Домашнее задание №1

Тема: статистическое оценивание параметров

Дедлайн: 15 марта 2024 г., 23:59

Задача 1. *(2.5 балла)* Пусть $X_1, X_2, \ldots, X_n$ — независимые случайные величины, имеющие распределение Бернулли с параметром $p \in (0,1)$. Для оценивания неизвестной дисперсии предлагаются две оценки:

$$\widehat{\theta}_1 := \frac{1}{n-1} T(X_1, X_2, \ldots, X_n) \quad \text{и} \quad \widehat{\theta}_2 := \frac{1}{n} T(X_1, X_2, \ldots, X_n),$$

где

$$T(X_1, X_2, \ldots, X_n) := \sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2, \qquad \overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

1. Докажите, что $\widehat{\theta}_1$ является несмещённой оценкой дисперсии. Существуют ли другие несмещённые оценки для дисперсии, обладающие меньшей дисперсией, чем $\widehat{\theta}_1$?

2. Определите значения параметра $p \in (0,1)$, при которых оценка $\widehat{\theta}_2$ будет лучше оценки $\widehat{\theta}_1$ относительно квадратичной функции потерь для любых значений $n \geq 2$.

   Подсказка: докажите, что

   $$\mathbb{E}\left[(T(X_1, X_2, \ldots, X_n))^2\right] = \frac{\sigma^4(n-1)}{n} \left(\frac{(n-1)\mathbb{E}\left[(X_1 - \mu)^4\right]}{\sigma^4} + n^2 - 2n + 3 \right),$$

   где $\mu = \mathbb{E}[X_1]$, $\sigma^2 = \mathrm{Var}(X_1)$.

Задача 2. *(1.5 балла)* Пусть доступна выборка из распределения с плотностью

$$p_\theta(x) = \lambda e^{-\lambda(x-\mu)} \cdot \mathbb{I}\{x \geq \mu\}, \quad \lambda, \mu > 0.$$

1. Найдите оценку параметра $\theta = (\lambda, \mu)$ методом максимального правдоподобия и методом моментов.

2. Смоделируйте случайную величину с данным распределением. Вычислите полученные в предыдущем пункте оценки по 100 выборкам объёма $n = 1000$ и сравните их точность, построив диаграммы размаха.

Задача 3. *(2 балла)* Пусть доступна выборка из гамма-распределения с плотностью

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \cdot \mathbb{I}\{x > 0\}, \quad \alpha, \beta > 0, \tag{1}$$

где $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$ — гамма-функция.

1. Предположим, что параметр α известен и равен 2, а параметр $\beta > 0$ требуется оценить. Докажите, что данное распределение принадлежит экспоненциальному семейству по параметру β. Используя только свойства экспоненциальных семейств, найдите оценку этого параметра методом максимального правдоподобия и методом моментов.

2. Предположим, что оба параметра $\alpha, \beta > 0$ неизвестны.

   (a) Докажите, что данное распределение принадлежит экспоненциальному семейству с двумерным параметром $\theta = (\alpha, \beta)$.

   (b) Предложите алгоритм нахождения оценок неизвестного параметра методом максимального правдоподобия и методом моментов с пробными функциями, основанными на достаточной статистике T. Возможно ли найти эти оценки в явном виде?

   (c) Зафиксируйте некоторые параметры $\alpha, \beta > 0$ и сгенерируйте выборку объёма $n = 1000$ из гамма-распределения. Численно найдите оценки параметров методом максимального правдоподобия и методом моментов (с базовыми пробными функциями).

Задача 4. *(2 балла)* Количество страховых заявок X, поступающих в страховую компанию, моделируется с помощью распределения Пуассона с параметром λ, который, в свою очередь, также является случайной величиной и имеет гамма-распределение с плотностью (1).

1. Покажите, что гамма-распределение является натуральным сопряжённым для распределения Пуассона. Предполагая, что доступна выборка из распределения X, найдите апостериорное распределение параметра λ и соответствующую байесовскую оценку.

2. Зафиксируйте параметры $\alpha_0, \beta_0 > 0$ априорного распределения λ и симулируйте выборку объёма $n = 10$ из распределения X. Отобразите на одном графике априорное и апостериорное распределения λ, а также функцию правдоподобия.

3. Используя одни и те же данные, найдите байесовскую оценку параметра λ для значений параметров α, β априорного распределения, взятых по сетке от от α_0, β_0 до $\alpha_0 + 50$, $\beta_0 + 50$ с шагом 10. Сравните полученные оценки со значением, посчитанным по истинному априорному распределению (с параметрами α_0, β_0). Как изменится результат, если увеличить объём выборки до $n = 100$, $n = 1000$, $n = 10000$?

   Подсказка: для выполнения пунктов 2 и 3 рекомендуется использовать функции `plot_gamma_poisson` и `summarize_gamma_poisson` пакета `bayesrules` в `R`.

Задача 5. *(2 балла)* Пусть доступна выборка из смеси K нормальных распределений

$$p(x) = \sum_{k=1}^{K} \alpha_k p_{(\mu_k, \sigma_k^2)}(x), \quad x \in \mathbb{R}, \tag{2}$$

где $\alpha_1, \alpha_2, \ldots, \alpha_K \geq 0$, $\sum_{k=1}^{K} \alpha_k = 1$, $p_{(\mu_k, \sigma_k^2)}(x)$ — плотность нормального распределения со средним μ_k и дисперсией σ_k^2.

1. Для случая $K = 2$ найдите в явном виде вектор $\vec{\theta} = (\alpha_1, \alpha_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ оценок EM-алгоритма, то есть решение оптимизационной задачи

$$\mathrm{E}_Y \left[\log L_{\vec{\theta}}(x_1, \ldots, x_n, Y_1, \ldots, Y_n) \,|X_1 = x_1, \ldots, X_n = x_n, \vec{\theta}^{(m-1)} \right] \to \max_{\vec{\theta}},$$

   где $X_1, X_2, \ldots, X_n$ — независимые случайные величины с плотностью (2), $Y_1, Y_2, \ldots, Y_n$ — латентные переменные, обозначающие принадлежность каждого наблюдения определённой компоненте смеси, L — совместная функция правдоподобия реализаций $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$ случайных величин $X_1, X_2, \ldots, X_n$ и $Y_1, Y_2, \ldots, Y_n$ соответственно, $\vec{\theta}^{(m-1)}$ — значение $\vec{\theta}$, полученное на предыдущем шаге.

2. Для случая $K = 3$ зафиксируйте значения параметров α_k, μ_k и σ_k^2, $1 \leq k \leq 3$, и смоделируйте выборку объёма $n = 1000$ с плотностью (2). Перебирая различные значения количества компонент (от 2 до 10), найдите оценку с наибольшим значением логарифма функции правдоподобия. Зависит ли полученный результат от начальных значений EM-алгоритма? Как изменится результат, если взять μ_1 и μ_2 очень близкими?