Прикладная статистика в машинном обучении

Домашнее задание #1

Часть 1

Дедлайн: 29 октября, 23:59 МСК

Правила игры

1. Домашнее задание состоит из двух частей. Часть 1 содержит 13 обязательных и две бонусных задачи и предполагает решение «от руки». Часть 2 содержит 3 обязательных задачи и предполагает программное решение.

2. Домашнее задание оценивается в 80 баллов. При этом часть 1 оценивается в 65 баллов, а часть 2 – в 15 баллов. По умолчанию за каждый пункт каждой задачи можно получить 1 балл. Однако за некоторые пункты некоторых задач можно получить другое количество баллов, которое явно указано в скобках рядом с меткой пункта.

3. Каждый пункт оценивается с промежутком 0.5. Например, если за пункт можно получить максимум 1 балл, то за полностью корректное решение ставится 1 балл, за решение с небольшими ошибками ставится 0.5 балла, за решение с серьёзными ошибками или неправильное решение ставится 0 баллов. Для пунктов, за которые можно получить максимум 2 балла, в зависимости от решения можно получить $2, 1.5, 1$ и т.д. баллов. При этом пункты проверяются независимо друг от друга: если пункт $t+1$ зависит от численных результатов пункта t, и в пункте t допускается ошибка, из-за которой в пункт $t+1$ приходят неверные входные данные, то при корректном решении пункт $t+1$ оценивается в максимальное количество баллов, которое можно за него получить.

4. Бонусные задачи X и Y приведены в конце части 1 и обозначены значком †. Эти задачи необязательны к решению и учитываются сверх установленных 80 баллов. Баллы за корректно решённые бонусные задачи прибавляются к набранным баллам, даже если в сумме получается больше 80 баллов (оценка за домашнюю работу в этом случае будет больше 10, и так и будет внесена в таблицу с оценками).

5. Весь код должен быть написан на Python, R, C или C++.

6. Решения принимаются до **29 октября 2021 года, 23:59 МСК** включительно. Работы, отправленные после дедлайна, проверяются, но **не оцениваются**.

7. Все решения нужно загрузить в личный репозиторий на GitHub Classroom.

8. Репозиторий должен содержать: PDF-файл с решениями задач части 1 и файл с кодом с решениями задач части 2. Решение задач части 1 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. Файлы должны быть названы по типу «name_surname_group_hw1_part1.pdf» и «name_surname_group_hw1_part2.ext», где вместо ext может быть .py, .ipynb, .R, .c, .cpp. Если решение части 2 разбивается на несколько файлов кода, то в репозиторий нужно загрузить все файлы, а в README.md подробно указать, что содержит каждый файл.

9. Разрешается использовать без доказательства любые результаты, встречавшиеся на лекциях или семинарах по курсу, если получение этих результатов не является вопросом задания. Разрешается использовать любые свободные источники с указанием ссылки на них.

10. Плагиат не допускается. При обнаружении случаев списывания, 0 за работу выставляется всем участникам нарушения, даже если можно установить, кто у кого списал.

Задача 1. Просто компания

Компания «Напиши-ка» производит три вида ручек: синие, красные и зелёные. Глава аналитического отдела компании Данил хочет понять, какая из ручек скорее всего «выстрелит», а какая не будет пользоваться успехом у покупателей. Для этого он анализирует выборку в 300 проданных ручек. Оказалось, что из них 150 синих, 100 красных и 50 зелёных ручек. Данил уверен, что ручки продаются независимо друг от друга, и вероятность того, что будет продана синяя ручка, равна p_1, а что красная p_2.

[а] Обозначим $p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$. Найдите $\hat{p}_{ML}$ интуитивно, не выписывая правдоподобие, и поясните, как вы это сделали.

[б] Выпишите функцию правдоподобия и найдите $\hat{p}_{ML}$ как точку её глобального максимума.

[в] Проверьте гипотезу

$$\begin{cases} H_0 : p_1 = 0.2, \\ H_A : p_1 \neq 0.2 \end{cases}$$

на уровне значимости 5% при помощи тестов LR и LM.

[г] Проверьте гипотезу

$$\begin{cases} H_0 : \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.2 \end{pmatrix}, \\ \\ H_A : \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \neq \begin{pmatrix} 0.3 \\ 0.2 \end{pmatrix} \end{cases}$$

на уровне значимости 5% при помощи тестов LR и W.

