ФКН ВШЭ, 3 курс, 3 модуль

Задание 3. Статистические решения. Последовательные тесты. Обнаружение разладки.

Вероятностные модели и статистика случайных процессов, весна 2017

Время выдачи задания: 15 марта (среда).

Срок сдачи: **27 марта (понедельник), 23:59.**

Среда для выполнения практического задания – PYTHON 2.x.

Правила сдачи

Инструкция по отправке:

1. Домашнее задание необходимо отправить до дедлайна на почту hse.cs.stochastics@gmail.com.

2. В письме укажите тему «[ФКН ССП17] Задание 3, Фамилия Имя».

3. Решения задач следует присылать единым файлом формата .pdf, набранным в LaTeX. Допускается отправка отдельных практических задач в виде отдельных файлов (ipython-тетрадок или исходных файлов с кодом на языке `python`).

Оценивание и штрафы:

1. **Каждая из задач имеет стоимость 2 балла**, при этом за задачу можно получить 0, 1 или 2 балла. Максимально допустимая

оценка за работу – 10 баллов. Баллы, набранные сверх максимальной оценки, считаются бонусными и влияют на освобождение от задач на экзамене.

2. Дедлайн жесткий. Сдавать задание после указанного срока сдачи нельзя.

3. Задание выполняется самостоятельно. «Похожие» решения считаются плагиатом и все задействованные студенты (в том числе те, у кого списали) не могут получить за него больше 0 баллов (подробнее о плагиате см. на странице курса). Если вы нашли решение какого-то из заданий (или его часть) в открытом источнике, необходимо указать ссылку на этот источник в отдельном блоке в конце Вашей работы (скорее всего вы будете не единственным, кто это нашел, поэтому чтобы исключить подозрение в плагиате, необходима ссылка на источник).

Необходимые теоретические сведения

1. Всюду в рассматриваемых задачах имеется две гипотезы $\mathbb{H}_0$ и $\mathbb{H}_1$ (иногда они обозначаются $\mathbb{H}_\infty$ и $\mathbb{H}_0$, соответственно), причем каждая из гипотез делает явные предположения о распределении или его параметрах.

2. Критерий Неймана-Пирсона предписывает принимать гипотезу исходя из значения величины

$$L_n(X_1, \ldots, X_n) = \frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)},$$

называемой отношением правдоподобия. А именно, пусть $\varphi(X_1, \ldots, X_n)$ – рандомизированное решающее правило, значение которого равно вероятности принять гипотезу $\mathbb{H}_1$. Тогда найдутся такие константы λ_a и h_a, что

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1, & L_n(X_1, \ldots, X_n) > h_a, \\ \lambda_a, & L_n(X_1, \ldots, X_n) = h_a, \\ 0, & L_n(X_1, \ldots, X_n) < h_a, \end{cases}$$

является наиболее мощным (т. е. с наименьшей вероятностью пропуска цели или ошибки 2 рода $\beta(\varphi)$) тестом среди тестов, вероятность ложной тревоги $\alpha(\varphi)$ (ошибки 1 рода) которых не выше a.

3. Последовательный тест отношения правдоподобия (sequential probability ratio test, SPRT) заключается в вычислении логарифма отношения правдоподобия $Z_n = \log L_n$ (см. выше; в случае независимых наблюдений формулы упрощаются) и сравнении этой величины в каждый момент времени с пороговыми значениями $A < 0, B > 0$, выбранными исходя из заданных вероятностей ошибок 1 и 2 рода. Наблюдения останавливаются в первый момент времени выхода статистики Z_n за «коридор» (A, B):

$$\tau_{A,B} = \inf\{n \geqslant 1 : Z_n \notin (A, B)\}.$$

При этом в каждый момент времени принимается одно из трех решений:

$$\begin{cases} \text{если } Z_n \leqslant A & \Longrightarrow \text{ верна гипотеза } \mathbb{H}_0, \\ \text{если } Z_n \geqslant B & \Longrightarrow \text{ верна гипотеза } \mathbb{H}_1, \\ \text{если } Z_n \in (A, B) & \Longrightarrow \text{ продолжить наблюдения.} \end{cases}$$

Построить последовательный тест – значит указать *момент остановки измерений* τ и *решающее правило* $\varphi(\cdot)$.

