Оценка параметров модели случайного процесса нагрузки сервера, обрабатывающего заявки

Владимир Руцкий, 5057/12

Санкт-Петербургский государственный политехнический университет

15 июня 2011

План презентации

Постановка задачи

Сервер

- Сервер обрабатывает поступающие заявки
- Для обработки заявки используются ресурсы сервера
    - Количество используемых ресурсов сервера — загрузка сервера — скалярная величина, например процессорное время

Задача

Дан лог загрузки сервера. Необходимо:

1. идентифицировать моменты поступления заявок,
2. оценить:
    - интенсивность поступления заявок,
    - загрузку сервера в фоновом режиме,
    - использование ресурсов сервера для обработки одной заявки

Математическая модель (1)

Загрузка сервера — случайный процесс $X(t)\colon \mathbb{R} \to \mathbb{R}$

Фоновая загрузка сервера

Сумма постоянной загрузки и винеровского процесса (шум):

$$B(t) = m + \sigma\mathcal{W}(t)$$

Загрузка сервера при обработке заявки

Заявка, поступившая в $t = t_c$, увеличивает загрузку сервера на:

$$K_{t_c}(t) = \mathcal{N}(m_c, \sigma_c^2) \cdot \mathrm{I}(t - t_c) \cdot e^{-\lambda_c(t-t_c)}$$

$\mathrm{I}(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geqslant 0 \end{cases}$ — Функция Хевисайда

Математическая модель (2)

$T_c = \{t_{c_1}, \ldots, t_{c_R}\}$ — множество моментов времени, когда поступили заявки (всего R заявок за время наблюдения за сервером)

Интенсивность поступления заявок

Интервал времени между поступлением двух последовательных заявок распределён экспоненциально:

$$(t_{c_i} - t_{c_{i-1}}) \sim \mathrm{Exp}(\lambda)$$

Общая загрузка сервера

$$X(t) = B(t) + \sum_{t_c \in T_c} K_{t_c}(t)$$

Перейдём к дискретному случайному процессу

$$X(t_i), \quad t_i = t_0 + i \cdot \Delta t \quad (t_0 \text{ и } \Delta t \text{ даны})$$

Входные данные

Лог загрузки сервера

Траектория $X(t_i)$: $\{x_i \mid i = 1, \ldots, N\}$

Считаем, что Δt достаточно мало и $t_{c_j} = t_0 + i \cdot \Delta t$ — заявки поступили в некоторые наблюдаемые моменты времени t_i.
Рассматривается случай $t_0 = 0$

Задача

Задача

По траектории $X(t_i)$

1. идентифицировать моменты поступления заявок,
2. оценить параметры модели:
   - m, σ — параметры фоновой загрузки,
   - λ — интенсивность поступления заявок,
   - m_c, σ_c, λ_c — параметры загрузки сервера при обработке заявок

Случай бесконечного времени обработки заявки

$\lambda_c \approx 0$

При поступлении заявки в момент времени t_c загрузка сервера увеличивается на $\Delta x \sim \mathcal{N}(m_c, \sigma_c)$

$$K_{t_c}(t) = \mathcal{N}(m_c, \sigma_c^2) \cdot \mathrm{I}(t - t_c)$$

Разностный аналог производной

d$X(t)$

Рассмотрим ненормированный разностный аналог производной:

$$\mathrm{d}X(t) = X(t) - X(t - \Delta t)$$

d$X(t)$ в момент времени поступления заявки t_c

$$\mathrm{d}X(t_c) = \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2)$$

(при условии, что в момент времени $(t_c - \Delta t)$ заявки не было)

d$X(t)$ в момент времени отсутствия заявок t

$$\mathrm{d}X(t) = \mathcal{N}(0, \sigma^2 \Delta t)$$

(при условии, что в момент времени $(t - \Delta t)$ заявки не было)

Итеративный метод идентификации заявок (1)

