COMPUTER DESIGN SYSTEMS. THEORY AND PRACTICE

Vol. 3, No. 1, 2021

https://doi.org/10.23939/cds2021.01.061

УДК 62-522.7
УДК 004.9

І. Фармага, В. Гадомський
Національний університет “Львівська політехніка”

СИСТЕМА ПРОВЕДЕННЯ АНАЛІЗУ, ДОСЛІДЖЕННЯ ТА ПЕРЕДБАЧЕННЯ ПОДІЙ У ПОСЛІДОВНОСТЯХ ДАНИХ ДИСКРЕТНОГО ЧАСУ

© Фармага І., Гадомський В., 2021

Розроблено програмне забезпечення для передбачення із використанням часових рядів за допомогою мови програмування Python. Для розроблення системи використано модель SARIMA.

Ключові слова: Python; SARIMA модель; часові ряди; передбачення із використанням часових рядів; AWS; AWS Lambda.

Вступ

У світі надзвичайно багато процесів, які залежать від часу. Тому часові передбачення, реалізовані із використанням часових рядів, сьогодні актуальні. Всі ми чули розмови людей про те, як змінюються ціни предметів із плином часу. Ціни то збільшуються, то знижуються на які завгодно предмети: чи продукти харчування, чи паливо, чи ювелірні вироби. Відсоток кредитів та депозитів також різниться із плином часу. Наскільки всі ці дані корисні та які вони взагалі? Всі описані в цій роботі дані відповідають вимогам часових рядів. Їх можна аналізувати для подальших прогнозів. Часові ряди можуть використовуватися як у природі, так і для обміну людей даними між собою, для описань і візуалізації певних даних. Взагалі час є фізичною величиною, але елементи, коефіцієнти, параметри та характеристики часових рядів – це математичні величини. Тому інтерпретація у реальному світі можлива і для часових рядів.

Актуальність використання часових рядів

Сьогодні в суспільстві виникає багато різнопланових завдань, для розв’язання яких потрібне передбачення щодо майбутнього. Для прогнозування тенденцій змін час для нас – важливий параметр. Наприклад, дуже цікаво спрогнозувати навантаження на метрополітен: яка кількість пасажирів за день, о котрій годині пік пасажиропотоку і чи змінюється показник у вихідні дні? Ось для таких подій можна використати прогнозування на основі методу часових рядів.

Важливість прогнозування подій за допомогою часових рядів

Передбачення майбутнього – дуже цікава тема, чи не так? Про це мріяв кожен ще дитиною, і прогнозування за допомогою часових рядів зможе наблизити нас до здійснення “дитячої” мрії. Прогнозування дає нам змогу “побачити” і досягти успіху в багатьох сферах. Саме таке завдання вирішується у цій роботі – досліджено питання про передбачення на основі часових рядів, їх корисність зокрема.

Передбачення за допомогою часових рядів

Поки ми накопичуємо дані, час залишається вирішальним фактором. По суті, в аналізі часових рядів час є важливим елементом даних.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Часовий ряд – це сукупність спостережень за конкретним процесом у певній часовій послідовності, із постійним інтервалом часу між кожним спостереженням (наприклад, тижні, місяці та роки), у деяких випадках допустимі невеликі відхилення у часових інтервалах. Дані тимчасових рядів корисні, коли ми хочемо вивчити, як цікавий для нас актив, цінний папір чи економічна змінна змінюються із часом.

Аналіз часових рядів

Аналіз часових рядів – це аналіз спостережень за точками даних, зібраними за певний період часу, тобто даних часових рядів. Аналізуючи часові ряди, аналітики даних записують спостереження даних із постійними інтервалами для набору періодів часу замість випадкового запису спостережень даних. Швидкість спостереження (тимчасовий інтервал) може становити від мілісекунд до кількох років.

Щоб перевірити, як змінні змінюються із часом, дані часового ряду описують досліджуване явище у певні моменти часу для аналізу коливань змінних із плином часу. Параметри, що нас цікавлять, можуть змінюватись залежно від домена, наприклад:

- значення, зареєстровані науковими приладами за день;
- кількість відвідувань деяких вебсайтів за день;
- щотижнева вартість акцій на фондовому ринку;
- кількість дощових днів за рік.

