README для задания 2

Пожилая саламандра

December 2019

Содержание

1. Теория временных рядов

1.1. Определения из курса ТВиМС

1.1.1. Теория вероятности

Случайная величина (с.в.) - числовая функция, заданная на некотором вероятностном пространстве $(\Omega, P) : X = X(\omega), \omega \in \Omega$.

Функция распределения с.в. - числовая функция числового аргумента, определяемая равенством:

$$F(x) = P(X \leq x), \quad x \in R \tag{1}$$

(R - множество действительных чисел).

Существует два класса с.в. - *дискретные* и *непрерывные*.

С.в. называется **дискретной**, если множество её значений конечно или счётно. Одно из представлений дискретной случайной величины:

$$X = (\overline{x}, \overline{p} : p_k = P(X = x_k)), \tag{2}$$

где k - не более, чем счётное.

С.в. называется **непрерывной**, если её функция распределения дифференцируема, т.е. существует производная $p(x) = F'(x)$, называемая **плотность распределения** с.в. X.

В частности, $F(x) = \int\limits_{-\infty}^{x} p(y)dy$.

Математическое ожидание (м.о.) *(среднее значение)* дискретной с.в. X, имеющей распределение (2) - есть по определению[1] ряд

$$E(X) = \sum_k x_k p_k \tag{3}$$

Для непрерывной случайной величины X с плотностью распределения $p(x)$ м.о. - это интеграл[2]

$$E(X) = \int\limits_{-\infty}^{+\infty} xp(x)dx \tag{4}$$

[1]При условии его абсолютной сходимости
[2]Также при условии,что он абсолютно сходится

Дисперсия с.в. X - числовая характеристика, отражающая степень «разброса» случайной величины относительно среднего значения. Она определяется равенством

$$Var(X) = E(X - EX)^2. \tag{5}$$

1.1.2. Математическая статистика

Случайной выборкой *объема n* называется последовательность наблюдений $X_1, \ldots, X_n$, если они получены как независимые реализации некоторой с.в. X с распределением $F(x)$.

Выборочными статистиками $X_1, \ldots, X_n$ называются следующие величины:

- выборочное среднее:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i; \tag{6}$$

- выборочная дисперсия:

$$Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2; \tag{7}$$

- размах:

$$d = \max_{1\le i\le n} X_i - \min_{1\le i\le n} X_i \tag{8}$$

  Если есть ещё одна случайная выборка $Y_1, \ldots, Y_n$, то определяются также:

- выборочная ковариация:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}); \tag{9}$$

- выборочная корреляция:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{10}$$

1.2. Временные ряды

Временной ряд - собранный в разные моменты времени статистический материал о значении каких-либо параметров (в простейшем случае одного) исследуемого процесса.

Стационарность временного ряда в узком смысле - ряд y_t^0 называется строго стационарным (*strictly stationarity*) или стационарным в узком смысле, если совместное распределение m наблюдений $y_{t_1}^0, \ldots, y_{t_m}^0$ не зависит от сдвига во времени, то есть совпадает с распределением $y_{t_1+t}^0 \ldots, y_{t_m+t}^0$ для любых $m, t, t_1, \ldots, t_m$.

Стационарность временного ряда в широком смысле - ряд y_t называется слабо стационарным (*weak stationarity*) или стационарным в широком смысле, если такие статистические характеристики временного ряда как его математическое ожидание (среднее), дисперсия (ср. кв. отклонение) и ковариация не зависят от момента времени:

$$E(y_t) = \mu < \infty, \quad Var(y_t) = \gamma_0, \quad Cov(y_t, y_{t-k}) = \gamma_k \tag{11}$$

Конечно, из строгой стационарности следует слабая стационарность (при условии конечности первого и второго моментов распределения). В дальнейшем мы будем везде под «стационарностью» понимать *слабую* стационарность.

Авторегрессионная (AR-) модель (англ. *autoregressive model*) — модель временных рядов, в которой значения временного ряда в данный момент линейно зависят от предыдущих значений этого же ряда. Авторегрессионный процесс порядка p (AR(p)-процесс) определяется следующим образом:

$$X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t, \tag{12}$$

где где $a_1, \ldots, a_p$ — параметры модели (коэффициенты авторегрессии), c — постоянная (часто для упрощения предполагается равной нулю), ε_t — белый шум.

Модель авторегрессии - скользящего среднего (АРСС) - *autoregressive moving average (ARMA)* - одна из математических моделей, использующихся для анализа и прогнозирования стационарных временных рядов в статистике. Модель ARMA обобщает две более простые модели временных рядов - модель авторегрессии (AR) и модель скользящщего среднего (MA).