Предположим, что в отрезке времени $[t_k, t_{k+n}]$ не поступило ни одной заявки.
Тогда $\mathrm{d}x_k, \ldots, \mathrm{d}x_{k+n}$ — наблюдения $\mathcal{N}(0, \sigma^2 \Delta t) = \mathrm{d}X(t)$.
Оценим σ^2 по $[x_k, x_{k+n}]$:

$$\widehat{\sigma}^2 \Delta t = \frac{1}{n-1} \sum_{i=1}^{n} ((x_{k+i} - x_{k+i-1}) - 0)^2$$

Гипотеза H_0

В отрезке времени $[t_{k+n}, t_{k+n} + \Delta t]$ не поступило ни одной заявки

Критерий принятия H_0 с уровнем значимости α

Разность значений наблюдений $(x_{k+n+1} - x_{k+n})$ лежит в $(1-\alpha)$ квантиле нормального распределения $\mathcal{N}(0, \widehat{\sigma}^2 \Delta t)$:

$$H_0 \text{ принимается } \iff (x_{k+n+1} - x_{k+n}) < \mathcal{N}_{1-\alpha}$$

Итеративный метод идентификации заявок (2)

Алгоритм итеративной идентификации первой заявки

$\textsc{Find-Request-Index}\left(D = \left\{\mathrm{d}x_{i+\text{offset}}\right\}, n, \alpha\right)$

for $k \leftarrow n$ **to** $length[D] - 1$

  do H_0*-criterion* $\leftarrow \textsc{Build-H0-Criterion}\left(D[1\,..\,k], \alpha\right)$

    if not H_0*-criterion*$(D[k+1])$

      then $\triangleright$ $(k+1)$-й выброс отвергает $H_0 \implies$ заявка

        return $k+1$

return 0 $\triangleright$ Заявка не обнаружена

Итеративный метод идентификации заявок (3)

Алгоритм итеративной идентификации всех заявок

$\textsc{Find-Requests-Iterative}\left(D = \{\mathrm{d}x_i\}, n, \alpha\right)$

$\widehat{T_c} \leftarrow \varnothing$

$idx \leftarrow 0$ ▷ Индекс последней идентифицированной заявки

while TRUE

  do ▷ Находим индекс следующей заявки

    $idx \leftarrow \textsc{Find-Request-Index}\left(D\left[idx + 1 \,..\, length[D]\right], n, \alpha\right)$

    if $idx \neq 0$

      then $\widehat{T_c} \leftarrow \widehat{T_c} \cup \{t_{idx}\}$

      else **return** $\widehat{T_c}$

Оценка параметров модели

Интенсивность поступления заявок

$$\widehat{\lambda} = \frac{1}{|\widehat{T_c}|} \sum_{i=1}^{|\widehat{T_c}|} (t_{c_{i+1}} - t_{c_i})$$

Параметры фоновой загрузки

$$\widehat{m} = \frac{1}{K-1} \sum_{i=1}^{K-1} x_i, \quad \widehat{\sigma}^2 = \frac{1}{K-2} \sum_{i=1}^{K-1} (x_i - \widehat{m})^2, \quad K = t_{c_1}/\Delta t$$

Параметры загрузки при обработке заявок

$$\widehat{m_c} = \frac{1}{|T_c|} \sum_{t_c \in T_c} \mathrm{d}x_{t_c/\Delta t}, \quad \widehat{\sigma}^2 \Delta t + \widehat{\sigma_c}^2 = \frac{1}{|T_c| - 1} \sum_{t_c \in T_c} \left(\mathrm{d}x_{t_c/\Delta t} - \widehat{m_c} \right)^2$$

EM-алгоритм. Скрытые случайные величины

Z_i

Введём скрытые случайные величины $Z_i, \quad i = 1, \ldots, N$

- $Z_i = 1$, если в момент времени t_i поступила заявка,
- $Z_i = 2$, если в момент времени t_i не поступила заявка

