Санкт-Петербургский государственный политехнический университет

# Лабораторная работа № 2

по курсу «Стохастические модели»

«Определение параметров распределения потока заявок
по наблюдениям нагруженности системы»


| | |
|---|---|
| Студент: | Руцкий В. В. |
| Группа: | 5057/2 |
| Преподаватель: | Иванков А. А. |


Санкт-Петербург 2011

# 1 Постановка задачи

Данной работе производится анализ лога загруженности процессора сервера при поступающих заявках на обработку информации.

В отсутствие заявок величина загруженности процессора представляет собой сумму некоторой постоянной величины загрузки $m$ и случайных отклонений:

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

где $\mathcal{W}(t)$ — это винеровский процесс.

Интенсивность поступления заявок подчиняются закону распределения Пуассона $\mathcal{P}(\lambda)$.

При поступлении одной заявки нагрузка на процессор мгновенно возрастает, а затем экспоненциально снижается до прежнего уровня. Увеличение загрузки процессора от одной заявки, поступившей в момент времени $t_c$ выражается следующим образом:

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

где $\mathrm{I}(x)$ — фунцкия Хевисайда.[1]

В логе загруженности процессора наблюдается общая загрузка процессора:

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

где $T_c$ — это моменты времени поступления заявок.

Необходимо по дискретным наблюдениям $x_i$ случайного процесса $X(t)$ в моменты времени $t_i, \quad i = 1, \ldots, N$

1. оценить моменты времени поступления заявок $T_c$,
2. оценить параметры модели $m$, $\sigma$, $\lambda$, $m_c$, $\sigma_c^2$, $\lambda_c$.

Наблюдения производятся через равные промежутки времени $\Delta t = t_{i+1} - t_i$.

Для упрощения решения $\lambda_c$ принимается равным величине близкой к нулю, т. е. каждая пришедшая заявка увеличивает загрузку процессора на некоторую фиксированную величину.

# 2 Решение

## 2.1 Идентификация моментов времени поступления заявок

Предположим, что в отрезке времени $[t_k, t_{k+n+1}]$ не пришло ни одной заявки. Тогда наблюдения $x_k, \ldots, x_{k+n+1}$ представляют собой наблюдения $B(t)$. Оценим по этим наблюдениям параметры $B(t)$.

Рассмотрим разности соседних наблюдений — они представляют собой наблюдения нормально распределённой случайной величины:

$$B(t_{i+1}) - B(t_i) = \sigma\mathcal{W}(t_{i+1}) - \sigma\mathcal{W}(t_i) = \sigma\mathcal{N}(0, \Delta t) = \mathcal{N}(0, \sigma^2\Delta t).$$

Построим точечную оценку $\widehat{\sigma}^2$ методом максимального правдоподобия:[2]

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

Обозначим гипотезу о том, что в промежутке времени $[t_{k+n+1}, t_{k+n+2}]$ не пришло ни одной заявки, как $H_0$. Тогда

$$\big(X(t_{k+n+2}) - X(t_{k+n+1}) | H_0\big) = \mathcal{N}(0, \sigma^2\Delta t).$$

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

[2] См. § 3.5 пункт 1 в [1].

В качестве критерия принятия гипотезы $H_0$ с уровнем значимости $\alpha < 0.5$ возьмём условие, что разность значений наблюдений $(x_{k+n+2} - x_{k+n+1})$ лежит между $\frac{\alpha}{2}$ и $\left(1 - \frac{\alpha}{2}\right)$ квантилями нормального распределения $\mathcal{N}(0, \sigma^2 \Delta t)$, обозначенные соответственно $\mathcal{N}_{\frac{\alpha}{2}}$ и $\mathcal{N}_{1-\frac{\alpha}{2}}$:

$$H_0 \text{ принимается } \iff \mathcal{N}_{\frac{\alpha}{2}} < (x_{k+n+2} - x_{k+n+1}) < \mathcal{N}_{1-\frac{\alpha}{2}}.$$

Алгоритм нахождения моментов времени поступления заявок $T_c$ состоит в следующем:

1. В предположении, что в первые $n+1$ наблюдений не пришло ни одной заявки, оценим $\widehat{\sigma}$ и построим критерий для отвержения $H_0$.
2. Будем добавлять к первым $n+1$ наблюдениям по одному наблюдению и проверять гипотезу $H_0$. Если $H_0$ не отвергается, то $\widehat{\sigma}$ и критерий для отвержения $H_0$ пересчитываются для добавленного наблюдения.
3. Как только встретиться наблюдение $n+1+l$, для которого гипотеза $H_0$ отвергается, то $t_{n+1+l} \in T_c$. Все наблюдения до $t_{n+1+l+1}$ отбрасываются и алгоритм начинается с шага 1 для поиска следующего момента времени прихода заявки.

## 2.2 Оценка интенсивности поступления заявок $\lambda$

Зная время прибытия заявок $T_c$ интенсивность поступления заявок можно оценить методом максимального правдоподобия: [3]

$$\widehat{\lambda} = \frac{1}{|T_c|} \sum_{i=0}^{|T_c|-1} (t_{c_{i+1}} - t_{c_i}).$$

## 2.3 Оценка параметров распределения величины нагрузки поступающих заявок

Рассмотрим ненормированный разностный аналог производной случайного процесса $X(t)$:

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

$\mathrm{d}\,X(t)$ в момент времени прихода заявки $t_c$ выражается следующим образом:

$$\begin{aligned} \mathrm{d}X(t_c) &= X(t_c) - X(t_c - \Delta t) = B(t_c) + K_{t_c}(t_c) - B(t_c - \Delta t) = \\ &= \sigma\mathcal{W}(t_c) - \sigma\mathcal{W}(t_c - \Delta t) + \mathcal{N}(m_c, \sigma_c^2) = \\ &= \sigma\mathcal{N}(0, \Delta t) + \mathcal{N}(m_c, \sigma_c^2) = \mathcal{N}(m_c, \sigma^2\Delta t + \sigma_c^2), \end{aligned}$$

предполагая, что в момент времени $t_c - \Delta t$ заявки не было.

