	$7*10^4$	$13*10^5$	0.0038	0.0140
5	$12*10^6$	$13*10^7$	0.0054	0.1200

Дані табл. 2 характеризують якість спільного використання запропонованого алгоритму з алгоритмом згладжування на базі методу найменших квадратів (МНК) при зміні характеру експериментальної виьбірки до кривої п'ятого порядку ($k=\overline{1\ldots5}$). Порівняння точності отримання кінцевої поліноміальної моделі проводилося з МНК, в якому фіксовано (без оптимального призначення) використовувався поліном другого порядку. Критерієм оцінювання ефективності запропонованого і традиційного підходів було обрано абсолютне відхилення математичного сподівання оцінок МНК від ідеальних значень обраної координати. У табл. 2 похибки розрахунку координат наведені для середньої і крайньої точки інтервалу спостереження, що мають відповідно найкращу і найгіршу методичну і випадкову точність згладжування.

Висновоки

Подані в табл. 2 результати показують, що із збільшенням порядку експериментальної кривої загальна (випадкова і динамічна) похибка згладжування збільшується, однак, швидкість збільшення похибки МНК оцінок, отриманих з використанням полінома фіксованого порядку, значно вища, ніж у випадку оптимального вибору поліноміальної моделі згідно із запропонованим підходом.

Значне збільшення похибки згладжування в традиційному підході викликане, головним чином, зростанням методичної похибки через неадекватне формування поліноміальної моделі. Збільшення ж похибки оцінок МНК при використанні розробленого алгоритму визначення порядку поліноміальної моделі зумовлене зростанням випадкової похибки згладжування і викликане додаванням в апроксимуючий поліном (2) додаткових коефіцієнтів, що містять випадкові похибки їх визначення.

Таким чином, запропонований підхід дозволяє реалізувати визначення порядку поліноміальної моделі за експериментальними даними, що забезпечує збільшення методичної і стохастичної точності кінцевого результату згладжування.

Література

1. *David Dietrich*. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data / David Dietrich, Barry Heller, Beibei Yang. – John Wiley & Sons, Inc., Indianapolis, Indiana, 2015. – 420 p.

2. *Sage Andrew, Melsa James,* Estimation Theory With Applications to Communications and Control. – McGraw-Hill Book Company, Inc.; First Edition, 1971. – 752 p.

Писарчук О.О., Корочкін О.В., Баран Д.Р.

ВИЗНАЧЕННЯ ПОРЯДКУ ПОЛІНОМІАЛЬНОЇ МОДЕЛІ ДЛЯ ПОБУДОВИ ЛІНІЇ ТРЕНДУ В ЗАДАЧАХ DATA SCIENCE

В роботі розглянуто проблема вдосконалення технологій data science, які сьогодні набули широке використання в багатьох галузях. Якість реалізації цих технологій значною мірою визначається точністю розрахунку параметрів трендових залежностей, що потребує адекватного визначення порядку поліноміальної моделі. Метою роботи є вдосконалення методів визначення порядку поліноміальної моделі для побудови лінії тренду в задачах data science.

Авторами запропоновано підхід до визначення порядку поліноміальної моделі для побудови лінії тренду в задачах data science, який базується на аналізі значень вищих похідних експериментальної кривої, враховуючи похибки виміру. Наведено результати оцінювання ефективності запропонованого підходу.

***Ключові слова**: data science, моделі для побудови лінії тренду, поліноміальна модель.*

Pysarchuk O.O., Korochkin O.V., Baran D.R.

DETERMINING THE ORDER OF A POLYNOMIAL MODEL FOR CONSTRUCTION OF TREND LINES IN DATA SCIENCE PROBLEMS

The work deals with the problem of improving data science technologies, which are now widely used in many industries. The quality of the implementation of these technologies is largely determined by the accuracy of the calculation of trend dependence parameters, which requires an adequate determination of the order of the polynomial model. The purpose of the work is to improve the methods of determining the order of the polynomial model for constructing a trend line in data science tasks.

The authors proposed an approach to determining the order of a polynomial model for building a trend line in data science tasks, which is based on the analysis of the values of higher derivatives of the experimental curve, taking into account measurement errors. The results of evaluating the effectiveness of the proposed approach are given.

***Keywords**: data science, models for building a trend line, polynomial model.*