Количество поступивших за время наблюдения заявок — случайная величина, распределённая по закону Пуассона $\mathcal{P}(\lambda)$.
$\mathcal{P}(\lambda)$ можно аппроксимировать Биномиальным законом $\mathcal{B}(N, \tau)$.
$\implies$ Z_i можно считать распределённой по закону Бернулли.
$\mathbf{P}(Z_i = 1) = \tau_1$, $\mathbf{P}(Z_i = 2) = \tau_2 = 1 - \tau_1$

$$
\begin{array}{lclcll}
\mathrm{d}X(t_i)|(Z_i = 1) & \sim & \mathcal{N}(\mu_1, \sigma_1^2) & = & \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2), & (t_i \in T_c), \\
\mathrm{d}X(t_i)|(Z_i = 2) & \sim & \mathcal{N}(\mu_2, \sigma_2^2) & = & \mathcal{N}(0, \sigma^2 \Delta t), & (t_i \notin T_c).
\end{array}
$$

Введём обозначения: $\theta = (\tau_1, \tau_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, $\mathrm{d}\mathbf{x} = (\mathrm{d}x_1, \ldots, \mathrm{d}x_N)$, $\mathbf{z} = (z_1, \ldots, z_N)$

EM-алгоритм. Построение функции правдоподобия

Функция правдоподобия

$$L(\theta; \mathrm{d}\mathbf{x}, \mathbf{z}) = \mathbf{P}(\mathrm{d}\mathbf{x}, \mathbf{z}|\theta) = \prod_{i=1}^{N} \sum_{j=1}^{2} \mathbb{I}(z_i = j)\, \tau_j\, f(\mathrm{d}x_i, \mu_j, \sigma_j^2),$$

$\mathbb{I}(\text{expr}) = \begin{cases} 0, & \text{expr} = \text{False} \\ 1, & \text{expr} = \text{True} \end{cases}$ — функция индикатор

$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ — плотность нормального распределения

В экспоненциальной форме:

$$L(\theta; \mathrm{d}\mathbf{x}, \mathbf{z}) = \exp\left\{ \sum_{i=1}^{N} \sum_{j=1}^{2} \mathbb{I}(z_i = j) \left[\log \tau_j - \frac{1}{2} \log(2\pi) - \log(\sigma_j) - \frac{(\mathrm{d}x_i - \mu_j)^2}{2\sigma_j^2} \right] \right\}$$

EM-алгоритм

EM-алгоритм

EM-Algorithm()

$\theta \leftarrow$ Estimate-Initial-Parameters()
while TRUE
    do $\theta_{\text{next}} \leftarrow$ M-Step(E-Step())
        if $|\theta - \theta_{\text{next}}| < \varepsilon$
            then return θ
            else $\theta \leftarrow \theta_{\text{next}}$

EM-алгоритм. Оценивание начальных параметров

τ_j

$$\tau_j^{(0)} = 0.5, \quad j = 1, 2$$

μ_j

Построим полигон частот $\mathrm{d}x_i$. В качестве $\mu_1^{(0)}$ возьмём последний локальный максимум частот, а в качестве $\mu_2^{(0)}$ — первый

σ_j

$$\sigma_j^{(0)} = \frac{1}{3}\left(\mu_1^{(0)} - \mu_2^{(0)}\right), \quad j = 1, 2$$

EM-алгоритм. E-шаг

Условная вероятность принадлежности i-го наблюдения

По Т. Байеса:

$$
\begin{aligned}
T_{j,i}^{(k)} &= \mathbf{P}\left(Z_i = j \middle| \mathrm{d}X(t_i) = \mathrm{d}x_i; \theta^{(k)}\right) = \\
&= \frac{\tau_j^{(k)} \cdot f\left(\mathrm{d}x_i; \mu_j^{(k)}, \sigma_j^{(k)}\right)}{\tau_1^{(k)} \cdot f\left(\mathrm{d}x_i; \mu_1^{(k)}, \sigma_1^{(k)}\right) + \tau_2^{(k)} \cdot f\left(\mathrm{d}x_i; \mu_2^{(k)}, \sigma_2^{(k)}\right)}
\end{aligned}
$$

Математическое ожидание логарифма функции правдоподобия

$$
\begin{aligned}
Q\left(\theta|\theta^{(k)}\right) &= \mathbf{E}\left[\log L(\theta; \mathrm{d}\mathbf{x}, \mathbf{z})\right] = \\
&= \sum_{i=1}^{N} \sum_{j=1}^{2} T_{j,i}^{(k)} \left[\log \tau_j - \frac{1}{2}\log(2\pi) - \log(\sigma_j) - \frac{(\mathrm{d}x_i - \mu_j)^2}{2\sigma_j^2}\right]
\end{aligned}
$$

