Департамент больших данных и
информационного поиска

21 января 2026 г.

Лекция 1

Структура курса.
Задачи моделирования временных рядов.

Матвей Зехов
mzekhov@hse.ru

Формула оценки

$$\text{Итог} = \text{Округление}(0.75 \cdot \text{Накоп.} + 0.25 \cdot \text{Экз.})$$

где:

$$\text{Накоп.} = \frac{2}{3} \cdot \text{ДЗ} + \frac{1}{3} \cdot \text{КР}$$

$$\text{ДЗ} = 0.25 \cdot \text{ДЗ}_1 + 0.25 \cdot \text{ДЗ}_2 + 0.25 \cdot \text{ДЗ}_3 + 0.25 \cdot \text{ДЗ}_4$$

- ДЗ — средняя оценка за домашние задания (4 домашних задания)
- КР — оценка за контрольную работу
- Экз. — оценка за экзамен
- Округление — арифметическое, применяется только к итоговой оценке
- Одно из домашних заданий является теоретическим

Автоматы

Всем студентам может быть автоматом выставлена оценка за экзамен, равная

$$\text{Экз.}_{\text{автомат}} = \min(7, \text{Накоп.})$$

Итоговая оценка будет рассчитана по стандартной формуле:

$$\text{Итог} = \text{Округление}(0.75 \cdot \text{Накоп.} + 0.25 \cdot \text{Экз.}_{\text{автомат}})$$

При явке на экзамен эта возможность аннулируется.

Дополнительные условия

- При невозможности выполнения любого из ДЗ по уважительной причине и при наличии соответствующей справки, студент вправе перенести вес ДЗ на Экзамен. Для этого необходимо передать справку в учебную часть, а также уведомить семинариста. Автомат в таком случае не может быть выставлен.
- При пропуске КР по уважительной причине вес КР переносится на Экзамен. Для этого необходимо передать справку в учебную часть, а также уведомить семинариста. Автомат в таком случае не может быть выставлен.

Домашние задания: Общие правила

- Домашние задания сдаются в Anytask
- Инвайт будет выслан в групповой чат
- Название основного файла должно быть в формате: `Surname_name_HW#.ipynb` (или Surname_name_HW#.pdf для теордз) где # - номер задания
- За несоответствие имени файла формату предусмотрен штраф 0.5 балла из 10.

Домашние задания: Дедлайны

- Мягких дедлайнов по ДЗ нет. Все дедлайны жёсткие.
- Студент имеет право два раза за курс просрочить дедлайн по любому из ДЗ (практическому или теоретическому) на 24 часа без штрафа
- Или можно просрочить одно ДЗ на 48 часов
- Для этого необходимо оставить соответствующий комментарий в anytask

Домашние задания: Плагиат

- При обнаружении плагиата оценки за домашнее задание обнуляются всем задействованным в списывании студентам
- Подается докладная записка в деканат
- При повторном списывании деканат имеет право отчислить студента

Домашние задания: Использование LLM

- Использование LLM в качестве помощника для домашнего задания не запрещается для небольших элементов кода
- Каждый сгенерированный автоматически элемент домашнего задания должен быть помечен явно
- Запрещается применение LLM для решения теоретического ДЗ и любых элементов практических ДЗ, требующих теоретических выводов
- Любые неразмеченные элементы, выполненные LLM, оцениваться не будут
- При попытке несамостоятельного выполнения существенной части ДЗ (на усмотрение лектора или семинариста) работа обнуляется, а также подается докладная записка в деканат

Контрольная работа

- Письменная
- Без чит-листов
- Очно
- Теоретические задачи
- Ориентировочно – первая неделя после весенней сессии

Экзамен

- Устный
- Без чит-листов
- Очно
- Теоретические вопросы и небольшие задачи
- Список теоретических вопросов появится не позднее чем за две недели до экзамена

Доклады по статьям

На предпоследней неделе курса лекция и семинар будут посвящены обзорам современных статей по прогнозированию и моделированию.

