Модуль 1 Основные понятие анализа Временных рядов.

**Введение:**

Модуль дает представление об основные понятия о ВР и методах его осмысления. Рассмотренный модуль затрагивает базовые понятия анализа временных рядов. Отметим, что показанный подход не единственный в анализе. Предлагаемый подход порожден работами в т.н. области эконометрика. Альтернативные подходы следуют, например, из статистической радиотехники, также могут быть и другие интерпретации. Подходы часто пересекаются, однако изученный подход является доминирующим и в тоже время наиболее простым сегодня.

В модуле будут изучены так вопросы, как:

* базовые методы предварительного анализа ВР,
* типы решаемых задач анализа ВР
* модели ВР
* Основные статистические свойства ВР
* Понятие о особых свойствах ВР, например:
  + АКФ,
  + Гетеросекдостичность
  + Белый Гауссов Шум,
  + особенности стационарного вида ВР.
* Виды простых предсказаний ВР, в т.ч.:
  + Наивные методы предсказаний,
  + Методы на основе скользящего среднего и экспоненциального сглаживания,
  + Особенности использования методов линейной регрессии в анализе ВР
  + Особенности использования методов нелинейной регрессии в анализе ВР
  + Особенности базовых методов разложения ВР

Также Вы познакомитесь с методами работы с ВР в рамках языка программирования Python, включая особенности:

* визуализации ВР
* определения базовых свойств ВР
* моделирования ВР
* особенности работы с фреймворком SKTime
* особенности выбора методов предсказания ВР

В результате прохождения модуля вы будете:

* Знать Основные понятия анализ временных рядов, основные свойства ВР
* Уметь предложить подход для решения той или иной задачи анализа временных рядов.
* Владеть методами предварительного анализа временных рядов

Модуль состоит из 2 лекций и 3 практик.

Лекции

1. Введение во ВР
2. Лекция 2 - простые методы предсказания ВР

Практики

1. Предварительный анализ ВР
2. Моделирование ВР
3. Знакомство с библиотекой SKTime и простые методы предсказания ВР

**В ходе изучения модуля ключевыми являются следующие понятия и термины:**

**Временные ряды —** это набор оцифрованных проиндексированных значений, в которых каждый отсчет связан с другим некоторой зависимостью, например, идут последовательно в заданном порядке.

**Анализ временных рядов** - это один или последовательный набор методов/алгоритмов для получения (или оценки/прогнозирования) некоторых параметров (или характеристик) ряда. Чаще всего анализ построен на статистическом подходе к модели временного ряда.

**Типичные задачи анализа временных рядов** – этопредсказание будущих значений, оценка параметров модели ВР, классификация сегментов ВР, выявлений особых значений или сегментов во ВР (аномалий) и другие, в том числе вспомогательные задачи.

**Модель временного ряда** - это специфическая (выбранная, формализованная) зависимость значений модели по своим отсчетам

**Тренд** – наиболее медленно меняющаяся часть зависимости временного ряда, может включать цикличность, смещение, зоны насыщения или точки перегиба, или исключать их как отдельную сущность. Тренд может быть, как детерминированным, так и стохастическим, например, типа броуновское движение.

**Сезонность** - некоторые «относительно быстро меняющиеся» периодические составляющие, как правило тут речь идет о регулярном периоде. Могут также представлять собой редкие, но регулярные события.

**Шум** - это некоторое случайное (стохастическое) искажение выходных значений - нерегулярные или случайные колебания (вариации). Шумы представляют собой остаточную часть от представления ВР моделью. Так в идеале шумы – это белый гауссов шум (**БГШ**, i.i.d), который в силу свойств не может быть объяснено. Однако, часто шумы могут включать необъясненные компоненты ВР.

**Автокорреляционная функция (ACF)** - это степень линейности временного ряда, связанная с отстающей (отложенной) версией самого себя.

**Стационарный ряд -** временной ряд называется стационарным, если его статистические свойства не меняются с течением времени.Другими словами, ряд имеет постоянное среднее значение и дисперсию, а ковариация зависит только от разницы индексов лагов.Это определение можно назвать стационарным в слабом смысле. Сильная стационарность означает, что ряд (и его распределение со статистической точки зрения) не меняются с течением времени. Отметим, что в ряде случаев также говорят о псевдостационарности – то есть регулярности воспроизведения определённых шаблонов поведения ВР. В зависимости от выбранных методов анализа ВР и, например, длины горизонта прогнозирования требования к стационарности ВР могут быть разными.

**Нестационарный ряд** - временной ряд считается нестационарным, если некоторые из его статистических свойств меняются с течением времени. Другими словами, он имеет переменную среднего времени, дисперсию или другие статистические характеристики. Стационарный временной ряд может быть описан только его средним значением и значениями дисперсии (и ковариации), для нестационарных временных рядов существует множество других статистических характеристик.

Для снижения нестационарности могут быть выполнены преобразования ВР. Например, его дифференцирование для устранения тренда, или **преобразования Бокса-Кокса** для устранения т.н. волатильности – то есть неравномерной дисперсии, она также называется **Гетеросекдостичность**.

