Компьютерный практикум. Задание 2.

Попова, Ситникова, Корнюхина. 311 группа.

Цели

1. Считать данные из training.xlsx. Проверить является ли ряд стационарным в широком смысле. Провести визуальную оценку, нарисовав ряд и скользящую статистику (среднее, стандартное отклонение). Построить график, на котором будет отображен сам ряд и различные скользящие. Сделать выводы из полученных результатов. Оценить достоверность статистики.
2. Разложить временной ряд на тренд, сезонность, 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.

Постановка задачи

Выборка – это часть объектов из совокупности, отобранных для изучения с целью получения информации обо всей совокупности. Пусть задано вероятностное пространство $(\Omega, \mathcal{F}, P)$, где Ω – это произвольное непустое множество, элементы которого называются элементарными событиями, исходами или точками; $\mathcal{F}$ – сигма–алгебра подмножеств Ω, называемых случайными событиями; P – вероятностная мера или вероятность.

Случайной величиной называется $\xi: \Omega \longrightarrow R$, т.ч. ξ измерима, т.е.

$$\xi^{-1}(B) = \{\omega \in \Omega : \xi(\omega) \in B\} \in \mathcal{F}.$$

Временной ряд – последовательность значений $X_1, X_2, \ldots, X_n$, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени $t_1, t_2, \ldots, t_n$. Совокупность таких случайных величин будем называть дискретным случайным или стохастическим процессом.

Математическое ожидание дискретной случайной величины – среднее значение случайной величины (распределение вероятностей стационарной случайной величины) при стремлении количества выборок или количества измерений её к бесконечности. EX

Отклонение – разность элемента выборки x_i и некоторого начального значения (например, среднего арифметического $\bar{x}$ или медианы).

Дисперсия случайной величины – мера разброса значений случайной величины относительно её математического ожидания (среднее арифметическое квадратов отклонений).

$$D = \frac{1}{n} \times \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Несмещённая дисперсия –

$$\tilde{S}^2 = \frac{1}{n-1} \times \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Стандартное отклонение по выборке – это корень из дисперсии.

Ковариация – мера линейной зависимости двух случайных величин.

$$cov(X, Y) = E[(X - EX)(Y - EY)$$

Автоковариационная функция – совокупность значений ковариаций при всевозможных значениях расстояния между моментами времени.

Корреляция – мера линейной зависимости двух случайных величин.

$$r(X, Y) = \frac{cov(X,Y)}{\delta_X \delta_Y}$$,где $\delta_X = \sum_{i=1}^{n} (x_i - \bar{x})$, $\delta_Y = \sum_{i=1}^{n} (y_i - \bar{y})$

Коэффициент корреляции может изменяться от –1 до +1.

Если ряд **стационарен в узком смысле,** то его математическое ожидание и дисперсия не зависят от времени, а автоковариационная и автокорреляционная функции зависят только от сдвига по времени и являются четными. Если автоковариационная и автокорреляционная функции не являются чётными, но остальные требования выполняются, то ряд **стационарен в широком смысле (слабо стационарен)**. Иначе ряд является **нестационарным.**

Анализ временных рядов предполагает, что данные состоят из систематической составляющей и случайного шума (ошибки). Большинство регулярных составляющих временных рядов принадлежит к двум классам: они являются либо трендом, либо сезонной составляющей.

Для анализа временного ряда принято выделять:

1) **тренд** (Т) – плавно изменяющаяся компонента, описывающая чистое влияние долговременных факторов (рост населения, изменение структуры возрастного состава и т.д.);

2) **сезонная компонента** (S) – состоит из последовательности почти повторяющихся циклов (объем продаж накануне Нового Года, объем перевозок пассажиров городским транспортом);
3) **случайный шум** (ε) – остается после полного вычленения закономерных компонент

ВР представляет собой

- либо сумму этих компонент $X = T + S + \varepsilon$ в **аддитивной модели**,
- либо произведение $X = T * S * \varepsilon$ в **мультипликативной модели**.

Цикл – изменения уровня ряда с переменным периодом. Выбор модели осуществляется после анализа структуры сезонных колебаний.
Если амплитуда этих колебаний постоянна, используют аддитивную модель временного ряда, в которой значения S предполагаются постоянными для различных циклов. Если амплитуда сезонных колебаний возрастает или уменьшается, используют мультипликативную модель временного ряда, то есть уровни ряда зависят от значений сезонной компоненты.