Формат:

- 7–8 докладов по 20 минут
- Один студент – одна статья

Оценка:

- Не входит в формулу оценки
- Дополнительные баллы к ДЗ или КР.
- 3 балла к любому ДЗ/1 балл к КР и 1 к ДЗ/1.5 балла к КР

Организация:

- Список статей будет опубликован после 3-го ДЗ
- По желанию можно предложить статью не из списка (по согласованию с лектором)
- Лектор и семинарист участвуют в дискуссии и модерируют

Дополнительные баллы

- Техать варианты предыдущих лет
- Техать решения задач в задачнике
- Искать ошибки в уже существующих решениях

Основные определения

Определение (Случайный процесс)

Случайный процесс — это семейство случайных величин $\{X_t, t \in T\}$, индексированных параметром t, который обычно интерпретируется как время. Здесь T — множество индексов (чаще всего $T \subseteq \mathbb{R}$ или $T \subseteq \mathbb{Z}$).

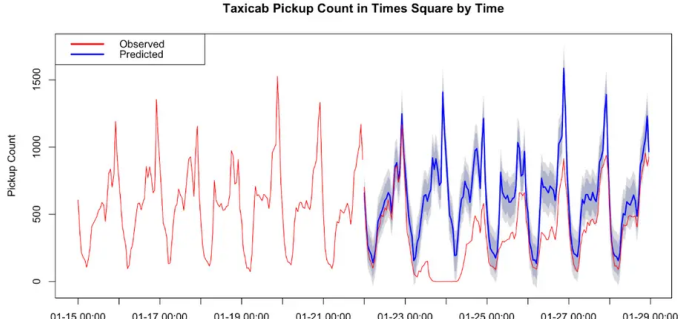
Определение (Временной ряд)

Временной ряд — это последовательность наблюдений $\{x_1, x_2, \ldots, x_n\}$, полученных в последовательные моменты времени, где каждое наблюдение x_t является реализацией случайной величины X_t из некоторого случайного процесса.

Таким образом, временной ряд — это выборка (реализация) случайного процесса.

Определение (Прогнозирование)

Прогнозирование — это процесс предсказания будущих значений временного ряда на основе наблюдаемых прошлых и настоящих значений.



Принципы прогнозирования

Использование исторических данных:

- Прогнозирование основано на анализе прошлых значений временного ряда
- Не предполагает знания будущих событий (no future peeking)
- Используются только доступные на момент прогноза данные

Сохранение паттернов:

- Предполагается, что некоторые закономерности сохраняются во времени
- Используются повторяющиеся паттерны (сезонность, тренды)
- Учитываются циклические и структурные зависимости

Выбор горизонта прогнозирования:

С увеличением горизонта прогнозирования неопределенность прогноза возрастает, что приводит к снижению его точности. Краткосрочные прогнозы обычно более точны, чем долгосрочные.

Типы прогнозирования

Прогнозирование можно классифицировать по различным критериям:

- По типу целевой переменной
- По горизонту прогнозирования
- По количеству прогнозируемых рядов
- По частоте обновления модели

Типы прогнозирования по таргету

По типу целевой переменной:

- Точечное прогнозирование (Point Forecasting)
- Прогнозирование квантилей (Quantile Forecasting)
- Интервальное прогнозирование (Interval Forecasting)
- Прогнозирование волатильности
- Прогнозирование плотности распределения
- Прогнозирование аномалий
- Сценарное прогнозирование

Прогнозирование по горизонту

Одношаговое прогнозирование (Single-step)

- Прогнозируется только следующее значение
- Используются доступные данные до текущего момента
- Высокая точность
- Простая модель

Многошаговое прогнозирование (Multi-step)

- Прогнозируется несколько будущих значений
- Может использовать предыдущие прогнозы
- Накопление ошибки
- Более сложная модель

Заполнение пропусков

- **Простые методы** — заполнение средним, медианой, переносом вперёд/назад (LOCF/BOCF).
- **Интерполяция** — линейная, полиномиальная, сплайновая, временная.
- **Статистические модели** — ARIMA, SARIMA, экспоненциальное сглаживание, модели пространства состояний.
- **Машинное обучение** — использование алгоритмов ML (Random Forest, XGBoost и др.) на основе лагов и признаков.
- **Иные методы** — глубокое обучение (LSTM, Transformer), вероятностные методы (Gaussian Processes), гибридные подходы, множественная импутация, методы на основе DTW, пространственно-временные модели, методы для категориальных и нерегулярных рядов.

Частота и периодичность временных рядов

Определение (Частота временного ряда)

Частота временного ряда — это количество наблюдений в единицу времени (например, ежедневные, еженедельные, ежемесячные данные).

