Временные ряды и модели ARIMA

Евгений Борисов

Временные ряды: основные определения и методы

временной ряд - признак, значения которого измеряется через постоянные интервалы времени.

если промежутки изменяются (случайно), то это уже не ВР но случайный процесс.

особенность датасета — неслучайная выборка, зависимость данных, будущее зависит от прошлого

Временные ряды: основные определения и методы

временной ряд - признак, значения которого измеряется через постоянные интервалы времени.

если промежутки изменяются (случайно), то это уже не ВР но случайный процесс.

особенность датасета - зависимость данных, будущее зависит от прошлого

задача прогнозирования

$\{y_1, \ldots, y_t\}$ $\qquad y_{t+h} = f(h | y_1, \ldots, y_t)$ $\qquad h \in \{1, 2, \ldots, H\}$

H — горизонт прогнозирования

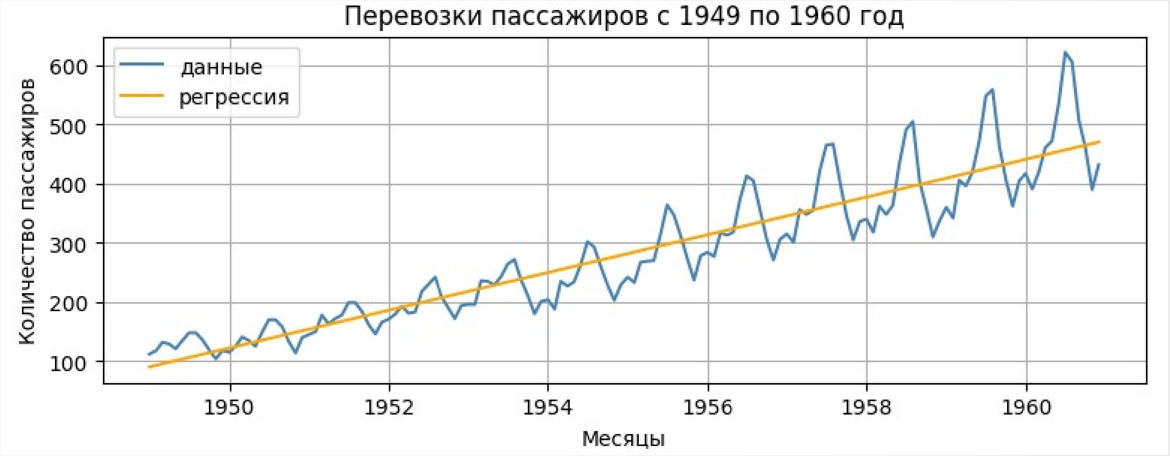
Временные ряды: основные определения и методы