Моделью ARMA(p,q), где p и q - целые числа, задающие порядок модели, называется следующий процесс генерации временного ряда:

$$x_k = c + \varepsilon_k + \sum_{i=1}^{p} a_i x_{k-i} + \sum_{i=1}^{q} b_i \varepsilon_{k-i} \tag{13}$$

где a_i и b_i - параметры модели (действительные числа, соответственно, авторегресссионные коэффициенты и коэффициенты скользящео среднего); c - константа, ε_k - белый шум.

ARIMA (autoregressive integrated moving average); модель Бокса - Дженкинса - расширение моделей ARMA для нестационарных рядов, которые можно сделать стационарными взятием разностей некоторого порядка от исходного временного ряда (так называемые интегрированные или разностно-стационарные ряды). Модель ARIMA(p, d, q) означает, что разности временного ряда порядка d подчиняются модели ARIMA(p, q).

2. Программная реализация на языке Python

2.1. Используемые библиотеки

- numpy
- pandas
- xlrd
- statistics
- statsmodels.
  - tsa.adfvalues
  - compat.python
  - graphics.tsaplots.plot_acf, plot_pacf
  - regression.linear_model.OLS
  - tsa.arima_model
  - tools.add_constant
- sklearn.metrics.mean_squared_error
- sklearn.metrics.r2_score
- warnings

2.2. Список реализованных функций

- diff_operator
- AIC_finder
- arima_optimizer_AIC
- arima_learn_forecast
- arima_learn_predict
- best_train_finder
- integral_definer
- avg_data
- white_noise
- get_lag

- df_test
- series_seasonal
- series_decompose_sum - разложение ряда по + модели
- series_decompose_mul - разложение ряда по × модели

3. Ход решения задачи

3.1. Проверка стационарности ряда

3.1.1. Тест Дики-Фуллера

Тест Дики-Фуллера используется для проверки ряда на стационарность, а также для проверки гипотезы о единичном корне (что, по сути, одно и то же), то есть гипотезы о равенстве коэффициента a в AR(1):

$y_t = a * y_{t-1} + \varepsilon_t,$

где y_t - временной ряд, а ε - ошибка.
Авторегрессионное уравнение AR(1) можно также представить в виде:

$\Delta y_t = b * y_{t-1} + \varepsilon_t,$

где $b = a - 1$, а $\Delta y_t = y_t - y_{t-1}$
Если ряд, полученный из первых разностей элементов исходного ряда, - стационарный, то гипотеза о единичном корне принимается.
Результатом теста является DF-статистика - t-статистика (не путать с распределением Стьюдента) для проверки значимости коэффициентов линейной регрессии. Если полученная статистика больше критического значения данной статистики, то ряд является нестационарным, иначе - стационарным.
На вход функции df_test() подается $NumPyArray$ со значениями ряда. Вычисляется максимальное значение лага, то есть временной сдвиг (количество предшествующих элементов, участвующих в разложении текущего). После получения массива с первыми разностями значений временного ряда строится полная карта лагов. Вычисляется стартовое значение лага, а затем оптимальное значение лага (при помощи функции $getlag()$, которая путем минимизации критерия Акаике подбирает оптимальные информационный критерий и лаг), после чего снова строится карта лагов. После вычисления необходимых данных строится методом наименьших квадратов аппроксимационная функция значений временного ряда, из которой далее получаем необходимую DF-статистику. После сравнения полученной статистики с

соответствующим критическим значением делаем вывод о стационарности или её отсутствии.

