Обнаружение разладки во временных рядах показов мобильной рекламы

К.В. Мерзляков, группа 622

Санкт-Петербургский Государственный Университет
Кафедра статистического моделирования

18.05.2019

Содержание

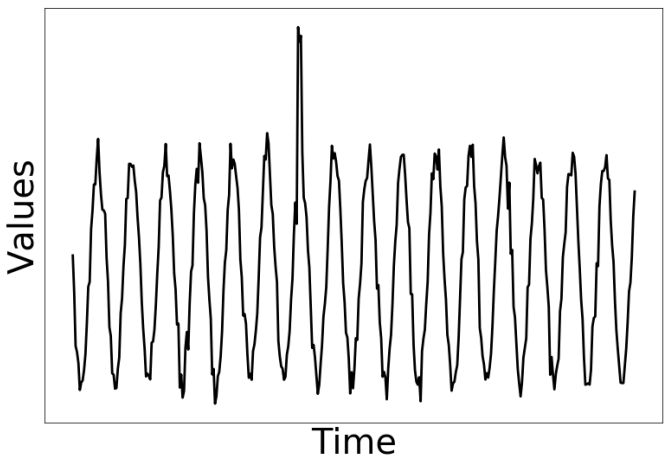
- Общие замечания
- Построение модели данных
- Методы обнаружения разладки
- Оценка качества
- Моделирование данных
- Применение моделей к смоделированным данным

Общие замечания

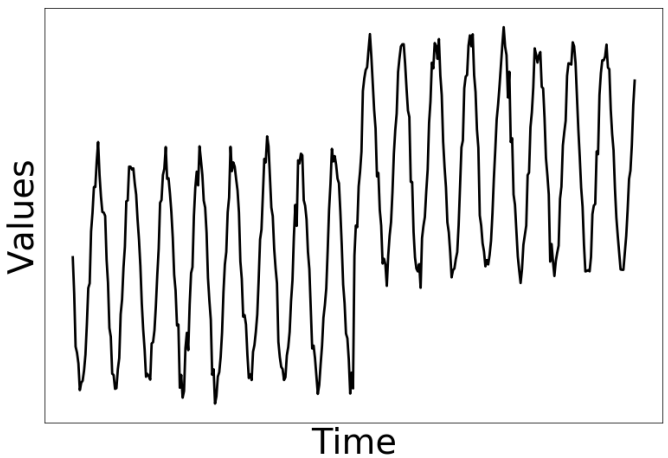
Разладка во временных рядах

- Разладкой во временных рядах называют момент времени, в который произошло существенное изменение в структуре временного ряда
- Методы обнаружения разладки — это группа методов, с помощью которых можно находить такие точки разладки
- Разладка может быть двух типов
  - Локальная — аномалия или выброс
  - Глобальная — изменение структуры ряда

Локальная разладка



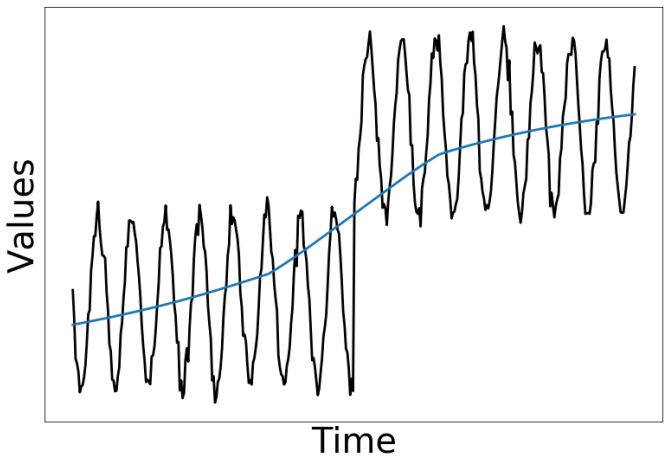
Глобальная разладка



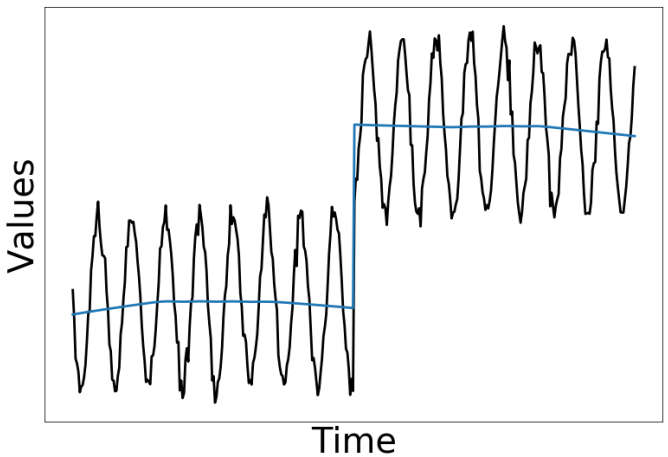
Мотивация

- Исторические данные
  - Прогнозирование
  - Извлечение тренда
  - Поиск проблем в исторических данных
- Текущие данные
  - Реакция на изменения своевременно