Крім того, аналіз часових рядів має справу із великою кількістю точок даних для забезпечення узгодженості та надійності. Величезний обсяг даних відображає порівняно великий розмір вибірки, який гарантує, що будь-яка виявлена тенденція чи закономірність не є викидом. Крім того, дані часових рядів можна використовувати для прогнозування майбутніх результатів на підставі попередніх даних.

Прогнозування часових рядів

Аналіз часових рядів [1] допомагає компаніям зрозуміти причини коливань тенденцій чи основних закономірностей із плином часом. Використовуючи різні методи візуалізації даних, організації можуть вивчати сезонні тенденції та вести додаткові дослідження, щоб зрозуміти причини цих тенденцій. Коли організації аналізують дані часових рядів через регулярні проміжки часу, вони використовують прогнозування часових рядів, щоб передбачити майбутні події.

Простіше кажучи, прогнозування часових рядів – це метод прогнозування майбутніх подій на підставі аналізу поведінки чи тенденцій минулих даних з урахуванням припущень про те, що майбутні тенденції будуть подібними на минулі тенденції.

Під час прогнозування даних часових рядів мета полягає у тому, щоб передбачити, як спостереження даних триватиме чи зміниться у майбутньому. Загалом часовий ряд дає змогу проаналізувати основні закономірності, такі як сезонність, циклічність тенденції та нерегулярність. Аналіз часових рядів використовують у різних галузях, наприклад, для розпізнавання образів, аналізу фондового ринку, прогнозування землетрусів, економічного прогнозування, аналізу даних перепису тощо.

Закономірності у часових рядах

Для часових рядів характерні тенденції, цикли та сезонність. На превеликий жаль, виникає багато плутанини щодо сезонності та циклів. І саме щоб її уникнути, потрібно проаналізувати кожен із цих факторів.

- Тенденції: коливання кількості даних упродовж певного проміжку часу.
- Сезонність. Зазвичай це відома та фіксована частота. Сезонна закономірність виникає, коли на часовий ряд впливають сезонні фактори, наприклад, пора року або день тижня.
- Цикл: у разі коливання даних виникає цикл, але якщо порівнювати із сезонними, то це не фіксована частота.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Вступ до SARIMA. Прогнозування часових рядів мовою програмування Python

Авторегресивне інтегроване змінне середнє, або ARIMA, – один із методів прогнозування, найширше використовуваних для одновимірного прогнозування даних часових рядів.

Хоча цей метод може обробляти дані із трендом, він не підтримує часові ряди із сезонною складовою.

Розширення ARIMA, яке підтримує пряме моделювання сезонної складової ряду, називається SARIMA.

Відкриємо для себе метод сезонного авторегресійного інтегрованого змінного середнього (або SARIMA), метод передбачення часових рядів, що охоплює тенденції та сезонність.

Що не так з ARIMA?

Авторегресійне інтегроване середнє змінне, або ARIMA, – це метод прогнозування одновимірних даних часових рядів.

Як випливає із назви, він підтримує елементи як авторегресії, так і змінного середнього. Інтегрований елемент відноситься до різниці, даючи змогу методу підтримувати дані часових рядів із трендом.

Проблема у тому, що ARIMA не підтримує сезонних даних. Це тимчасовий ряд із циклом, що повторюється.

ARIMA очікує дані, які або не є сезонними, або в них видалено сезонний компонент, наприклад, з урахуванням сезонних коливань за допомогою таких методів, як сезонна різниця.

Що таке SARIMA?

Сезонне авторегресійне інтегроване середнє змінне, SARIMA (також відоме як Seasonal ARIMA), є додатком до ARIMA, який підтримує дані одновимірних часових рядів із сезонною складовою.

Він додає три нові гіперпараметри, щоб вказати авторегресію (AR), різницю (I) та ковзне середнє (MA) для сезонного компонента ряду, а також додаткові параметри для періодів сезонності.

SARIMA утворюється методом додавання додаткових членів до ARIMA. Сезонна частина моделі складається із компонентів, які походять від несезонних компонентів моделі, але охоплюють зміщення назад сезонного періоду [2].

