

Оглавление

1	ЕМ-алгоритм: ещё ММП	2
1.1	Введение	2
1.2	Определения	2
1.3	Возможные постановки задач	3
1.4	Алгоритмы сглаживания	6
1.5	STL	6
1.6	Стратегии прогнозирования табличных данных	7

Введение

Анализ временных рядов в большинстве своих приложений является частным случаем стандартной задачи регрессии или классификации. Вместо независимых наблюдений в кросс-секционных данных мы рассматриваем одну или несколько последовательностей точек. Подобные данные встречаются во многих областях.

Пример 1.1. Группа компаний Y6 хочет уметь для каждого своего магазина оценивать спрос по всем товарам, чтобы оптимизировать поставки. Из кассовой системы известны продажи за предыдущие дни. Требуется построить модель, которая по каждому товару для каждого магазина будет выдавать прогноз на фиксированный горизонт. Хорошая модель оценки спроса поможет ритейлеру оптимизировать цепочки поставок и избежать *эффекта хлыста*.

Пример 1.2. ВыньДаПоложьБанк хочет спрогнозировать свою выручку. На основе данных о выданных кредитах, хранящихся депозитах, сроках их погашения и т.п. банк может по-отдельности спрогнозировать эти компоненты и с помощью простой иерархической модели построить прогноз для общей выручки.

Пример 1.3. Иннокентий Чёрнодыров получает сигнал от исследовательского спутника, который собирает данные о солнечной активности. Так как в космосе присутствует излучение, вносящее помехи в работу датчиков, Иннокентию придётся внимательно изучать временной ряд на наличие выбросов, аномалий и пропусков в наблюдениях.

Введём определение случайного процесса и временного ряда. Наши определения будут достаточно нестроги, но мы осознанно пойдём на это и оставим коллегам с профильных курсов право сделать это за нас.

1.1 Определения

Определение 1.1.1. Случайный процесс – это *последовательность* случайных величин Y_t , где t – некоторая дискретная шкала времени.

Мы специально сконцентрируемся только на дискретных последовательностях с вещественными значениями, так как для большинства задач этого достаточно. Непрерывные модели в приложениях встречаются редко из-за сложностей в оценке и в целом необходимости их оценки. Следовательно, временным рядом мы будем называть некоторую *реализацию* случайного процесса, y_t . Иногда эту реализацию ещё называют траекторией. Также можно встретить определение, что временной ряд это и есть случайный процесс (последовательность случайных величин), но с практической точки зрения это немного не интуитивно и мы постараемся этого избегать.

На практике обычно мы имеем дело с последовательностями с конечным числом элементов $(y_t)_{t=1}^T$, где T – количество наблюдений. Иногда в учебных целях иногда будем затрагивать последовательности с бесконечным числом элементов: $(y_t)_{t=1}^{t=+\infty}$ или $(y_t)_{t=-\infty}^{t=+\infty}$.

В классических моделях машинного обучения мы предполагали наблюдения в обучающей выборке независимыми и одинаково распределёнными: $X = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$. Однако от этих предпосылок нам придётся отказаться. Почти всегда элементы последовательности будут зависимы между собой и нашей задачей будет выяснить характер этой связи. Грубо говоря, нам необходимо восстановить характеристики случайного процесса по сгенерированной траектории. Да, наличие структуры в данных не позволяет нам напрямую использовать стандартные техники машинного обучения, но в то же время из этой структуры можно выделить много дополнительной и полезной для прогнозирования информации.

Можем ли мы в принципе быть уверены, что сможем его восстановить? В какой-то мере ответ на это даёт Теорема Дуба о разложении [1]. Говоря в очень упрощённых терминах, она говорит о том, что почти любой "хороший" процесс можно разложить на прогнозируемую (детерминированную) и принципиально непрогнозируемую части. Следовательно, мы никогда не сможем восстановить процесс идеально. Но тем не менее, часть процессов более склонна к детерминированному поведению, а часть – менее. Например дневная температура является очень сезонной величиной, то есть имеет паттерн, удобный для прогнозирования. Мы будем учиться обнаруживать и выделять такие паттерны в данных. С другой стороны, котировки акций наиболее близки к хаотичному движению и весьма трудно поддаются прогнозированию. Этим занимаются скорее в области количественных финансов. В нашем курсе котировки акций могут быть рассмотрены скорее из-за удобства и качества данных, но не более чем для иллюстрации. Исключение составит тема прогнозирования волатильности.

1.2 Возможные постановки задач

Задачи на временных рядах можно рассматривать с двух сторон. Во-первых, их можно свести к стандартным методам машинного обучения с минимальными оговорками в подготовке данных. Во-вторых, можно рассматривать это направление как развивавшееся независимо в контексте эконометрических задач с уклоном в логику описания данных. Мы кратко поговорим про первый подход и более подробно про второй.