Извлечение тренда без анализа разладок

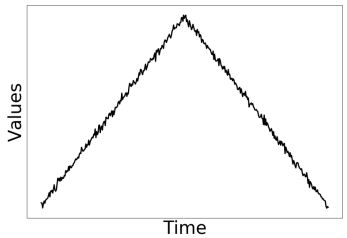


Извлечение тренда с анализом разладок

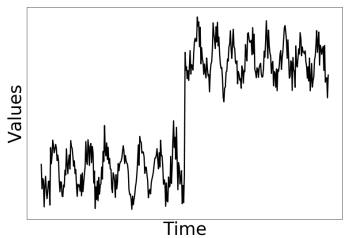


Виды разладок

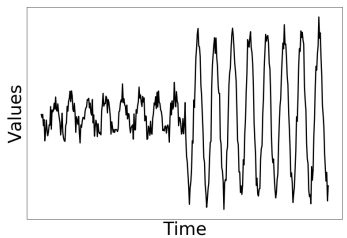
Изменение в тренде



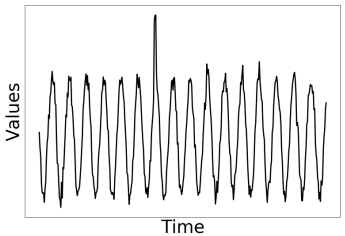
Изменение в среднем



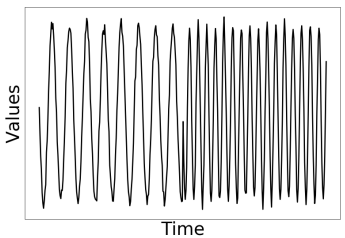
Изменение в амплитуде колебаний



Локальное изменение



Изменение в периодике



Структура исследования

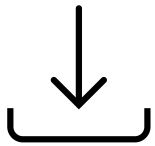
- Смоделировать данные, близкие к реальным
- Применить к смоделированным данным набор методов
- Оценить и сравнить качество примененных методов