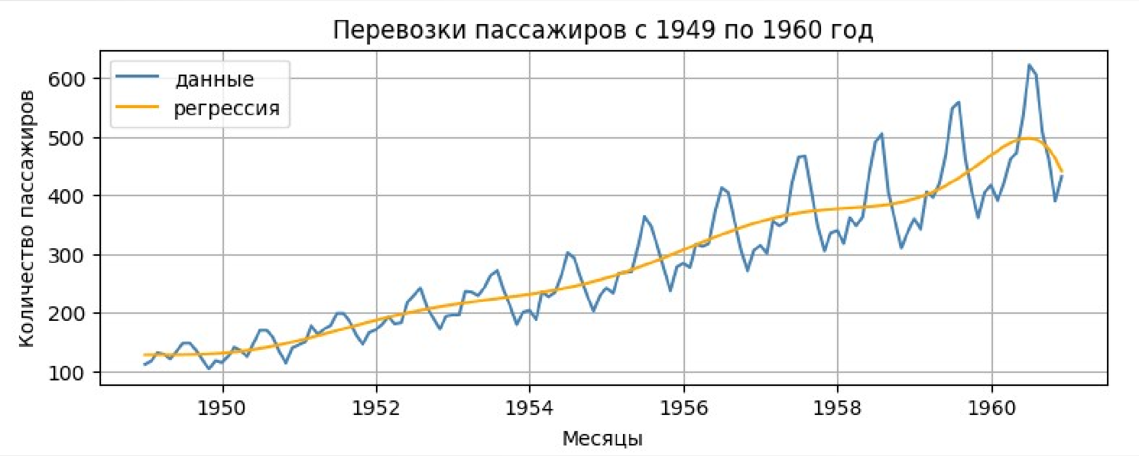
особенность датасета

неслучайная выборка

зависимость данных, будущее зависит от прошлого

простая регрессия плохо справляеться с задачей

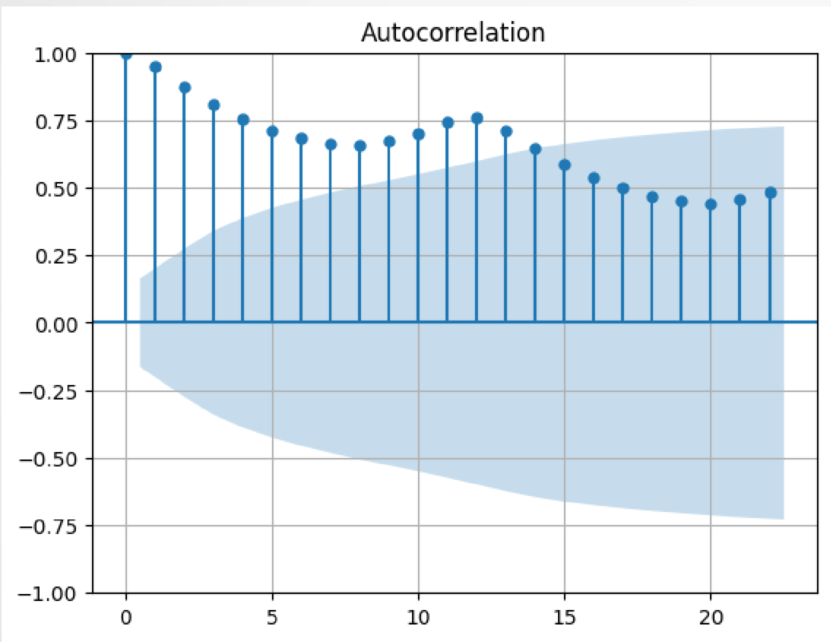




Временные ряды: основные определения и методы

Как определить, что данные неслучайные
но упорядоченные и зависимые (будущее зависит от прошлого) ?

автокорреляция - корреляция Пирсона ряда с тем же рядом сдвинутым на t шагов (*лаг автокорреляции*)

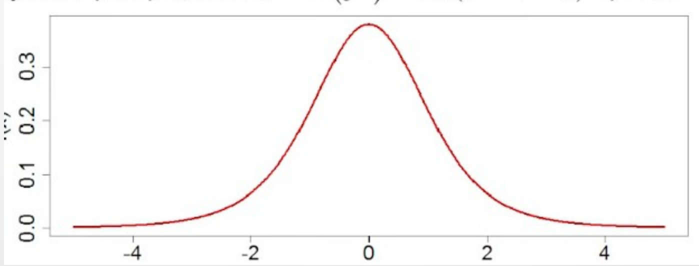


палки — значение автокореляции для разных лагов ;

синяя область — коридор значимости,
палки вне коридора значимо отличаються от нуля ;

значимость автокореляции для разных лагов
проверяем с помощью критерия Стьюдента

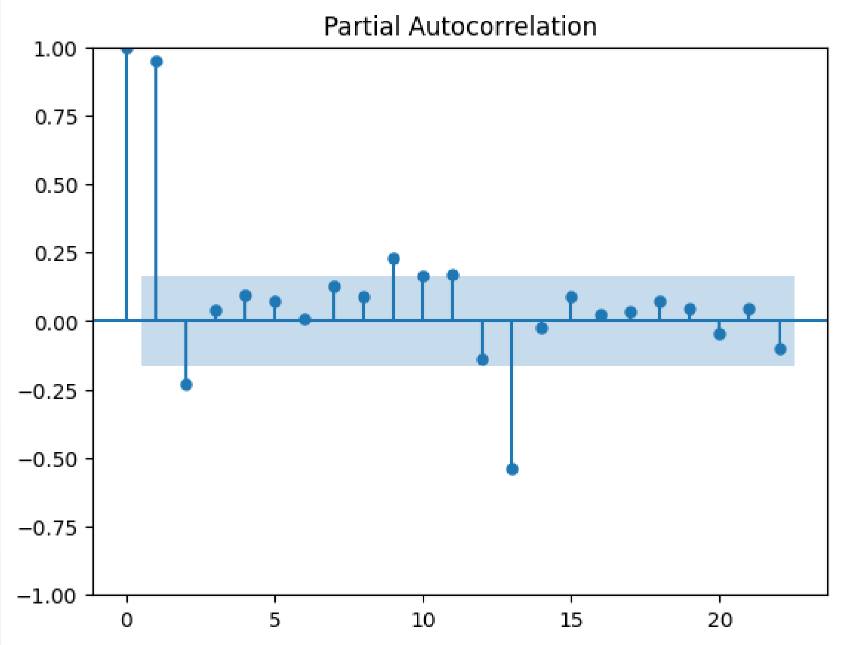
временной ряд: $y^T = y_1, \dots, y_T$;
нулевая гипотеза: $H_0 : r_\tau = 0$;
альтернатива: $H_1 : r_\tau \neq 0$;
статистика: $T\left(y^T\right) = \frac{r_\tau \sqrt{T-\tau-2}}{\sqrt{1-r_\tau^2}}$;
нулевое распределение: $T\left(y^T\right) \sim St\left(T-\tau-2\right)$ при H_0.



