Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Мерзляков Климент Викторович

Обнаружение разладки во временных рядах показов мобильной рекламы

Отчет о научно-исследовательской работе

Научный руководитель:
к. ф.-м. н., доцент Н. Э. Голяндина

Санкт-Петербург
2018

Оглавление

Введение

Рекламной сетью называют некоторую площадку или систему, которая является посредником между рекламодателями и собственниками рекламных мест — владельцев сайтов, мобильных приложений и каких-либо других пространств, где можно размещать рекламу.

В интернет-рекламе взаимодействие рекламной сети с пользователем можно описать следующей последовательностью событий. При выполнении некоторых условий (например, пользователь открыл мобильное приложение) с устройства пользователя отправляется запрос на показ рекламы. Если запрос удовлетворяется, то происходит событие „показ“, то есть пользователь непосредственно видит рекламу. После этого может произойти событие „клик“ и далее какое-либо целевое действие. В мобильной интернет-рекламе „показ“ является одним из ключевых событий, поскольку он отражает количество рекламы доставленное до конечного пользователя.

Рекламные интернет-сети являются интересным объектом для исследования с точки зрения обнаружения разладки, поскольку все показатели отслеживаются с точностью до секунды, происходит большое количество событий, а так как рекламные сети, как правило, работают на международном рынке, то существует возможность тестировать гипотезы на большом количестве различных временных рядов.

Одной из текущих проблем, стоящих перед рекламными сетями — это низкая скорость реагирования на любые резкие изменения текущего состояния. Такие изменения безусловно отражаются в данных в виде аномальных значений, резких всплесков и внезапных изменений тренда. Однако проблема заключается в том, что показателей требующих отслеживания могут быть десятки, при этом на каждый показатель может влиять большое количество факторов. Поэтому зачастую, чтобы локализовать и устранить проблему требуется просмотреть сотни графиков. Отсюда следует, что наличие качественного метода обнаружения разладки каждого показателя по каждому измерению позволило бы не только существенно сэкономить ресурсы, но и в целом повысить эффективность бизнеса. Поэтому целью данной работы является разработка методики обнаружения разладки. В работе будут использоваться фактические, данные одной из работающих рекламных сетей.

Глава 1

Обнаружение разладки во временных рядах

1.1. Моделирование данных

Реальные данные интернет-рекламы имеют стабильную дневную периодичность (на рисунке 1.1 приведен пример типичной динамики в рамках дня). По более длинному ряду изображенному на рисунке 1.2 видно, что в данных время от времени возникают разладки разных видов, при этом сам ряд имеет мультипликативный характер (с изменением среднего уровня ряда пропорционально меняется и амплитуда колебаний). В реальных временных рядах достаточно сложно разметить наличие разладок — зачастую сложно отделить разладку от шума. Поэтому вместо разметки реальных рядов мы будем моделировать искусственные ряды похожие на ряды данных интернет-рекламы с определенным шумом и разладками в известных местах.

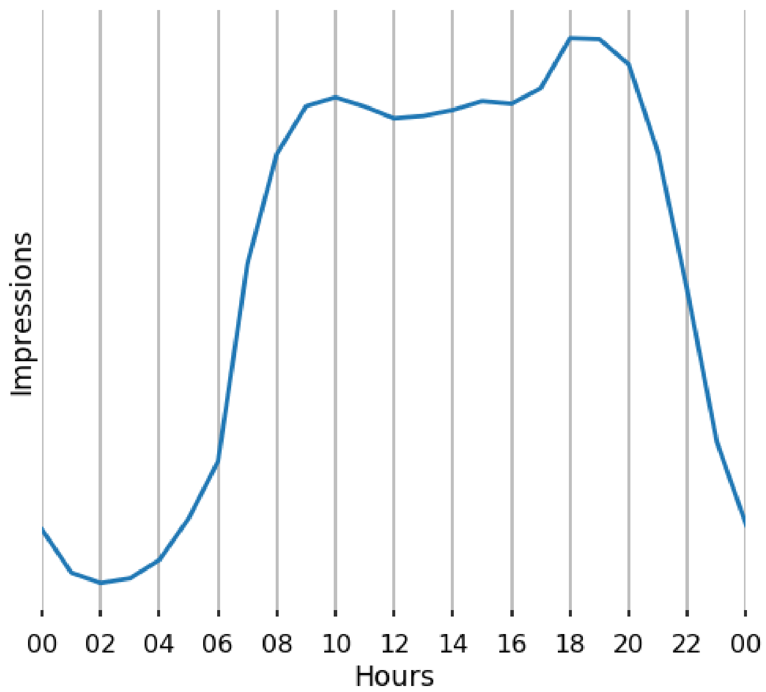

Рис. 1.1. Пример показов рекламы за сутки

Заменить рисунок с длиным рядом на рисунок с изменением среднего.

