Оглавление

1 Введение **2**

1.1 Определения 2

1.2 Возможные постановки задач 3

2 Алгоритмы декомпозиции **7**

2.1 Алгоритмы сглаживания 8

2.2 STL 8

2.3 MSTL 8

3 Сведение к табличной задаче **9**

3.1 Формирование признаков 9

3.2 Стратегии прогнозирования табличных данных 9

4 ETS **13**

4.1 Модели экспоненциального сглаживания 13

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 велечин), 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: $X = \{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}$. 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 тракетории. 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. Следственно, мы никогда не сможем восстановить процесс идеально. Но тем не менее, часть процессов более склонна к детерминированному поведению, а часть – менее. Например дневная температура яввляется 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 стандратных 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. Обычно выделяют четыре: тренд, сезонность, цикличность и остаток. Отдельные компоненты можно спрогнозировать более простыми моделями, а потом собрать прогноз воедино.

Детекция разладки

Пусть мы оценили прогнозную модель для какого-либо ряда. Однако с появлением новых наблюдений модель можно переоценить, чтобы учесть всю актуальную информацию. Однако характеристики наблюдаемых процессов могут меняться со временем. Это означает, что предпосылки, на основе которых мы оценивали исходную модель могут стать не актуальными. Задача определения этого факта, а также точки во времени, когда он произошёл, и есть задача детекции разладки.

Агрегация и дезагрегация

Агрегация – это переход от высокой частоты к низкой. Например, перевод часовых данных в дневные. Для этого перехода необходима функция агрегации, выбор которой должен быть согласован с логикой данных. Обычно это стандартные статистики: сумма, среднее, минимум, максимум и т.п.

Дезагрегация – это переход от низкой частоты к высокой. Данная операция уже существенно более сложная и не имеет какой-то стандартной процедуры или очевидной логики. Обычно для проведения дезагрегации требуются прокси-ряды.

Пример 1.4. Данные по ВВП выпускаются поквартально. Однако для оперативного экономического анализа в течение квартала хотелось бы получать представления о его динамике. Для этого можно воспользоваться компонентами ВВП, которые можно оценивать по крайней мере ежемесячно. Например, потребление или государственные расходы.

§2.2 Случай набора рядов

Для большого количества рядов можно решать все те же задачи, что и для отдельных рядов. Например, прогнозировать отдельно каждый показатель. Однако задача может усложняться, если ряды связаны между собой.

Классификация

Каждый временной ряд можно расценивать как отдельный объект, который можно классифицировать.

Пример 1.5. Ваши умные часы или телефон постоянно собирают статистику: данные гироскопа, температуры, давления и т.п. Фитнес-трекер классифицирует эти данные по характеру колебаний и определяет, чем вы сейчас занимаетесь.

Кластеризация

Аналогично табличным задачам, после классификации следует упомянуть кластеризацию. Каких-то явных приложений и широкого применения я не встречал, но как вспомогательная задача выявления похожих рядов на неразмеченном массиве имеет место быть.

Выявление связи между рядами

Задача выявления связи чуть более тонкая чем "подобрать фичи для лучшего прогноза". Требуется максимально точно восстановить характер взаимосвязи переменных. Такой подход больше распространён в экономических приложениях и эконометрических задачах в частности. Мы немного поговорим об этом в главе про SVAR-модели.

Алгоритмы декомпозиции

Наиболее широко используется аддитивная декомпозиция ряда. Во многом это связано с её простотой и наличием устойчивых алгоритмов декомпозиции.

$$y_t = t_t + s_t + c_t + e_t,$$

где t_t – тренд, s_t – сезонность, c_t – цикличность, а e_t – остаток.

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*.

2.1 Алгоритмы сглаживания

§1.1 Moving average

§1.2 OLS

§1.3 LOESS

2.2 STL

2.3 MSTL

Сведение к табличной задаче

3.1 Формирование признаков

§1.1 Использование времени

§1.2 Использование зависимой переменной

§1.3 Использование внешних переменных

3.2 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:

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

§2.1 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

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	?

записать следующим образом.

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

$$\tilde{y}_t = \begin{cases} y_t & \text{t<=T} \\ \hat{y}_t & \text{t>T} \end{cases}$$

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.

§2.2 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.

§2.3 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. Стратегий 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.

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

$$\tilde{y}_t = \begin{cases} y_t & \text{t<=T} \\ \hat{y}_t & \text{t>T} \end{cases}$$

§2.4 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.

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

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.