Построение модели данных

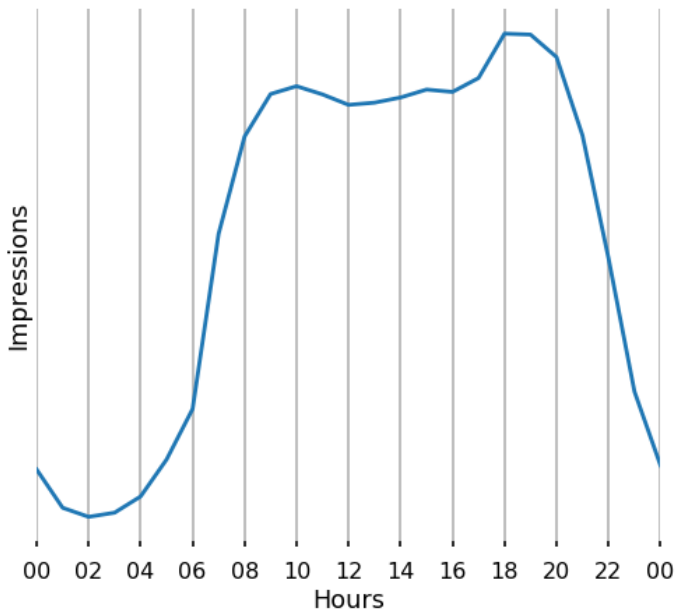
Данные мобильной рекламы

Запрос > Показ > Клик > Конверсия

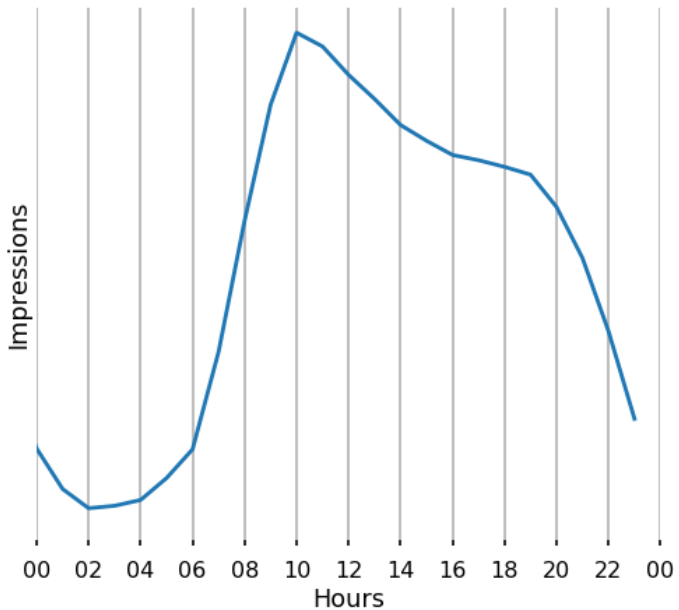


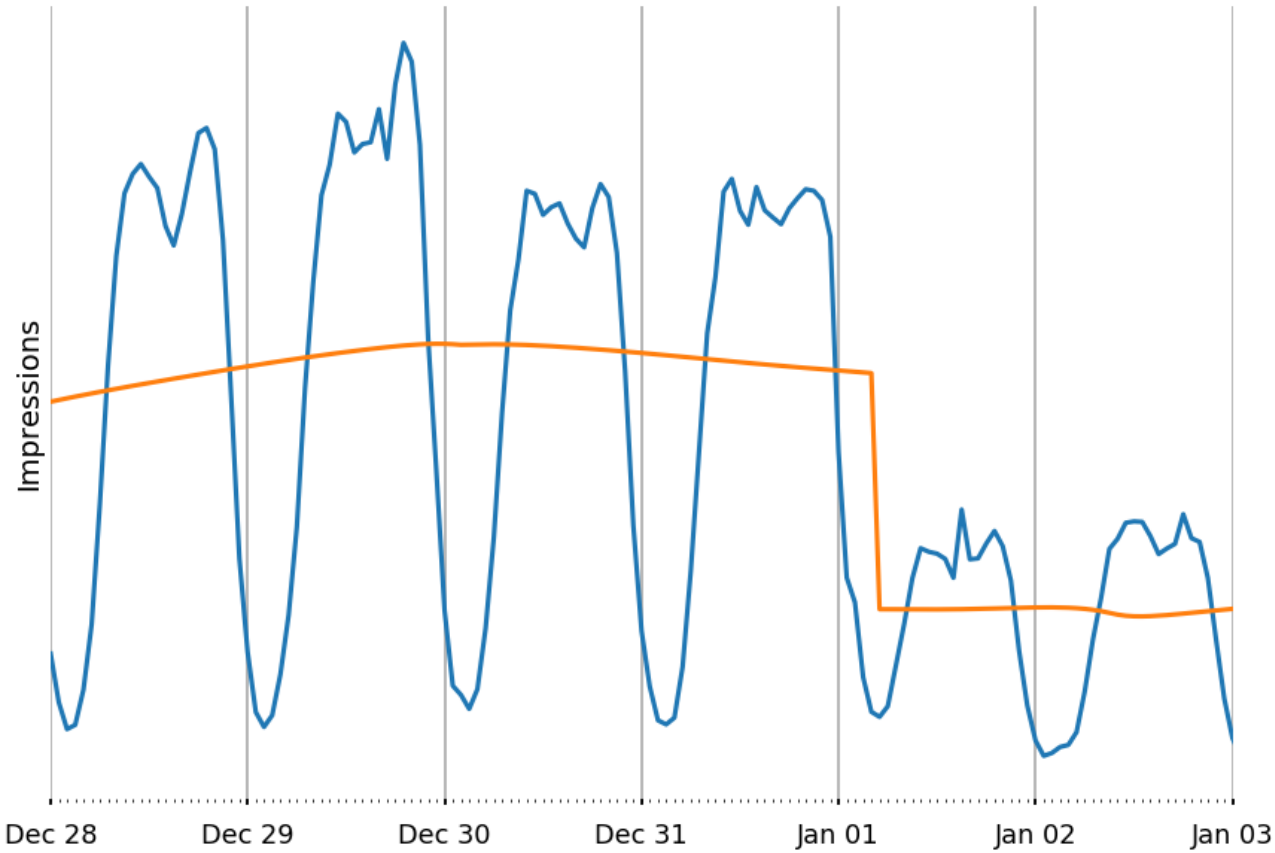
Типичный будний день



Типичный выходной день



Изменение в среднем



Построение модели ряда

- Обозначим временной ряд $Y = (y_1, \ldots, y_n)$
- Наблюдаемые значения можно представить в виде $Y = T + S + E$, где $T = (t_1, \ldots, t_n)$ компонента-тренд, $S = (s_1, \ldots, s_n)$ периодическая компонента, $E = (\epsilon_1, \ldots, \epsilon_n)$ остатки или шум
- Для каждой из этих компонент требуется построить модель

Построение модели ряда

Модель можно задать следующим образом:

$$t_i = c, \quad i = 1, \ldots, n,$$

$$s_i = \sum_{j=1}^{J} A_j \cos\left(\frac{2\pi}{a_j} i + \phi_j\right), \quad i = 1, \ldots, n,$$

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n,$$

где i индекс элемента ряда; j индекс косинуса в периодической компоненте; J — количество косинусов в периодической компоненте; c — константа; A_j — амплитуда j-го косинуса; a_j — период j-го косинуса; ϕ_j — фаза j-го косинуса.