Временные ряды: основные определения и методы

Как определить, что данные неслучайные
но упорядоченные и зависимые (будущее зависит от прошлого) ?

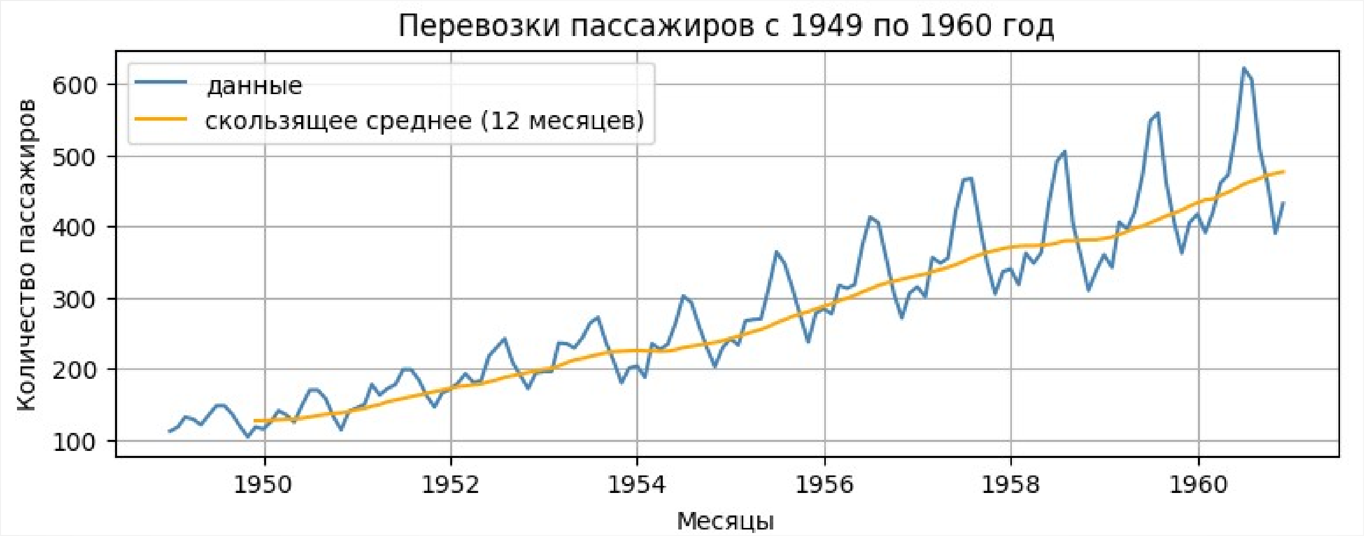
Частичная автокорреляция - дополнительно удаляет линейную зависимость между сдвинутыми рядами



