Московский государственный университет
имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

Отчет по заданию №2

Выполнили:
Долгая Л. В. 311гр.
Шумилин Я. Т. 312гр.
Рзянина А. Т. 312гр.

Москва
2019

Содержание

Постановка задачи

1. Считать данные из training.xlsx. Ответы на тестовой выборке testing.xlsx не следует использовать ни в каких экспериментах, кроме финального. Проверить является ли ряд стационарным в широком смысле.
Это можно сделать двумя способами:

- Провести визуальную оценку, отрисовав ряд и скользящую статистику (среднее, стандартное отклонение). Постройте график на котором будет отображен сам ряд и различные скользящие
- Провести тест Дики - Фуллера.

Сделать выводы из полученных результатов. Оценить достоверность статистики.
2. Разложить временной ряд на тренд, сезональность, остаток в соответствии с аддитивной, мультипликативной моделями.
Визуализировать их, оценить стационарность получившихся рядов, сделать выводы.
3. Проверить является ли временной ряд интегрированным порядка k. Если является, применить к нему модель ARIMA, подобрав необходимые параметры с помощью функции автокорреляции и функции частичной автокорреляции. Выбор параметров обосновать. Отобрать несколько моделей. Предсказать значения для тестовой выборки. Визуализировать их, посчитать r2 score для каждой из моделей. Произвести отбор наилучшей модели с помощью информационного критерия Акаике. Провести анализ получившихся результатов.
Помимо этого обязательные пункты к выполнению:

- соблюдение PEP8
- использование для визуализации библиотек bokeh или seaborn
- оформление файла readme.pdf
- прохождение интервью

Метод решения

Вводные понятия

Временной ряд – это последовательность значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени, обычно через равные промежутки.

Анализ временных рядов - совокупность математико-статистических методов анализа, предназначенных для выявления структуры временных рядов и для их прогнозирования.

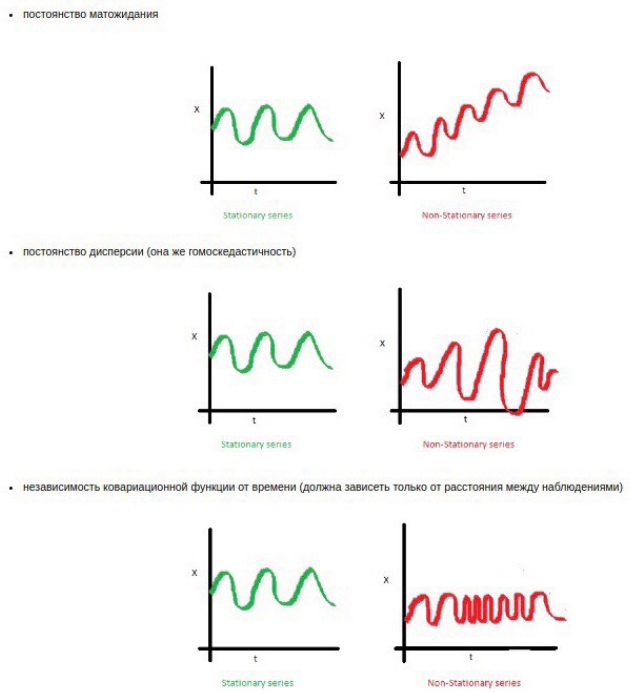
Тренд соответствует медленному изменению, проходящему в некотором определенном направлении, которое сохраняется в течение значительного промежутка времени.

Сезонные колебания соответствуют изменениям, которые происходят регулярно в течение года, недели или суток. Они связаны с сезонами и ритмами человеческой активности.

Дисперсия выборки - это среднее арифметическое квардратов отклонений.

Отклонение - это разность числа и некоторой точки отсчёта, чаще всего это среднее арифметическое или медиана.

Стационарность - свойство процесса не менять свои характеристики со временем. Временной ряд стационарен, если его свойства не зависят от времени.



Ковариация - в теории вероятностей и математической статистике мера линейной зависимости двух случайных величин.

Анализ временных рядов предполагает, что данные содержат систематическую составляющую (обычно включающую несколько компонент) и случайный шум (ошибку), который затрудняет обнаружение регулярных компонент.

Большинство регулярных составляющих временных рядов принадлежит к двум классам: либо они являются трендом, либо сезонной составляющей. Таким образом, каждый уровень временного ряда может формироваться из трендовой (T), циклической или сезонной компоненты (S), а также случайной (E) компоненты.

Модели, где временной ряд представлен в виде суммы перечисленных компонент, называется аддитивными, если в виде произведения - мультипликативными.

Аддитивная модель имеет вид: Y = T + S + E

Мультипликативная модель имеет вид: Y = T * S * E

Скользящая статистика - общее название для семейства функций, значения которых в каждой точке определения равны среднему значению исходной функции за предыдущий период. Скользяая статистика обычно используется с данными временных рядов для сглаживания краткосрочных колебаний и выделения основных тенденций и циклов.