Модель разладки. Изменение в среднем

Модель разладки можно задать следующим образом:

- Разладка только в одной точке ряда;
- Разладка заключается в сдвиге.

τ — точка (индекс) разладки, тогда тренд с разладкой $\tilde{T} = (\tilde{t_1}, \ldots, \tilde{t_n})$, где

$$\tilde{t_i} = \begin{cases} t_i, & i < \tau, \\ t_i + \delta^{(mean)}, & i \geqslant \tau, \end{cases}$$

$\delta^{(mean)}$ — значение разладки. Чтобы разладка была заметна, введем ещё минимальное допустимое значение разладки $\delta_{min}^{(mean)}$, так что:

$$\delta^{(mean)} = \max(\delta^{(mean)*}, \delta_{min}^{(mean)}),$$
$$\delta^{(mean)*} \sim N(\mu^{(\text{cp\_mean})}, \sigma^{2(\text{cp\_mean})}).$$

Модель разладки. Локальная

Отличие от предыдущего типа разладки в том, что в локальной разладки разладка влияет только на одну точку ряда.

$$\tilde{t_i} = \begin{cases} t_i, & i \neq \tau, \\ t_i + \delta^{(local)}, & i = \tau, \end{cases}$$

$$\delta^{(local)} = \max(\delta^{(local)*}, \delta_{min}^{(local)}),$$

$$\delta^{(local)*} \sim N(\mu^{(\text{cp\_local})}, \sigma^{2(\text{cp\_local})}).$$

Остальное остается идентичным предыдущему варианту.

Методы обнаружения разладки

Общая канва

- У временного ряда есть некоторая структура (сигнал)
- Сигнал может быть описан моделью
- Идея подхода: около точки разладки модель плохо описывает временной ряд
- Используя меру ошибки мы можем измерять насколько хорошо описывает выбранная модель реальные данные
- Как только ошибка (отклонение модели от реальных данных) превышает заданный порог, метод сигнализирует о разладке

Можно выделить два типа методов в данном подходе:

- Методы на основе прогнозирования
- Методы на основе аппроксимации

Моделируемый ряд и функции рядов в методе разладки

Исходная модель ряда одна:

$$Y = T + S + E = c + \sum_{j=1}^{J} A_j \cos\left(\frac{2\pi}{a_j} i + \phi_j\right) + \epsilon_i, \quad i = 1, \ldots, n,$$

Моделей сигнала для обнаружения разладки может быть много. Мы будем использовать следующие:

- $f(x|b) = b$
- $f(x|P, p, \chi, b) = P\cos(\frac{2\pi}{p}x + \chi) + b$
- $f(x|\{P_j, p_j, \chi_j\}, b) = \sum_{j=1}^{J} P_j \cos(\frac{2\pi}{p_j}x + \chi_j) + b$

Пусть l — ширина окна. При этом $1 < l < n$, l чётное. С помощью ширины окна из исходного ряда образуется последовательность отрезков $W = \{w_j\}_{j=1}^{k}$, где $k = n - l + 1$ — количество таких отрезков; а $w_j = (y_j, \ldots, y_{j+l-1})$ — j-ый отрезок. Каждый отрезок w_j в свою очередь делится на два отрезка одинаковой длины:
$W^{(\text{left})} = \{w_j^{(\text{left})}\} = \{(y_j, \ldots, y_{j+\frac{l}{2}-1})\}$ и
$W^{(\text{right})} = \{w_j^{(\text{right})}\} = \{(y_{j+\frac{l}{2}}, \ldots, y_{j+l-1})\}$.
Таким образом, для каждого ряда W можно сформировать тройки рядов:

$$W^{(\text{all})} = \{w_j^{(\text{all})}\}_{j=1}^{k} = \{(w_j; w_j^{(\text{left})}; w_j^{(\text{right})})\}_{j=1}^{k}.$$

Аппроксимация

Пусть есть функция ошибки $e(\cdot)$, такая что:

$$e(X) = \min_{\theta} \sum_{p=1}^{m} (x_p - f(x_p|\theta))^2,$$

где $X = (x_1, \ldots, x_m)$ — вещественный временной ряд длины m, а $f(x|\theta)$ — модель сигнала этого временного ряда с параметрами θ.

