Случайные процессы ПМИ

Прикладной поток

Семинар 12

ФИВТ МФТИ

1. Модели типа ARIMA

Модель ARMA(p, q)

$$
\begin{aligned}
y_t &= \alpha + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} \\
&\quad + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}
\end{aligned}
$$

Эквивалентная запись:

$$
a(L) y_t = \alpha + b(L)\varepsilon_t \qquad \text{или} \qquad y_t = \alpha + \frac{b(L)}{a(L)} \varepsilon_t,
$$

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

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

L — оператор сдвига: $L y_t = y_{t-1}, L\varepsilon_t = \varepsilon_{t-1}$.

Модель ARIMA

Модель ARIMA(p, d, q) для y_t — модель ARMA(p, q) для ряда разностей порядка d исходного ряда. Позволяет учесть нестационарности, в частности, тренд.

Что такое разность порядка 1? $y_t - y_{t-1} = (1 - L)y_t$

Получаем формулу модели ARIMA:

$$a(L)(1-L)^d y_t = \alpha + b(L)\varepsilon_t \qquad \text{или} \qquad (1-L)^d y_t = \alpha + \frac{b(L)}{a(L)}\varepsilon_t.$$

То есть многочлен $\widetilde{a}(z) = a(z)(1-z)^d$ имеет d единичных корней.

Оценка коэффициентов в ARIMA

Пусть p, d, q фиксированы.

Предполагаем, что ε_t — гауссовский белый шум.

Даны значения временного ряда $y_1, ..., y_T$ и их совместная плотность согласно модели ARIMA(p, d, q) равна $f(a_1, ..., a_T)$.

Тогда $f(y_1, ..., y_T)$ — функция правдоподобия.

В качестве оценок коэффициентов берется оценка максимального правдоподобия.

Почему p, d, q нельзя оценивать с помощью ОМП?

Оценка p и q

Частичная автокорреляция (PACF) — корреляция ряда с самим собой после снятия линейной зависимости от предыдущих элементов ряда.

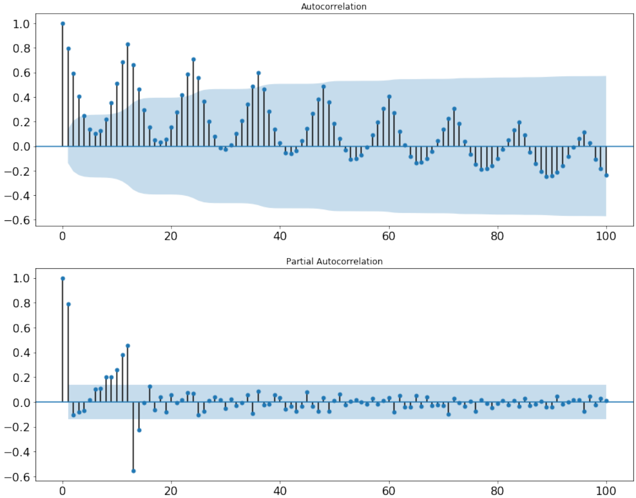
$$\varphi_h = \begin{cases} r(y_{t+1}, y_t), & h = 1; \\ r\left(y_{t+h} - y_{t+h}^{h-1}, y_t - y_t^{h-1}\right), & h \geqslant 2, \end{cases}$$

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

$$y_t^{h-1} = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_{h-1} y_{t-(h-1)}$$

$$y_{t+h}^{h-1} = \beta_1 y_{t+h-1} + \beta_2 y_{t+h-2} + \dots + \beta_{h-1} y_{t+1}$$

Оценка p и q


Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
20
40
60
80
100
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
20
40
60
80
100

Оценка p и q

Начальное приближение q: последний значимый пик у ACF

Начальное приближение p: последний значимый пик у PACF

Далее используем поиск по сетке вокруг подобранных значений,

минимизируя информационный критерий:

$AIC = -2L + 2(p + q + 1)$ — критерий Акаике;

$AIC_c = -2L + \frac{2(p+q+1)(p+q+2)}{T-p-q-2}$ — критерий Акаике (короткие ряды);

$BIC = -2L + (\log T - 2)(p + q + 1)$ — критерий Шварца,

где L — логарифм функции правдоподобия (для ОМП).