[д] Постройте график логарифма правдоподобия в трёхмерной плоскости. Покажите на графике $\hat{p}_{ML}$ визуальную интерпретацию тестов LR и W для гипотезы из предыдущего пункта.

[е] Постройте 95%-ый доверительный интервал для p_3.

[ж] Постройте 99%-ый доверительный интервал для $p_1 + p_2$.

[з] Постройте 90%-ый доверительный интервал для $\hat{p}_1$.

Подсказка: помните, что мы работаем в рамках частотного подхода.

[и] Приведите разумное интерпретируемое определение того, что ручка «выстрелила».

[к] Пользуясь определением из предыдущего пункта, сформулируйте гипотезу о том, что «выстрелит» ручка синего цвета и проверьте её при помощи любого из тестов LR, LM или W на уровне значимости 5%.

Задача 2. Анекдоточная

Станислав знает, что хороший анекдот должен быть не очень коротким, но и не слишком длинным. Время, за которое Станислав произносит один анекдот, – это непрерывная случайная величина с плотностью

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

где b – некоторый параметр. Станислав собрал случайную выборку по продолжительности рассказанных им анекдотов: $X_1, X_2, \ldots, X_n$, где $n = 10^6$. Оказалось, что $\sum X_i^2/n = 20$, $\sum X_i/n = 2$.

[а] Найдите $\hat{b}_{ML}$.

[б] Проверьте гипотезу

$$\begin{cases} H_0 : b = 3, \\ H_A : b \neq 3 \end{cases}$$

на уровне значимости 5% при помощи теста LR.

[в] Рассчитайте LM-статистику для проверки гипотезы

$$\begin{cases} H_0 : b = 1, \\ H_A : b \neq 1 \end{cases}$$

Чему приблизительно равно соответствующее p-value?

[г] Проверьте гипотезу из предыдущего пункта, построив соответствующий доверительный интервал для b.

Задача 3. «Я не дерево. Я энт».

Исследователь Матвей подбрасывает монетку с вероятностью орла p до тех пор, пока не выпадет два орла (всего, не обязательно подряд). Он сыграл четыре игры, и оказалось, что первая завершилась за 3 хода, вторая – за 3 хода, третья – за 2 хода, четвёртая – за 4 хода. Будем считать, что подбрасывания в течение одной игры независимы. Также предположим, что игры происходили независимо друг от друга.

[а] **(2 балла)** Найдите $\hat{p}_{ML}$.

[б] Найдите $\hat{a}_{ML}$ для нового параметра $a = (p^2 + 3p^3 - 1)$.

[в] **(2 балла)** Покажите, что $\hat{p}$ является состоятельной оценкой p.

Подсказка: для решения **Задачи X** потребуется доказать, что если M – число ходов, за которое завершится игра, то $\mathbb{E}(M) = \frac{2}{p}$. В этой задаче можно пользоваться этим утверждением без доказательства.

Задача 4. Полезное утверждение

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 $\phi \to \theta$:

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

[а] **(2 балла)** Докажите утверждение Гермионы либо для случая функций плотности, либо для случая функций вероятности.

[б] **(2 балла)** 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 5. Модель для зелий

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:

$$y_i = (\beta_1)^2 e^{-\beta_2 x_i} u_i,$$

где β_1 и β_2 – неизвестные положительные коэффициенты, u_i – случайная ошибка, причём $\ln u_i \sim \mathcal{N}(0, 2)$.

[а] **(2 балла)** Введите любые разумные ограничения на переменные. Найдите $\hat{\beta}_1$ и $\hat{\beta}_2$ методом максимального правдоподобия.

[б] **(2 балла)** Полумна собрала выборку, для которой оказалось, что

$$\sum_{i=1}^{n} y_i = 100,\ \sum_{i=1}^{n} x_i = 50,\ \sum_{i=1}^{n} x_i y_i = 200,$$

$$\sum_{i=1}^{n} x_i^2 = 2500,\ \sum_{i=1}^{n} y_i^2 = 10000,\ \sum_{i=1}^{n} e^{-\hat{\beta}_2 x_i} = 1,$$

$$n = 500.$$

Проверьте гипотезу

$$\begin{cases} H_0 : \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \\ \\ H_A : \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \neq \begin{pmatrix} 1 \\ 2 \end{pmatrix} \end{cases}$$

на уровне значимости 5% при помощи теста W.

Подсказка: $\ln Y \sim N(\mu, \sigma^2) \Rightarrow \mathbb{E}(Y) = e^{\mu + \frac{\sigma^2}{2}}$.

Задача 6. Функции правдоподобия

Пусть $X_1, \ldots, X_n$ – выборка независимых одинаково распределённых величин из распределения с параметром $p \in [0, 1]$. Известно, что $n = 100$, $\bar{X} = 20$, $(\bar{X^2}) = 400$. Найдите $\hat{p}_{ML}$ для следующих функций (можно либо вывести в явном виде, либо использовать математический анализ):

[а]

$$\ell(p) = \frac{\sqrt{X_1 + \ldots + X_n}}{50 - p} + \frac{\ln p}{X_1^2 + \ldots + X_n^2}.$$

[б]

$$\ell(p) = \frac{(p^2 - \ln p)\bar{X}^2}{\bar{X}}.$$

Задача 7. Дивергент

Рассмотрим распределения $p = \mathcal{N}(1, 2)$, $q = \mathrm{Exp}(1)$, $r = \mathrm{Bin}(3, 0.5)$. Для каждого пункта приведите математическое обоснование ответа.

[а] Найдите $D_{KL}(p\|q)$.

[б] Найдите $D_{KL}(q\|p)$.

[в] Найдите $D_{KL}(p\|r)$.

[г] Найдите $D_{KL}(q\|r)$.

[д] Возможно ли применить линейное преобразование к p, q или r так, чтобы ответ на хотя бы один из пунктов выше изменился?

Задача 8. Между молотом и наковальней

Одной из симметричных альтернатив KL-дивергенции является *взаимная информация*: для случайных величин X и Y она определяется как

$$I(X, Y) = H(X) - H(X|Y),$$

где $H(X|Y) = -\int p(x, y) \ln \frac{p(x, y)}{p(y)}$.

[а] Покажите, что $I(X,Y) = I(Y,X)$.

[б] **(2 балла)** Покажите, что $I(X,Y) = D_{KL}(p(x,y) \| p(x) \times p(y))$.

[в] Поясните интуитивную интерпретацию $I(X,Y)$.

Задача 9. Хорошая задача на экзамен

Случайная величина X принимает значение 0 с вероятностью p, значение 1 с вероятностью $1/3$ и значение 2 с вероятностью $2/3 - p$.

[а] Постройте график зависимости $H(X)$ как функцию от p.

[б] При каком p энтропия будет максимальна? Поясните полученный результат.

Задача 10. Порисуем!

Рассмотрим модель множественной регрессии $y = X\beta + u$, которая оценивается при помощи МНК. Число наблюдений равно $n = 400$, число регрессоров равно $k = 10$, включая константный. Все регрессоры ортогональны друг другу.

[а] Долорес Амбридж строит регрессию по константному и следующим за ним четырём регрессорам. Корнелиус Фадж строит регрессию по константному и оставшимся пяти регрессорам. Покажите на *единой* картинке МНК $\hat{y}$, TSS, ESS, RSS и R^2 в их регрессиях.

[б] Альбус Дамблдор строит регрессию по всем 10 регрессорам. Покажите на той же картинке МНК $\hat{y}$, TSS, ESS, RSS и R^2 в его регрессии.

[в] **(2 балла)** Гарри Поттер хочет сравнить регрессии Амбридж и Дамблдора при помощи F-теста. Напомним, что

$$F = \frac{(RSS_R - RSS_{UR})/(k_{UR} - k_R)}{RSS_{UR}/(n - k_{UR})}.$$

Покажите на картинке МНК RSS_R, RSS_{UR} и угол, квадрату тангенса которого пропорциональна F-статистика.

[г] Приведите геометрическую интерпретацию F-теста.

Задача 11. Подпространства

Рассмотрим пространство $\mathbb{R}^3$ и два подпространства в нём

$$W = \{(x_1, x_2, x_3) | 3x_1 + 2x_2 - x_3 = 0\}$$

и

$$V = \text{Lin}[(1,1,1)^T].$$

[а] Найдите $\dim V$, $\dim W$, $\dim(V \cap W)$, $\dim V^\perp$, $\dim W^\perp$.

[б] Найдите проекцию произвольного вектора u на V, W, $V \cap W$, $V^\perp$, $W^\perp$. Найдите квадрат длины каждой проекции.