Определение (Периодичность временного ряда)

Периодичность временного ряда — это регулярность появления наблюдений:

- **Периодические ряды** — данные поступают через равные промежутки времени (например, ежедневные продажи).
- **Спорадические ряды** — данные поступают нерегулярно (например, транзакции в интернет-магазине).

Периоды сезонности

Определение (Период сезонности)

Период сезонности — это длина временного интервала, после которого повторяется определённый паттерн в данных.

Примеры периодов сезонности:

- **Ежедневные данные**: период сезонности может составлять 7 дней (недельная сезонность) или 365 дней (годовая сезонность).
- **Ежемесячные данные**: период сезонности может составлять 12 месяцев (годовая сезонность).
- **Ежечасные данные**: период сезонности может составлять 24 часа (дневная сезонность) или 168 часов (недельная сезонность).

Агрегация временных рядов

Определение (Агрегация временного ряда)

Агрегация временного ряда — это процесс объединения данных временного ряда на более высоком уровне детализации (например, суммирование или усреднение данных по дням для получения месячных данных).

Примеры агрегации:

- Агрегация ежедневных продаж в месячные для стратегического планирования
- Суммирование часовых данных потребления электроэнергии для получения суточных показателей
- Усреднение минутных данных цен на акции для получения часовых графиков

Дезагрегация временных рядов

Определение (Дезагрегация временного ряда)

Дезагрегация временного ряда — это процесс разбиения данных временного ряда с более высокого уровня агрегации на более низкий уровень детализации (например, разбиение месячных данных на дневные).

Примеры дезагрегации:

- Дезагрегация годового бюджета в месячные планы для операционного управления
- Распределение квартальных прогнозов продаж по месяцам
- Разбиение недельных данных трафика на дневные показатели

Детекция выбросов в временных рядах

Определение (Выброс)

Выброс — это отдельное наблюдение в временной последовательности, которое значительно отличается от соседних значений и не соответствует общему паттерну ряда.

Причины возникновения выбросов:

- Ошибки измерения или сбои в системе сбора данных
- Редкие события или аномальные ситуации
- Внешние воздействия или шоки

Методы детекции выбросов:

- Статистические методы: Z-оценка, межквартильный размах (IQR)
- Методы на основе скользящего окна: сравнение со средним значением в окне
- Методы машинного обучения: изолирующий лес (Isolation Forest)

Фильтрация временных рядов

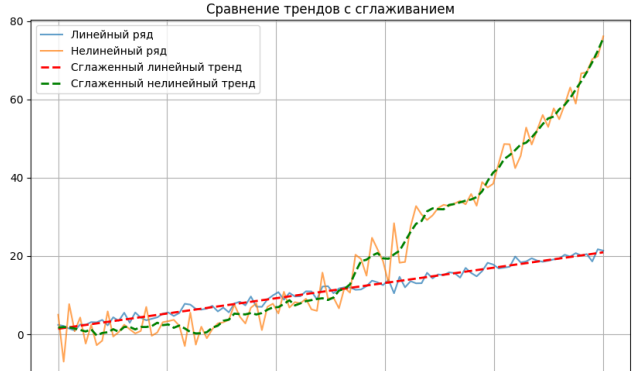
Фильтрация временных рядов — это процесс удаления шума из данных для выделения полезного сигнала и улучшения качества анализа.

Основные подходы к фильтрации:

- **Скользящее среднее** — усреднение значений в скользящем окне
- **Медианный фильтр** — замена значений медианой в окне, эффективен для удаления выбросов
- **Фильтр Калмана** — рекурсивный алгоритм оптимальной фильтрации, использующий модель системы и статистические характеристики шума

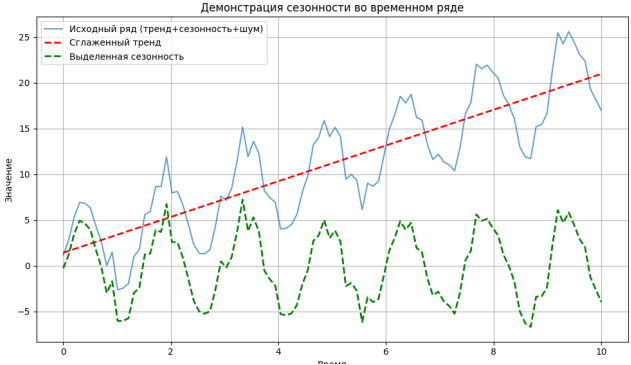
Декомпозиция временного ряда: Trend (Тренд)