Обозначим временной ряд $Y = (y_1, \ldots, y_n)$. Наблюдаемые значения ряда можно пред-

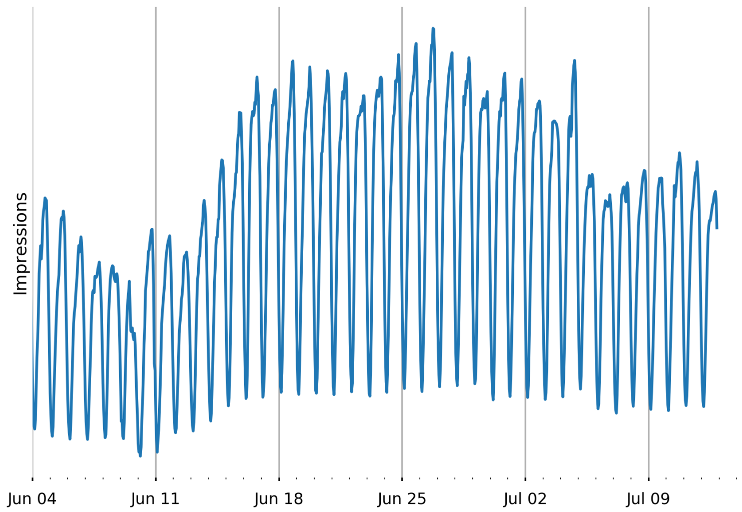

Рис. 1.2. Показы рекламы в одной стране за один месяц

ставить в виде суммы компонент:

$$Y = T + S + E$$

, где $T = (t_1, \dots, t_n)$ компонента-тренд, $S = (s_1, \dots, s_n)$ периодическая компонента, $E = (\epsilon_1, \dots, \epsilon_n)$ остатки или шум. Каждую из этих компонент требуется промоделировать. Это можно сделать, например, следующим образом:

$$t_i = const$$

$$s_i = A\cos(\frac{2\pi}{a}i + \phi).$$

, где i индекс ряда; A — амплитуда; a — период; ϕ — фаза

$$\epsilon_i \sim N(\mu, \sigma^2)$$

Ряды, которые мы хотим смоделировать имеют мультипликативность (амплитуда колебаний меняется пропорционально изменению тренда). При моделировании, такого эффекта можно достичь взяв экспоненту от моделируемого ряда:

$$Y = e^{T+S+E}$$

С экспонентой пока получилось нескладно. Сначала Y это одно, теперь другое. Пока двигаюсь дальше, позже вернусь к этому месту.

Далее нам нужно промоделировать разладку. Модель разладки имеет следующий вид:

- Разладка только в одной точке ряда
- Разладка только в тренде
- При этом разладка может отсутствовать с некоторой вероятностью

Формально это можно описать так: пусть τ — точка (индекс) разладки, тогда тренд с разладкой будет иметь вид:

$$\tilde{t}_i = \begin{cases} t_i, & i < \tau \\ t_i + \delta, & i \geqslant \tau \end{cases}$$

, где δ — значение разладки.

Тренд с разладкой тогда будет выглядеть следующим образом:

$$\tilde{T} = (\tilde{t}_1, \ldots, \tilde{t}_n)$$

Значение разладки является случайной величиной, которую можно моделировать по тому или иному закону. В данной работе мы будем моделировать значение разладки из нормального распределения, с некоторой вероятностью возникновения ρ:

$$\delta = \begin{cases} \delta^* \sim N(\mu^{(cp)}, \sigma^{2(cp)}), & \rho \\ 0, & 1 - \rho \end{cases}$$

Переписать это как смесь распределений

При этом точка разладки τ тоже является случайной величиной с равномерным распределением на $[n_0, \ldots, n]$, где n_0 — самая первая возможная точка разладки, которая задается параметром. n_0 введена намеренно, чтобы разладка при моделировании не возникала в первых точках ряда.