[в] Как распределён квадрат длины проекции в каждом случае, если дополнительно известно, что вектор u имеет многомерное стандартное нормальное распределение?

Задача 12. Парная регрессия

Исследователь Борис работает с обычной парной регрессией

$$y_i = \beta_0 + \beta_1 X_i + u_i,$$

которую он оценивает при помощи МНК:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

[а] Просто для удобства выпишите RSS в этой регрессии и условия первого порядка в задаче минимизации.

[б] Докажите, что $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$.

[в] Докажите, что $\bar{y} = \bar{\hat{y}}$.

[г] Докажите, что точка $(\bar{x}, \bar{y})$ лежит на линии оценённой регрессии.

[д] Докажите, что $\sum_{i=1}^{n} x_i(y_i - \hat{y}_i) = 0$.

Задача 13. Гипотезы в линейной регрессии

Линейная регрессионная модель задаётся в следующем виде:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i.$$

Предположим, что $u \sim \mathcal{N}(0, \sigma^2 I)$. Известно, что

$$X = \begin{pmatrix} 1 & 1 & 3.1 \\ 1 & 12 & 2.2 \\ 1 & -3 & 0.1 \\ 1 & 2 & 0.5 \\ 1 & 0 & 11.3 \end{pmatrix}, y = \begin{pmatrix} 1.1 \\ 2.5 \\ 2.2 \\ 4 \\ 1 \end{pmatrix}$$

В процессе решения используйте калькулятор, все числа округляйте до сотых.

[а] Найдите $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$.

[б] Найдите $\hat{y}$.

[в] Найдите TSS, ESS, RSS и R^2.

[г] Найдите $\hat{\sigma}$.

[д] Найдите $\widehat{\mathrm{Var}}(\hat{\beta})$.

[е] На уровне значимости 5% проверьте гипотезу

$$\begin{cases} H_0 : \beta_1 = 1, \\ H_1 : \beta_1 \neq 1. \end{cases}$$

[ж] На уровне значимости 10% проверьте гипотезу

$$\begin{cases} H_0 : \beta_2 = 1, \\ H_1 : \beta_2 < 1. \end{cases}$$

[з] Проверьте регрессию на значимость в целом.

[и] На уровне значимости 5% проверьте гипотезу

$$\begin{cases} H_0 : \beta_1 = \beta_2, \\ H_1 : \beta_1 \neq \beta_2. \end{cases}$$

[к] Постройте 95%-ый доверительный интервал для β_1.

[л] Пусть $x_{1,6} = 10$, $x_{2,6} = 7$. Найдите $\hat{y}_6$.

[м] Постройте 95%-ый доверительный интервал для $\mathbb{E}(y_6 | x_{1,6}, x_{2,6})$.

Задача X†. Методы моментов и первого шага

Альтернативой методу максимального правдоподобия является *метод моментов*, суть которого заключается в том, чтобы приравнять теоретические моменты как функции от оцениваемых параметров к их выборочным аналогам, и из полученной системы найти оценки.

[а] Пункт для тренировки. Рассмотрим выборку X_1, X_2, $X_3 \sim i.i.d.\ \mathcal{N}(\mu, 1)$. Оказалось, что $X_1 = 1$, $X_2 = 2$, $X_3 = 3$. Найдите $\hat{\mathbb{E}}(X_1)_{MM}$.

[б] **(6 баллов)** Исследователь Матвей подбрасывает монетку с вероятностью орла p до тех пор, пока не выпадет два орла (всего, не обязательно подряд). Оказалось, что среднее число ходов, за которое завершится игра, равно 40. Найдите $\hat{p}_{MM}$.

Подсказка: докажите, что если M – число ходов, за которое завершится игра, то $\mathbb{E}(M) = \frac{2}{p}$.

Задача Y†. Известное неравенство

(6 баллов)

Рассмотрим линейную модель

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i,$$

где $u_i \sim \mathcal{N}(0, 1)$, а все предпосылки ТГМ выполнены. Исследователь Вадим тестирует гипотезу вида

$$\begin{cases} H_0 : \beta_1 = C \\ H_A : \beta_1 \neq C, \end{cases}$$

где C – некоторая константа, при помощи тестов LR, LM и W. Докажите, что в такой постановке всегда верно, что $LM \leqslant LR \leqslant W$.