мФТиАД ФКН ВШЭ, 1 курс, 3 модуль

Задание 5. Проверка статистических гипотез. Последовательные тесты. Марковские процессы

Прогнозирование временных данных и случайных процессов, весна 2018

Время выдачи задания: 19 апреля (четверг).

Срок сдачи: **6 мая (воскресенье), 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. 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. Случайный процесс $X = (X_t)_{t \geqslant 0}$ с дискретным множеством E значений и матрицей P переходных вероятностей называется дискретной марковской цепью, если:

(a) $P = (p_{ij}), \qquad \sum\limits_{j \in E} p_{ij} = 1, \qquad p_{ij} \geqslant 0, \qquad i, j \in E,$

(b) $\mathrm{P}(X_n = j | X_{n-1} = i) = p_{ij}, \quad i, j \in E, \quad n \in \mathbb{N},$

(c) $\mathrm{P}(X_{n+1} = j | X_0 = i_0, \ldots, X_n = i_n) = \mathrm{P}(X_{n+1} = j | X_n = i_n).$

Вариант 1

1. (2 балла) 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. (2 балла) 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. (3 балла) 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. (4 балла) Провести моделирование для сравнения критерия Неймана-Пирсона и последовательного критерия отношения правдоподобия в задаче 3. В этом моделировании:

   (a) (1 балл) 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) (1 балл) Проделать то же самое для последовательного критерия отношения правдоподобия.

(c) (1 балл) Привести графическое сравнение зависимости $n(\alpha)$ объема требуемых данных n от требуемого уровня значимости α для двух критериев, сделать выводы.

(d) (1 балл) 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. (2 балла) Пусть $\xi_1, \xi_2, \dots$ – независимые одинаково распределенные случайные величины, причем

$$\mathrm{P}(\xi_t = 1) = p, \quad \mathrm{P}(\xi_t = -1) = 1 - p.$$

Проверить, являются ли цепями Маркова следующие последовательности случайных величин:

(a) $\eta_t = \xi_t \xi_{t+1}$,

(b) $\eta_t = \xi_1 \xi_2 \dots \xi_t$.

Для цепей Маркова найти вероятности перехода за один шаг.

6. (3 балла) Движение частицы по целым точкам отрезка $[0, N]$ описывается цепью Маркова с $N+1$ состояниями и вероятностями перехода за один шаг $p_{00} = p_{NN} = 1$; $p_{i,i+1} = p, p_{i,i-1} = 1 - p = q$; $i = 1, \dots, N-1$ («случайное блуждание с поглощающими стенками»). Пусть ξ_t – положение частицы в момент t, и $\tau = \min\{t : \xi_t = N\}$ – первый момент времени частицей верхней стенки. Показать, что

$$m_k = \mathrm{E}[\tau \mid \xi_0 = k] = \begin{cases} \frac{2pq}{(p-q)^2}\left(\left(\frac{q}{p}\right)^N - \left(\frac{q}{p}\right)^k\right) - \frac{N-k}{q-p}, & \text{если } p \neq q, \\ (N-k)(N+k), & \text{если } p = q. \end{cases}$$

Вариант 2

1. (2 балла) По выборке $(X_1, \dots, X_n)$ 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. (2 балла) В последовательности $\xi_1, \dots, \xi_n$ 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. (3 балла) Пусть гипотезы $\mathbb{H}_0$ и $\mathbb{H}_1$ имеют вид

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

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

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

   (b) (1 балл) Проделать то же самое для последовательного критерия отношения правдоподобия.

(c) (1 балл) Привести графическое сравнение зависимости $n(\alpha)$ объема требуемых данных n от требуемого уровня значимости α для двух критериев, сделать выводы.

(d) (1 балл) 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. (2 балла) Пусть $\xi_1, \xi_2, \ldots$ – независимые одинаково распределенные случайные величины, причем

$$\mathrm{P}(\xi_t = 1) = p, \quad \mathrm{P}(\xi_t = -1) = 1 - p = q.$$

Положим $\eta_0 = 0, \eta_{t+1} = \eta_t + \xi_{t+1}$. Выяснить, является ли последовательность η_t цепью Маркова. Найти $\mathrm{P}(\eta_t = m), m = 0, 1, \ldots$

6. (3 балла) Движение частицы по целым точкам отрезка $[0, N]$ описывается цепью Маркова с $N+1$ состояниями и вероятностями перехода за один шаг $p_{00} = p_{NN} = 1$; $p_{i,i+1} = p, p_{i,i-1} = 1 - p = q$; $i = 1, \ldots, N-1$ («случайное блуждание с поглощающими стенками»). Пусть ξ_t – положение частицы в момент t, и $p_{ij}(t) = \mathrm{P}(\xi_t = j | \xi_0 = i)$. Найти вероятности поглощения частицы в точках 0 и N:

$$\pi_k^{(0)} = \lim_{t\to\infty} p_{k0}(t), \qquad \pi_k^{(N)} = \lim_{t\to\infty} p_{kN}(t).$$