http://www.machinelearning.ru/wiki/index.php?title=Частичная_автокорреляция

Временные ряды: основные определения и методы

Свойства временого ряда



тренд - долгосрочное (плавное) изменение ряда

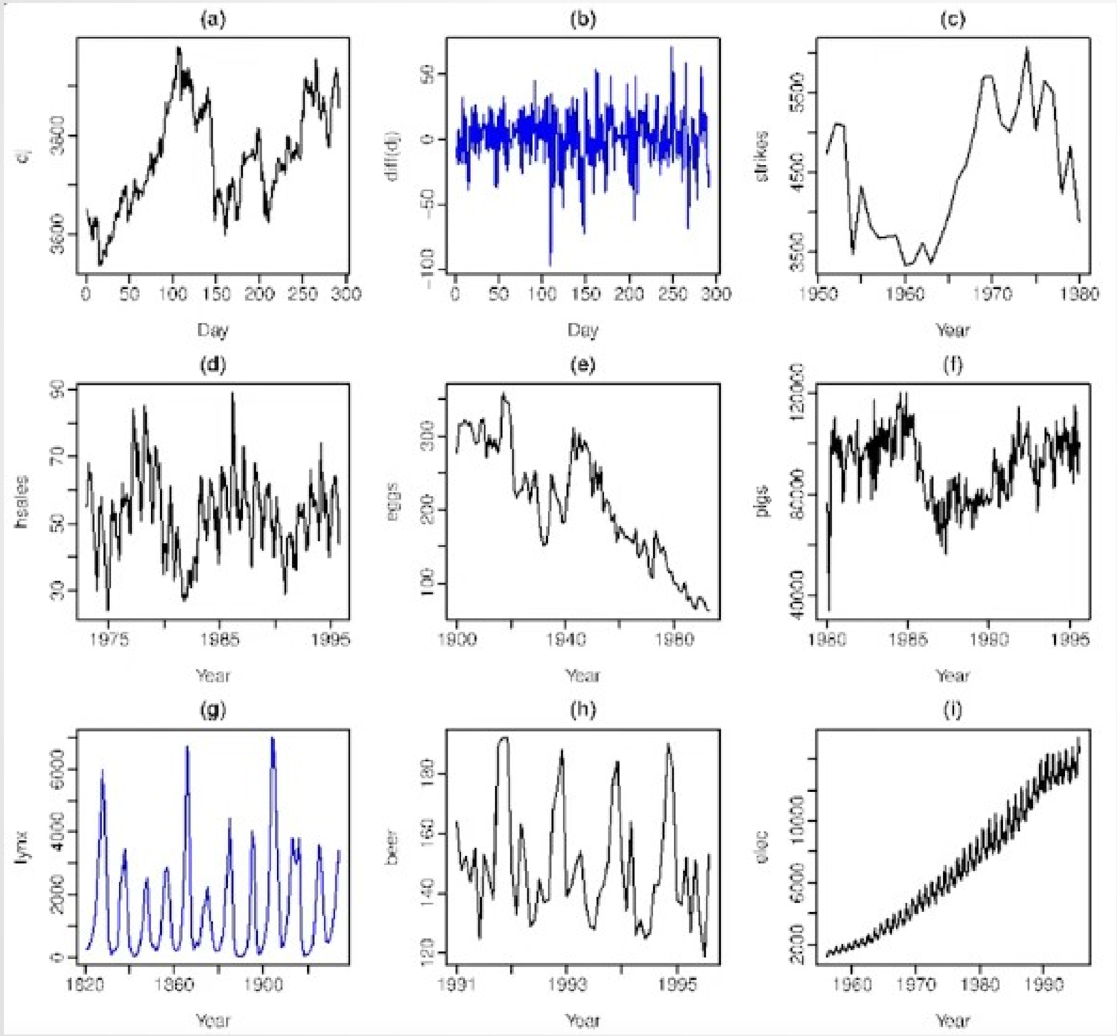
сезонность - циклические изменения уровня ряда с **постоянным** периодом (зима/лето)

цикл - циклические изменения уровня ряда с **НЕпостоянным** периодом (экономические циклы)

ошибка - непрогнозируемая случайная компонента ряда

Временные ряды: основные определения и методы

стационарность - распределение в любом временном окне ряда одинаковое



Пример: **синие** — стационарны ;