Итог: прогнозирование с помощью ARIMA

1. Анализ выбросов: замена нерелевантых выбросов на NA или "усреднение" по соседним элементам.

2. Стабилизация дисперсии (преобразования).

3. Дифференцирование, если ряд не стационарен.

4. Выбор пилотных p и q по ACF и PACF.

5. Вокруг этих параметров подбираем оптимальную модель по AIC/AIC_c.

6. Если для полученной модели не выполняются необходимые свойства остатков, модель можно улучшить.

Итог: прогнозирование с помощью ARIMA

7. Построение прогноза:
   - для $t \leqslant T$: $\varepsilon_t \Longrightarrow \widehat{\varepsilon}_t = y_t - \widehat{y}_t$;
   - для $t > T$: $\varepsilon_t \Longrightarrow 0$;
   - для $t > T$: $y_t \Longrightarrow \widehat{y}_t$.
8. Построение предсказательного интервала:
   - если остатки модели нормальны и гомоскедастичны, то строим теоретический предсказательный интервал;
   - иначе интервалы строятся с помощью симуляции.

Слишком просто...

Давайте усложнять!

Модель SARIMA

ARIMA(p, d, q):

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

Пусть s — период сезонности ряда. Добавим в модель ARIMA(p, d, q) компоненты на значения в предыдущие сезоны...

SARIMA$(p, d, q) \times (P, D, Q)_s$:

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

где $a(z) = 1 - \varphi_1 z - \ldots - \varphi_p z^p, \qquad b(z) = 1 + \theta_1 z + \ldots + \theta_q z^q,$

$$A(z) = 1 - \varphi_1^s z - \ldots - \varphi_P z^P, \qquad B(z) = 1 + \theta_1^s z + \ldots + \theta_Q^s z^Q.$$

Еще усложним?

Модель ARIMAX

ARIMA(p, d, q):

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

Пусть $x_t \in \mathbb{R}^n$ — процесс регрессоров, *известный до начала прогноза*.

Простой вариант:

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

Общий случай:

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

Модель SARIMAX

ARIMA(p, d, q):

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

Соединим...

SARIMAX$(p, d, q) \times (P, D, Q)_s$:

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

Пакет forecast в R

Построение модели:

```
auto.arima(y, d = NA, D = NA, max.p = 5, max.q = 5, max.P
= 2, max.Q = 2, max.order = 5, max.d = 2, max.D = 1,
start.p = 2, start.q = 2, start.P = 1, start.Q = 1,
stationary = FALSE, seasonal = TRUE, ic = c(''aicc'', ''aic'',
''bic''), stepwise = TRUE, trace = FALSE, approximation =
(length(x) > 150 | frequency(x) > 12), truncate = NULL,
xreg = NULL, test = c(''kpss'', ''adf'', ''pp''), seasonal.test
= c(''ocsb'', ''ch''), allowdrift = TRUE, allowmean = TRUE,
lambda = NULL, biasadj = FALSE, parallel = FALSE,
num.cores = 2, x = y, ...)
```

Пакет forecast в R

Прогнозирование:

```
forecast(object, h = ifelse(frequency(object) > 1, 2 *
frequency(object), 10), level = c(80, 95), fan = FALSE,
robust = FALSE, lambda = NULL, find.frequency = FALSE,
allow.multiplicative.trend = FALSE, model = NULL, ...)
```

2. Модели типа экспоненциального сглаживания

Простое экспоненциальное сглаживание

Наивный прогноз: $\hat{y}_{T+1|T} = y_T$

Прогноз средним: $\hat{y}_{T+1|T} = \sum_{t=t_0}^{T} y_t$

Прогноз взвешенным средним с экспоненциально убывающими весами:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \ldots =$$

$$= \alpha y_T + (1-\alpha)\hat{y}_{T|T-1}$$

- $\alpha \approx 1 \to$ больший вес последним точкам: $\hat{y}_{T+1|T} \approx y_T$
- $\alpha \approx 0 \to$ большее сглаживание: $\hat{y}_{T+1|T} \approx \bar{y}$