§2.1 Случай одного ряда

Предположим, что весь наш набор данных состоит из одного временного ряда $(y_t)_{t=1}^T$. На основе него можно сформулировать следующие задачи.

Прогнозирование

Эта задача наиболее популярна. Нам необходимо на основе истории наблюдений и, возможно, каких-то дополнительных данных о внешнем мире предсказать будущие значения ряда. Точку T в таком случае называют *forecast origin*. Пусть мы хотим построить прогноз на h шагов вперёд относительно T . Тогда h называется горизонтом прогнозирования, *forecast horizon*. Ещё можно встретить в некоторых библиотеках (например, `sktime`) понятие абсолютного и относительного горизонта. $T + h$ мы будем называть абсолютным горизонтом, а h относительным.

Прогнозы могут быть различными по форме, но по сути своей они обычно пытаются приблизить некоторую статистику от распределения y_{t+h} . Основных можно выделить три:

1. Точечный прогноз

Самый простой случай. Мы просто хотим узнать конкретное значение показателя в зависимости от периода. Самые популярные модели обычно приближают математическое ожидание или квантиль распределения. Например, ARIMA, пытается приближать условное математическое ожидание $\mathbb{E}(y_{T+h} | (y_t)_{t=1}^T)$

2. Прогноз изменчивости

Обычно под этим подразумевается прогноз дисперсии и некоторые производные от этого. Мы будем более подробно говорить про это в разделе про прогнозирование риска.

3. Интервальный прогноз

Это некоторая комбинация двух предыдущих случаев. Требуется предсказать интервал, в который попадёт y_{t+h} с некоторой заданной вероятностью. Например, 95%. Для явного вычисления требуется знать закон распределения $y_{t+h} | (y_t)_{t=1}^T$ или по крайней мере иметь способ расчёта квантилей. Мы разберём такие примеры в разделе про ETS-модели.

Существуют также способы приближённого вычисления доверительных интервалов с помощью симуляций. Мы обсудим это в разделе про прогнозирование риска.

Заполнение пропусков

В стандартных кросс-секционных данных наблюдения с пропусками иногда не критичны. Например, если их мало, то несколько наблюдений можно удалить, или если пропуски сами по себе являются признаком, то можно их учесть. Стандартные модели временных рядов основываются на том, что в данных нет пропусков и наблюдения расположены через равные промежутки времени. Следовательно, заполнение пропусков может быть вспомогательной задачей прогнозирования. Оно может быть и независимой задачей, если нам необходимо восстановить какие-либо зависимости в прошлом.

Декомпозиция

Благодаря наличию темпоральной структуры временные ряды обладают большим количеством паттернов. Существует несколько различных методов разложения одного ряда на составляющие его компоненты. В общей постановке можно представить временной ряд в виде $y_t = t_t + s_t + c_t + e_t$, где t_t – тренд, s_t – сезонность, c_t – цикличность, а e_t – остаток, не относящийся ни к одной из компонент. Аддитивность разложения вовсе необязательна, каждая из компонент технически может быть мультипликативной. Например, из модели ETS мы сможем достать мультипликативные компоненты, но всё же по большей части используют аддитивные подходы из-за простоты интерпретации.

Обсудим смысл каждой из компонент. Дать им строгие формальные определения довольно затруднительно, поэтому они будут скорее интуитивными. *Трендом* мы будем называть долгосрочное изменение уровня ряда. Можно условно подразделить тренды на восходящие, нисходящие и изменяющие своё направление. Нас же будет больше интересовать природа ряда. В разделе про нестационарные модели мы подробно обсудим, что тренды могут быть порождены как детерминированными функциями, так и стохастическими. *Сезонность* это периодические колебания с фиксированным периодом. Например, продажи мороженого будут стабильно расти летом и падать зимой, а пассажиропоток в метро имеет довольно конкретные часы-пик, почти не изменяющиеся день ото дня. Но здесь важно не угождать нашему антропоцентризму и помнить, что, например, в астрономических задачах периоды сезонности могут не совпадать с земными. *Цикличность* отличается от сезонности только нестабильным периодом и обычно большей длительностью колебаний. Хорошим примером могут быть циклы в любой крупной экономике, где полный период может занимать десятилетия.

Декомпозиция может помочь для разных задач. С помощью неё удобно смотреть на временной ряд в разрезе и проводить эксплоративный анализ данных. Также в некоторых задачах требуется очистить ряд от той или иной компоненты. Например, макроэкономические ряды часто очищают от сезонности для анализа и прогнозирования. Наконец, прогнозировать ряд по частям может быть более удобно и надёжно. Так как все компоненты кроме остатка по построению довольно простые, их можно прогнозировать тривиальными моделями. Если тренд устойчивый, то его несложно экстраполировать линейной или экспоненциальной функцией, а с сезонностью неплохо справляются наивные модели. Цикличность тоже можно экстраполировать моделью сглаживания. Самое сложное обычно кроется в остатках, так как в этой компоненте будут зашиты нетривиальные зависимости. Именно на ряд остатков придётся строить сложную модель с дополнительными признаками и продвинутыми методами. Далее прогнозы всех компонент суммируются (или комбинируются по-другому если разложение не было аддитивным), и сумма будет прогнозом исходного ряда. Такие модели называют *sandwich*.

