Временные ряды

Лисицкая Елизавета
Муринов Андрей
Смаева Елизавета

Санкт-Петербургский политехнический университет Петра Великого

17 ноября 2025 г.

1. Временные ряды
2. Аналитика временных рядов
3. Модель вида ARIMA

Временные ряды

Определение

Временной ряд — значения меняющихся во времени признаков, полученные в некоторые моменты времени.

Задача прогнозирования

Пусть $(y_t, t \in \mathbb{N})$ - временной ряд, для которого известны значения $y_1, \ldots, y_T$. Требуется построить прогноз - функцию f, такую что величина $\widehat{y}_{T+h} = f(y_1, \ldots, y_T, h)$ как можно лучше приближает значение y_{T+h}, где h - количество шагов на которое нужно построить прогноз.
Иногда требуется построить доверительный интервал (d_{T+h}, u_{T+h}), т.ч.

$$\mathbb{P}(d_{T_h} \leq y_{T+h} \leq u_{T+h}) \geq \alpha$$

Состовляющие временного ряда

В общем случае модель временного ряда имеет вид:

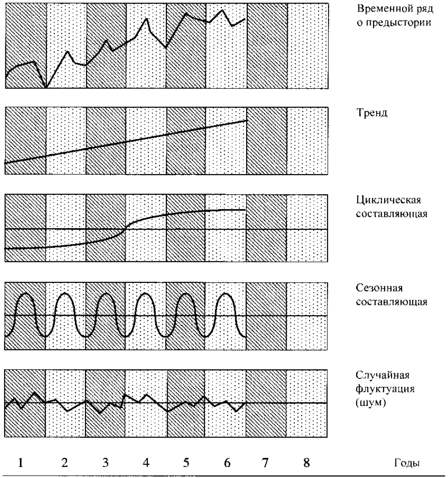
$$y_t = F(d_t, \varepsilon_t),$$

где d_t - систематическая составляющая ряда, ε_t - случайная составляющая ряда, с нулевым математическим ожиданием

Определения

- Тренд - плавное долгосрочное изменение временного ряда.
- Сезонность и цикличность – повторяющиеся изменения временного ряда с постоянным периодом (как правило сезонном считают период меньше года, циклом - больше года)
- Шум - непрогнозируемая случайная компонента ряда.

Декомпозиция ВР



Пусть T_t - тренд, S_t - сезонность, R_t - шум.

- Аддитивная декомпозиция: $y_t = T_t + S_t + R_t$
- Мультипликативная декомпозиция: $y_t = T_t \cdot S_t \cdot R_t$

Декомпозиция на основе скользящего среднего

Пусть s - известный период сезонности.

- Тренд:

$$T_t = \frac{1}{s} \sum_{i=t-s/2}^{t+s/2} y_i$$

- Сезонность:

$$y_t := y_t - T_t$$

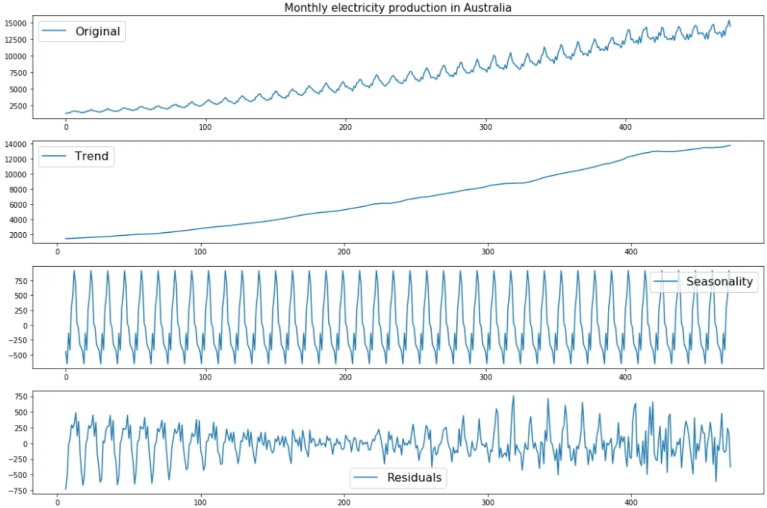
$$G_i = \{y_i, y_{i+s}, \ldots, y_{i+ks}\}, \text{ где } i \in [1:s]$$

$$S_t = \overline{G}_{(t \mod s)}$$

- Ошибка:

$$R_t = y_t - T_t - S_t$$

Пример декомпозиции



Автокорреляционная функция

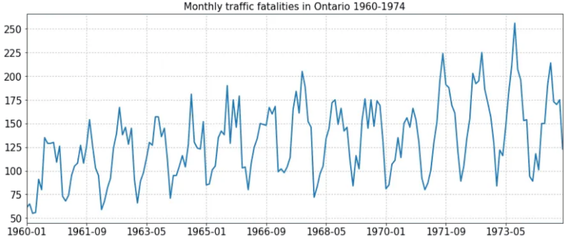
Временной ряд может содержать зависимые между собой признаки.