§2.5 DIRMO

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.

§2.6 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".

ETS

4.1 Модели экспоненциального сглаживания

§1.1 Простое экспоненциальное сглаживание

Предположим, что мы хотим спрогнозировать некоторый временной ряд $(y_t)_{t=1}^T$. Также предположим, что данный ряд не имеет выраженной сезонности или тренда. Самой простой моделью прогнозирования можно считать наивную:

$$\hat{y}_{T+1|T} = y_T \tag{4.1.1}$$

Данная модель хорошо подходит для бенчмарка, но в большинстве случаев (не всегда!) слишком проста для прогнозирования. Как минимум, она никак не учитывает историю до y_T. Попробуем это исправить. Например, можно добавить усреднение всей истории.

$$\hat{y}_{T+1|T} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{4.1.2}$$

Мы добавили зависимость от истории, однако перестарались. Все наблюдения в таком случае будут иметь одинаковый вес. Логично предположить, что наблюдения, близкие к моменту времени T должны иметь больший вес. Например, если в далёком прошлом, близко к моменту времени 1 временной ряд имел выбросы или структурные сдвиги, не хотелось бы придавать этому большой вес. Сама собой напрашивается геометрическая прогрессия с убывающими весами. Зададим параметр $\alpha \in [0, 1]$ как вес наблюдения y_T и будем уменьшать его на вес q.

$$\hat{y}_{T+1|T} = \sum_{i=0}^{T-1} \alpha q^i y_{T-i} \tag{4.1.3}$$

Найдём веса q. Для простоты предположим, что их сумма должна равняться 1.

$$\alpha + q\alpha + q^2\alpha + \ldots + q^{T-1}\alpha = 1 \tag{4.1.4}$$

Однако полученное для суммы этой прогрессии уравнение будет зависеть от T и решать его не очень удобно:

$$\frac{\alpha(q^T - 1)}{q - 1} = 1 \tag{4.1.5}$$

Для упрощения предположим, что T велико и воспользуемся бесконечно убывающей геометрической прогрессией:

$$\frac{\alpha}{q-1} = 1 \Rightarrow q = 1 - \alpha \tag{4.1.6}$$

Конечно, у нас в реальном мире не бесконечное количество данных и это будет приближением, но довольно точным. Таким образом мы получим финальную форму модели:

$$\hat{y}_{T+1|T} = \sum_{i=0}^{T-1} \alpha(1-\alpha)^i y_{T-i} \tag{4.1.7}$$

Веса $\alpha(1-\alpha)^{t-1}$ убывают экспоненциально с ростом t, откуда и получила название модель простого экспоненциального сглаживания. Многошаговый прогноз такой модели будет плоским и будет просто повторять одношаговый прогноз:

$$\hat{y}_{T+h|T} = \sum_{i=0}^{T-1} \alpha(1-\alpha)^i y_{T-i} \tag{4.1.8}$$

Параметр α можно подобрать, численно решив следующую задачу оптимизации:

$$\sum_{t=1}^{T} (y_t - \hat{y}_{t|t-1})^2 \to \min_{\alpha}$$

Для дальнейшего анализа будет полезно рассмотрерь несколько дополнительных форм модели экспоненциального сглаживания.

Модель коррекции ошибок

Сгруппируем последнее выражение относительно α:

$$\begin{aligned}
\hat{y}_{T+1|T} &= \alpha y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \ldots \\
&= \alpha y_t + (1-\alpha)[\alpha y_{t-1} + \alpha(1-\alpha) y_{t-2} + \ldots] \\
&= \alpha y_T + (1-\alpha)\hat{y}_{T|T-1} \\
&= \alpha(y_T - \hat{y}_{T|T-1}) + \hat{y}_{T|T-1} \\
&= \alpha e_T + \hat{y}_{T|T-1}
\end{aligned} \tag{4.1.9}$$

Из этой записи следует, что прогноз можно представить как коррекцию предыдущего прогноза на его ошибку относительно истинного значения с некоторым коэффициентом. Сейчас нам этот результат интересен скорее как занимательный факт, но в дальнейшем похожая идея будет использоваться в модели VECM (Vector Error Correction Model).