Таким образом, моделируемый ряд с разладкой будет иметь следующий вид:

$$Y = e^{\tilde{T} + S + E}$$

В результате модель временного ряда имеет пять параметров: A, a, ϕ, μ, σ, а модель разладки имеет еще пять параметров: $\delta, \mu^{(cp)}, \sigma^{(cp)}, \rho, n_0$.

1.2. Методы обнаружения разладки

Опишем один из подходов к обнаружению разладки. Данный подход не является единственным, хотя заключает в себе широкое разнообразие методов. Как правило, у временного ряда есть некоторая структура (сигнал), котора может быть описана той или иной моделью. Идея подхода заключается в том, что около точки разладки модель плохо описывает временной ряд. Используя некоторую функцию потерь мы можем измерять то, насколько хорошо или плохо описывает выбранная модель реальные данные в некотором диапазоне. Можно выделить два типа методов в данном подходе:

- Методы на основе предсказания
- Методы на основе аппроксимации

Разберемся как работают методы на основе аппроксимации.

Пусть l четное вещественное число, называемое длиной окна. При этом $1 < l < n$. С помощью длины окна из исходного ряда образуется последовательность подрядов $W = \{w_j\}_{j=1}^{k}$, где $k = n - l + 1$ — количество таких подрядов; а $w_j = (y_j, \ldots, y_{j+l+1})$ — j-ый подряд. Каждый подряд w_j в свою очередь делится на два подряда одинаковой длины (это возможно поскольку l четное по условию): $W^{(left)} = \{w_j^{(left)\}} = (y_j, \ldots, y_{\frac{j+l+1}{2}})$ и $W^{(right)} = \{w_j^{(right)}\} = (y_{\frac{j+l+1}{2}}, \ldots, y_{j+l+1})$.

Таким образом, для каждого ряда W можно сформировать тройки рядов:

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

Пусть есть функция потерь $c(\cdot)$, такая что:

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

, где x некоторый вещественный временной ряд длины m, а $f(x|\theta)$ некоторая модель этого временного ряда с параметрами θ.

Пока не разобрался как обозначить то что мы функции $c()$ на самом деле подаем набор рядов.

.

Функция $f(x|\theta)$ может быть линейной ($\theta = (const)$):

$$f(x|const) = const$$

Либо другой подходящей под наш ряд функцией, например:

$$f(x|A, a, \phi) = A \cos(\frac{2\pi}{a} x + \phi)$$

.

Введем функцию разладки:

$$F(W^{(all)}) = \frac{c(W) - c(W^{(left)}) - c(W^{(right)})}{h}$$

, где h — значение нормировки.

Расчет нормы является открытой проблемой, поскольку имеются разные варианты её расчета со своими плюсами и минусами. Например, можно рассчитывать её как значение функции потерь на первом отрезке ряда (предполагая, что на этом отрезке не происходило разладок):

$$h = f(w_1^{(all)}) = c(w_1) - c(w_1^{(left)}) - c(w_1^{(right)})$$

1.3. Оценка качества

В рамках данной работы мы разрабатываем систему своевременного оповещения о разладках во временных рядах. При такой постановке задачи важны две характеристики: точность обнаружения разладки и скорость обнаружения разладки. Поскольку мы используем моделированные данные, то мы точно знаем в каких из смоделированных нами рядов произошла разладка, а в каких разладки не было. Более того, мы точно знаем момент разладки. Благодаря этому мы можем строить матрицы сопряжённости и считать метрики качества.

Разберем четыре возможных варианта:

- Разладка произошла и метод обнаружил точку разладки **после** фактической точки. Такая ситуация попадает под категорию True positive

- Разладка произошла и метод не обнаружил точку разладки, либо обнаружил ее до фактической разладки. False negative

- Разладки не было, но метод обнаружил разладку. False positive

- Разладки не было и метод не обнаружил разладку. True negative

Заключение

Список литературы

1. Голяндина Н.Э. Метод „Гусеница“-SSA: прогноз временных рядов. Учебное пособие. СПб., 2003.
2. Armstrong. Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Publishers, 2001.
3. Dagum, Estela, Bianconcini. Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation, 2016.
4. N. Golyandina, V. Nekrutkin, A. Zhigljavsky. Analysis of Time Series Structure - SSA and Related Techniques, 2001.
5. R. Hyndman, G. Athanasopoulos. Forecasting: Principles and Practice. 2013.
6. R. Hyndman. Moving averages. 2009.