3.1.2. Выводы и оценка достоверности статистики

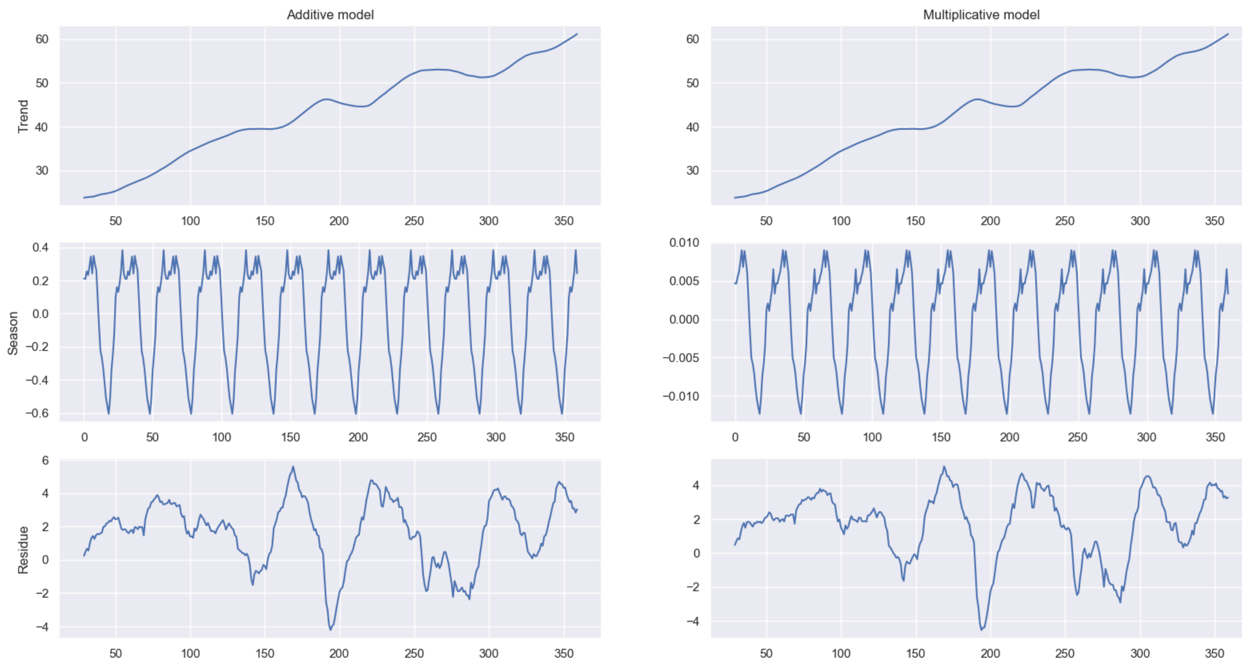
Ряд нестационарен, достоверность оценивается отрисовкой графика.

3.2. Разложение временного ряда

3.2.1. Тренд, сезональность, остаток (по $+, \times$ -ым моделям)

Существует 2 модели временных рядов - аддитивная и мультипликативная. Рассмотрим первую из них.
В аддитивной модели временной ряд может быть представлен в виде:
$Y = T + S + E$,
где T - трендовая компонента, Y - циклическая (сезонная) компонента, а E - остаточная часть.
Трендовая составляющая находится через скользящее среднее окно, то есть элементы, начиная с индекса $window-1$ выражаются через среднее арифметическое среднее предыдущих window элементов. В нашем случае $window = 30$, то есть это месячное катящееся среднее. Устраняя трендовую компоненту из исходного ряда и находя среднее арифметическое элементов, находящихся на позициях $i + window$, где i пробегает значения от 0 до $window - 1$ (то есть суммируем каждые window элементов, делим сумму на их количество, сдвигаемся на 1 вправо и повторяем предыдущие операции до самого конца), получаем сезонну компоненту ряда. Далее, вычитая из исходного ряда трендовую и сезонную компоненты, получаем остаточную часть.
В мультипликативной модели временной ряд может быть представлен в виде:
$Y = T \times S \times E$,
где T - трендовая компонента, Y - циклическая (сезонная) компонента, а E - остаточная часть.
Алгоритм нахождения составляющих такой же, как и для аддитивной модели, только перед нахождением сезонной составляющей мы делим исходный ряд на тренд и далее работаем с полученным рядом.
Функции $seriesdecomposesum()$ и $seriesdecomposemul()$ получают на вход исходный ряд и размер временного промежутка. Далее в них происходит алгоритм, описанный выше. Функция series_seasonal() получает на вход временной ряд без тренда, размер окна и находит сезонную составляющую ряда, которая потом "размножается"на весь временной промежуток.

3.2.2. Анализ и визуализация



На графиках видно, что тренд присутствует и возрастающий, но сезонная компонента крайне незначительна и колеблется от -0.6 до 0.4 на аддитивной модели, а на мультипликативной и то меньше.