Як налаштувати SARIMA?

Для налаштування SARIMA необхідно вибрати гіперпараметри як трендових, так і сезонних елементів ряду.

Елементи тренду

Є три елементи тенденції, які потрібно налаштувати.

Вони такі самі, як модель ARIMA, а саме:

p: порядок авторегресії тренду;

d: порядок різниці трендів;

q: порядок ковзного середнього тренду.

Сезонні елементи

Необхідно налаштувати чотири сезонні елементи, що не є частиною ARIMA, а саме:

P: сезонний авторегресійний порядок;

D: порядок сезонної різниці;

Q: порядок середньої сезонної змінної;

m: кількість тимчасових кроків одного сезонного періоду.

У сукупності позначення для моделі SARIMA набувають такого вигляду:

SARIMA(*p,d,q*)(*P,D,Q*)*m*, де вказано спеціально вибрані гіперпараметри для моделі. Наприклад:

SARIMA (3,1,0)(1,1,0)12

Важливо, що параметр *m* впливає на параметри *P, D* і *Q*. Наприклад, значення *m*, що дорівнює 12 для місячних даних, означає річний сезонний цикл.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Якщо P = 1, буде використано перше спостереження за сезонним зміщенням у моделі, наприклад, $t - (m \times 1)$ або t–12. А якщо P = 2, використовуватимуться два останніх сезонних зміщення спостережень $t - (m \times 1)$, $t - (m \times 2)$.

Аналогічно D, що дорівнює 1, обчислює сезонну різницю першого порядку, а Q = 1 використовує помилки першого порядку в моделі (наприклад, змінне середнє).

Сезонна модель ARIMA використовує диференціацію із лагом, що дорівнює кількості сезонів, щоб вилучити додаткові адитивні сезонні ефекти. Як і у разі диференціації із лагом 1 для усунення тенденції, диференціація лаг s вводить змінне середнє. Сезонна модель ARIMA містить терміни авторегресії та змінного середнього із лагом s [3].

Елементи тенденції можна вибрати, ретельно проаналізувавши графіки ACF та PACF, з огляду на кореляції останніх кроків часу (наприклад, 1, 2, 3).

Аналогічно, графіки ACF і PACF можна проаналізувати, щоб встановити значення для сезонної моделі, розглянувши кореляцію на сезонних кроках часу затримки.

Сезонні моделі ARIMA потенційно можуть містити велику кількість параметрів і комбінацій термінів. Тому доцільно випробувати широкий спектр моделей під час пристосування до даних і вибрати найкращу модель за відповідним критерієм [4].

Крім того, можна використовувати пошук у сітці за тенденцією та сезонними гіперпараметрами.

Алгоритм використання SARIMA в Python:

Метод прогнозування часових рядів SARIMA підтримується у Python через бібліотеку Statsmodels [5].

Щоб використовувати SARIMA, потрібно виконати три кроки:

1. Визначити модель.
2. Підігнати визначену модель.
3. Виконати прогноз за допомогою відповідної моделі.

Алгоритм передбачення за допомогою часових рядів

Як приклад, для демонстрації працездатності та тестування розробленої системи використаємо дані про погоду. Для аналізу та прогнозування використано інформацію про погоду в місті Київ з 2020 до 2021 рр. Ці дані враховують новітні фактори зміни погоди: сезонність, тенденції та цикли. Хоч вибрані дані відображають погоду, використати розроблену систему можна і для даних, пов’язаних із продажами, навантаженням тощо. Ми можемо це зробити, адже передбачення залежать насамперед від часу, від часових рядів, на основі яких ми і будуємо передбачення.

Щоб отримати більше інформації про вхідні дані, бібліотека Python Pandas надає функцію “опис”, щоб показати кількість, середнє, стандартне відхилення, мінімальне/максимальне значення та квантилі нашого набору даних. Деякі відмінні закономірності з’являються, коли ми візуалізуємо дані (рис. 1). Для часового ряду характерна сезонність, наприклад температура завжди низька на початку і висока в середині року.