Взвешенное среднее

Воспользуемся результатом из уравнения 4.1.9.

$$\begin{aligned}\hat{y}_{t+1|t} &= \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \dots \\ &= \alpha y_t + (1-\alpha)[\alpha y_{t-1} + \alpha(1-\alpha)y_{t-2} + \dots] \\ &= \alpha y_t + (1-\alpha)\hat{y}_{t|t-1}\end{aligned} \tag{4.1.10}$$

Получается, что наш прогноз можно представить как взвешенное среднее наблюдаемого значения y_T и его прогноза, полученного на предыдущем шаге $\hat{y}_{T|T-1}$. Однако надо заметить, что для корректности такой формы нужно ввести один дополнительный параметр l_0, инициализирующий последовательность. Далее станет ясно, почему в качестве имени мы взяли именно l_0.

$$\hat{y}_{2|1} = \alpha y_1 + (1-\alpha)l_0$$

Тогда прогнозное уравнение также изменится.

$$\hat{y}_{T+1|T} = \frac{1}{T}\sum_{i=0}^{T-1}\alpha(1-\alpha)^i y_{T-i} + (1-\alpha)^T l_0$$

Вес последнего слагаемого будет быстро убывать при больших T, и модель будет эквивалентна стандартной постановке. Для полной эквивалентности можно просто положить $l_0 = 0$.

Параметр l_0 можно найти из той же задачи оптимизации:

$$\sum_{t=1}^{T}(y_t - \hat{y}_{t|t-1})^2 \to \min_{\alpha, l_0}$$

Компонентный вид

Вы часто думаете о Римской империи? Для дальнейших выводов воспользуемся древним римским принципом "разделяй и властвуй". Разобьём наше уравнение на два:

$$\begin{aligned}&\text{Уравнение прогноза} && \hat{y}_{t+1|t} &&= l_t \\ &\text{Уравнение сглаживания} && l_t &&= \alpha y_t + (1-\alpha)l_{t-1}\end{aligned} \tag{4.1.11}$$

Эта формулировка эквивалентна предыдущим. Она удобна технически для того, чтобы впоследствии добавлять уравнения и новые компоненты в уравнение прогноза. Здесь мы также заострим внимание на том, что l_t в такой постановке можно интерпретировать как сглаженный уровень ряда.

§1.2 Трендированные модели

Предыдущая модель подходит только для данных без ярко выраженных трендов. Для добавления большей динамики введём ещё один показатель. b_t будет означать локальную скорость роста за один период модели. Грубо говоря, этот параметр будет отвечать за приращения компоненты l_t. Обновлённая система уравнений будет выглядеть следующим образом.

$$\begin{array}{lll} \text{Уравнение прогноза} & \hat{y}_{t+h|t} & = l_t + hb_t \\ \text{Уравнение сглаживания} & l_t & = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}) \\ \text{Уравнение тренда} & b_t & = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \end{array} \tag{4.1.12}$$

В последнем уравнении мы усредняем оценку тренда на основе приращений ($l_t - l_{t-1}$) и предыдущую оценку b_{t-1}. Добавление уравнения также увеличивает количество параметров, к уже имеющемуся списку прибавим β и b_0.

Важно отметить, что несмотря на долгую историю своего существования, этот метод не является сакральным эталоном и он довольно эвристичен. Никто не мешает модифицировать эти формулы в зависимости от вашего процесса или внутренних убеждений. Например, приращения можно оценивать на основе истинных данных ($y_t - y_{t-1}$), а не сглаженных. Всё на ваше творческое усмотрение, мы лишь описываем модели, некогда оказавшиеся удачными.

Можно заметить, что прогноз из плоского стал линейным. Подобная линейная экстраполяция может быть плоха по ряду причин. Во-первых, тренды могут менять направление на прогнозном горизонте. С этим ничего не поделать силами такой простой модели, но от неё это и не требуется. Во-вторых, эмпирически установлено, что модели линейного тренда склонны переоценивать тренд на больших горизонтах. Проще говоря, на практике тренды, близкие к линейным, склонны затухать. Если ваш ряд растёт экспоненциально, то скорее всего затухать будет его логарифм. Предлагается штрафовать модель на небольшой коэффициент $\varphi \in [0,1]$ за каждый последующий шаг.

$$\begin{aligned} \hat{y}_{t+h|t} &= l_t + (\varphi + \varphi^2 + \ldots + \varphi^h)b_t \\ l_t &= \alpha y_t + (1-\alpha)(l_{t-1} + \varphi b_{t-1}) \\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)\varphi b_{t-1} \end{aligned} \tag{4.1.13}$$

Для далёких горизонтов значение прогноза будет выходить на константу. Однако в целом не рекомендуется слишком сильно полагаться на модель на больших горизонтах, так как дисперсия прогнозов растёт довольно быстро. Это нельзя увидеть явно без введения вероятностной модели, но вскоре мы до этого доберёмся.

$$\lim_{h\to\infty} \hat{y}_{t+h|t} = l_t + \frac{\varphi b_t}{1-\varphi} \quad \text{при} \quad \varphi \in (0,1)$$

Параметр φ также можно оценить с помощью задачи оптимизации.

