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 3

Глава 1. Обнаружение разладки во временных рядах 4

1.1. Моделирование данных 4

1.2. Методы обнаружения разладки 5

1.3. Оценка качества 6

Заключение 8

Список литературы 9

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.

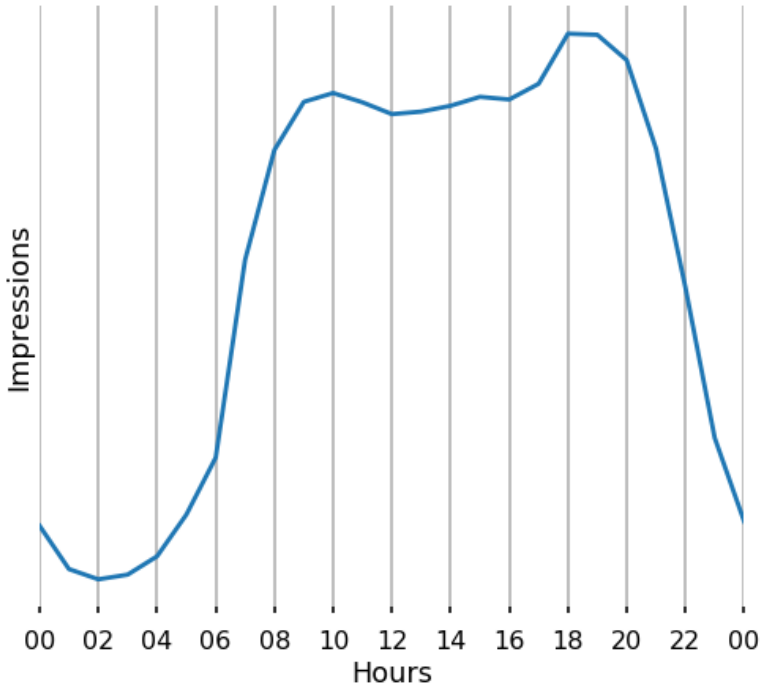

Рис. 1.1. Пример показов рекламы за сутки

Сигнал дневной периодичности будем моделировать с помощью косинуса с периодом 24 и амплитудой :

$$S_t = \cos(\frac{2\pi}{24}t + \frac{\pi}{2}).$$

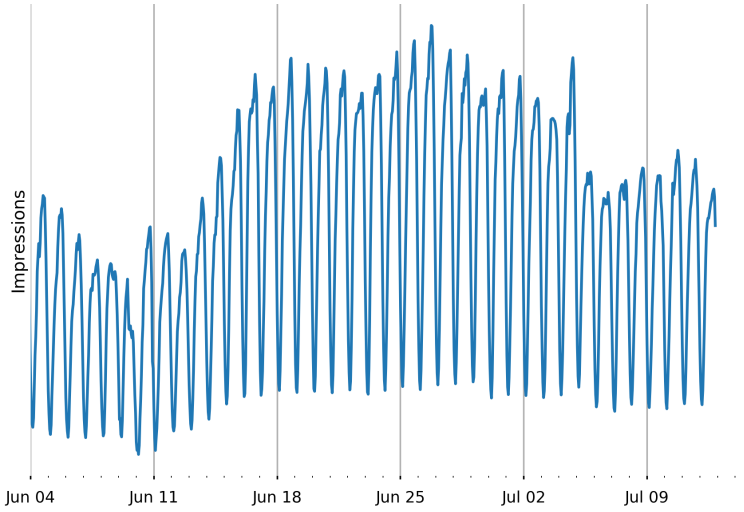

Рис. 1.2. Показы рекламы в одной стране за один месяц

Шум будем генерировать как нормальное распределение:

$$R_t \sim N(0, 0.5)$$

Разладку будем генерировать только в одной точке ряда, как изменение среднего. При этом будем брать небольшой отступ от начала ряда (чтобы разладка не происходила в начале ряда). Разладка может произойти в любой из точек ряда равновероятно (не считая начальных, в отступе). Значение разладки:

$$CP = \begin{cases} N(0, 3), & \rho = 0.8. \\ 0, & \rho = 0.2 \end{cases}$$

Таким образом, временной ряд принимает следующий вид:

$$X = S + R + CP$$

Поскольку реальные данные имеют мультипликативный характер, то мы можем преобразовать наш ряд в мультипликативный просто взяв экспоненту:

$$X = e^{S+R+CP}$$

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 ту или иную 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. Задается некоторое окно w, в рамках которого производятся расчеты. Это окно делится пополам на w_l и w_r. Далее мы скользим этим окном по временному ряду рассчитывая значения функции разладки ($F(t)$) в каждой точке ряда.

$$F(t) = \frac{C(W) - C(w_{left}) - C(w_{right})}{N}$$

, где C — функция потерь, а N некоторая норма, чтобы привести значения функции потерь к некоторому общему виду.

В качестве функции потерь можно использовать сумму квадратичных отклонений фактических значений ряда от модельных:

$$C(y) = \sum_{i=1}^{m}(y_i - \hat{y})^2$$

, где y - временной ряд, y_i одно значение временного ряда, а $\hat{y}$ модельное значение временного ряда.

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).

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.