4. Разладкой процесса $X = (X_n)_{n=1,\,2,\,\ldots}$ называется ситуация, в которой траектория процесса генерируется двумя (или в общем случае несколькими) независимыми вероятностными мерами P_∞ и P_0, причем наблюдения имеют структуру

$$X_n = \begin{cases} X_n^\infty, & \text{если } 1 \leqslant n < \theta, \\ X_n^0, & \text{если } n \geqslant \theta, \end{cases}$$

где $X^\infty = (X_n^\infty)_{n=1,\,2,\,\ldots}$ — процесс, соответствующий мере P_∞, и $X^0 = \big(X_n^0\big)_{n=1,\,2,\,\ldots}$ — процесс, соответствующий мере P_0. Момент $\theta \in [0, \infty]$ называется моментом разладки, причем ситуация $\theta = 0$ соответствует тому, что с самого начала идут наблюдения от «разлаженного» процесса X^0, а ситуация $\theta = \infty$ заключается в том, что разладка не появляется никогда. Таким образом, траектория процесса X выглядит следующим образом:

$$\underbrace{X_1^\infty, X_2^\infty, \ldots, X_{\theta-1}^\infty}_{\text{мера } \mathrm{P}^\infty}, \underbrace{X_\theta^0, X_{\theta+1}^0, \ldots}_{\text{мера } \mathrm{P}^0}$$

5. Статистика кумулятивных сумм.

- Вводятся статистики $\gamma = (\gamma_n)_{n=1,2,\dots}$ и $\gamma = (\gamma_n)_{n=1,2,\dots}$

$$\gamma_n = \sup_{\theta \geqslant 0} \frac{f_\theta(X_1, \dots, X_n)}{f_\infty(X_1, \dots, X_n)} \qquad \text{и} \qquad T_n = \log \gamma_n$$

- Если случайные величины $X_1, \dots, X_n$ независимы, то

$$\gamma_n = \max\left\{1, \max_{1 \leqslant \theta \leqslant n} \prod_{k=\theta}^{n} \frac{f_0(X_k)}{f_\infty(X_k)}\right\},$$

$$T_n = \max\left\{0, \max_{1 \leqslant \theta \leqslant n} \sum_{k=\theta}^{n} \log \frac{f_0(X_k)}{f_\infty(X_k)}\right\} = \max\left\{0, \max_{1 \leqslant \theta \leqslant n} \sum_{k=\theta}^{n} \zeta_k\right\}$$

- Статистика T_n обладает свойством $T_n = \max(0, T_{n-1} + \zeta_n)$ и называется статистикой кумулятивных сумм (CUmulative SUMs, CUSUM).

- Момент остановки

$$\tau_{\text{CUSUM}} = \inf\{n \geqslant 0 : T_n \geqslant B\},$$

построенный по статистике кумулятивных сумм, оптимален (т. е. обладает наименьшей задержкой в обнаружении разладки) в классе

$$\mathcal{M}_T = \{\tau : \mathrm{E}_\infty \tau \geqslant T\}$$

тех моментов остановки, для которых среднее время до ложной тревоги не меньше T.

6. Статистика Ширяева-Робертса.

- Вводится статистика

$$R_n = \sum_{\theta=1}^{n} \frac{f_\theta(X_1, \dots, X_n)}{f_\infty(X_1, \dots, X_n)}$$

- Если случайные величины $X_1, \dots, X_n$ независимы, то

$$R_n = \sum_{\theta=1}^{n} \prod_{k=\theta}^{n} \frac{f_0(X_k)}{f_\infty(X_k)} = \sum_{\theta=1}^{n} \prod_{k=\theta}^{n} l_k.$$

- Статистика R_n обладает свойством $R_n = (1 + R_{n-1})l_k$ и называется статистикой Ширяева-Робертса (Shiryaev-Roberts, SR).
- Момент остановки

$$\tau_{\mathrm{SR}} = \inf\{n \geqslant 0 : R_n \geqslant B\},$$

построенный по статистике Ширяева-Робертса, оптимален (т. е. обладает наименьшей задержкой в обнаружении разладки) в классе

$$\mathcal{M}_T = \{\tau : \mathrm{E}_\infty \tau \geqslant T\}$$

тех моментов остановки, для которых среднее время до ложной тревоги не меньше T.