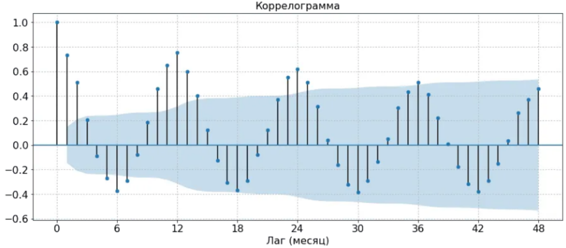
Коэффициент корреляции Пирсона ряда с лагом τ

$$r_\tau = \widehat{corr}(y_t, y_{t+\tau}) = \frac{\sum\limits_{t=1}^{T-t}(y_t - \overline{y})(y_{t+\tau} - \overline{y})}{\sum\limits_{t=1}^{T}(y_t - \overline{y})^2},$$

где $\overline{y}$ - среднее всего ряда.

Коррелограмма. Пример.





Стационарные временные ряды

Определение

Временной ряд считается стационарным, если выполняются три условия:

1. Постоянное математическое ожидание:

$$E(y_t) = \mu \quad \text{для всех } t$$

2. Постоянная дисперсия:

$$Var(y_t) = \sigma^2 \quad \text{для всех } t$$

3. Ковариация зависит только от сдвига:

$$Cov(y_t, y_{t+k}) = \gamma_k \quad \text{для всех } t \text{ и любого } k$$

Кросс-валидация для временных рядов

Схемы кросс-валидации

1. Обучаем модель на первых t значениях, прогнозируем следующие Δt значений
2. Обучаем модель на $t+\Delta t$ значениях, прогнозируем значения $y_{\Delta t+t},\ldots,y_{t+2\Delta}$
3. ...
4. На каждой итерации считаем ошибки и усредняем

1. Обучаем модель на первых t значениях, прогнозируем следующие Δt значений
2. Обучаем модель на значениях $y_{\Delta t+1},\ldots,y_{\Delta t+t}$, прогнозируем значения $y_{t+\Delta t+1},\ldots,y_{t+2\Delta}$
3. ...
4. Считаем ошибки и усредняем

Приведение временных рядов к стационарности

Методы преобразования

- Дифференцирование:

$$y'_t = y_t - y_{t-1}$$

  Устраняет линейный тренд

- Сезонное дифференцирование:

$$y'_t = y_t - y_{t-m}$$

  Устраняет сезонность (m - период)

Преобразование Бокса-Кокса:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

Стабилизирует дисперсию

Модель скользящего среднего (MA)

Значение временного ряда в момент t описывается как линейная комбинация прошлых ошибок прогноза (“белого шума”).

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

- Где:
  - y_t — значение ряда в момент t
  - μ — константа (среднее значение ряда)
  - $\varepsilon_t, \varepsilon_{t-1}, \ldots$ — ошибки (белый шум), $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
  - $\theta_1, \ldots, \theta_q$ — параметры модели
  - q — порядок модели (MA(q))

Модель авторегрессии (AR)

Значение временного ряда в момент t описывается как линейная комбинация его собственных прошлых значений.

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

- Где:
  - $y_t, y_{t-1}, \ldots$ — значения ряда в моменты времени $t, t-1, \ldots$
  - c — константа
  - $\phi_1, \ldots, \phi_p$ — параметры модели
  - ε_t — ошибка (белый шум) в момент t
  - p — порядок модели (AR(p))
- Отметим, что, вообще говоря, для стационарности нужны некоторые условия на коэффициенты

Стационарность AR модели

Характеристическое уравнение

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

Условие стационарности для AR(p)

Для AR(p) модели: $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$
Модель является стационарной, если все корни характеристического уравнения лежат вне единичного круга.

Модель авторегрессии и скользящего среднего (ARMA)

Комбинирует подходы AR и MA, чтобы использовать преимущества обеих моделей. Значение ряда зависит как от своих прошлых значений, так и от прошлых ошибок прогноза.

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

- Где:
  - p — порядок авторегрессионной части (AR)
  - q — порядок части скользящего среднего (MA)
- Обозначается как ARMA(p, q).
- Стационарность модели будет определяться только его AR(p) компонентой

Модель ARIMA(p, d, q)

Определение:

Модель ARIMA(p, d, q) — это расширение моделей типа ARMA на нестационарные временные ряды, которые однако могут стать стационарным после примененией процедуры дифференцирования ряда.

Формула:

$$a(L)(1 - L)^d y_t = \alpha + b(L)\varepsilon_t,$$

или

$$(1 - L)^d y_t = \mu + \frac{b(L)}{a(L)}\varepsilon_t.$$

Замечание: многочлен $\tilde{a}(z) = a(z)(1 - z)^d$ имеет d единичных корней => модель позволяет учесть нестационарности, в частности, тренд.

