мФТиАД ФКН ВШЭ, 1 курс, 3 модуль

Задание 2. Обнаружение разладок временных рядов.

Прогнозирование временных данных и случайных процессов, весна 2018

Время выдачи задания: 22 февраля (четверг).

Срок сдачи: **8 марта (четверг), 23:59.**

Среда для выполнения практического задания – PYTHON 2.x.

Правила сдачи

Выполнение работы в команде

1. Домашнее задание допускается выполнять в команде от 1 до 4 человек.

2. Командное решение достаточно загрузить в AnyTask только один раз. При этом в посылке следует указать состав команды.

3. Баллы, набранные командой, выставляются всем членам команды одинаковыми. Бонусные баллы выставляются всем членам команды одинаковыми. Это означает, что каждый член команды получает баллы, набранные его командой, независимо от его вклада в решение работы.

Инструкция по отправке:

1. 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. 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 командой независимо от других команд. «Похожие» решения считаются плагиатом и все студенты обеих команд (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. 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:

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

2. Статистика кумулятивных сумм.

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

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

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

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

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

- 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.

3. Статистика Ширяева-Робертса.

- Вводится статистика

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

- Если случайные величины $X_1, \ldots, X_n$ 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

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

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. (3 балла) 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 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?

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

   (a) (2 балла) При использовании статистики $\gamma = (\gamma_n)_{n=1,2,\ldots}$ 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) (2 балла) С помощью метода Монте-Карло подсчитать и дать в виде графика значения величины

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

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

3. 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 $X_1, \dots, X_{1000}$ 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) (1 балл) Предложите какие-либо модели временных рядов X_n^0 и X_n^∞, адекватно описывающие наблюдения обучающей выборки.

(b) (1 балл) 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) (1 балл) Реализуйте этот алгоритм в программном коде.

(d) (1 балл) Проверьте его работу на обучающих данных, нарисуйте траекторию статистики этого алгоритма, сравните ее с индикатором разладки.

(e) (1 балл) 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. (3 балла) 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 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?

2. 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) (2 балла) 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) (2 балла) 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.

3. 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 $X_1, \ldots, X_{1000}$ 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) (1 балл) Предложите какие-либо модели временных рядов X_n^0 и X_n^∞, адекватно описывающие наблюдения обучающей выборки.

(b) (1 балл) 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) (1 балл) Реализуйте этот алгоритм в программном коде.

(d) (1 балл) Проверьте его работу на обучающих данных, нарисуйте траекторию статистики этого алгоритма, сравните ее с индикатором разладки.

(e) (1 балл) 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.