Вариант 1

1. По выборке $(X_1, \ldots, X_n)$ из биномиального распределения $\mathrm{Bin}(k, p)$ построить критерий Неймана-Пирсона для проверки гипотезы $\mathbb{H}_0 : p = p_0$ против альтернативы $\mathbb{H}_1 : p = p_1$, где $0 < p_0 < p_1 < 1$.

2. Дана выборка $(X_1, \ldots, X_n)$ из нормального $\mathcal{N}(\mu, \sigma^2)$ распределения. Построить критерий проверки гипотезы $\mathbb{H}_0 : \mu = 0$ против альтернативы $\mathbb{H}_1 : \mu = 0.1$ и определить наименьший объем выборки, при котором вероятности ошибок 1 и 2 родов не превышают 0.01.

3. Необходимо произвести выбор между двумя гипотезами о возможных значениях p_0 и p_1 вероятности события A $(p_0 < p_1)$. В этих целях осуществляется последовательность независимых опытов, в каждом из которых определяется, происходит или не происходит событие A. Построить последовательный критерий отношения вероятностей при заданных значениях α и β вероятностей ошибок первого и второго рода.

4. Провести моделирование для сравнения критерия Неймана-Пирсона и последовательного критерия отношения правдоподобия в задаче 3. В этом моделировании:

   (a) Для заданных уровня значимости $\alpha_i = i\Delta, \Delta = 0.01, i = 1, \ldots, 99$, и вероятности ошибки второго рода $\beta_i = \alpha_i$ подсчитать объем наблюдений, требуемый в критерии Неймана-Пирсона для достижения этих характеристик.

   (b) Проделать то же самое для последовательного критерия отношения правдоподобия.

   (c) Привести графическое сравнение зависимости объема требуемых данных от требуемого уровня значимости $n(\alpha)$ для двух критериев, сделать выводы.

(d) Изменяется ли соотношение между требуемыми объемами выборок при изменении отношения $\gamma = p_0/p_1$ в рассматриваемых гипотезах? Построить зависимости $n(\gamma)$ для двух критериев при некотором фиксированном уровне значимости α.

5. Процесс $X = (X_n)_{n=1,2,\dots}$, наблюдаемый в режиме реального времени, задается нормально распределенным белым шумом (с нулевым средним и единичной дисперсией), т. е.

$$X_n = \varepsilon_n, \qquad n = 1, 2, \dots$$

В неизвестный момент времени $\theta \geqslant 1$ происходит разладка (изменение статистических свойств) процесса X_n, которая состоит в том, что для $n \geqslant \theta$ процесс X задается уравнением типа AR(1), то есть

$$X_n = \alpha_0 + \alpha_1 X_{n-1} + \varepsilon_n, \qquad n \geqslant \theta,$$

где $|\alpha_1| < 1$.

Построить процедуру обнаружения разладки, основанную на статистике кумулятивных сумм, для обнаружения момента θ. Параметры α_0, α_1 процесса считать известными. Привести формулы для отношения правдоподобия, а также для одного шага итеративного алгоритма кумулятивных сумм. В какой момент следует поднимать тревогу об обнаружении разладки?

6. Провести моделирование для определения оперативных характеристик процедуры обнаружения разладки, разработанной в задаче 5. Считать заданными параметры $\alpha_0 = 0, \alpha_1 = 0.8$.

(a) При использовании статистики $\gamma = (\gamma_n)_{n=1,2,\dots}$ прежде всего необходимо подобрать значение порога $B = B_T$ в зависимости от значения параметра T так, чтобы $\tau(B_T; \{\gamma_n\}) \in \mathcal{M}_T$. Требуется подсчитать (с помощью метода Монте-Карло) и дать в виде графика значения величины

$$\mathbb{T}_{\text{CUSUM}}(B) = \mathrm{E}_\infty \tau(B; \{\gamma_n\})$$

для разных значений B (и малых и больших).

(b) С помощью метода Монте-Карло подсчитать и дать в виде графика значения величины

$$\mathbb{R}_{\text{CUSUM}}(B) = \mathrm{E}_0 \tau(B; \{\gamma_n\}).$$

для разных значений B (и малых и больших). Графики нарисовать для достаточно частых значений B.

