Прикладная статистика в машинном обучении

Домашнее задание #1

Дедлайн: 11 октября, 23:59 МСК

Правила игры

1. Домашнее задание оценивается в 10 баллов.
2. Решения принимаются до **11 октября 2020 года, 23:59 МСК** включительно. Работы, отправленные после дедлайна, оцениваются следующим образом:
   - До 7:00 МСК 12 октября: максимум 8 баллов.
   - До 23:59 МСК 12 октября: максимум 7 баллов.
   - До 23:59 МСК 13 октября: максимум 6 баллов.

   Работы, отправленные после 13-го октября, будут проверены без оценки.
3. Все решения нужно загрузить в личный репозиторий на GitHub Classroom.
4. Репозиторий должен содержать:

   либо: PDF-файл с решениями теоретических задач и .ipynb-файл с решениями экспериментальных задач. Решение теоретических задач можно набрать в любом электронном редакторе или написать от руки, а затем сделать качественный скан. Все решения должны быть расположены в правильном порядке в одном файле. Если экспериментальная задача является частью теоретической, то в .ipynb-файле нужно явно указать номер теоретической задачи.

   PDF-файл должен иметь название name_surname_hw1.pdf, а .ipynb-файл должен называться name_surname_hw1_code.pdf

   либо: один .ipynb-файл с решениями и практических, и теоретических задач, оформленных в ячейках Markdown.

   .ipynb-файл должен иметь название name_surname_hw1.pdf
5. Весь код должен быть написан на Python.
6. Разрешается использовать без доказательства любые результаты, встречавшиеся на лекциях или семинарах по курсу, если получение этих результатов не является вопросом задания.
7. Разрешается использовать любые свободные источники с указанием ссылки на них.
8. Плагиат не допускается. При обнаружении случаев списывания, 0 за работу выставляется всем участникам нарушения, даже если можно установить, кто у кого списал.

Задача 1. Полезное утверждение (3 балла)

Гарри никак не может понять, почему при большой информации Фишера оценки максимального правдоподобия лежат к истинному параметру ближе, чем при малой информации Фишера. Гермиона решает продемонстрировать аналитическую интуицию, стоящую за этим утверждением:

«Если взять выборку независимых одинаково распределённых случайных величин $Y_1, \ldots, Y_N$, каждая из которых имеет функцию плотности или функцию вероятности $f(y|\theta)$, и предположить, что выполнены все необходимые условия регулярности, то при $\phi \to \theta$:

$$D_{KL}\left[f(y|\theta), f(y|\phi)\right] = \frac{1}{2} I(\theta)(\phi-\theta)^2 + O((\phi-\theta)^3)\text{».}$$

a) Докажите утверждение Гермионы либо для случая функций плотности, либо для случая функций вероятности.

b) Поясните Гарри, почему при большей информации Фишера ML-оценки лежат ближе к истинному параметру.

Подсказка: $H(f) = \mathbb{E}(\ln f)$, аналогично для кросс-энтропии.

(По мотивам: Williams, Weighing the Odds)

Задача 2. Непростительные заклинания (2 балла)

Рон убеждён, что время, которое уходит на произнесение одного непростительного заклинания – это непрерывная случайная величина с плотностью

$$f(x|q) = \begin{cases} \frac{2x}{q} e^{\frac{-x^2}{q}}, & \text{если } x > 0, \\ 0, & \text{иначе,} \end{cases}$$

где $q > 0$. Предположим, что все непростительные заклинания произносятся за одинаковое время. Рон длительное время наблюдал за тёмными волшебниками, а потому собрал случайную выборку $X_1, \ldots X_N$, где X_i – время произнесения непростительного заклинания, а N очень велико.

a) Найдите $\hat{q}_{ML}$.

b) Найдите $\widehat{(q^2)}_{ML}$

c) Постройте 95%-ый доверительный интервал для q.

d) Повторите предыдущие пункты на компьютере, численно оптимизировав функцию правдоподобия. Подробно опишите, с какими трудностями вы столкнулись в процессе и как вы их преодолели. Если трудностей не возникло, также напишите об этом.

Задача 3. Палочки и волшебники (2 балла)

Дамблдор уверен, что среди первокурсников встречаются только обладатели палочек из вишни, дуба и вяза. Воспользовавшись своими способностями в легилименции на пиру в честь начала нового учебного года, Дамблдор узнаёт, что из 150 первокурсников 75 имеют палочки из вишни, 30 – из дуба и 45 – из вяза. Дамблдор считает, что палочки выбирают волшебников независимо друг от друга, и вероятность того, что у волшебника окажется вишнёвая палочка, равна p_1, а что дубовая, равна p_2.

a) Обозначим $p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$. Найдите $\hat{p}_{ML}$.

b) Проверьте гипотезу $H_0 : p_1 = 0.7$ против $H_A : p_1 \neq 0.7$ на уровне значимости 5% при помощи тестов LR и LM.

c) Проверьте гипотезу

$$H_0 : \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}$$

против

$$H_A : \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \neq \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}$$

на уровне значимости 5% при помощи теста W.

d) Постройте 95%-ый доверительный интервал для $p_1 - p_2$.

Задача 4. Модель для зелий (3 балла)

Полумна хочет построить предсказательную модель, которая бы описывала зависимость популярности зелья y_i от силы его положительного влияния x_i. Обе величины являются количественными непрерывными переменными на $\mathbb{R}$. Предположим, что Полумна знает, как измерить популярность и силу влияния и верит, что искомая зависимость имеет следующий вид:

$$y_i = \beta_1 e^{\beta_2 x_i} u_i,$$

где β_1 и β_2 – неизвестные коэффициенты, не равные нулю, u_i – случайная ошибка, причём $\ln u_i \sim \mathcal{N}(0{,}1)$.

a) Является ли данная зависимость линейной по β_1? А по β_2?

b) Найдите $\hat{\beta}_1$ и $\hat{\beta}_2$ методом максимального правдоподобия.

c) Симулируйте 300 наблюдений (x_i, y_i) таким образом, что $x_i \sim \mathcal{N}(0, 1)$, $y_i = 3e^{x_i} u_i$, $\ln u_i \sim \mathcal{N}(0, 1)$.

d) По полученным данным найдите $\hat{\beta}_1$ и $\hat{\beta}_2$ в числах.

e) Проверьте гипотезу

$$H_0 : \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

против

$$H_A : \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \neq \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

на уровне значимости 5% при помощи теста LR.