3.3. Проверка на интегрированность порядка k

Разностный оператор:

$$\nabla = 1 - B \tag{14}$$

$$X_t \nabla = X_t - X_{t-1} \tag{15}$$

Полином $AR:\quad \phi(z) = 1 - \phi_1 z \ldots \phi_p z^p$

Оператор $AR:\quad \phi(B) = 1\phi_1 B \ldots \phi_p B^p$

Полином $MA:\quad \theta(\mathrm{z}) = 1 + \theta_1 z + \ldots + \theta_q z^q$

Оператор $MA:\quad \theta(\mathrm{B}) = 1 + \theta_1 B + \ldots + \theta_q B^q$

$ARMA(p, q)$: X_t - это $ARMA(p, q)$ процесс, если он стационарен и имеется белый шум Z_t, :$\phi(B) X_t = \theta(\mathrm{B}) \mathrm{Z}_t$

$ARIMA(p, d, q)$: X_t - это $ARIMA(p, d, q)$ процесс, если $\theta^d \mathrm{X}_t$ - это $ARMA(p, q)$: $\phi(B) \nabla^d X_t = \theta(B) Z_t$

Ряд называется интегрируемым порядка k, если его разности порядка $k - 1$ включительно нестационарны, а k-я разность — стационарна.

3.4. Применение модели ARIMA

3.4.1. Подборка необходимых параметров с помощью функций автокорреляции, частичной автокорреляции

Параметры модели $AR(p)$ подбираются через график $PACF$ (функции частичной автокорреляции). Выбор параметров происходит выбор наиболее выбивающихся точек из синей области графика.

Аналогично происходит выбор параметров модели $MA(q)$.

3.4.2. Отбор моделей

Отбор моделей происходит иттерационным методом "пробега"по возможным комбинациями ARIMA(p,d,q) модели, где p - возможные параметры AR(p), d - порядок интегрированности, а q - возможные параметры MA(q).

Выбор наилучшей модели происходит по наименьшему показателю AIC для итогового теста.

3.5. Работа с тестовой выборкой

3.5.1. Предсказание значений

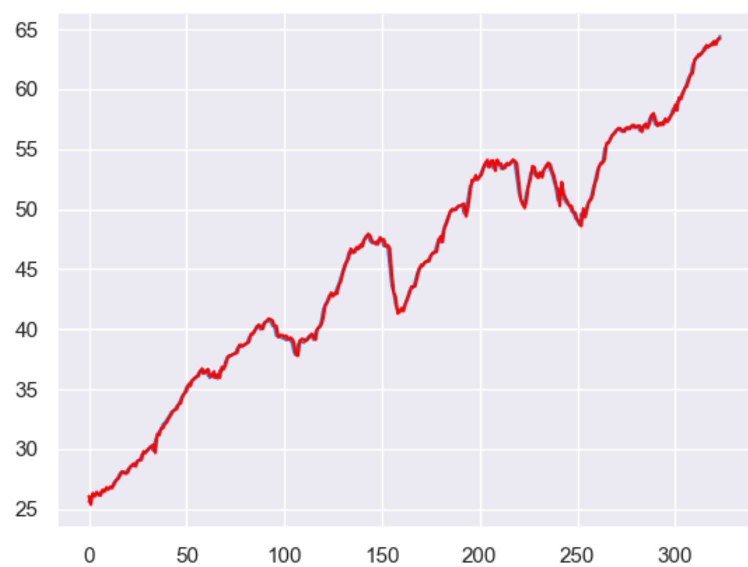
В нашей программе реализованно две различных функции предсказания значений с помощью $ARIMA(p, d, q)$.

1) forecast() - для реализации One-Step Out-of-Sample Forecast, когда $ARIMA(p, d, q)$ дообучается на каждом шаге прохождения тестовой таблицы.

2) $predict()$ - для реализации полного предсказания n шагов вперёд.

Очевидно, что показатели AIC (информационный критерий Акаике) и r^2score лучше у 1 реализации.

3.5.2. Визуализация, подсчёт r^2score



Визуализация проводится через $mathplotlib$ с использованием $seaborn$, а также r^2score берётся из библиотеки $sklearn$

r^2score - коэфициент детерминации - это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. От меньше 1 и может быть отрицательным.

3.5.3. Отбор наилучших моделей с помощью информационного критерия Акаике

Информационный критерий Акаике - AIC - критерий, применяющийся исключительно для выбора из нескольких статистически моделей.

В общем случае:

$AIC = 2k - 2ln(L)$, где k — число параметров в статистической модели, L — максимизированное значение функции правдоподобия модели. Чем AIC меньше, тем лучше подобрана модель.

4. Список участников и их вклад в проект

- Денисов Никита - тест Дики-Фуллера, разложение ряда
- Ловягин Андрей - $ARIMA$
- Иванков Михаил - компиляция README, графики, матчасть

5. Использованная литература

- Эконометрика. Начальный курс. Магнус Я.Р., Катышев П.К., Пересецкий А.А.: 6-е изд., перераб. и доп. - М.: Дело, 2004.
- Лекции по временным рядам. Канторович Г.Г.
- wikipedia.org, англ. и рус. версия, различные статьи
-