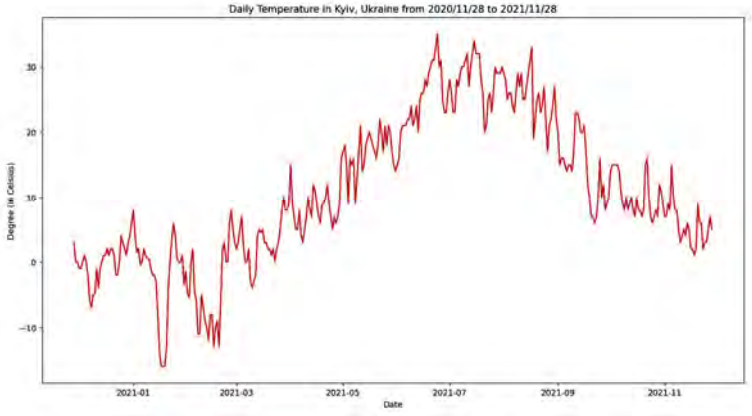


Рис. 1. Візуалізація вхідних даних

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Наведений вище графік трішки “зашумлений”, оскільки містить всі добові температури. Однак, уважно переглянувши точки даних, бачимо лише незначну зміну температури між поточною датою та наступною.

Як побудувати графік, що відображає лише зміну температури, яка перевищує добову, і, отже, виглядає акуратніше?

Ось інструмент: змінне середнє (рис. 2) переважно використовують із даними часових рядів, щоб фіксувати короткострокові коливання, зосереджуючись на триваліших тенденціях.

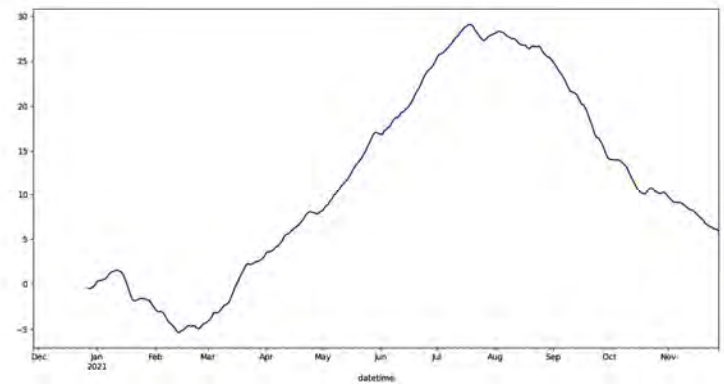


Рис. 2. Змінне середнє часового ряду

Ми також можемо візуалізувати наші дані за допомогою методу, який називається декомпозицією часових рядів і дає нам змогу розкласти цей часовий ряд на три окремі компоненти: тенденцію, сезонність і шум.