Мера ошибки позволяет нам рассчитать, насколько хорошо аппроксимируется отрезок ряда с помощью выбранной модели. Однако, необходимо еще ввести функцию разладки:

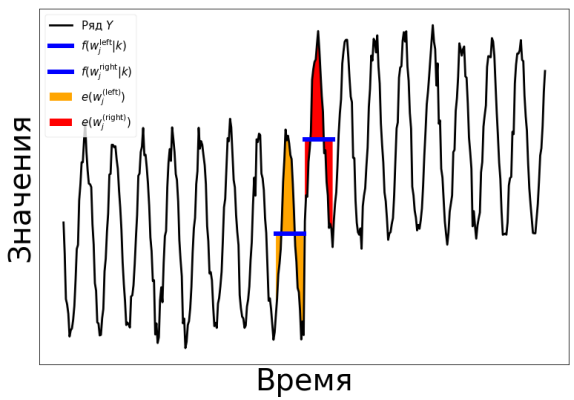
$$f_j = F(w_j^{(\text{all})}) = \frac{e(w_j) - e(w_j^{(\text{left})}) - e(w_j^{(\text{right})})}{h},$$

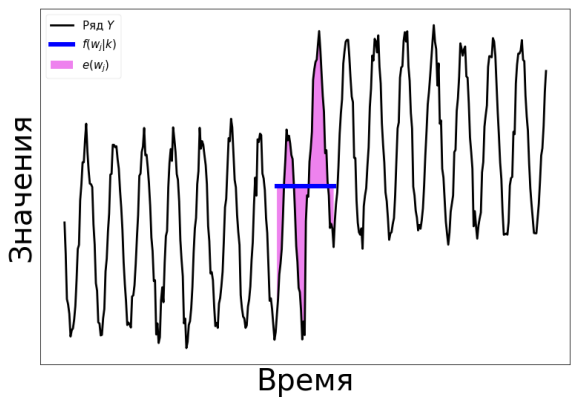
где h — значение нормировки, $j = 1, \ldots, k$.
Синхронизация: f_1 соответствует y_l, а f_k соответсвует y_n. Введем синхронизированную функцию разладки :

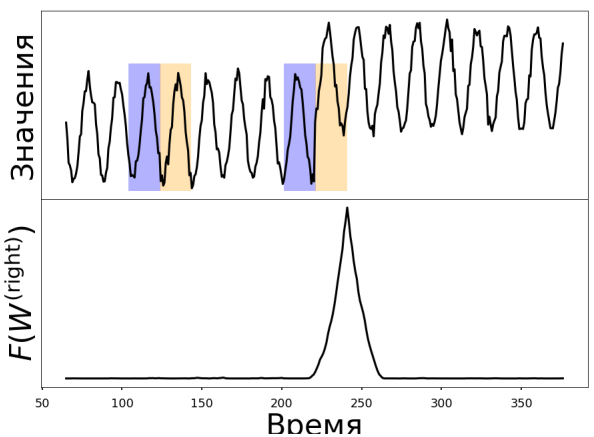
$$q_i = \begin{cases} f_{i-l+1}, & i \geq l, \\ 0, & i < l. \end{cases}$$

Аппроксимация

- Итого, взяв ряд Y, мы «скользим» по нему окном ширины l
- Рассчитываем значения функции разладки $F()$ для каждого из получаемых отрезков $W^{(\text{all})}$
- Функция разладки начинает расти в окрестности точки разладки τ,
- Следовательно можно задать порог γ, такой что при превышении функции разладки этого порога в какой-то точке $\hat{\tau}$, разладка будет обнаружена