Прогноз плоский: $\hat{y}_{T+h|T} = \hat{y}_{T+1|T}$

Простое экспоненциальное сглаживание

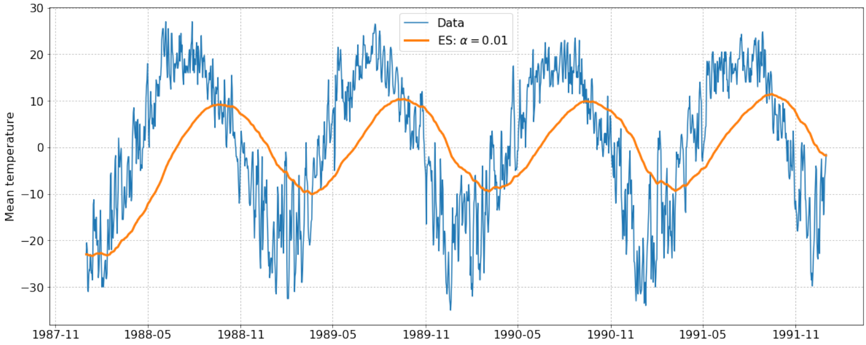
Оптимальное α^*:

$$\sum_{t=t_0}^{T} (\widehat{y}_t(\alpha) - y_t) \to \min_{\alpha}$$

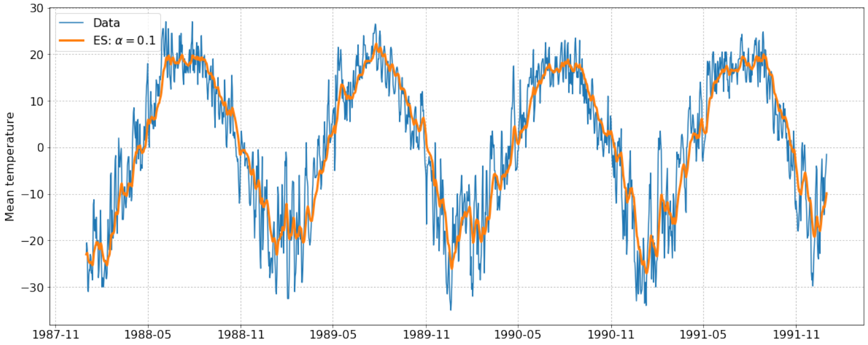
Эмпирические правила:

- если $\alpha^* \in (0, 0.3)$ то ряд стационарен, можно применять экспоненциальное сглаживание;
- если $\alpha^* \in (0.3, 1)$ то ряд нестационарен, нужно применять модель тренда.

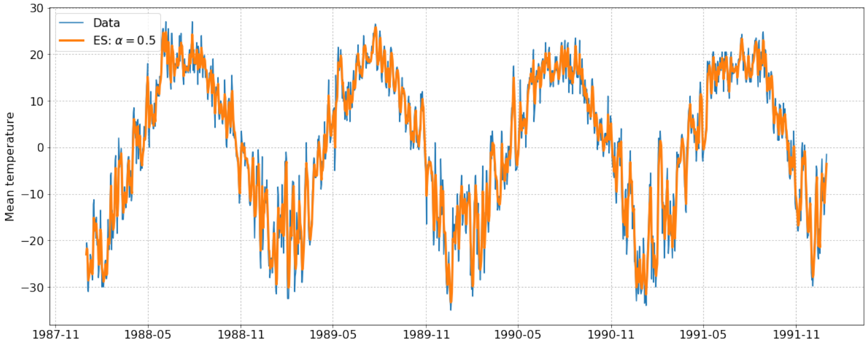
Простое экспоненциальное сглаживание



Простое экспоненциальное сглаживание



Простое экспоненциальное сглаживание



Учет тренда

Аддитивный линейный тренд (метод Хольта):

$$\widehat{y}_{t+d|t} = l_t + db_t,$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

