Моделирование временных рядов
Лекция 1
Введение

Борис Демешев, Матвей Зехов

Временные ряды по сути своей лишь частный случай стандартной задачи регрессии или классификации. Они возникают в большом количестве областей. Например, любая компания из ретейла будет прогнозировать количество товаров, которые необходимо поставить в магазин или даркстор. Также обычно сразу приходит на ум котировки на фондовой бирже. Из физики и инженерных задач возникает анализ сигналов от датчиков. Введём определение временного ряда. Перед этим мы должны ввести (очень нестрого) понятие случайного процесса.

1 Определения

Случайный процесс – это некоторая *последовательность* случайных величин Y_t. Мы специально сконцентрируемся только на дискретных последовательностях с вещественными значениями, так как для большинства задач этого достаточно. 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.

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.

2.1.1 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.

2.1.2 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.

2.1.3 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*.

2.1.4 Детекция разладки

2.1.5 Агрегация и дезагрегация

§2.2 Случай набора рядов

2.2.1 Задачи случая одного ряда

2.2.2 Классификация

2.2.3 Кластеризация

2.2.4 Выявление связи между рядами

3 Алгоритмы сглаживания

§3.1 Moving average

§3.2 OLS

§3.3 LOESS

4 STL

Список литературы

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