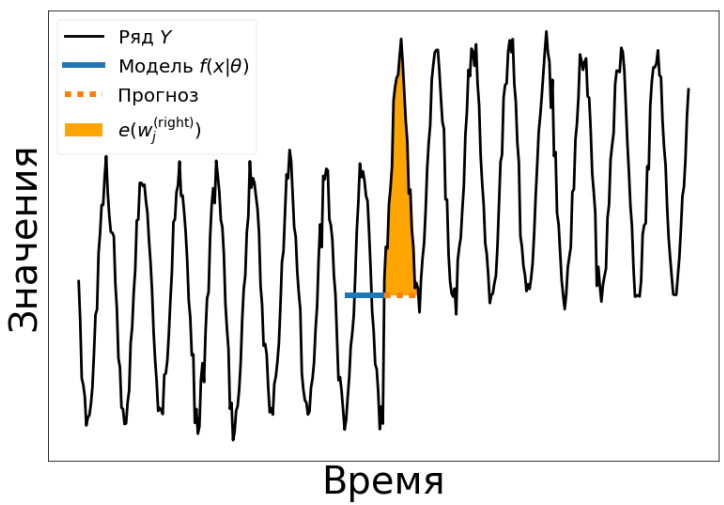


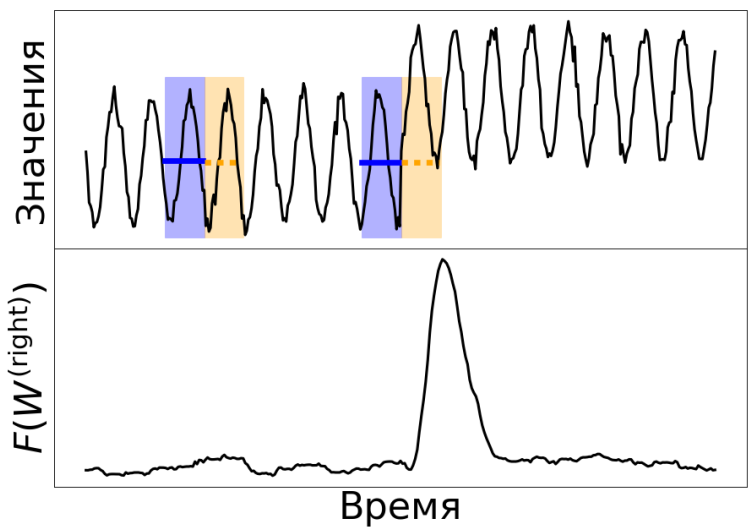
- Строим прогноз на несколько точек ряда вперед и считаем отклонение фактических значений от прогнозных
- В случае, если отклонение выше заданного порога, метод обнаруживает разладку
- Формально, оставаясь в тех же обозначениях, есть та же ширина окна l
- Есть последовательность отрезков $W = \{w_j\}_{j=1}^{k}$
- Каждый отрезок w_j делится в этом методе на два ряда не обязательно одинаковой длины
- Введем индекс g, который будет указывать в какой точке ряда w_j он будет разделен на два
- формируется набор из пар рядов: $W^{(\text{left})} = \{w_j^{(\text{left})}\} = (y_j, \ldots, y_{j+g})$ и $W^{(\text{right})} = \{w_j^{(\text{right})}\} = (y_{j+g}, \ldots, y_{j+l})$

Прогнозирование

Ключевое отличие от методов аппроксимации: вместо расчета меры ошибки на том же ряду на котором подбирались параметры модели, мы оцениваем параметры θ модели $f(x|\theta)$ на ряде $w_j^{(\text{left})}$, делаем прогноз на $l - g$ точек и рассчитываем функцию ошибки $e(\cdot)$ на ряде $w_j^{(\text{right})}$.
Функция разладки принимает следующий вид:

$$f_j = F(w_j^{(\text{right})}) = \frac{e(w_j^{(\text{right})})}{h}.$$





Оценка качества

Допущения

В рамках данной работы мы разрабатываем систему своевременного оповещения о разладках во временных рядах.

- Нам важны две характеристики: точность и скорость обнаружения разладки
- Нам точно известны ряды с разладками и без
- Можем строить матрицы ошибок классификации и считать метрики качества
- Мы фиксируем точку разладки τ и приемлемую задержку d

Матрица ошибок классификации

Таким образом, у нас имеется приемлемая задержка, в рамках которой мы хотим обнаружить разладку. При этом, за пределами приемлемой задержки нас не интересует что происходит с рядом. Исходя из этого возможны четыре варианта:

- Разладка произошла и метод обнаружил точку разладки в диапазоне $(\tau, \cdots, \tau + d)$. Такая ситуация попадает под категорию True positive.
- Разладка произошла и метод не обнаружил точку разладки в диапазоне $(\tau, \cdots, \tau + d)$. Это случай False negative.
- Метод обнаружил разладку в диапазоне $(\tau, \cdots, \tau + d)$ в ряде без разладки. Это ситуация False positive.
- Разладки не было и метод не обнаружил разладку в диапазоне $(\tau, \cdots, \tau + d)$. Это случай True negative.

Задача классификации

В рамках данной работы мы будем строить классифицирующее правило:

$$a(Y) = \begin{cases} 1, & f_j \geq \gamma \text{ и } \tau - l \leq j \leq \tau - l + d, \\ 0, & f_j < \gamma \text{ или } j < \tau - l \text{ или } j > \tau - l + d. \end{cases}$$

Таким образом, метод будет определять, либо не определять разладку в заданном диапазоне в зависимости от задаваемого порога γ.

