Prognozowanie sezonowe - porównanie metod

Teoria Automatów i Obliczeń

Piotr Widomski

22.12.2021

1 Cel projektu

Celem projektu jest porównanie wydajności prognozowania sezonowych szeregów czasowych za pomocą maszyny `SARIMA` (`Seasonal ARIMA`) oraz liniowej kombinacji maszyn `ARIMA`, działających na co n-tym elemencie szeregu.

2 Użyte dane

Do porównania metod użyte zostały dane przedstawiające średnią miesięczną temperaturę powietrza mierzoną na dublińskim lotnisku od roku 1941. Dane zostały pobrane z portalu inicjatywy Open Data prowadzonej przez rząd irlandzki. Pomiary wykonane zostały przez Met Éireann - irlandzki narodowy serwis meteorologiczny i udostępnione na licencji CC Attribution 4.0.

Na potrzeby projektu użyty został wycinek danych zawierających pełne lata, od 1942 do 2020 roku, czyli 79 lat pomiarów. Dane te zostały podzielone na treningowe oraz testowe w stosunku około 80%, gdzie dane treningowe zawierają 63 lata, od 1942 do 2004 roku, a dane testowe - 16 lat, od 2005 do 2020 roku.

3 Opis metod

3.1 ARIMA

`ARIMA` jest klasą modelów, które opisują szereg na podstawie poprzednich wartości. Model `ARIMA` składa się z trzech procesów:

- `AR` (autoregresyjny) - każda wartość jest liniową kombinacją pewnej liczby poprzednich wartości. Liczbę poprzednich wartości użytych przy obliczeniu następnej oznaczamy przez p, a a proces autoregresyjny z rzędem regresji p - `AR(p)` i przedstawiamy następująco:

  $$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

  gdzie y_{t-i} jest i-tą poprzednią wartością ciągu w chwili t, ϕ_i jest współczynnikiem i-tej poprzedniej wartości, ϵ_t zaburzeniem w momencie t, a α - wartością podstawową.

- `MA` (średniej ruchomej) - każda wartość jest zależna od zaburzenia w chwili obecnej oraz wcześniejszych. Liczbę poprzednich zaburzeń uwzględnionych przy obliczeniu następnej wartości oznaczamy przez q, a proces średniej ruchomej rzędu q - `MA(q)` i przedstawiamy:

  $$y_t = \alpha + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

  gdzie ϵ_t jest zaburzeniem w momencie t, θ_i - współczynnikiem i-tego poprzedniego zaburzenia, a α - wartością podstawową.

- `I` (integracja) - w celu modelowania ciągów niestacjonarnych, model `ARIMA` może wykonywać różnicowanie pewnego stopnia, w celu operowania na bardziej stacjonarnym ciągu. W takim wypadku, zamiast następnej wartości, model prognozuje różnicę wartości. Jednak dzięki znajomości poprzedniej wartości możemy wyliczyć na podstawie różnicy następną wartość. Stopień różnicowania oznacza się przez d.

Łącząc powyższe procesy, otrzymujemy ogólny wzór modelu `ARIMA`:

$$y_t^* = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i}^* + \sum_{i=1}^{q} + \theta_i \epsilon_{t-i} + \epsilon_t$$
$$y_t^* = \Delta^d y_t$$

Poszczególne modele `ARIMA` definiuje się za pomocą trzech zmiennych całkowitych p, d, q i zapisuje `ARIMA(p, d, q)`.

3.2 SARIMA

`SARIMA`, lub `Seasonal ARIMA`, jest rozszerzeniem modelu `ARIMA` wspierającym ciągi czasowe wykazujące sezonowość, czyli takie, które posiadają powtarzające się z pewną określaną częstością zmiany wartości. Częstość tę oznaczymy przez S.

Rozszerzenie to dodaje trzy nowe, analogiczne do modelu `ARIMA` procesy:

- Sezonową autoregresję - stopień oznaczamy przez P
- Sezonową średnią ruchomą - stopień oznaczamy przez Q
- Sezonową integrację - stopień oznaczamy przez D

W celu zdefiniowania wzoru modelu `SARIMA` zdefiniujmy operator poprzednika B:

$$By_t = y_{t-1}$$
$$B^i y_t = y_{t-i}$$

Uwzględniając nowe procesy oraz operator B, otrzymujemy ogólny wzór modelu `SARIMA`:

$$\prod_{i=1}^{p}(1-\phi_i B^i)\prod_{i=1}^{P}(1-\Phi_i B^{Si})\prod_{i=1}^{d}(1-B^i)\prod_{i=1}^{D}(1-B^{Si})y_t = \prod_{i=1}^{q}(1+\theta_i B^i)\prod_{i=1}^{Q}(1+\Theta_i B^{Si})\epsilon_t$$

Poszczególne modele `SARIMA` definiuje się za pomocą zmiennych całkowitych niesezonowych p, d, q, zmiennych sezonowych P, D, Q oraz częstości sezonowej S i zapisuje `SARIMA(p, d, q) x (P, D, Q)S`.

3.3 Kombinacja liniowa maszyn ARIMA

Model ten składa się z grup maszyn `ARIMA` działających na co n-tym elemencie ciągu z ostatnich S elementów, gdzie S jest częstością maszyny `SARIMA`. Każda z maszyn w grupie operuje na pewnym podciągu ciągu bazowego zawierającego co n-ty element ciągu. Podciągi te są parami rozłączne, oraz ich suma równa jest ciągowi bazowemu. Każdy podciągów w grupie jest równej długości. Z tego powodu narzucamy ograniczenie na n: $n|S$. Maszyny

wewnątrz grupy są uporządkowane według kolejności podciągów, czyli za pierwszą uznajemy maszynę, której podciąg zaczyna się od pierwszego elementu ciągu, drugą - od drugiego, i tak dalej.

Weźmy grupy maszyn spełniające powyższe warunki i ponumerujmy je $1, 2, \ldots, n$. Oznaczmy przez $y_{i,t}$ element otrzymany z grupy i w momencie t. Wartość tą możemy uzyskać stosując wzór modelu `ARIMA` z maszyną z grupy i, której indeks wewnątrz grupy jest równy $tmodi$. Stosując kombinację liniową maszyn `ARIMA` otrzymujemy następujący wzór na następny element:

$$y_t = \sum_{i=1}^{n} \alpha_i y_{i,t}$$

gdzie α_i jest współczynnikiem i-tej grupy.

4 Implementacja

4.1 Biblioteka modeli

Do budowy modeli została użyta biblioteka `statsmodel` zapewniająca implementacje modeli `ARIMA` oraz rozszerzenia `SARIMA`. Dodatkowo biblioteka ta umożliwia automatyczne dobranie odpowiednich współczynników na podstawie danych wejściowych oraz parametrów modelu. Dodatkowo udostępnia metody umożliwiające prognozowanie kolejnych elementów ciągu.

Dodatkowo, w celu dobrania optymalnych parametrów modeli, użyta została biblioteka `pmdarima`, która umożliwia wyszukanie optymalnych parametrów modelu na podstawie ustalonych kryteriów, po czym zwraca maszynę utworzoną z dobranych przez siebie parametrów w implementacji kompatybilnej z tą, która znajduje się w bibliotece `statsmodels`.