Детекция разладки

Агрегация и дезагрегация

§2.2 Случай набора рядов

Задачи случая одного ряда

Классификация

Кластеризация

Выявление связи между рядами

1.3 Алгоритмы сглаживания

§3.1 Moving average

§3.2 OLS

§3.3 LOESS

Литература

[1] https://wikichi.ru/wiki/Doob_decomposition_theorem

1.5 Стратегии прогнозирования табличных данных

Основные подходы к моделированию временных рядов делятся на два направления. Первый предполагает рассматривать ряд именно как последовательность и пытаться моделировать породивший его случайный процесс в виде стохастического разностного или дифференциального уравнения. Под это направление попадают все основные классические методы: ETS, ARIMA, GARCH и т.д. Второй подход предлагает свести задачу к табличному виду и попытаться решить её классическими моделями машинного обучения. В текущей главе мы рассмотрим как раз второй подход и различные стратегии прогнозирования, которые он позволяет использовать.

Предположим, что мы хотим решить задачу одношагового прогнозирования. Используем модель авторегрессии порядка k и одной внешней переменной x_t :

$$\hat{y}_{t+1} = \hat{f}(y_t, \dots, y_{t-k+1}, x_t, \dots, x_{t-p+1}), t \in \overline{1, T}$$

Под функцией f будем подразумевать любую стандартную модель регрессии. Подготовим обучающую выборку для $k = 3, p = 1$ и $T = 8$. Так как момент времени $t = 8$ уже наступил, нам известны значения y_8 и x_8 . Оценим эту модель. Построить прогноз на один шаг не составит труда:

t	\hat{y}_t	y_t	y_{t-1}	y_{t-2}	y_{t-3}	x_{t-1}
4	-	4	3	2	1	15
5	-	5	4	3	2	20
6	-	6	5	4	3	12
7	-	7	6	5	4	17
8	-	8	7	6	5	30
9	9.5	-	8	7	6	50
10	-	-	?	8	7	?

Однако для прогноза \hat{y}_{10} у нас недостаточно данных, неизвестны y_9 и x_9 . Рассмотрим подходы, которые помогут обойти эту проблему.

§5.1 Рекурсивная стратегия

Предположим, что наша модель f оценивается очень долго и мы не можем себе позволить оценивать дополнительные модели. Значит необходимо обойтись уже оценённой одношаговой моделью. Попытаемся аппроксимировать неизвестный y_9 наиболее очевидным образом, то есть полученным на предыдущем шаге прогнозом. Стратегию можно записать следующим образом.

$$\hat{y}_{t+h} = \hat{f}(\tilde{y}_{t+h-1}, \dots, \tilde{y}_{t+h-k+1}, \tilde{x}_{t+h}, \dots, \tilde{x}_{t+h-p+1}), t \in \overline{1, T}$$

$$\tilde{z}_t = \begin{cases} z_t & t \leq T \\ \hat{z}_t & t > T \end{cases}$$

Из определения следует, что для внешней переменной x_t тоже необходимо каким-то образом получать прогнозы. Обычно для этого строят несложную вспомогательную модель.

t	\hat{y}_t	\hat{x}_t	y_t	y_{t-1}	y_{t-2}	y_{t-3}	x_{t-1}
4	-	-	4	3	2	1	15
5	-	-	5	4	3	2	20
6	-	-	6	5	4	3	12
7	-	-	7	6	5	4	17
8	-	-	8	7	6	5	30
9	9.5	55	-	8	7	6	50
10	10.5	40	-	9.5	8	7	55

Среди преимуществ рекурсивной стратегии можно назвать относительно низкий разброс. В случае, если у нас мало внешних регрессоров и мы не оцениваем на них большое количество вспомогательных моделей, итоговый прогнозный алгоритм получается довольно простым. Как следствие, оценить модель и построить прогноз можно довольно быстро. К недостаткам можно отнести высокое смещение. Сама модель \hat{f} по построению обучается как одношаговая. Рекурсивные подстановки прогнозов будут очень быстро накапливать ошибку. Эта стратегия довольно плохо подходит для прогнозов на далёкие горизонты. Однако она может хорошо послужить для краткосрочного прогнозирования или в качестве бенчмарка с небольшим количеством регрессоров.