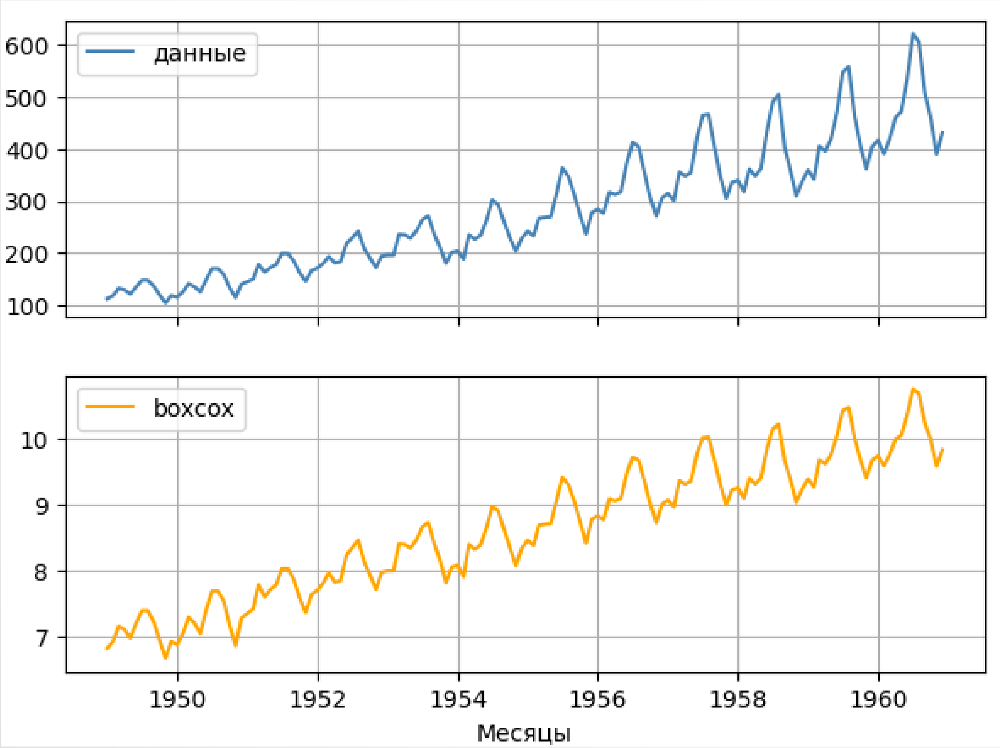
(тренд и сезонность нестационарны)

критерии нестационарности ряда Дики-Фуллера

Временные ряды: основные определения и методы

Преобразование ряда в стационарный

стабилизация дисперсии



преобразование Бокса-Кокса
(BoxCox transform)

логарифмирование ряда,

сглаживает нарастающую амплитуду колебаний

значения должны быть > 0.
сдвигаем на константу,
выполняем преобразование,
сдвигаем обратно

для выдачи окончательного прогноза необходимо
применить обратное преобразование

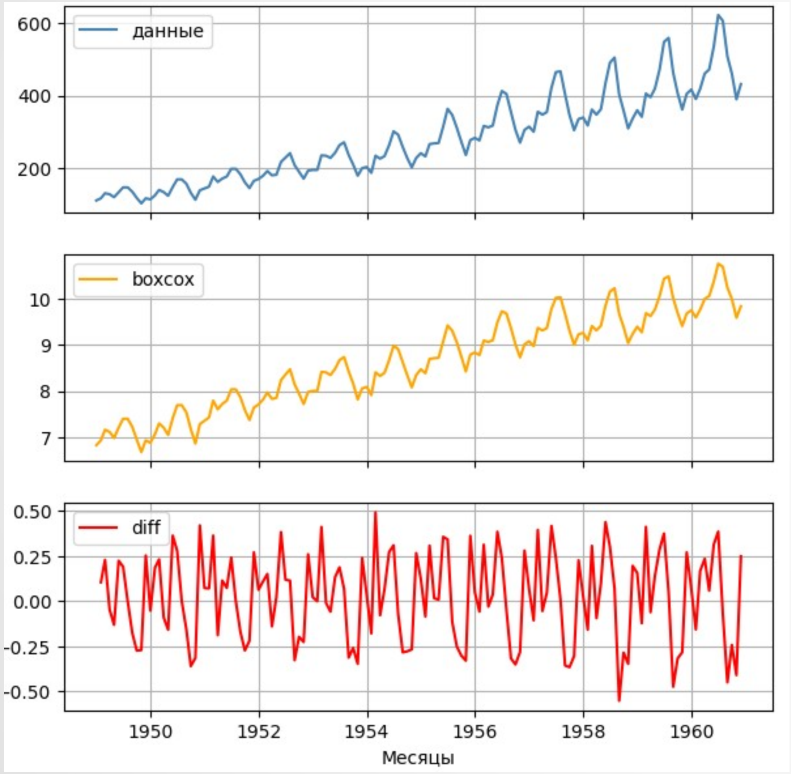
$$y_i^{\lambda} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(y_i), & \text{if } \lambda = 0. \end{cases}$$