Во время отсутствия заявок $\mathrm{d}X(t)$ выражается как:

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

Значит в каждый отдельно взятый момент времени $t$ случайная величина $\mathrm{d}X(t)$ представляет собой смесь двух нормально распределённых случайных величин, причем параметры случайных величин со временем не меняются. Оценим их параметры EM-алгоритмом (на основе примера из [2]).

Введём скрытые случайные величины $Z_i, \quad i = 1, \ldots, N$, принимающие значения 1 или 2, в зависимости от того, пришла ли заявка в момент времени $t_i$ или нет соответственно, а $z_i$ — наблюдения $Z_i$ в момент времени $t_i$.

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

Пусть $\mathbf{P}(Z_i = 1) = \tau_1$ и $\mathbf{P}(Z_i = 2) = \tau_2 = 1 - \tau_1$.

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

[3] `http://en.wikipedia.org/wiki/Poisson_distribution#Maximum_likelihood` или в общем случае в § 3.5 пункт 1 в [1].

Построим функцию правдоподобия:

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

где $\mathbb{I}(\mathrm{expr})$ — функция индикатор,[4] а $f(x,\mu,\sigma^2)$ — это функция плотности распределения, в данном случае нормального:

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

Перепишем функцию правдоподобия в экспоненциальной форме:

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

Пусть имеется начальная оценка параметров $\theta$: $\theta^{(0)}$ (использованный способ вычисления $\theta^{(0)}$ будет описан ниже). Последовательно выполняя E- и M-шаги будем уточнять оценку $\theta^{(k)}$, пока она не сойдётся к какой-то величине $\theta^{(k)} \xrightarrow[k\to\infty]{} \theta$, которую и примем за результат.

**E-шаг** Имея текущую оценку параметров $\theta^{(k)}$, вычислим по теореме Байеса условную вероятность принадлежности $i$-го наблюдения $j$-му нормальному распределению:

$$T_{j,i}^{(k)} = \mathbf{P}(Z_i=j|X(t_i)=x_i;\theta^{(k)}) = \frac{\tau_j^{(k)}f(x_i;\mu_j^{(k)},\sigma_j^{(k)})}{\tau_1^{(k)}f(x_i;\mu_1^{(k)},\sigma_1^{(k)}) + \tau_2^{(k)}f(x_i;\mu_2^{(k)},\sigma_2^{(k)})}.$$

Построим функцию — математическое ожидание логарифма функции правдоподобия:

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

**M-шаг** Теперь найдём параметры $\theta^{(k+1)}$ максимизирующие $Q\left(\theta|\theta^{(k)}\right)$:

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

В соответствии с вычислениями в [2]:

$$\tau_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{N}T_{j,i}^{(k)},\quad \mu_j^{(k+1)} = \frac{\sum\limits_{i=1}^{N}T_{j,i}^{(k)}x_i}{\sum\limits_{i=1}^{N}T_{j,i}^{(k)}},\quad \sigma_j^{(k+1)} = \frac{\sum\limits_{i=1}^{N}T_{j,i}^{(k)}(x_i-\mu_j^{(k+1)})^2}{\sum\limits_{i=1}^{N}T_{j,i}^{(k)}}.$$

**Вычисление** $\theta_0$ $\tau_j^{(0)}$ вычислим из информации о $T_c$, полученной в пункте 2.1:

$$\tau_1^{(0)} = \frac{|T_c|}{N},\quad \tau_2^{(0)} = 1 - \frac{|T_c|}{N}.$$

Для вычисления $\mu_j^{(0)}$ построим полигон частот[5] $\mathrm{d}x_i$: в качестве $\mu_1^{(0)}$ возьмём последний локальный минимум частот, а в качестве $\mu_2^{(0)}$ — первый (т. к. $\mathbf{E}\left[B(t)-B(t-\Delta t)\right] = 0$, а $\mathbf{E}\left[B(t_c)+K_{t_c}(t_c)-B(t-\Delta t)\right] = m_c > 0$).

В качестве $\sigma_j^{(0)}$ возьмём $\frac{1}{3}(\mu_1^{(0)}-\mu_2^{(0)})$, $\quad j=1,2$.

# 3 Результаты работы

[4]Функция индикатор: $\mathbb{I}(\mathrm{expr}) = \begin{cases} 0, & \mathrm{expr}=\mathrm{False} \\ 1, & \mathrm{expr}=\mathrm{True}\end{cases}$.

[5]См. § 2.1 пункт 4 в [1].

## Список литературы

[1] Г.И. Ивченко and Ю.И. Медведев. *Введение в математическую статистику*. М: Издательство ЛКИ, 2010.

[2] Wikipedia: Expectation-maximization algorithm. `http://en.wikipedia.org/w/index.php?title=Expectation-maximization_algorithm&oldid=423422317`.