§1.3 Сезонные модели

Добавим в нашу римско-имперскую модель ещё одно уравнение на сезонность: s_t. Для этого нужно выбрать период сезонности m. В базовом варианте модель предполагает одну сезонность.

Для начала создадим само уравнение сезонности. Сезонность предполагается цикличной и мало изменяющейся компонентой. Как и все предыдущие уравнения, будем строить его как взвешенное среднее между модельной и наблюдаемой величинами.

Наблюдаемую сезонность можно выделить как разность $(y_t - l_{t-1} - b_{t-1})$. Так как сезонность предполагается цикличной, в качестве модельной величины возьмём просто предыдущее значение s_{t-m}.

Также модифицируем уравнение сглаживания. По нашим предпосылка l_t является некоторым сглаженным уровнем ряда y_t. Значит в фактической части уравнения нужно очистить y_t от сезонности.

Наконец, модифицируем уравнение прогноза. Для прогноза будем использовать последний сезонный цикл тренировочных данных. Так, если мы прогнозируем месячные данные на декабрь, в качестве сезонной компоненты прогноза используем последний декабрь из выборки. Формула в данном случае оказывается сложнее идеи: $s_{t+h-m(k+1)}$, где k =

$$\begin{array}{lll} \text{Уравнение прогноза} & \hat{y}_{t+h|t} & = l_t + hb_t + s_{t+h-m(k+1)} \\ \text{Уравнение сглаживания} & l_t & = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \\ \text{Уравнение тренда} & b_t & = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \\ \text{Уравнение сезонности} & s_t & = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \end{array} \tag{4.1.14}$$

Также для уравнения сезонности часто используется аналогичная формулировка:

$$s_t = \gamma^*(y_t - l_t) + (1-\gamma^*)s_{t-m} \tag{4.1.15}$$

Подставив в это уравнение l_t, легко убедиться, что $\gamma = \gamma^*(1-\alpha)$.

Для сезонной модели необходимо сделать два важных замечания.

1. Для уравнения сезонности также необходимо ввести стартовые параметры. Заметьте, что для корректной оценки необходимо ввести m параметров: $s_1, s_2, \ldots, s_m$

2. Так как сезонность – цикличная компонента, и притом в нашей постановке аддитивная, введём следующее ограничение:

$$s_1 + \ldots + s_m = 0$$

3. Ограничение из предыдущего пункта позволяет выразить один параметр через все остальные. Следовательно, для задачи достаточно определить не m, а $m-1$ параметров.

§1.4 Мультипликативные модели

§1.5 ETS

Все рассмотренные до этого модели не были стохастическими. Для них можно было просто выписать функционал на основе MSE и оптимизировать градиентным спуском. Однако на все описанные модели путём нескольких простых переходов можно навесить стохастику. Приведём пример для самой простой версии экспоненциального сглаживания:

$$\begin{array}{lll} \text{Уравнение прогноза} & \hat{y}_{t+1|t} & = l_t \\ \text{Уравнение сглаживания} & l_t & = \alpha y_t + (1-\alpha) l_{t-1} \end{array} \tag{4.1.16}$$

Вспомним, что уравнение сглаживания можно переписать в форме error correction:

$$l_t = \alpha y_t + (1-\alpha) l_{t-1} = l_{t-1} + \alpha(y_t - l_{t-1}) = l_{t-1} + \alpha e_t, \tag{4.1.17}$$

где $e_t = y_t - l_{t-1} = y_t - \hat{y}_{t|t-1}$

Иходя из определения e_t:

$$y_t = l_{t-1} + e_t \tag{4.1.18}$$

Достаточно естественно в данной постановке моделировать e_t через распределение. Например, $e_t = \varepsilon_t = \mathcal{N}(0, \sigma^2)$.

Таким образом, итоговая модель:

$$\begin{array}{lll} \text{Уравнение прогноза} & \hat{y}_{t+1|t} & = l_{t-1} + \varepsilon_t \\ \text{Уравнение сглаживания} & l_t & = l_{t-1} + \alpha\varepsilon_t, \end{array} \tag{4.1.19}$$

Литература

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