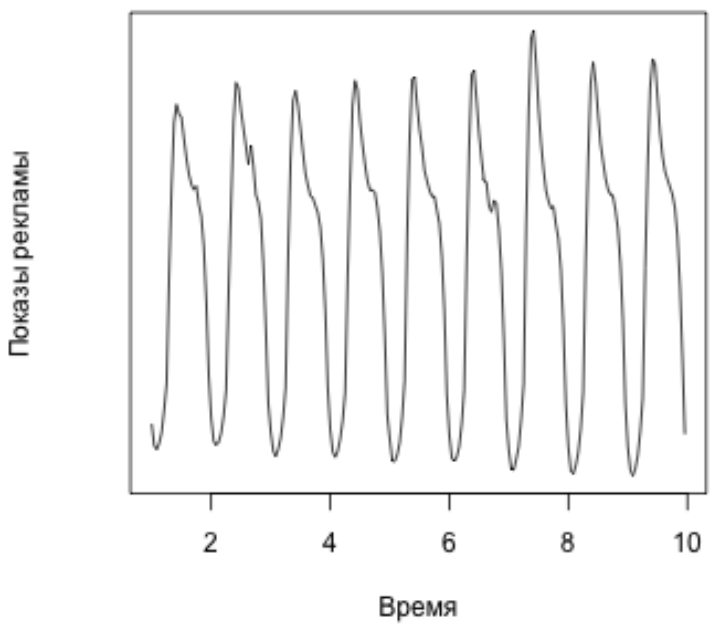
ROC-кривая

- Можно строить ROC-кривые (изменяя порог γ) для разных методов обнаружения разладки, сравнивая как работают те или иные методы в контролируемой среде эксперимента.
- ROC-кривая — график, позволяющий оценить качество бинарной классификации. Он отображает соотношение между долей верно-положительно классифицированных наблюдений от общего количества положительных классов, и долей ложно-отрицательно классифицированных наблюдений от общего количества отрицательных наблюдений при варьировании порога γ.
- Другими словами, ROC кривая это график, где по оси ординат откладывается TPR (англ. True Positive Rate), а по оси абсцисс откладывается FPR (англ. False Positive Rate). При этом каждая точка является значением TPR и FPR для какого-то конкретного значения порога.
- $TPR(\gamma) = \frac{\sum \text{Верно-положительные классификации}}{\sum \text{Все положительные наблюдения}}$,
  $FPR(\gamma) = \frac{\sum \text{Ложно-отрицательные классификации}}{\sum \text{Все отрицательные наблюдения}}$
- Для сравнения качества методов мы будем пользоваться метрикой ROC-AUC, которая является площадью под ROC-кривой.

Моделирование данных

Моделировать ряд будем как сумму тренда, периодики и шума. Тренд будем брать за константу, а периодику зададим как сумму косинусов с определенными периодичностями, амплитудами и фазами.

Пример реального ряда



Оценка параметров

Длина ряда с предыдущего слайда 216 (то есть 9 суток). Применим к этому ряду метод SSA с окном 96. И оценим параметры периодичности по первым 10 компонентам (исключив тренд).

Периоды	Фазы	Фазы (примерные)	Амплитуды
24	2.78	$8\pi/9$	1.00
12	1.55	$\pi/2$	0.39
8	-1.56	$-\pi/2$	0.13
6	-2.95	$-15\pi/16$	0.11

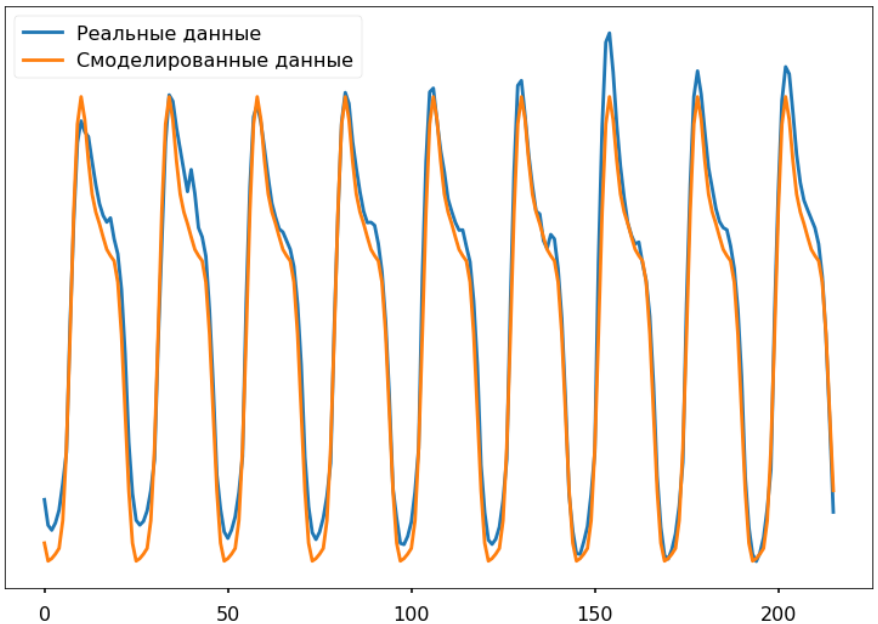
Таким образом, модель периодической составляющей s_i нашего ряда можно записать в следующем виде:

$$s_i = \cos(\frac{2\pi}{24}i+\frac{8\pi}{9})+0.39\cos(\frac{2\pi}{12}i+\frac{\pi}{2})+0.13\cos(\frac{2\pi}{8}i-\frac{\pi}{2})+0.11\cos(\frac{2\pi}{6}i-\frac{15\pi}{16}),$$