Визуальная оценка.
Если временные ряды содержат значительную ошибку, то первым шагом построения модели является сглаживание. **Сглаживание** – некоторый способ локального усреднения данных, при котором несистематические компоненты взаимно погашают друг друга.
Самый общий метод сглаживания – **скользящее среднее** ($MA(k)$), в котором каждый член ряда заменяется простым или взвешенным средним n соседних членов, где k – ширина "окна".

$$X_t = \sum_{i=0}^{k} \beta_i \varepsilon_{t-i} ,$$

где β_t – параметры модели, ε_t – шум. Процесс белого шума формально считаем скользящим средним нулевого порядка – $\mathrm{MA}(0)$, где $\beta_0 = 1$. Т.е. текущее наблюдение ряда представляет собой сумму случайной компоненты в данный момент и линейной комбинации случайных воздействий в предыдущие моменты времени.
Скользящее среднее и стандартное отклонение от него называются **скользящими статистиками.**
Ширина окна (интервал сглаживания) берётся равная периоду сезонности. После применения этого метода сезонность и белый шум будут отделены от временного ряда.
Когда ошибка измерения очень большая, используется метод сглаживания методом наименьших квадратов, взвешенных относительно расстояния или метод отрицательного

экспоненциально взвешенного сглаживания. Все эти методы отфильтровывают шум и преобразуют данные в относительно гладкую кривую.

После применения метода скользящих средних сезонная составляющая будет исключена из ряда.

Визуальная оценка проводится следующим образом: выявим наличие у графика тренда. Тренд есть – ряд нестационарный, тренда нет – ряд стационарен.

Разложение на тренд, сезонность и остаток.

1) Выравнивание ряда X_t с помощью скользящей средней, полученный ряд назовём Y_t.
2) Расчёт сезонности S_t: вычитание из X_t ряда Y_t (или деление X_t на Y_t).
3) Приближение временного ряда Y_t методом наименьших квадратов, получение тренда T_t и ошибок ε_t.

Метод наименьших квадратов основан на минимизации суммы квадратов отклонений функции от значений, предсказанных моделью. Более конкретно, оценки наименьших квадратов (НК) параметра получаются минимизацией функции:

$$\varepsilon_t = \sum_{i=0}^{n} [Y_i - T_i]^2$$

Модель ARIMA

Интегрированный временной ряд – нестационарный временной ряд, разности некоторого порядка которого являются стационарным временным рядом.

Для начала необходимо определить понятие TS–ряда (ряда, называемого стационарным относительно тренда. Ряд Y_t называется TS–рядом, если существует некоторая детерминированная функция $f(t)$ такая, что разность $Y_t - f(t)$ является стационарной. К TS–рядам относятся все стационарные ряды.

Временной ряд Y_t называется интегрированным рядом порядка k, если разности k–го порядка $\Delta^k Y_t$ являются стационарными, а разности меньшего порядка стационарными не являются.

Количество единичных корней временного ряда совпадает с порядком его интегрированности.

- $\Delta Y_k = Y_{k+1} - Y_k$
- $\Delta^2 Y_k = \Delta Y_{k+1} - \Delta Y_k = Y_{k+2} - 2 * Y_{k+1} + Y_k$
- $\Delta^m Y_k = \Delta^{m-1} Y_{k+1} - \Delta^{m-1} Y_k$, где Δ – разностный оператор

Обозначение: $Y_k \sim I_k$ – интегрированный временной ряд порядка k

Регрессия – это метод, используемый для моделирования и анализа отношений между переменными, а также для того, чтобы увидеть, как эти переменные вместе влияют на получение определенного результата.

AR (AutoRegressive) – авторегрессионная модель временных рядов, в которой значение временного ряда в данный момент зависит от предыдущих значений этого же ряда.
MA (Moving Average) – модель скользящего среднего, в которой моделируемый уровень временного ряда можно представить как линейную функцию прошлых ошибок, т.е. разностей между прошлыми фактическими и теоретическими уровнями.
ARMA (AutoRegressive Moving Average) объединяет модели авторегрессии (AR) и скользящего среднего (MA). ARMA(p, q), где p – порядок авторегрессии, q – порядок скользящего среднего.

$$X_t = c + \varepsilon_t + \sum_{i=0}^{k} \alpha_i X_{t-i} \sum_{i=0}^{k} \beta_i \varepsilon_{t-i}$$

ARIMA (AutoRegressive Integrated Moving Average) – интегрированная модель авторегрессии скользящего среднего.
ARIMA использует три основных параметра (p, d, q), которые выражаются целыми числами. Потому модель также записывается как ARIMA(p, d, q). Вместе эти три параметра учитывают сезонность, тенденцию и шум в наборах данных:

- p – порядок авторегрессии (AR), который позволяет добавить предыдущие значения временного ряда. Этот параметр можно проиллюстрировать утверждением «завтра, вероятно, будет тепло, если в последние три дня было тепло».
- d – порядок интегрирования (I; т. е. порядок разностей исходного временного ряда). Он добавляет в модель понятия разности временных рядов (определяет количество прошлых временных точек, которые нужно вычесть из текущего значения). Этот параметр иллюстрирует такое утверждение: «завтра, вероятно, будет такая же температура, если разница в температуре за последние три дня была очень мала».
- q – порядок скользящего среднего (MA), который позволяет установить погрешность модели как линейную комбинацию наблюдавшихся ранее значений ошибок.