**Непараметрические методы предсказания** – это методы, базирующиеся на основе аналитической модели ВР, не требующей излишне прецизионной настройки некоторых гиперпараметров. Как правило при наличии некоторых допущений о его свойствах (например, шумы – это БГШ). Как правило в условиях своей работы такие модели наиболее точные, легко интерпретируемые и наиболее быстрые.

**Параметрические методы предсказания** – это методы, базирующиеся на основе параметрической модели ВР. Такая модель является аналитической, но требует явного указания ряда своих параметров для точной работы. Такие параметры могу быть выбраны из некоторых эвристических предположений или настроены при помощи внешних критериев. Благодаря этому при корректности модель имеет более широкую область применения, нежели непараметрические модели.

**Модели, управляемые данными или подход машинного обучения** – это методы, базирующиеся на основе идеи о замене некоторой модели ВР на примеры входных данных и результатов на выходе модели. Такие модели предназначены для тех случаев, когда использовать предыдущие подходы не удаётся. Среди таких подходов в т.ч. подходы на основе глубокого обучения нейронных сетей. Различия классических подходов и подходов на основе глубокого обучения нейронных сетей в формализации признаков. Последние предназначены быть использованными тогда, когда признаки не могут быть выделены в ВР. Например, для продолжительных многомерных временных рядов со сложной структурой.

**Наивное предсказание –** это некоторая базовая оценка, точностью меньше которого любой другой алгоритм не должен обладать.

**Скользящее среднее** – это некоторая группа методов, основанных при предположении о самоподобие ВР, то есть предположении о том, что следующие его значения связаны с рядом предыдущих. Причем ВР должен изменяться со скоростью медленней, чем число усредняемых выборок.

**Экспоненциальное сглаживание** – это набор методов,предполагающий, что истинные значения следующих выборок временного ряда частично связаны со всеми предыдущими истинными и текущими зашумлёнными значениями ВР. Причем чем дальше предыдущие значения, тем экспоненциально меньше эта связь. Это позволяет модели с достаточной степенью гибко реагировать на новейшие изменения в данных, сохраняя при этом информацию об историческом поведении временного ряда. Среди таких методов могут быть выделены как простые методы, так и методы с раздельным учетом связанности для тренда, уровня и сезонности. Могут быть и другие, более сложные модели. Общая группа таких моделей часто называется Error-Trend-Seasonality (ETS).

**Подход линейной регрессии** – это подход на основан на представлении ВР в виде некоторой линейной модели, это может быть, как просто линейная модель тренда, так и модель сводимая к линейной путем некоторых преобразований. Благодаря таким преобразованиям даже весьма сложные ВР могут быть описаны моделями такого класса. Решения таких моделей в условиях БГШ наиболее точные.

Проблемами, на которые указывает регрессионный подход являются т.н. проблемы **плохой обусловленности данных** - относительно высокая изменчивость результатов оценки, вызванная небольшим возмущением (изменением) данных. Решение такой проблемы может быть достигнуто путем **регуляризации**.

**Разложение ВР** - это набор методов, позволяющих представить (факторизовать) компоненты ВР по отдельности. Достаточно грубые методы разложения как правило предполагают выделения тренда и сезонности. Методы могут быть поострены на свойствах регрессии или скользящего среднего.

**Подход нелинейной регрессии** – это подход на основан на представлении ВР в виде некоторой линейной модели, решение которой, однако достигается общими с регрессией методами. В том числе используются приемы гладкой оптимизации, например, метод градиентного спуска, или решения с регуляризацией. К таким моделям относятся, например, нейронные сети, а также, к примеру, обобщенная адаптивная регрессия. Одной из практических реализацией последней является модель **Prophet.** Эта модель адаптирована для решения ряда бизнес задач и позволяет достигать на них высокой точности.

**Остаток модели (невязка)** – это некоторая составляющая ВР, необъясненная моделью. Такая часть может быть БГШ в идеале. Однако на практике это не всегда так. Поэтому остатки требуют анализа. Анализ может быть выполнен визуально или при помощи статистических тестов.

**Заключение:**

Модуль «Основные понятие анализа Временных рядов» является важной частью курса, однако, показывающей лишь только основные – наиболее простые подходы к анализу этой модальности данных. Изученные материалы будут полезны при постановке и проверке базовых гипотез об анализируемом ВР и в рамках поставленных задач.

Изученный программные модули SKTime, а также особые модули Pandas, пожалуй, являются на сегодня практически базовыми фреймворками для анализа ВР различного уровня сложности.

Теоретические материалы, изученные в модуле, носят характер дискурса и раскрывают основные нарративы изучаемой дисциплины. В т.ч. модуль дает понимание основных концепций, такие как:

* Задачи анализа ВР,
* Тренд ВР,
* сезонность,
* представление о статистическом анализе ВР,
* анализ остаточных частей ВР,
* методы предсказания ВР.

Все эти концепции, хотя и могут быть переосмыслены в рамках иных подходов, объединившись создают область знаний, дающую передовые на сегодня методы решения релевантных задач.

В следующих модулях вы узнаете о более сложных подходах к анализу ВР, однако изученные в данном модуле технологии останутся базовыми (т.н. baseline), не позволяющим у ходить слишком далеко в априорных-синтетических суждениях более высокого порядка.