§5.2 Прямая стратегия

Теперь предположим, что модель f оценивать не очень дорого. Чтобы избежать рекурсивного прогнозирования, построим для каждого горизонта отдельную модель.

$$\hat{y}_{t+h} = \hat{f}_h(y_t, \dots, y_{t-k+1}, x_t, \dots, x_{t-p+1}), t \in \overline{1, T}$$

Для этого необходимо сформировать h обучающих выборок. Для $h = 2$ выборки будут выглядеть следующим образом:

t	y_t	y_{t-1}	y_{t-2}	y_{t-3}	x_{t-1}
4	4	3	2	1	15
5	5	4	3	2	20
6	6	5	4	3	12
7	7	6	5	4	17
8	8	7	6	5	30

t	y_{t+1}	y_{t-1}	y_{t-2}	y_{t-3}	x_{t-1}
4	5	3	2	1	15
5	6	4	3	2	20
6	7	5	4	3	12
7	8	6	5	4	17

Легко заметить, что мы просто сдвигаем вектор целевой переменной на один шаг вниз. С каждой итерацией мы теряем одно наблюдение, но на достаточно больших датасетах это не критично. Таким образом модель для горизонта h изначально обучается для прогнозов на h шагов вперёд. Прогноз для каждой модели будет рассчитываться из последней доступной точки:

t	h	\hat{y}_{t+h}	y_t	y_{t-1}	y_{t-2}	x_t
8	1	9.5	8	7	6	50
8	2	10.5	8	7	6	50
8	3	11	8	7	6	50

Прямая стратегия хорошо справляется с прогнозированием на далёкие горизонты. Особенно это удобно если промежуточные значения не важны для прогноза. Эта модель обычно характеризуется большим разбросом, так как приходится оценивать много моделей, но при этом сильно меньшим смещением. Основная проблема такой стратегии заключается в независимости точек прогнозов между собой. Да, падает смещение, но мы утрачиваем или слишком неявно задаём взаимосвязь процесса на участке прогноза.

§5.3 DirRec

Существует довольно много различных промежуточных вариантов между прямой и рекурсивной стратегиями. Например, DirRec. Нам бы хотелось уметь прогнозировать сразу на далёкий горизонт, но при этом не строить дополнительных моделей на признаки и учитывать корреляции между значениями прогнозов. Стратегия предлагает оценивать на каждый горизонт свою модель, но при этом для каждого последующего горизонта

добавлять по одному признаку: прогнозу предыдущего шага. Таким образом в каждой модели будут присутствовать разные наборы параметров.

$$\hat{y}_{t+h} = \hat{f}_h(\tilde{y}_{t+h-1}, \dots, \tilde{y}_{t-k+1}, x_t, \dots, x_{t-p+1}), t \in \overline{1, T}$$

$$\tilde{z}_t = \begin{cases} z_t & t \leq T \\ \hat{z}_t & t > T \end{cases}$$

§5.4 MIMO

Multi-Input Multi-Output (MIMO) немного отличается от рассмотренных нами вариантов. Вместо того, чтобы моделировать горизонты отдельными скалярными моделями, можно попытаться построить векторную модель.

$$[y_{t+h}, \dots, y_{t+1}] = F(y_t, \dots, y_{t-k+1}) + w, \quad t \in \overline{k, T-h}, \quad F: \mathbb{R}^k \rightarrow \mathbb{R}^h, \quad w \in \mathbb{R}^h$$

Из статистических моделей в пример можно привести VARMA, а из нейросетевых подходов MLP и всё семейство рекуррентных сетей. Такой подход позволяет отдать на откуп самой модели учёт корреляций вектора прогнозов и при этом позволяет оценивать всего одну модель. Однако из недостатков можно отметить тот факт, что одна модель может быть недостаточно гибкой для длинных горизонтов. Вариация стратегии, решающая эту проблему, рассматривается ниже.

§5.5 DIRMО

Чтобы управлять гибкостью модели, разобьём горизонт прогнозирования на несколько блоков. Для простоты предположим их равными, но для конкретных задач можно сделать их различной длины. Пусть мы рассматриваем $m = \frac{h}{s}$ участков горизонта, каждый длины s . На каждом из этих участков будем оценивать MIMO-модель. Таким образом при $s = 1$ мы получим прямую стратегию. При $s = h$ мы получим базовую MIMO-стратегию. Параметр s позволяет нам делать выбор между степенью учёта корреляций прогноза и общей гибкостью модели.

§5.6 Комбинации

В литературе можно найти множество стратегий для конкретных задач, но не исключено, что вам потребуется построить свою собственную. Например, близкие горизонты прогнозировать рекурсивно, а далёкие – прямой стратегией. Главное в этом подходе, и во временных рядах в частности, руководствоваться здравым смыслом и выбирать наиболее простую модель из сопоставимых по качеству. Бритва Оккама может вам замечательно помочь: "Не множьте сущее без необходимости".