Обучающая выборка (training sample) — выборка, по которой производится настройка (оптимизация параметров) модели зависимости.
Тестовая (или контрольная) выборка (test sample) — выборка, по которой оценивается качество построенной модели.

1) Оценивается стационарность ряда. Выявляется порядок интегрированности временного ряда. Далее при необходимости ряд преобразуется и для преобразованной модели строится ARMA–модель, поскольку

предполагается, что полученный процесс является стационарным, в отличие от исходного нестационарного процесса.

2) Чтобы итерировать различные комбинации параметров, используется сеточный поиск. Для каждой комбинации параметров можно подобрать новую сезонную модель ARIMA и оценить ее общее качество. Оптимальным набором параметров будет тот, в котором нужные критерии наиболее производительны. При оценке и сравнении статистических моделей, соответствующих различным параметрам, учитывается, насколько та или иная модель соответствует данным и насколько точно она способна прогнозировать будущие точки данных.

Информационный критерий – применяемая в эконометрике мера относительного качества эконометрических моделей, учитывающая степень "подгонки" модели под данные с корректировкой на используемое количество оцениваемых параметров. Т.е. критерии основаны на некотором компромиссе между точностью и сложностью модели. Критерии различаются тем, как они обеспечивают этот баланс. Обычно чем меньше значения критериев, тем выше относительное качество модели.

3) Для оценки используется значение **AIC (Akaike Information Criterion).** AIC оценивает, насколько хорошо модель соответствует данным, принимая во внимание общую сложность модели. Чем меньше функций использует модель, чтобы достичь соответствия данным, тем выше её показатель AIC. Поэтому нужно найти модель с наименьшим значением AIC. После проверки каждой модели ARIMA() на экране появится рейтинг значений AIC. Поскольку некоторые комбинации параметров могут выдавать неточный результат, нужно явно отключить предупреждающие сообщения, чтобы избежать перегрузки предупреждений. Эти неточности также могут приводить к ошибкам, поэтому комбинации параметров, которые вызывают подобные проблемы, нужно игнорировать. Наименьший показатель AIC соответствует оптимальной модели.

AIC(an information criterion) – информационный критерий Акаике – критерий, применяющийся исключительно для выбора из нескольких статистических моделей.

$$AIC = 2k - 2ln(L),$$

где k – число параметров в статистической модели, и $L = \frac{rss}{n}$ – максимизированное значение функции правдоподобия модели, n – количество наблюдений, res – остаток.

4) Выбранные модели нужно проанализировать более подробно. Сравним прогнозируемые значения моделей с 1988 года с реальными значениями временного ряда testing и оценим точность прогнозов. Визуализируем

данные и обращаем внимание на то, что прогнозы, в целом, соответствуют истинным значениям, демонстрируя общий тренд на увеличение. Затем оценим их точность с помощью коэффициента детерминации (R2_score). В итоге выбрали наилучшую модель – ARIMA (8,1,6).

Коэффициент детерминации (R_2) – доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости (квадрат коэффициента корреляции выборки). Лучший результат, который может дать R_2 score – это 1.0. Однако возможны как отрицательные значения, так и значения, превышающие 1.0. Это говорит о неточности предсказания модели.

И сгенерированные прогнозы, и связанный с ними интервал можно использовать для дальнейшего анализа и прогнозирования временных рядов. Чем дальше строится прогноз, тем менее точны его значения. Это отражается на интервалах, генерируемых моделью (чем дальше прогноз, тем больше интервал).

Описание программы

Программа состоит из нескольких модулей:

- Папка mypackage состоит из 3 файлов: __init__.py – файла, определяющего, что папка используется как пакет; visualize.py – файла, содержащего функции визуализации и скользящих статистик; seasonal_dec – файла, содержащего функции, используемые для разложения временного ряда на тренд, сезонность и остаток
- main.ipynb – основной файл, в котором показаны результаты работы программы на данных training.xlsx

Используемые библиотеки и модули

pandas – библиотека для обработки и анализа данных
numpy – библиотека для поддержки высокоуровневых математических функций, предназначенных для работы с многомерными массивами.
matplotlib.pyplot – библиотека для визуализации данных
seaborn базируется на matplotlib, но оптимизирован для визуализации статистических моделей, используется для отображения простых временных распределений
statsmodels – библиотека, предоставляющая предоставляет инструменты для статистического моделирования.
datetime – модуль для работы с датой и временем в python
itertools предоставляет методы работы с итераторами

warnings – библиотека предупреждений–исключений