ДИСЦИПЛИНА	**Проектирование интеллектуальных систем (часть 1/2)**
ИНСТИТУТ	**информационных технологий**
КАФЕДРА	**вычислительной техники**
ВИД УЧЕБНОГО МАТЕРИАЛА	**Материалы для практических/семинарских занятий**
ПРЕПОДАВАТЕЛЬ	**Холмогоров Владислав Владиславович**
СЕМЕСТР	**6, 2023-2024**

Практическая работа № 5

«Анализ и прогнозирование временных рядов»

по дисциплине «Проектирование интеллектуальных систем (часть 1/2)»

Цели: приобрести навыки анализа и прогнозирования временных рядов.

Задачи:

1) Провести анализ и прогнозирование наборы данных временных рядов, выполнив следующие пункты:

– определить предметную область решаемой задачи, ею могут выступать прогнозирование будущих тенденций (например, продаж, валют и других экономических показателей), прогнозирование течения процессов во времени (заболеваний, передвижения, стоимости на рынке, погоды), обнаружение закономерностей и аномалий, анализ рисков, стратегическое планирование, прогноз погоды, отраслевые прогнозы, автоматизация процессов и т.д.;

– найти или сгенерировать набор данных, включающий в себя изменение признаков в зависимости от показателя времени (см. Примечание 1);

– выполнить (хотя бы) минимальную предобработку данных временного ряда (см. Примечание 2) и анализ на стационарность ряда (см. Примечание 3);

– выполнить анализ временного ряда любым методом с помощью декомпозиции и создания модели (см. Примечание 3);

– выполнить прогнозирование временного ряда (см. Примечание 3);

– выполнить визуализацию результатов (см. Примечание 4)

2) В качестве дополнительного задания реализовать хотя бы один из следующих пунктов (один на оценку 4, все – на оценку 5):

– реализовать не менее двух (на каждую из задач) методов анализа и прогнозирования временных рядов, сравнить и проанализировать полученные результаты (см. Примечание 3);

– рассчитать метрики качества прогнозирования модели (хотя бы одну);

– выполнить классификацию и/или сегментацию данных временного ряда (см. Примечание 5).

ПРИМЕЧАНИЕ:

1) Главным критерием при выборе набора данных для анализа временных рядов является наличие в наборе показателя времени с равными промежутками и числовых характеристик, зависимость которых как раз таки будут подвержены анализу и прогнозированию. Стоит учитывать, что помимо самих данных временных рядов существуют перекрёстные данные (они же данные «поперечного сечения», они же cross-sectional), включающие с себя признаки разных объектов в один и тот же период времени (обычно подвергаются регрессионному анализу), и панельные данные (они же «объединенные данные», они же panel или pooled data), сочетающие в себе все предыдущие варианты и используемые как для каждого из видов анализов отдельно, так и совместно для «лонгитюдного» анализа.

2) Наиболее распространёнными проблемами при анализе временных рядов являются неупорядоченные метки времени, пропущенные значения (или метки времени), выбросы и шумы в данных, при этом заменить пропущенные значения традиционными методами (вроде удаления значений, замены статистическими или специальными данными) нельзя, так это может слишком сильно изменить характер из-за изменения порядка, поэтому для этого применяют методы интерполяции (по времени, сплайн и линейную), шумы обрабатываются скользящим средним (где показатели скользящей модели основаны на предыдущих значениях) и преобразованием Фурье, обнаружение и обработка выбросов проводится с помощью методов скользящих (Rolling Statistical Bound based approach), изолирующего леса (Isolation Forest) и кластеризации.

3) Цель анализа данных временных рядов состоит в том, чтобы построить модель, которая фиксирует основные закономерности и структуры для прогнозирования будущих значений ряда, в контексте данных – способ изучения характеристики относительно времени как независимой переменной. Анализ и прогнозирование временного ряда являются разными вещами, так как первый позволяет декомпозировать данные на составные части – уровень