EM-алгоритм. M-шаг

Максимизация $Q\left(\theta|\theta^{(k)}\right)$

$$\theta^{(k+1)} = \operatorname*{argmax}_{\theta} Q\left(\theta|\theta^{(k)}\right).$$

$$\tau_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{N} T_{j,i}^{(k)}, \quad \mu_j^{(k+1)} = \frac{\sum\limits_{i=1}^{N} T_{j,i}^{(k)}\,\mathrm{d}x_i}{\sum\limits_{i=1}^{N} T_{j,i}^{(k)}},$$

$$\sigma_j^{(k+1)} = \frac{\sum\limits_{i=1}^{N} T_{j,i}^{(k)}\left(\mathrm{d}x_i - \mu_j^{(k+1)}\right)^2}{\sum\limits_{i=1}^{N} T_{j,i}^{(k)}}$$

EM-алгоритм. Получение оценок параметров модели

Оценка σ^2, m_c, σ_c

$$\widehat{m_c} = \mu_1, \quad \widehat{\sigma}^2 = \frac{\sigma_2^2}{\Delta t}, \quad \widehat{\sigma_c}^2 = \sigma_1^2 - \sigma_2^2$$

Оценка времени поступления заявок

$$\widehat{T_c} = \left\{ t_i \quad i = 1, \ldots, N \mid T_{1,i}^{(k)} > T_{2,i}^{(k)} \right\}$$

Оценки m и λ получаем так же, как в итеративном методе

Случай конечного времени обработки заявки

$\lambda_c \gg 0$

$$
\begin{aligned}
X(t) &= B(t) + \sum_{t_c \in T_c} K_{t_c}(t) = \\
&= m + \sigma \mathcal{W}(t) + \sum_{t_c \in T_c} \mathcal{N}(m_c, \sigma_c^2) \cdot \mathrm{I}(t - t_c) \cdot e^{-\lambda_c (t - t_c)}
\end{aligned}
$$

Считаем, что заявки поступают достаточно редко и обрабатываются достаточно быстро: для каждой заявки предыдущие заявки практически не влияют на уровень загруженности

Разностный аналог производной

$\mathrm{d}X(t)$ в момент времени поступления заявки t_c

$$\begin{aligned} \mathrm{d}X(t_c) &= X(t_c) - X(t_c - \Delta t) = \\ &= \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2). \end{aligned}$$

(при условии, что в момент времени $(t_c - \Delta t)$ заявки не было)

$\mathrm{d}X(t)$ в момент времени отсутствия заявок t

$$\begin{aligned} \mathrm{d}X(t) &= X(t) - X(t - \Delta t) = \\ &= \mathcal{N}(0, \sigma^2 \Delta t) + \mathcal{N}(m_c, \sigma_c^2) \cdot \left(e^{-\lambda_c (t - t_c)} - e^{-\lambda_c (t - \Delta t - t_c)} \right). \end{aligned}$$

(при условии, что в момент времени $(t - \Delta t)$ заявки не было)

Идентификация моментов времени поступления заявок

1. Построим вариационный ряд для $\mathrm{d}x_i$:

$$\mathrm{d}x_{(1)} \leqslant \mathrm{d}x_{(2)} \leqslant \ldots \leqslant \mathrm{d}x_{(N)}.$$

   Приходящие заявки вносят существенно большее изменение уровня загруженности ресурсов сервера, чем фоновая нагрузка $\Longrightarrow$ все наблюдения $\mathrm{d}X(t_c)$ находятся в правой части ряда.