7. Вам выданы файлы `sig1.train` (обучающий) и `sig1.test.public` (валидационный) (третий файл `sig1.test.private` имеется у лектора). Обучающий файл содержит два столбца, причем первый столбец — это реализация $X_1, \ldots, X_{1000}$ некоторого случайного процесса, полученная следующим образом:

$$X_n = \begin{cases} X_n^\infty, & \text{если } n \notin [\theta, \theta + \Delta], \\ X_n^0, & \text{если } n \in [\theta, \theta + \Delta], \end{cases}$$

а второй столбец — это индикатор действия процесса X_n^0, т. .е. процесс

$$Y_n = \mathbb{1}_{[\theta, \theta+\Delta]}(n) = \begin{cases} 0, & \text{если } n \notin [\theta, \theta + \Delta], \\ 1, & \text{если } n \in [\theta, \theta + \Delta]. \end{cases}$$

Сечения процесса X могут быть как зависимы, так и независимы.

(a) Предложите какие-либо модели временных рядов X_n^0 и X_n^∞, адекватно описывающие наблюдения обучающей выборки.

(b) Используя предложенные модели и рассмотренные на лекциях и семинарах подходы (полезно также рассматривать и их композиции), предложите алгоритм обнаружения разладки процесса X. Этот алгоритм должен работать в режиме реального времени, т. е. для вынесения решения о разладке в момент n он не может использовать всю доступную траекторию процесса X, а может использовать лишь наблюдения до момента n включительно. (Тем не менее, для построения алгоритма можно использовать все доступные данные).

(c) Реализуйте этот алгоритм в программном коде.

(d) Проверьте его работу на обучающих данных, нарисуйте траекторию статистики этого алгоритма, сравните ее с индикатором разладки.

(e) Нарисуйте траекторию статистики этого алгоритма на тестовых данных, вставьте в отчет рисунок. Сохраните эту траекторию в текстовый файл (по одному значению на строку) и пришлите вместе с исходным кодом, реализующим метод обнаружения разладки.

Вариант 2

1. По выборке $(X_1, \ldots, X_n)$ из пуассоновского распределения $\Pi(\lambda)$ построить критерий Неймана-Пирсона для проверки гипотезы $\mathbb{H}_0 : \lambda = \lambda_0$ против альтернативы $\mathbb{H}_1 : \lambda = \lambda_1$, где $0 < \lambda_0 < \lambda_1$.

2. В последовательности $\xi_1, \ldots, \xi_n$ независимых испытаний, выполненных согласно схеме Бернулли, $\mathrm{P}(\xi_i = 1) = p, \mathrm{P}(\xi_i = 0) = 1 - p$. Построить критерий проверки гипотезы $\mathbb{H}_0 : p = 0$ против альтернативы $\mathbb{H}_1 : p = 0.01$ и определить наименьший объем выборки, при котором вероятности ошибок 1 и 2 родов не превышают 0.01.

3. Пусть гипотезы $\mathbb{H}_0$ и $\mathbb{H}_1$ имеют вид

$$\begin{aligned} \mathbb{H}_0 : \ & f(x) = \theta_0^{-1} \exp(-x/\theta_0), \qquad x > 0; \\ \mathbb{H}_1 : \ & f(x) = \theta_1^{-1} \exp(-x/\theta_1), \qquad x > 0, \quad \theta_1 = 2\theta_0; \end{aligned}$$

   Построить процедуру последовательного критерия отношения правдоподобия различения гипотез $\mathbb{H}_0$ и $\mathbb{H}_1$ при заданных величинах вероятностей ошибок первого и второго рода $\alpha = \beta \leqslant 0.05$.

4. Провести моделирование для сравнения критерия Неймана-Пирсона и последовательного критерия отношения правдоподобия в задаче 3. В этом моделировании:

   (a) Для заданных уровня значимости $\alpha_i = i\Delta, \Delta = 0.01, i = 1, \ldots, 99$, и вероятности ошибки второго рода $\beta_i = \alpha_i$ подсчитать объем наблюдений, требуемый в критерии Неймана-Пирсона для достижения этих характеристик.

   (b) Проделать то же самое для последовательного критерия отношения правдоподобия.

(c) Привести графическое сравнение зависимости объема требуемых данных от требуемого уровня значимости $n(\alpha)$ для двух критериев, сделать выводы.

(d) Изменяется ли соотношение между требуемыми объемами выборок при изменении отношения $\gamma = \theta_0/\theta_1$ в рассматриваемых гипотезах? Построить зависимости $n(\gamma)$ для двух критериев при некотором фиксированном уровне значимости α.