- Медленно изменяющийся уровень ряда, обычно выделяемый через сглаживание
- Может быть линейным или нелинейным



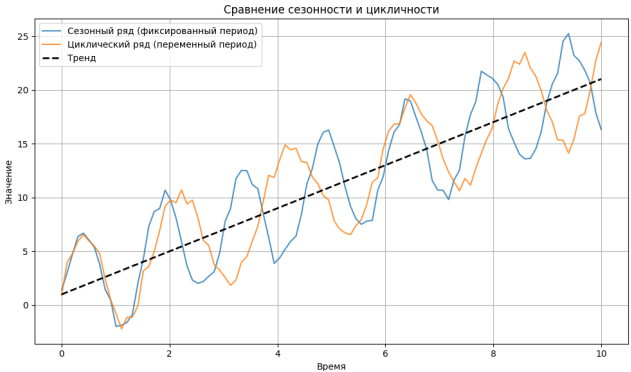
Декомпозиция временного ряда: Seasonality (Сезонность)

- Повторяющиеся колебания значений временного ряда
- Возникают через фиксированные промежутки времени

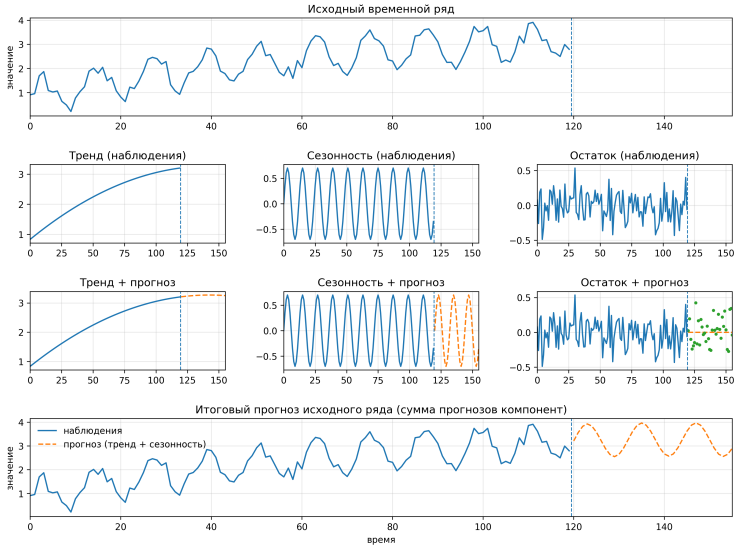


Декомпозиция временного ряда: Cyclicity (Цикличность)

- Колебания значений временного ряда, не имеющие фиксированного периода
- Возникают через нерегулярные промежутки времени
- Часто связаны с экономическими или бизнес-циклами



Декомпозиция временного ряда



Связи между задачами временных рядов

Различные задачи временных рядов тесно связаны между собой:

Детекция аномалий и прогнозирование:

- Обнаружение выбросов улучшает качество прогнозов
- Аномальные значения могут искажать модель прогнозирования
- Предварительная очистка данных повышает точность

Фильтрация и все последующие задачи:

- Удаление шума улучшает результаты всех методов анализа
- Фильтрация является важным этапом предобработки
- Повышает стабильность моделей

Декомпозиция и выбор моделей:

- Понимание компонентов ряда помогает выбрать подходящие модели
- Разные компоненты могут требовать разных подходов
- Упрощает интерпретацию результатов

Классификация временных рядов

Определение (Классификация временных рядов)

Классификация временных рядов — это задача машинного обучения, в которой временной ряд относится к одной из заранее определённых категорий или классов.

Примеры применения:

- Классификация типов ЭКГ сигналов
- Определение типа потребительского поведения по временным рядам покупок
- Классификация жестов по данным с акселерометра

Основные подходы:

- Извлечение признаков с последующим применением классических алгоритмов (SVM, Random Forest)
- Использование глубокого обучения (CNN, RNN)
- Методы на основе расстояний между рядами (DTW)

Кластеризация временных рядов

Определение (Кластеризация временных рядов)

Кластеризация временных рядов — это задача unsupervised learning, в которой временные ряды группируются по схожести их паттернов без заранее определённых меток классов.