2. Используя итеративный метод для идентификации заявок Find-Request-Index, найдём границу k наблюдений $\mathrm{d}X(t_c)$ в вариационном ряде:

$$\mathrm{d}x_{(i)} \sim \mathrm{d}X(t_c), \quad i \geqslant k.$$

3. По границе k выделим наблюдения из $\widehat{T_c}$

Метод наименьших квадратов (1)

В моменты времени между поступлениями заявок $[t_k, t_{k+n}]$

$$
\begin{aligned}
\mathrm{d}X(t) &= \mathcal{N}(0,\sigma^2\Delta t) + \mathcal{N}(m_c,\sigma_c^2)\cdot\left(e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)}\right) \\
&= \left(\mathcal{N}(0,\sigma^2\Delta t) + \mathcal{N}(0,\sigma_c^2)\cdot\left(e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)}\right)\right) + \\
&\quad m_c\cdot\left(e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)}\right)
\end{aligned}
$$

Метод наименьших квадратов (2)

Метод наименьших квадратов

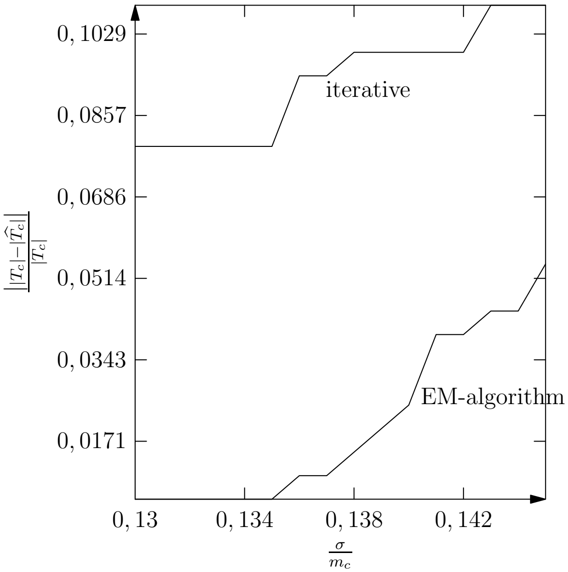
$$
\begin{aligned}
(m_c, \lambda_c) &= \operatorname*{argmin}_{m_c,\ \lambda_c} S(m_c, \lambda_c) = \\
&= \operatorname*{argmin}_{m_c,\ \lambda_c} \sum_{i=1}^{n} \left(\mathrm{d}x_i - m_c \cdot \left(e^{-\lambda_c (t_i - t_k)} - e^{-\lambda_c (t_i - \Delta t - t_k)} \right) \right)^2
\end{aligned}
$$

$S(m_c, \lambda_c)$ выпукла вниз на интересущей для анализа области — используем численный метод минимизации (метод покоординатного спуска)

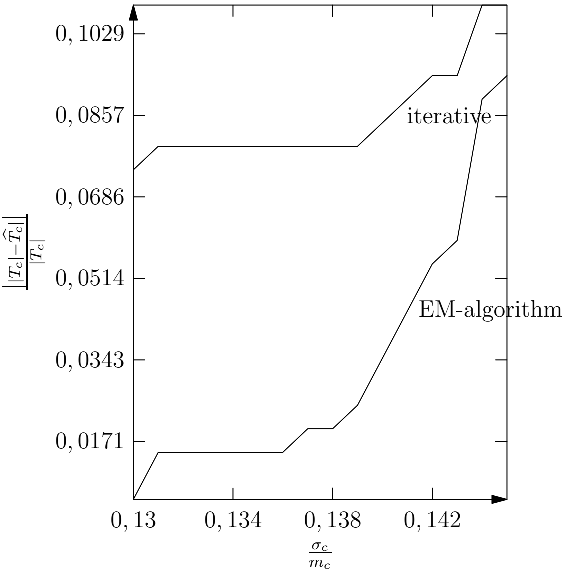
Оценка параметров фоновой загрузки

m и σ оцениваются так же, как в итеративном методе

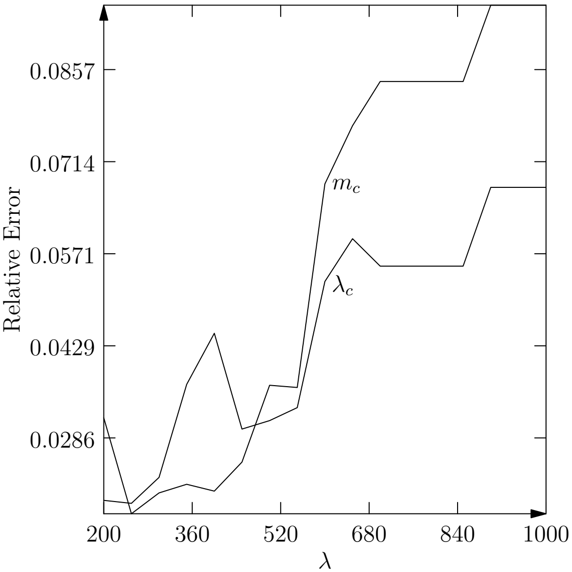
Результаты работы (1)