однопараметрическое преобразование Бокса-Кокса с параметром λ

Временные ряды: основные определения и методы

Преобразование ряда в стационарный

стабилизация дисперсии
стабилизация тренда



дифференцирование ряда - переход к попарным разницам:

преобразует ряд в стационарный, прибивает тренды ;
можно применять несколько раз последовательно ;

(применяем после стабилизации дисперсии с помощью BoxCox)

$$dy_t = y_t - y_{t-1}$$

обратное преобразование:

$$y_t = dy_t + y_{t-1}$$

сезонное дифференцирование (вычитаем из текущего декабря предыдущий декабрь)
т.е. переход к попарным разностям в соседних сезонах
для прибивания сезонности

Временные ряды: модели для прогнозирования

Модели прогноза

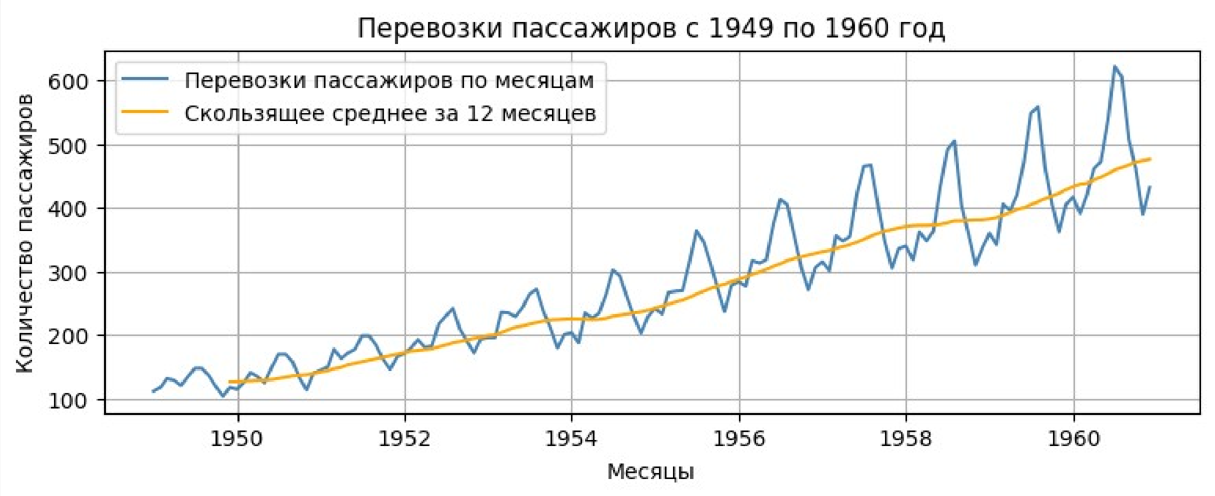
- тривиальный прогноз

- авторегрессия (AR)

- шум и скользяцее среднее (MA)

- комбинированные модели (SARIMAX)

Временные ряды: модели для прогнозирования



тривиальный прогноз

прогноз равен текущему значению

$$f(y_1, \ldots, y_t) = y_t$$

работает для прогноза погоды

Временные ряды: модели для прогнозирования

авторегрессия - регрессия для предсказывания следующего значения по предыдущим

$$AR(p): y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$$

$\alpha - const;$
$y_{t-i} - \text{значение в момент } t-1;$
$\phi_i - \text{коэффициент};$
$\epsilon_t - \text{гаусов шум, с нулевым мат. ожиданием и некоторой постоянной дисперсией } \sigma_\epsilon^2;$

может описывать стационарный ряд

Временные ряды: модели для прогнозирования

скользящее среднее - авторегрессия на шум

моделируем значение через гаусовский шум

$$MA(q): y_t = \alpha + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$$

$\alpha - const;$

$\theta_i - коэффициент;$

$\epsilon_t - гаусов\, шум, с\, нулевым\, мат.\, ожидаением\, и\, некоторой\, постоянной\, дисперсией\, \sigma_\epsilon^2;$