Частичная автокорреляционная функция (PACF)

Определение

Частичная автокорреляция (PACF) — корреляция ряда с собой после снятия линейной зависимости от промежуточных значений ряда.

Цель — учесть опосредованное влияние промежуточных значений и оценить непосредственное влияние $y_{t-\tau}$ на y_t.

Формула

$$\gamma_\tau = \begin{cases} \mathrm{corr}(y_{t+1}, y_t), & \tau = 1; \\ \mathrm{corr}(y_{t+\tau} - y_{t+\tau}^{\tau-1}, y_t - y_t^{\tau-1}), & \tau \geq 2, \end{cases}$$

где $y_t^{\tau-1}$ — линейная регрессия на $y_{t-1}, y_{t-2}, \ldots, y_{t-(\tau-1)}$:

- $y_t^{\tau-1} = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_{\tau-1} y_{t-(\tau-1)}$
- $y_{t+\tau}^{\tau-1} = \varphi_1 y_{t+\tau-1} + \varphi_2 y_{t+\tau-2} + \ldots + \varphi_{\tau-1} y_{t+1}$

Оценка коэффициентов в ARIMA

Пусть гиперпараметры p, d, q фиксированы, ε_t — гауссовский белый шум, выпишем функцию правдоподобия:

$$L_y(\theta, \varphi, \alpha) = p_{\theta,\varphi,\alpha}(y_1, \dots, y_T)$$

где $p_{\theta,\varphi,\alpha}(y_1, \dots, y_T)$ — совместная плотность.

В качестве оценок параметров берется оценка максимального правдоподобия.

Для поиска начальных приближений p и q:

- p: последний значимый пик у PACF
- q: последний значимый пик у ACF

Далее используется поиск по сетке, минимизируя информационный критерий:

- $AIC = -2\ell^* + 2(p+q+1)$ — критерий Акаике
- $AIC_c = -2\ell^* + \frac{2(p+q+1)(p+q+2)}{T-p-q-2}$ — для коротких рядов
- $BIC = -2\ell^* + (\log T - 2)(p+q+1)$ — Байесовский информационный критерий

где $\ell^* = \ln L_y(\hat{\theta}, \hat{\varphi}, \hat{\alpha})$ — логарифм функции правдоподобия.

План прогнозирования с помощью модели ARIMA

1. Анализ выбросов: замена нерелевантных выбросов на NA или усреднение по соседним элементам
2. Стабилизация дисперсии с помощью преобразований
3. Дифференцирование, если ряд нестационарен
4. Выбор пилотных p и q по PACF и ACF
5. Подбор оптимальной модели по AIC/AIC$_c$ вокруг этих параметров
6. Пошаговое построение прогноза:
   - для $t \leq T$: $\varepsilon_t = y_t - \hat{y}_t$
   - для $t > T$: $\varepsilon_t = 0$
   - для $t > T$: $y_t = \hat{y}_t$
7. Построение предсказательного интервала:
   - если остатки модели нормальны и гомоскедастичны (дисперсия постоянна), то строится теоретический предсказательный интервал;
   - иначе интервалы строятся с помощью бутстрепа.

Модель SARIMA

Пусть s — известная сезонность ряда. Добавим в модель ARIMA(p,d,q) компоненты, отвечающие за значения в предыдущие сезоны.
SARIMA $(p, d, q) \times (P, D, Q)_s$:

$$(1 - L)^d (1 - L^s)^D y_t = \mu + \frac{b(L) B(L^s)}{a(L) A(L^s)} \varepsilon_t,$$

где

$$a(z) = 1 - \varphi_1 z - \cdots - \varphi_p z^p,$$

$$b(z) = 1 + \theta_1 z - \cdots - \theta_q z^q,$$

$$A(z) = 1 - \varphi_1^s z - \cdots - \varphi_P^s z^P,$$

$$B(z) = 1 + \theta_1^s z - \cdots - \theta_Q^s z^Q.$$

Модель ARIMAX

ARIMAX — обобщение модели ARIMA, которая учитывает некоторые экзогенные факторы. Пусть $x_t \in \mathbb{R}^n$ — ряд регрессоров, известный до начала прогноза.

Простой вариант:

$$(1-L)^d y_t = \mu + \sum_{i=1}^{n} \frac{\beta_i}{a(L)} x_t^i + \frac{b(L)}{a(L)} \varepsilon_t$$

Общий случай:

$$(1-L)^d y_t = \mu + \sum_{i=1}^{n} \frac{u_i(L)}{v_i(L)} x_t^i + \frac{b(L)}{a(L)} \varepsilon_t$$

Источники

Использованные материалы

[1] Учебник по машинному обучению.
Яндекс. Хэндбук.
Обзор основных концепций и алгоритмов машинного обучения.

[2] Сурина А.В. Анализ временных рядов.
СПбПУ. Научно-технические ведомости.
Методы и модели для анализа и прогнозирования временных рядов.

Спасибо за внимание!