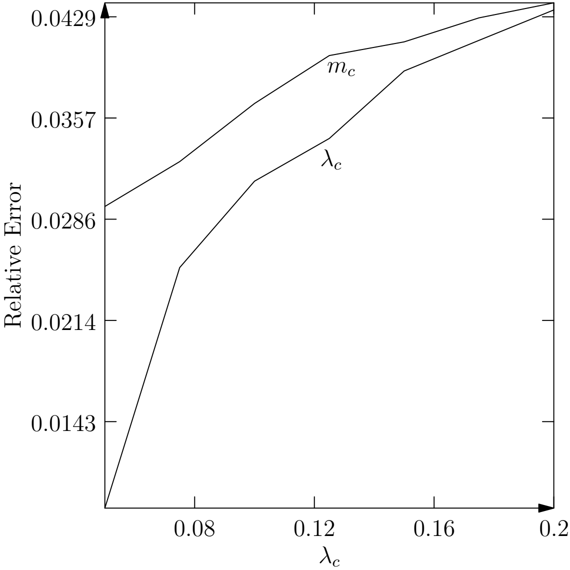
Результаты работы (2)



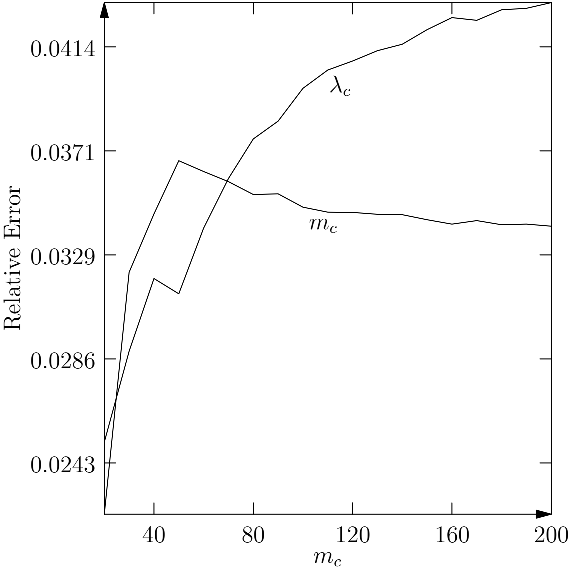
Результаты работы (3)



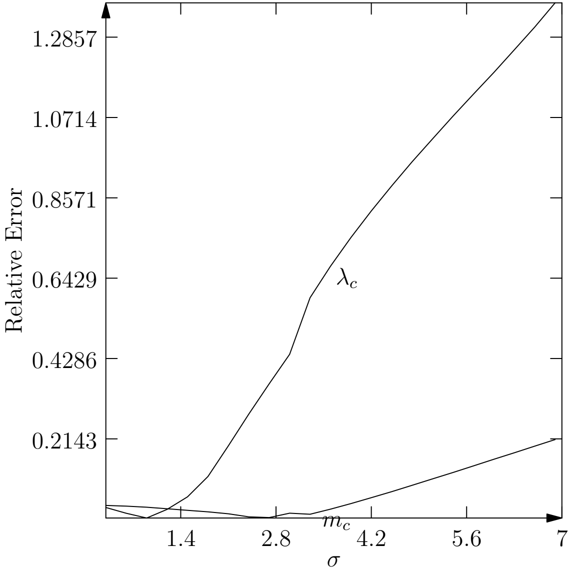
Результаты работы (4)



Результаты работы (5)



Результаты работы (6)



Результаты работы (7)