5. Процесс $X = (X_n)_{n=1,2,\ldots}$, наблюдаемый в режиме реального времени, задается нормально распределенным белым шумом (с нулевым средним и единичной дисперсией), т. е.

$$X_n = \varepsilon_n, \qquad n = 1, 2, \ldots$$

В неизвестный момент времени $\theta \geqslant 1$ происходит разладка (изменение статистических свойств) процесса X_n, которая состоит в том, что для $n \geqslant \theta$ процесс X задается уравнением типа ARCH(1), то есть

$$X_n = \sigma_n \varepsilon_n, \qquad \sigma_n^2 = \alpha_0 + \alpha_1 X_{n-1}^2, \qquad n \geqslant \theta,$$

где $|\alpha_1| < 1$.

Построить процедуру обнаружения разладки, основанную на статистике Ширяева-Робертса, для обнаружения момента θ. Параметры α_0, α_1 процесса считать известными. Привести формулы для отношения правдоподобия, а также для одного шага итеративного алгоритма Ширяева-Робертса. В какой момент следует поднимать тревогу об обнаружении разладки?

6. Провести моделирование для определения оперативных характеристик процедуры обнаружения разладки, разработанной в задаче 5. Считать заданными параметры $\alpha_0 = 0.146, \alpha_1 = 0.107$.

(a) При использовании статистики $\gamma = (\gamma_n)_{n=1,2,\dots}$ прежде всего необходимо подобрать значение порога $B = B_T$ в зависимости от значения параметра T так, чтобы $\tau(B_T; \{\gamma_n\}) \in \mathcal{M}_T$. Требуется подсчитать (с помощью метода Монте-Карло) и дать в виде графика значения величины

$$\mathbb{T}_{\mathrm{SR}}(B) = \mathrm{E}_\infty \tau(B; \{\gamma_n\})$$

для разных значений B (и малых и больших).

(b) С помощью метода Монте-Карло подсчитать и дать в виде графика значения величины

$$\mathbb{R}_{\mathrm{SR}}(B) = \mathrm{E}_0 \tau(B; \{\gamma_n\}).$$

для разных значений B (и малых и больших). Графики нарисовать для достаточно частых значений B.

7. Вам выданы файлы `sig2.train` (обучающий) и `sig2.test.public` (валидационный) (третий файл `sig2.test.private` имеется у лектора). Обучающий файл содержит два столбца, причем первый столбец — это реализация $X_1, \dots, X_{1000}$ некоторого случайного процесса, полученная следующим образом:

$$X_n = \begin{cases} X_n^\infty, & \text{если } n \notin [\theta, \theta + \Delta], \\ X_n^0, & \text{если } n \in [\theta, \theta + \Delta], \end{cases}$$

а второй столбец — это индикатор действия процесса X_n^0, т. .е. процесс

$$Y_n = \mathbb{1}_{[\theta, \theta+\Delta]}(n) = \begin{cases} 0, & \text{если } n \notin [\theta, \theta + \Delta], \\ 1, & \text{если } n \in [\theta, \theta + \Delta]. \end{cases}$$

Сечения процесса X могут быть как зависимы, так и независимы.

(a) Предложите какие-либо модели временных рядов X_n^0 и X_n^∞, адекватно описывающие наблюдения обучающей выборки.

(b) Используя предложенные модели и рассмотренные на лекциях и семинарах подходы (полезно также рассматривать и их композиции), предложите алгоритм обнаружения разладки процесса X. Этот алгоритм должен работать в режиме реального времени, т. е. для вынесения решения о разладке в момент n он не может использовать всю доступную траекторию процесса X, а может использовать лишь наблюдения до момента n включительно. (Тем не менее, для построения алгоритма можно использовать все доступные данные).

(c) Реализуйте этот алгоритм в программном коде.

(d) Проверьте его работу на обучающих данных, нарисуйте траекторию статистики этого алгоритма, сравните ее с индикатором разладки.

(e) Нарисуйте траекторию статистики этого алгоритма на тестовых данных, вставьте в отчет рисунок. Сохраните эту траекторию в текстовый файл (по одному значению на строку) и пришлите вместе с исходным кодом, реализующим метод обнаружения разладки.