Примеры применения:

- Группировка клиентов по схожести временных рядов покупок
- Выявление типичных паттернов потребления электроэнергии
- Сегментация временных рядов цен на акции

Основные подходы:

- Кластеризация на основе признаков (k-means, DBSCAN)
- Кластеризация на основе расстояний между рядами (k-means с DTW)
- Использование автоэнкодеров для снижения размерности с последующей кластеризацией

Поиск похожих временных рядов

Определение (Поиск похожих временных рядов)

Поиск похожих временных рядов — это задача нахождения временных рядов из базы данных, которые имеют схожие паттерны с заданным временным рядом.

Примеры применения:

- Поиск похожих паттернов на финансовых рынках
- Рекомендация товаров на основе схожести временных рядов покупок
- Поиск аномалий через сравнение с нормальными паттернами

Основные подходы:

- Использование расстояний между рядами (Euclidean, DTW, LCSS)
- Поиск на основе признаков (хэширование, индексирование)
- Использование представлений временных рядов (embeddings) для поиска в векторном пространстве

Особенности временных рядов

Ограниченность данных:

- Временных рядов в целом не очень много
- Собираются хуже, чем текстовые данные
- Часто отсутствуют большие датасеты

Разнородность структуры:

- Сильно различаются по структуре
- Разные частоты, длины, сезонности
- Требуют индивидуальной настройки

Краткость рядов:

- Часто короткие, особенно для низких частот
- Недостаток исторических данных
- Требуются специальные методы для малых выборок

Классические статистические модели (часть 1)

Классические статистические модели включают в себя:

- ETS (Exponential Smoothing State Space Model)
- SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors)
- GARCH (Generalized AutoRegressive Conditional Heteroskedasticity)
- Иные линейные и нелинейные модели (TAR, ARFIMA, ...)

Когда применять:

- Мало данных для обучения
- Требуется высокая скорость работы модели
- Необходимо получить бенчмарк для сравнения с более сложными моделями
- Интерпретируемость модели важнее точности

Классические статистические модели (часть 2)

Преимущества:

- Хорошо изучены и теоретически обоснованы
- Быстрая обучаемость и предсказание
- Интерпретируемость параметров

Недостатки:

- Ограниченная гибкость в моделировании сложных зависимостей
- Требуют определенных предположений о структуре данных
- Для некоторых моделей может потребоваться ручная предобработка

Классические методы машинного обучения (часть 1)

Классические методы машинного обучения на табличных данных:

- Линейная регрессия (Linear Regression)
- Градиентный бустинг (XGBoost, LightGBM, CatBoost)
- Случайный лес (Random Forest)
- Метод опорных векторов (SVM) и другие

Когда применять:

- Среднее или большое количество данных
- Требуется гибкость в моделировании сложных зависимостей
- Необходимо учитывать внешние факторы (экзогенные переменные)
- Есть достаточно времени и ресурсов на ресёрч

Классические методы машинного обучения (часть 2)

Преимущества:

- Высокая гибкость в моделировании зависимостей
- Хорошо работают с признаками, извлеченными из временных рядов
- Относительно простая интерпретация

Недостатки:

- Требуют ручного извлечения признаков
- Могут иметь проблемы с долгосрочным прогнозированием
- Не всегда эффективны при сложных временных зависимостях

Модели глубокого обучения (часть 1)

Модели глубокого обучения включают в себя:

- Рекуррентные нейронные сети (RNN)
- Долгая краткосрочная память (LSTM)
- Gated Recurrent Units (GRU)
- Трансформеры (Transformers)
- Сверточные нейронные сети (CNN) для временных рядов

Когда применять:

- Большое количество данных
- Высокая частота временных рядов
- Сложные нелинейные зависимости и паттерны
- Однородные наборы данных

Модели глубокого обучения (часть 2)

Преимущества:

- Высокая точность на больших наборах данных
- Автоматическое извлечение признаков
- Хорошо работают с многомерными временными рядами

Недостатки:

- Требуют много данных для обучения
- Сложность в очистке и подготовке данных
- Часто нестабильны и требуют тонкой настройки
- Низкая интерпретируемость модели