Мультипликативный линейный тренд:

$$\widehat{y}_{t+d|t} = l_t b_t^d,$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} b_{t-1})$$

$$b_t = \beta \frac{l_t}{l_{t-1}} + (1-\beta)b_{t-1}$$

Учет сезонности

Аддитивная сезонность с периодом длины m

(метод Хольта-Уинтерса):

$$\widehat{y}_{t+d|t} = l_t + db_t + s_{t-m+(d \bmod m)},$$

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

Учет сезонности

Мультипликативная сезонность:

$$\widehat{y}_{t+d|t} = (l_t + db_t)s_{t-m+(d \bmod m)},$$

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

$$s_t = \gamma \frac{y_t}{l_{t-1} - b_{t-1}} + (1-\gamma)s_{t-m}$$

Адаптивное экспоненциальное сглаживание

Пусть $\widehat{\varepsilon}_t = y_t - \widehat{y}_t$ — ошибка прогноза, сделанного на шаге $t-1$.

$E_t = \gamma\widehat{\varepsilon}_t + (1-\gamma)E_{t-1}$ — среднее значение ошибки

$A_t = \gamma\,|\widehat{\varepsilon}_t| + (1-\gamma)A_{t-1}$ — средний разброс ошибки

Берем

$$\alpha_t = \min\left(\frac{|E_t|}{|A_t|}, 1\right)$$

Обычно берут $\gamma \in (0.05, 0.1)$.

3. Качество моделей

Метрики качества

Средняя квадратичная ошибка

$$MSE = \frac{1}{T-R+1}\sum_{t=R}^{T}(\widehat{y}_t - y_t)^2 \,.$$

Средняя абсолютная ошибка

$$MAE = \frac{1}{T-R+1}\sum_{t=R}^{T}|\widehat{y}_t - y_t| \,.$$

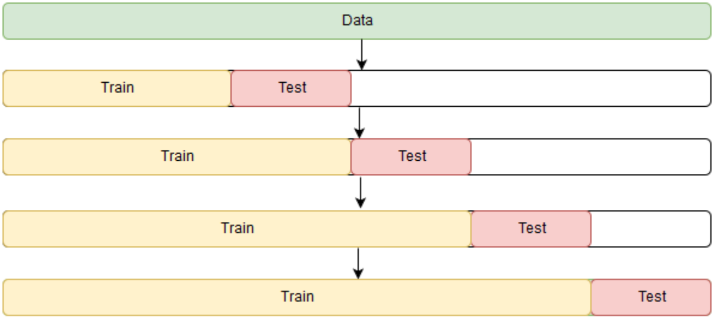
Средняя абсолютная ошибка в процентах

$$MAPE = \frac{100}{T-R+1}\sum_{t=R}^{T}\left|\frac{\widehat{y}_t - y_t}{y_t}\right| \,.$$

Симметричная средняя абсолютная ошибка в процентах

$$SMAPE = \frac{200}{T-R+1}\sum_{t=R}^{T}\left|\frac{\widehat{y}_t - y_t}{\widehat{y}_t + y_t}\right| \,.$$

Кросс-валидация для временных рядов



Кросс-валидация для временных рядов

Как выбрать лучшую модель среди тех, которые обладают хорошими свойствами?

1. Считаем качество прогнозов $\widehat{y}_{t_0+1|t_0}, \dots, \widehat{y}_{t_0+\Delta t|t_0}$
2. Считаем качество прогнозов $\widehat{y}_{t_0+\Delta t+1|t_0+\Delta t}, \dots, \widehat{y}_{t_0+2\Delta t|t_0+\Delta t}$
3. ...
4. Считаем качество прогнозов $\widehat{y}_{t_0+k\Delta t+1|t_0+k\Delta t}, \dots, \widehat{y}_{t_0+(k+1)\Delta t|t_0+k\Delta t}$
5. Усредняем полученные значения качества.
6. Выбираем модель, которая показывает наилучшее усредненное качество.

4. Практическое задание 9*