(level, среднее значение во временном ряду), тренд (долгосрочное движение или направленность данных с течением времени), сезонность (периодические колебания или закономерности, которые происходят через регулярные промежутки времени), циклические вариации (долгосрочные колебания, которые не имеют фиксированного периода, подобного сезонности) и нерегулярности (они же «шумы», могут быть результатом случайных событий, ошибок измерения или других непредвиденных факторов) для выявления аномалий или сдвигов в структуре с течением времени, а второй использует исторические данные для составления прогнозов относительно будущих событий. Необходимо анализировать данные на стационарность, так как нестационарные данные содержат тенденции и сезонности, однако они требуют таких операций, как, например, устранение тренда и дифференциации, в стационарных данных статистические свойства остаются постоянными с течением времени, в них отсутствуют какие-либо заметные тенденции, сезонность или закономерности, а значит и меньше случайная ошибка, сначала методами вроде теста ADF или KPSS проверяется стационарность ряда, и в случае обнаружения нестационарных данных производится преобразование в стационарный ряд с помощью таких методов, как удаление тренда (Detrending), дифференцирования (Differencing), трансформация (Transformation) и методы скользящего среднего (SMA, CMA, и EMA). Анализ данных временных рядов выполняется с помощью описанных выше декомпозиции ряда на подынтегральные величины (тренд, сезонность и т.д.) и анализа на стационарность, после чего следует представление модели в аддитивной (когда тренд временного ряда представляет собой линейную зависимость между подынтегральными величинами, т.е. частота (ширина) и амплитуда (высота) ряда одинаковы, y(t) or x(t) = level + trend + seasonality + noise) или мультипликативной методологии (когда временной ряд не является линейной зависимостью между подынтегральными элементами, сезонные колебания увеличиваются со временем с экспоненциальным или квадратичным характером, y(t) or x(t) = Level * Trend * Seasonality * Noise), а

также автокорреляционный анализ, функции частичной автокорреляции (PACF), анализ тенденций, анализ сезонности, спектральный анализ, анализ скользящей корреляции и кросс-корреляционный анализ. Прогнозирование данных временных рядов основывается на анализе закономерностей и выполняется с помощью авторегрессионной модели (AR), авторегрессионного скользящего среднего (ARMA), авторегрессионного интегрированного скользящего среднего (ARIMA), ARIMAX (X – значит, что выполняется с помощью добавления экзогенных переменных, которые не объясняются другими переменными в модели), сезонного авторегрессионного интегрированного скользящего среднего (SARIMA), SARIMAX, модели векторной авторегрессии (VAR), метода экспоненциального сглаживания (SES), обобщенной аддитивной модели (GAM), случайного леса, градиентного бустинга, модели пространства состояний, динамической линейной модели (DLM), скрытой Марковской модели и рекуррентных нейронных сетей.

4) При визуализации результатов анализа временных рядов обычно используют следующие типы графиков:

- линейные графики: отображают точки данных с течением времени, позволяя наблюдать тенденции, циклы и колебания;
- сезонные графики: разбивают данные временных рядов на сезонные компоненты, визуализируя закономерности в течение определенных периодов времени;
- гистограммы и графики плотности: показывают распределение значений данных во времени, предоставляя такие характеристики данных, как асимметрия и эксцесс (меры отклонения графика влево или вправо и его высоты);
- графики автокорреляции и частичной автокорреляции: визуализируют корреляцию между временным рядом и его запаздывающими значениями, помогая определить сезонность и запаздывающие взаимосвязи;

– спектральный анализ: визуализируют частотные характеристики, что необходимо для определения периодичности и цикличностей;
– графики декомпозиции: разбивают временной ряд на трендовые, сезонные и остаточные компоненты, помогая понять лежащие в основе закономерности.

5) Классификация данных временных рядов заключается в том, чтобы связать все данные конкретного временного ряда с конкретной меткой класса (обычно для временных рядов используются деревья решений, классификаторы ближайших соседей и модели глубокого обучения), то есть, например, на основе суточной статистики работы сервера определить, к какому из следующих классов она относится: «стабильная работа», «нестабильная работа» или «нерабочий сервер». Сегментация данных временного ряда подразумевает их разбиение на отдельные области (методы сегментации могут быть либо нисходящими, когда весь ряд делится на сегменты, либо восходящими, когда отдельные точки данных объединяются в сегменты), каждая из которых несёт полезную информацию и каждой из которых присваивается метка определённого класса, в контексте примера с сервером данные о суточной работе могут быть разбиты на почасовые, где каждой области данных будут присвоены классы, а по результатам будет выведена общая статистика за весь период (ряд целиком).

ОСОБЫЙ БОНУС (доступен только в том случае, если выполнены пункты 2-го задания):

Выполнить на одном и том же наборе данных предсказание выбранного признака на основе другого признака (или нескольких признаков при сравнительном анализе, выборе и генерации признаков, либо понижении размерности, признак должен быть косвенно связан с показателем времени, например, возраст, история роста, количество точек продаж и т.д., иначе сравнение будет невозможно) с помощью регрессии (практическая работа №4) и предсказание этого же признака на основе данных временного ряда.