Временные ряды: модели для прогнозирования

ARMA - авторегрессионная модель скользящего среднего

комбинация AR(p) и MA(q)

$$ARMA(p,q): y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$$

$y_{t-i} - \text{стационарный ВР};$

$\alpha - const;$

$\phi_i - \text{коэффициент};$

$\theta_i - \text{коэффициент};$

$\epsilon_t - \text{гаусов шум, с нулевым мат. ожидаением и некоторой постоянной дисперсией } \sigma_\epsilon^2;$

теорема Вольда: стационарный ряд может быть описан моделью ARMA с любой точностью

Временные ряды: модели для прогнозирования

SARMA - ARMA с учётом сезонности периода S

$$SARMA(p,q)x(P,Q): y_t=\alpha+\sum_{i=1}^{p}\phi_i y_{t-i}+\sum_{i=1}^{q}\theta_i\epsilon_{t-i}+\epsilon_t+\sum_{i=1}^{P}\phi_{i\dot{S}} y_{t-i\dot{S}}+\sum_{i=1}^{Q}\theta_{i\dot{S}}\epsilon_{t-i\dot{S}}$$

y_{t-i} − стационарный ВР;

α − const;

ϕ_i − коэффициент;

θ_i − коэффициент;

ϵ_t − гаусов шум, с нулевым мат. ожидаением и некоторой постоянной дисперсией σ_ϵ^2;

дополнительные компоненты с шагом S

Временные ряды: модели для прогнозирования

ARIMA - autoregressive integrated moving average

ARIMA(p,d,q) - к d раз продифференцированному ВР применяем ARMA(p,q)

Временные ряды: модели для прогнозирования

SARIMA - добавление сезонности к ARIMA

SARIMA(p,d,q) x (P,D,Q)

применяем d раз обычное дифференцирование
и D раз - сезонное (с шагом S)

Временные ряды: модели для прогнозирования

SARIMAX - дополнительные признаки к SARIMA

например: бинарный признак "в этот день праздник"

Временные ряды: оценка результата

оценка результата - метрика AIC (критерий Акаике)

остаток - разница между фактом и прогнозом

остатки должны быть

- несмещённые т.е. среднее должно быть близко к нулю

- стационарными т.е. не зависеть от времени

- неавтокоррелироваными

Временные ряды: оценка результата

Подбор гиперпараметров модели SARIMA (p,d,q),(P,D,Q)

d,D (количество дифференцирований) выбираем минимально необходимое для стационарности ряда

строим график автокорреляции

q - номер последнего НЕсезонного лага, при котором автокорреляция значима, но меньше длинны сезона ;
Q - номер последнего сезонного лага, при котором автокорреляция значима ;

строим график частичной автокорреляции

p - номер последнего НЕсезонного лага, при котором автокорреляция значима, но меньше длинны сезона ;
P - номер последнего сезонного лага, при котором автокорреляция значима ;

Временные ряды: схема применения моделей

Общая схема построения модели ARIMA

- визуальная оценка графика ряда
- стабилизируем дисперсию (Бокс-Кокс)
- преобразуем в стационарный, подбираем порядок дифференцирования d,D (критерий Дики-Фуллера)
- анализируем автокорреляции подбираем p,q и P,Q
- обучаем модель
- выполняем обратные преобразования (дифференцирование,логарифмирование) для выдачи прогноза
- оцениваем по AIC (критерий Акаике)
- оцениваем остатки (разница между прогнозом и тестовыми данными)

эвристики для улучшения результата

- суммы за месяц можно делить на количество дней в месяце
- если в начале ряда имеем аномалию, то её можно обрезать
- выбросы оказывают значительное влияние на результат ARIMA, их стоит выкинуть (заменить на N/A)

Временные ряды: литература

Борисов Е.С. Методы машинного обучения. 2024
https://github.com/mechanoid5/ml_lectorium_2024_I

Дмитрий Макаров   Временные ряды
https://www.dmitrymakarov.ru/intro/time-series-20/

Евгений Рябенко   Прогнозирование временных рядов
https://www.youtube.com/watch?v=u433nrxdf5k

Рон Хайндман и Джордж Атанасопулос
Прогнозирование: принципы и практика / пер. с англ. А. В. Логунова. – М.:
ДМК Пресс, 2023. – 458 с.: ил.
ISBN 978-5-93700-151-1