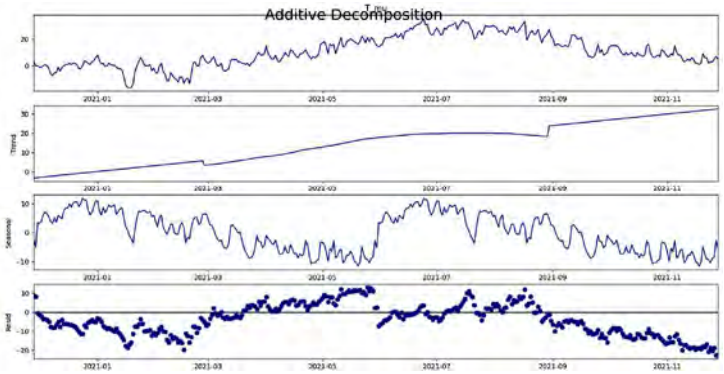


Рис. 3. Графік тенденції, сезонності та шуму

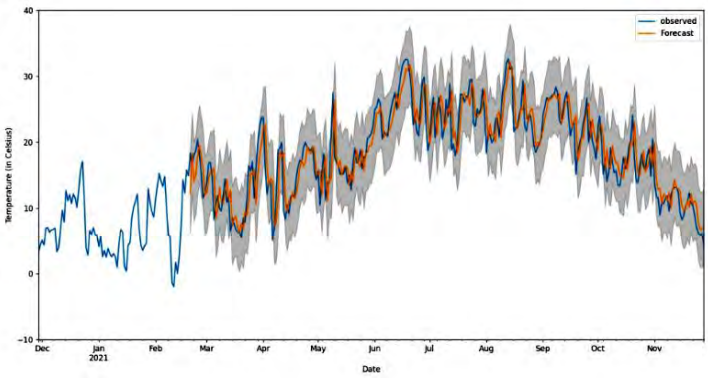


Рис. 4. Графік прогнозу

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Із наведеного вище графіка (рис. 3) чітко видно, що температура нестабільна, а також очевидна її сезонність.

Проаналізувавши ряд, тепер можемо відкласти реальні та прогнозовані значення середньої добової температури (рис. 4), щоб оцінити, наскільки добре ми виконали завдання. Зверніть увагу, як ми збільшили кінець часового ряду, розрізавши індекс дати.

Загалом, наші прогнози дуже добре відповідають істинним значенням, показуючи сезонний цикл тривалістю 365 днів. Також корисно кількісно оцінити точність наших прогнозів. Скористаємося MSE (середньою квадратичною помилкою), в якій для кожного передбаченого значення обчислюємо його відстань до справжнього значення та підносимо результат у квадрат.

Стосовно прогнозу погоди похибка прогнозу в 2,97 градуса здається перспективною та достатньою, оскільки існує багато інших факторів, які сприяють зміні температури, ураховуючи, крім всього іншого, швидкість вітру, тиск повітря тощо.

Розгортання системи передбачення у хмарному середовищі AWS

Головна наша мета – розробити систему передбачення за допомогою часових. Відповідно основна логіка (передбачення) повинна реалізуватися в окремому компоненті. Для цього було вибрано сервіс AWS Lambda. AWS Lambda – це сервіс для обчислення, який може запускати програмний код майже для будь-яких типів програм або серверних служб. І все це може здійснюватися без необхідності адміністрування. AWS Lambda може виконати всі процедури адміністрування за користувача. Нам потрібно лише надати свій код на одній із мов, яку підтримує Lambda.

Коли створено функцію, виникає проблема – як викликати її. Сьогодні це можна зробити лише вручну. Допомогти нам можуть тригери. AWS Lambda виконується під час кожного запуску виконання у відповідь на тригер повідомлення про подію. Відповідно, створивши тригер, ми матимемо точку доступу до самої функції. Для цього ми використали сервіс Amazon API Gateway. Він допомагає створити інтерфейс доступу до AWS Lambda функції. Також варто зазначити, що було додано приватний ключ, яким функція захищена. Отже, лише авторизовані користувачі зможуть її виконувати.

Висновки

Для передбачення за допомогою часових рядів розроблено систему із використанням мови програмування Python. Система основана на моделі SARIMA, яку детально описано. Здійснено також аналіз та візуалізацію вхідних даних для наочності виконання алгоритму.

Література

1. Часовий ряд.URL: https://uk.wikipedia.org/wiki/ Часовий_ряд.

2. Hyndman, R. J. G. Athanasopoulos. Forecasting: principles and practice. Otexts.com., 2013, 242.

3. Cowpertwait, P. S. P., Metcalfe, A. V. Intoductory Time Series with R (Use R!). Springer, 2009, 142. https://doi.org/10.1007/978-0-387-88698-5_1.

4. Cowpertwait, P. S. P., Metcalfe, A. V. Intoductory Time Series with R (Use R!). Springer, 2009, 143–144. https://doi.org/10.1007/978-0-387-88698-5_1.

5. Statsmodels. https://www.statsmodels.org/stable/index.html

I. Farmaha, V. Hadomskyi
Lviv Polytechnic National University

SYSTEM FOR ANALYSIS, RESEARCH AND FORECAST OF EVENTS IN DISCRETE TIME DATA SEQUENCES

© Farmaha I., Hadomskyi V., 2021

This paper is devoted develop software for time series forecasting using Python programming language. SARIMA model was used to develop the system.

Key words: Python; SARIMA model; time series; time series forecast; AWS; AWS Lambda.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua