Методичка по временным рядам

Игорь Евсин

5 июля 2025 г.

Оглавление

Глава 1

Поехали

https://otexts.com/fpp3/intro.html

1.1 Как мы построили данный курс

У нас собралось большое количество разнообразных материалов на тему временных рядов из различных областей науки. Они все довольно разнообразны, при этом между областями слабое пересечение. Из-за этого полноценное понимание сути временных рядов не удаётся. Мы поставили задачу связать все эти области воедино. Мы не претендуем на полноту и точность знаний: курс во многом авторский. Мы хотим, чтобы к его окончанию у студентов развилась интуиция в плане работы с временными рядами. Мы не ставим задачу, чтобы студенты помнили каждую формулу из рассмотренных нами, и не планируем рассматривать каждую модель временного ряда, известную человечеству.

Дело в том, что временные ряды включают элементы следующих систем:

1. ТВиМС
2. Случайные процессы
3. Технические методы временных рядов
4. Эконометрические методы временных рядов
5. Обработка сигналов
6. Технический анализ
7. ML
8. DL

Стратегия: надо накидать структуру курса и затем включать в неё все имеющиеся материалы. Если что-то не вмещается в структуру, надо обновить структуру.

Тактика: выложить на GitLab методичку и лекции. Идти последовательно: я и ассистенты создаём лекции по одной. Когда заканчиваем одну, делаем другую. Пишем ноутбуки и сами же разрабатываем домашние задания (не слишком усложняя их).

Формат курса:

1. Методичка. Источником вдохновения является методичка по ПЛА. Со ссылками на дополнительные источники: всё равно всё покрыть невозможно, главное — дать базу.

2. Ноутбуки. 16 ноутбуков, соответствующих 15 коммодитис из индекса. Цель каждого задания — спрогнозировать ряд с наилучшей метрикой качества, используя изученные методы.

3. Очень хочется обойтись без слайдов.

1.2 Что стоит включить

What can be forecast? Взять 4 пункта и написать, что делать, если они не выполняются. Для judgmental forecasting дать ссылку на суперфоркастинг и, возможно, на темы, которые там не покрываются (если такие есть). Футурология.

Описать, с чем мы будем работать — случаи, когда данные всё-таки есть.

Чек-лист построения (ML) модели. Взять за основу разделы 1.3, 1.4, 1.6. Зачем нужны чек-листы? Не для механического построения моделей, а чтобы не забывать о важных шагах.

Связь случайных процессов с временными рядами. Глава 1.7, но более математизированная. Возможно, упомянуть факты вроде того, что условное матожидание — лучший прогноз для МСК.

Мартингал. Наивный прогноз. Задание: построить наивный прогноз для простого ряда. Можно ненавязчиво включать разные официальные источники данных в курс.

Определение 1.1

Пусть T - произвольное множество, $(\Omega, \mathcal{F}, \mathbb{P})$ - вероятностное пространство.

Тогда отображение

$$X : T \times \Omega \to \mathbb{R}$$

называется случайной функцией, если для любого $t \in T$ функция

$$X(t, \omega) := X_t(\omega) = X_t, \quad \omega \in \Omega,$$

является случайной величиной на $(\Omega, \mathcal{F}, \mathbb{P})$.

Наиболее важные виды случайных функций:

1. случайные процессы (stochastic processes, random processes): $T \subset \mathbb{R}$
   (a) случайные процессы с дискретным временем: $T = \mathbb{Z}_+$ (иногда $T = \mathbb{Z}$)
   (b) случайные процессы с непрерывным временем: $T = [0, \infty)$ (иногда $T = \mathbb{R}$)
2. Случайные поля: $T \subset R^k, \quad k \geq 2$.

Определение 1.2

Траекторией случайного процесса X_t называется отображение $t \mapsto X_t(\omega)$ при фиксированном ω. Конечномерным распределением случайного процесса X_t называется распределение вектора $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ для фиксированного набора моментов времени $t_1, t_2, \ldots, t_n$.

Определение 1.3

Временным рядом Y_t называется набор значений траектории случайного процесса, измеренных в наблюдаемые моменты времени. Он представляет собой дискретизированную или наблюдаемую версию траектории случайного процесса.

Эквидистантный временной ряд — это последовательность измерений, сделанных через равные промежутки времени. Однако на практике интервалы между измерениями часто бывают неравными, особенно в экономике или при наблюдении за природными явлениями. Например, данные могут собираться только в рабочие дни, что приводит к пропускам на выходные и праздники. Эти пропуски нельзя игнорировать, особенно если временные лаги играют важную роль (например, при доставке скоропортящихся товаров). Такие увеличенные интервалы не всегда можно просто выбросить, особенно если речь идет о непрерывных производствах или фиксированных временных лагах. Например, если для производства товара А нужен скоропортящийся товар Б, на доставку которого надо 3 дня, то рост производства Б в понедельник и вторник отобразится в товаре А на третий рабочий день, а рост производства Б в пятницу - уже в понедельник (а это следующая точка, если выходные отброшены).

Главное отличие временного ряда от других типов данных — важность временного порядка. Не только само значение, но и момент его измерения имеют значение.

Проводя наблюдения за каким-то природным явлением, мы вовсе не извлекаем получаемые значения из одной и той же генеральной совокупности. Даже если настройки прибора и положение датчиков не менялись, состояние измеряемого объекта в каждый новый момент времени будет другое. Попросту говоря, это будет уже другая случайная величина. Серию измерений, выполненных одним и тем же прибором, даже неподвижно стоящим под одной и той же горой, нельзя рассматривать, как серию выборок из одного и того же пространства элементарных событий. Для описания случайного процесса, в отличие от случайной величины, недостаточно задать его функцию распределения один раз. Просто потому, что в разные моменты времени t она может быть разной. А еще для случайного процесса надо определить функцию совместного распределения вероятностей для моментов времени t и t+dt и так далее.

Чтобы оценить эти функции, наблюдая за случайным процессом, нужна не одна реализация, а целый ансамбль. Ну, хотя бы десяток реализаций. Причем, это обязательно должны быть реализации одного и того же случайного процесса. Тогда и только тогда для каждого момента времени у нас будет несколько измерений одной и той же случайной величины. Как их обрабатывать дальше, мы уже знаем из школьного вузовского курса статистики.

Это часто игнорируется при применении стандартных статистических методов, что может привести к ошибочным выводам, например, к ложной корреляции между числом пиратов и глобальным потеплением.

Но что же делать, если у нас есть только одна Земля? Как изучать взаимосвязи между процессами, каждый из которых мы наблюдаем в единственном экземпляре?! Проблема эта невероятно сложная, практически нерешаемая. Мы еще вернемся к ней, когда будем говорить о свойствах случайных процессов, в частности о стационарности и эргодичности. Оказывается, что для некоторых классов случайных процессов, все характеристики которых неизменны во времени, наличие ансамбля не обязательно! То есть, нам не потребуется десять реализаций, чтобы оценить какую-нибудь статистику. Вместо этого достаточно некоторое время понаблюдать за одной! Например, чтобы оценить коэффициент корреляции между X и Y, достаточно иметь одну реализацию X и еще одну – Y. Что, собственно, все мы и делаем, когда вычисляем коэффициент корреляции между потеплением и пиратами.

Но что же делать с остальными рядами, теми которые не имеют данных приятных свойств. Рядами с трендами, сезонными и суточными циклами, и т.д.? Как искать связь между ними и оценивать ее значимость? На этот вопрос мы в том числе попытаемся найти ответ в данном учебнике.

Пока же следует запомнить, что общая проблема алгоритмов обработки временных рядов — это отсутствие математической строгости. Ведь даже когда мы используем для анализа экспериментальных сигналов алгоритмы со строгим обоснованием и доказанной оптимальностью, у нас всегда остается открытым вопрос о том, не нарушены ли условия применимости таких алгоритмов? Ведь любой алгоритм всегда начинается с преамбулы (требования), что исходные данные должны обладать вполне определенными свойствами. Но когда мы имеем дело с экспериментальным сигналом, доказать выполнение этих требований почти невозможно. Поэтому было бы наивно думать, что корректность результата можно гарантировать строгостью метода. На практике использование строгих методов не дает никаких преимуществ, если с тем же уровнем строгости не доказана адекватность модели данных, в рамках которой сформулирован метод. При режимных наблюдениях это почти невозможно. В лучшем случае можно только предполагать, что «базовая модель» сигнала вполне адекватна реальным данным. В худшем (и, к сожалению, более типичном) случае, наоборот, имеются видимые несоответствия между требованиями теоретической модели и экспериментальным сигналом. Но если у нас нет уверенности в адекватности используемой модели данных, это ставит под сомнение и все результаты, полученные в рамках такой модели.

С одной стороны, такая постановка проблемы может обескуражить - кажется, что не существует универсального рецепта построения идеальной модели временного ряда.

С другой стороны, все это дает право на гибкость и вольность мысли. При изложении материала мы бы хотели, чтобы читатели постоянно держали в голове связь между временным рядом и свойствами породившего его случайного процесса. Мы верим в то, что именно понимание специфических свойств временного ряда способно привести к пониманию, как построить наиболее точный и устойчивый прогноз. Как следствие, мы не хотим делать слишком большой упор на разъяснение механизмов работы конкретных моделей временных рядов или алгоритмов машинного обучения. Вместо этого мы бы хотели, чтобы у читателей сформировалось понимание, какие особенности исходных данных пытались решить создатели того или иного алгоритма. Сфокусироваться на исходном problem space, а не на текущем solution space (тем более, что он постоянно меняется, появляются все новые алгоритмы предобработки данных, все новые алгоритмы машинного обучения, схемы обучения и т.д.).

[1]

Список литературы

1. *adeshere*. Корреляция между временными рядами: что может быть проще? / Хабр. — 16.02.2021. — URL: https://habr.com/ru/articles/542638/ (дата обр. 04.07.2025).

Глава 2

Свойства случайных процессов

Часть III

Свойства случайных процессов

2.1 Стационарность

Определение 9.1

Случайный процесс X_t называется стационарным (stationary, стационарным в узком смысле), если все его конечномерные распределения инвариантны относительно сдвигов, т.е. для любых наборов моментов времени $t1, \dots, tn$, любых вещественных $x1, \dots, xn$ и любого $h > 0$:

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \mathbb{P}(X_{t_1+h} \le x_1, \dots, X_{t_n+h} \le x_n),$$

Определение 9.2

Случайный процесс X_t называется стационарным стационарным в широком смысле (wide sense stationary, weakly stationary, covariance stationary, second-order stationary), если $m(t)$ является постоянной величиной (не зависящей от t), и кроме того, для любых $h > 0$, s, $t \in R$ выполнено

$$K(t+h, s+h) = K(t, s).$$

Другими словами, процесс с постоянным математическим ожиданием является стационарным в широком смысле тогда и только тогда, когда существует функция $\gamma : \mathbb{R} \to \mathbb{R}$ такая, что $K(t, s) = \gamma(t-s)$. для любых t, s. Функция $\gamma(\cdot)$ называется автоковариационной функцией (autocovariance function) и обладает следующими свойствами.

Утверждение 9.3. Пусть γ — автоковариационная функция некоторого стационарного в широком смысле случайного процесса. Тогда

(i) $\gamma(0) \ge 0$.

(ii) $|\gamma(u)| \le \gamma(0)$.

(iii) γ является четной функцией.

2.2 Эргодичность

Понятие эргодичности мотивировано законом больших чисел. Рассмотрим процесс X_t, наблюдаемый в дискретные моменты времени $t = 1, 2, \ldots, T$ и зададимся вопросом, сходится ли процесс

$$M_T = \frac{1}{T}\sum_{t=1}^{T} X_t$$

при устремлении горизонта времени $T \to \infty$.

Определение 11.1. Процесс X_t с дискретными временами $t = 1, 2, \ldots$ называется эргодическим, если

$$M_T \xrightarrow{p} \mu, \quad \text{при} T \to \infty$$

где μ — некоторая константа.

Давайте вернемся к тому, с чего начинали. Во введении мы говорили о необходимости иметь множество реализаций случайного процесса для того, чтобы иметь возможность адекватно посчитать по ним статистики. Оказывается, что для эргодических случайных процессов, все характеристики которых неизменны во времени, наличие ансамбля не обязательно! То есть, нам не потребуется десять реализаций, чтобы оценить какую-нибудь статистику. Вместо этого достаточно некоторое время понаблюдать за одной! Например, чтобы оценить коэффициент корреляции между X и Y, достаточно иметь одну реализацию X и еще одну – Y. Что, собственно, все мы и делаем, когда вычисляем коэффициент корреляции между потеплением и пиратами.

Иногда различают эргодичность по среднему, по дисперсии и т.д. При анализе наблюдений очень часто априори считается, что исследуемые процессы являются эргодическими. Иногда это даже не оговаривается специально.

Но если мы хотим избежать грубейших ошибок, то нельзя забывать, что гипотеза эргодичности – это только гипотеза. Подавляющее большинство долговременных наблюдений продолжается конечное время (вы поняли, это такая шутка), а на выходе получается единственный ряд. Доказать эргодичность такого процесса в принципе невозможно. Поэтому, начиная анализ данных, мы чаще всего просто постулируем ее явным образом или неявно. А что еще остается делать, если в наличии куча данных и руки чешутся начальник требует срочно использовать всю мощь безупречного, многократно проверенного теоретиками статистического инструментария для достижения практических целей?

На самом деле, подавляющее большинство временных рядов вовсе не являются эргодическими. И если доказать эргодичность процесса достаточно сложно (практически нереально), то вот опровергнуть ее часто можно без особых усилий. Достаточно просто вспомнить, что практически все экспериментальные временные ряды существенно нестационарны. Огромный массив накопленных экспериментальных данных однозначно свидетельствует, что априорная "базовая модель"почти любого природного процесса – это вовсе не белый шум (для которого действительно можно заигрывать с эргодичностью). Нет, спектры большинства реальных сигналов имеют степенной вид (Точнее, они обычно становятся степенными после удаления доминирующих периодичностей - сезонной, суточной и т.д.)

Но если ряд не стационарен, то он заведомо не может рассматриваться, как последовательность измерений одной и той же случайной величины. Для него совершенно бессмысленно оценивать те статистики, которые вводятся и исследуются при анализе случайных величин.