$$i = 1, \ldots, n.$$

Оценка параметров

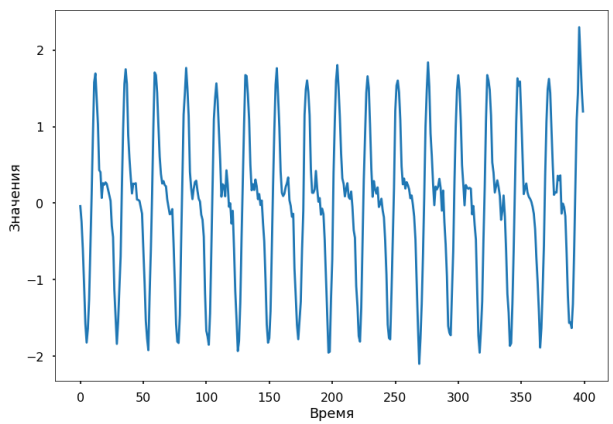
В результате, модель сигнала (без шума) получилась внешне достаточно похожая на исходные реальные данные:



Прочие параметры модели

- Длину ряда зафиксируем $n = 400$
- Значение тренда выберем нулевым: $c = 0$, то есть $t_i = 0, i = 1, \cdots, n$
- Параметры шума возьмем $\mu = 0, \sigma = 0.1$
- Величины разладки $\delta^{(mean)*} \sim N(\mu = 0, \sigma = 0.2)$, $\delta^{(local)*} \sim N(\mu = 0, \sigma = 1)$;
- Минимальные допустимые значения разладок: $\delta^{(mean)*}_{min} = 0.3$, $\delta^{(local)*}_{min} = 0.5$;
- Место возникновения разладки зададим в середине ряда ряда $\tau = 216$
- Задержек выберем несколько $d = (4, 24, 48)$

Пример сгенерированного ряда с шумом и разладкой



Применение методов

Моделирование рядов

Попробуем применить, описанные выше модели к смоделированным данным.

- Смоделируем 50 рядов
- У каждого ряда начало периодической компоненты выбирается случайно (то есть первый ряд может начинаться с нулевого часа, второй с пятого и т.п.). Это сделано, чтобы нивелировать влияние периодичности на оценку качества метода.
- Параметры методов выбраны следующие. Длина окна l принимает значения 2, 4, 24, 48, 96.
- Разладка возникает двух типов: локальная, разладка в среднем

Методы

И для подхода с аппроксимацией и для подхода с прогнозированием мы будем использовать следующие модели:

- Среднее $f(x|b) = b$
- Четыре косинуса с периодами из модели генерации ряда $f(x|P_i, p_i, \chi_i, b) = \sum_{i=1}^{4} P_i \cos(\frac{2\pi}{p_i}x + \chi_i) + b$, где $p_1 = 24, p_2 = 12, p_3 = 8, p_4 = 6$
- Один косинус с периодом 24 + тренд $f(x|P, 24, \chi, b) = P\cos(\frac{2\pi}{24}x + \chi) + b$
- Среднее + тренд $f(x|b) = b + cx$

Таким образом, у нас есть 7 методов с одной стороны, и решетка параметров из 5 вариантов (длина окна l) с другой. Мы будем оценивать качество всех методов для комбинаций методов и параметров.

Обратите внимание

Методы, в которых лежит модель, отличная от среднего, бессмысленно применять для окон l менее 48. Поскольку невозможно оценить какие либо параметры синуса, если длина ряда менее одного периода.

В таблице приведены сводные результаты ROC-AUC для экспериментов на 50 временных рядах.

Результаты применения методов к смоделированным данным

Тип разладки	локальная															в среднем													
Задержка (d)	4					24					48					4					24					48				
Длина окна (l)	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96	2	4	24	48	96
Место разладки	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216	216
Метод																														
approximation_mean	0,98	0,76	0,50	0,77	0,67	0,95	0,64	0,56	0,83	0,71	0,92	0,67	0,63	0,86	0,71	0,67	0,58	0,52	0,82	0,72	0,56	0,54	0,92	1,00	1,00	0,60	0,53	0,92	1,00	1,00
approximation_sin				0,93	0,82				0,95	0,81				0,94	0,85				0,95	0,85				1,00	1,00				1,00	1,00
approximation_sin_insight				0,97	0,88				0,94	0,91				0,98	0,92				0,99	0,86				1,00	1,00				1,00	1,00
approximation_trend				0,58	0,50				0,53	0,56				0,56	0,52				0,50	0,54				0,55	0,52				0,77	0,52
prediction_mean	0,95	0,80	0,59	0,69	0,70	0,93	0,66	0,57	0,82	0,71	0,93	0,67	0,61	0,75	0,73	0,73	0,58	0,54	0,67	0,59	0,51	0,47	0,86	0,99	0,94	0,55	0,48	0,81	0,96	0,98
prediction_sin				0,87	0,77				0,94	0,92				0,90	0,76				0,70	0,65				1,00	0,99				1,00	1,00
prediction_sin_insight				1,00	0,98				0,98	0,99				0,99	0,98				0,99	0,96				1,00	1,00				1,00	1,00

Выводы