Суть метода

1. Скользящая статистика:

- Среднее (Simple Moving Average)
  В общем случае, взвешенные скользящие средние вычисляются по формуле:

$$SMA_t = \frac{1}{n} * \sum_{i=1}^{n-1} p_{t-i} \tag{1}$$

- Стандартное отклонение (Standart Deviation)

$$SD_t = \sqrt{\frac{1}{n} * \sum_{i=1}^{n-1} (p_{t-i} - ME)^2} \tag{2}$$

  где n - окно скользящей статистики, а ME - среднее

2. Автокорреляционная функция (АКФ): Автокорреляционная функция - зависимость взаимосвязи между функцией (сигналом) и ее сдвинутой копией от величины временного сдвига. Для детермиированных сигналов АКФ сгнала f(t) определяется интегралом:

$$\int_{-\infty}^{\infty} f(t) * f(t - \tau) * dt \tag{3}$$

3. ARIMA Модель ARIMA(p, d, q) для нестационарного временного ряда X_t имеет вид:

$$\Delta^d * X_t = c + \sum_{i=1}^{p} a_i * \Delta^d * X_{t-1} + \sum_{j=1}^{q} b_i * \varepsilon_{t-j} + \varepsilon_t \tag{4}$$

где ε_t - стационарный временной ряд
c, a_i, b_i - параметры модели
Δ^d - оператор разности временного ряда порядка d (последовательное взятие d раз разностей первого порядка - сначала от временного ряда, затем от полученных разностей первого порядка, затем от второго порядка и т.д.)

Описание программы

Строится график временного ряда и осуществляется проверка временного ряда на стационарность. С помощью функций Series.rolling.mean() и Series.rolling.std() находятся среднее и стандартное отклонения соответственно, строятся их графики вместе с изначальным с разным окном. Определяем сезонность визуально по графику.

Разлагаем ряд на тренд, сезонность и остаток с помощью самостоятельно реализованной функции my_seasonal_decompose по аддитивной и мультипликативной модели. Далее идет проверка на стационарность с помощью проведения визуальной оценки. Определяем, что сезонность и остатки стационарны, а тренд нет, так как растет со временем.

Ряд проверяется на интегрируемость порядка k. Так как матожидание ряда растет со временем, ряд не стационарен. Для нахождения порядка k берется разность рядов. Если проведенный тест подтвердил предположения о не стационарности ряда, для нахождения k берется разность рядов. Если первые разности ряда стационарны, то он называется интегрированным рядом первого порядка. В нашем случае k = 1 и ряд интегрируем, значит с помощью функций автокорреляции и частичной автокорреляции подбираются параметры для модели ARIMA. Для построения модели нужно знать ее порядок, состоящий из трех параметров: p - порядок компоненты AR, d - порядок интегрированного ряда, q - порядок компоненты MA.

Параметр d равен 1, осталось определить p и q. Для их определения надо изучить автокорреляционную (ACF) и частично автокорреляционную (PACF) функции для ряда первых разностей. ACF поможет определить q, т. к. по ее коррелограмме можно определить количество автокорреляционных коэффициентов сильно отличных от 0 в модели MA. PACF поможет определить p, т. к. по ее коррелограмме можно определить максимальный номер коэффициента сильно отличного от 0 в модели AR. Чтобы построить соответствующие коррелограммы, в пакете statsmodels имеются функции: plot_acf() и plot_pacf(). Они выводят графики ACF и PACF, у которых по оси X откладываются номера лагов, а по оси Y значения соответствующих функций. В первой модели ACF экспоненциально затухает, начиная с первого лага, причем затухание может носить монотонный или колебательный характер. PACF затухает экспоненциально, монотонно или колебательно. Это означает, что p = 1, а q = 3. Вторая модель является наилучшей по AIC с учетом ограничений $q+p \leq 2$, а третья без учета этих ограничений. Четвертая модель не является оптимальной по AIC, но имеет меньшее значение r2 score.

Далее строятся модели ARIMA и осуществляется прогноз. Строится график, на котором изображены данные из файла testing.xlsx и построенный прогноз для каждой из моделей. Для каждой модели считается R2 - коэффициент детерминации, чтобы понять какой процент наблюдений описывает данная модель, и критерий Акаике (AIC), выбирающий наилучшую модель.

Вывод

Под стационарностью понимают свойство процесса не менять своих статистических характеристик с течением времени, а именно постоянство математического ожидания, постоянство дисперсии и независимость ковариационной функции от времени (должна зависеть только от расстояния между наблюдениями). Оригинальный ряд и трендовая составляющая для обеих моделей не стационарны, т.к. стандартное отклонение и скользящее среднее зависят от времени.

Сезонная составляющая и остаток (для обеих моделей) стационарны, их стандартное отклонение и скользящее среднее не зависят от времени. Если r2 score близок к 1, то условная дисперсия модели достаточно мала и весьма вероятно, что модель неплохо описывает данные. Если же он сильно меньше 1, то с большеи˘ долей уверенности модель не отражает реальное положение вещей. Значения наших моделей равны: -3.32, -3.28, -4.01 и 0.001. Это говорит о том, что они не являются точными. Считается, что наилучшей будет модель с наименьшим значением критерия Акаике. Однако погрешность этой модели велика, следовательно мы считаем модель с минимальным r2 score наилучшей в силу меньшей погрешности вычислений.

Необходимые компоненты

- Библиотеки

  – matplotlib - пакет, используемаый для отрисовки графиков.

  – statsmodels - пакет Python, который позволяет пользователям исследовать данные, оценивать статистические модели и выполнять статистические тесты. Он дополняет модуль статистики SciPy. Мы используем ее для построения модели ARIMA.

  – sklearn - пакет для машшиного обучения. Мы используем ее для оценки r2_value.

  – pandas - библиотека, предназначенная для хранения таблиц. Так же содержит огромное количество универсальных функций для их комфортной обработки.

  – pylab - большой универсальный пакет питон. Мы используем для задания параметров отрисовки.

  - seaborn - библиотека для визуализации.

- Программы

  Jupiter Notebook

Участники

- README.pdf - Долгая Л. В. Рзянина А. Т. Шумилин Я. Т.
- Задание №1, №2, №3 - Долгая Л. В. Шумилин Я. Т. Рзянина А. Т.