

**ПСМО ФКН ВШЭ, 3 курс, 2 модуль**

## Задание 4. Регрессия. Моделирование выборок. Бутстреп.

**Прикладная статистика в машинном обучении, осень 2019**

Время выдачи задания: 9 декабря.

Срок сдачи: **19 декабря (четверг), 23:59.**

Среда для выполнения практического задания – PYTHON 2.x/PYTHON 3.x.

## Правила сдачи

### Инструкция по отправке:

1. Решения задач следует присылать единым файлом формата **pdf**, набранным в **L<sup>A</sup>T<sub>E</sub>X**, либо в составе **ipython**-тетрадки в форматах **ipynb** и **html** (присылайте оба формата, т.к. AnyTask из-за высокой загрузки иногда не рендерит тетрадки в формате **ipynb** – а если мы не увидим ваши задачи, мы их не проверим). Отправляйте практические задачи в виде отдельных файлов (**ipython**-тетрадок или исходных файлов с кодом на языке **python**).

### Оценивание и штрафы:

1. Максимально допустимый набранный балл за все задания – 17 баллов.
2. Оценка за домашнее считается как  $\min(10, \text{набранный балл})$ .

3. Бонусные баллы считаются как  $\max(0, \text{набранный балл} - 10)$ .
4. Дедлайн жесткий. Сдавать задание после указанного срока сдачи нельзя.
5. Задание выполняется каждым студентом индивидуально и независимо от других студентов. «Похожие» решения считаются плагиатом и все студенты (в том числе те, у кого списали) не могут получить за него больше 0 баллов, причем обнуляются и бонусные баллы. 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).

## Основные задачи

1. (4 балла) Воспроизведите часть результатов построения доверительных интервалов для оценивания интенсивности пуассоновского потока событий в наличии мешающего параметра. Пусть

$$N \sim \text{Poisson}(\theta + \nu)$$

наблюдаемая случайная величина (число реализаций пуассоновского потока событий при наличии *сигнала* в данных), где  $\theta$  – интересующая нас интенсивность пуассоновского потока сигнала, а  $\nu$  – интенсивность *мешающего* пуассоновского потока *фоновых* событий. Для определения интенсивности  $\nu$  делается дополнительное измерение

$$K \sim \text{Poisson}(\tau\nu)$$

с известной постоянной  $\tau$ . Задача – построить доверительные интервалы для  $\theta$ .

- Рассмотрите ряд значений  $\theta_i = \Delta i$ , где  $\Delta = 0.1$ ,  $i = 1, \dots, 200$ . Возьмите фиксированные значения  $\nu = 1$  и  $\tau = 1$ .
- Смоделируйте выборку объема 100 измерений сигнала (случайной величины  $N$ ) и 1000 измерений фона (случайной величины  $K$ ).
- Рассмотрите следующие методы построения доверительных интервалов: Exact likelihood ratio test inversion, Asymptotic LR Test Inversion, Simple Percentile Bootstrap, Automatic Percentile Bootstrap.
- Для каждого из рассматриваемых методов и для каждого значения  $\theta_i$  постройте 95% доверительный интервал для оценки  $\hat{\theta}$ .

- В ответе приведите для каждого метода описание процедуры расчета доверительных интервалов и графики построенных доверительных интервалов в сравнении с истинным 95% доверительным интервалом для среднего значения  $\mathbb{E}\theta$ .

2. (2 балла) Пусть  $T_n = \bar{X}_n^2$ ,  $\mu = \mathbb{E}(X_1)$ ,  $\alpha_k = \int (x - \mu)^k dF(x)$  и  $\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^k$ . Докажите, что оценка дисперсии функционала  $T_n$  с помощью бутстрепа равна:

$$v_{boot} = \frac{4\bar{X}_n^2\hat{\alpha}_2}{n} + \frac{4\bar{X}_n\hat{\alpha}_3}{n^2} + \frac{\hat{\alpha}_4}{n^3} + \frac{\hat{\alpha}_2^2(2n-3)}{n^3}$$

То есть математическое ожидание берётся по бутстрапной (эмпирической) плотности распределения.

3. (2 балла) Пусть есть выборка из 11 элементов:  $x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)} < x_{(5)} < x_{(6)} < x_{(7)} < x_{(8)} < x_{(9)} < x_{(10)} < x_{(11)}$ . Оцениваемая статистика  $\theta$  – медиана.

1. Покажите что для оценки  $\hat{\theta}$  по бутстрапной выборке верно следующее:

$$P(\hat{\theta} > x_{(i)}) = \sum_{j=0}^5 Bi\left(j, n, \frac{i}{n}\right),$$

где  $Bi(j; n, p) = C_n^j p^j (1-p)^{n-j}$ .

2. Покажите что оценка  $\hat{\theta}$  по бутстрапной выборке равна  $x_{(i)}$  с вероятностью:

$$P(\hat{\theta} = x_{(i)}) = \sum_{j=0}^5 \left( Bi\left(j; n, \frac{i-1}{n}\right) - Bi\left(j, n, \frac{i}{n}\right) \right),$$

3. Используя результат пункта (1) выведите 90% бутстрепный доверительный интервал для медианы (подсказка: подсчитайте  $p(\hat{\theta} \geq x_{(3)})$  и  $p(\hat{\theta} \geq x_{(9)})$ ).
4. (3 балла) Рассмотрим данные по адресу <https://vincentarelbundock.github.io/Rdatasets/csv/MASS/galaxies.csv>, описывающие скорости 82 галактик из созвездия Северной Короны. Мы хотим узнать, есть ли пустоты или суперкластеры в данной части вселенной. Одним из свидетельств является мультимодальность распределения скоростей галактик. Другими словами, нам необходимо проверить гипотезу унимодальности распределения, т.е.:

$$H_0 : n_{mode}(p) = 1 \text{ vs } H_a : n_{mode}(p) > 1$$

Плотность распределения будем оценивать напараметрическим ядерным методом:

$$\hat{p}_{K,h}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

- (a) По данным найдите минимальное  $\hat{h}_{uni}$  при котором распределение ещё унимодально.

Найденная  $\hat{h}_{uni}$  является оценкой по данным для реальной  $h_{uni}$ . Если окажется, что  $h_{uni} > \hat{h}_{uni}$ , то это значит что в реальности мод больше одной. Т.е. нулевая гипотеза отвергается на уровне значимости  $\alpha$ :

$$P(\text{multimodal}) = P(h_{uni} > \hat{h}_{uni}) \leq \alpha$$

- (b) Используя бутстреп оцените следующую величину:

$$\hat{P}(h_{uni} > \hat{h}_{uni}) \approx \frac{1}{B} \sum_{b=1}^B \left( \hat{h}_{uni}^b \geq \hat{h}_{uni} \right)$$

Сэмплирование делайте из  $\hat{p}_{K, \hat{h}_{uni}}(x)$ , т.е.  $X^* \sim X + \hat{h}_{uni}N(0, 1)$ , где  $X$  – случайный элемент изначальной выборки.

Н.В.: так как сэмплирование делается не из оригинальной эмпирической выборки, а из сглаженной, то дисперсия стала выше. Подумайте как нужно скорректировать предложенную схему сэмплирования, чтобы дисперсия не изменилась?

Подсказка: для схемы сэмплирования  $X^* \sim a + b(X + \hat{h}_{uni}N(0, 1))$ , где  $X$  – случайный элемент изначальной выборки, найдите такие  $a$  и  $b$ , что первый и второй моменты этого распределения и эмпирического распределения совпадают.

(с) С каким уровнем значимости отвергается нулевая гипотеза?

5. (2 балла) Рассмотрим датасет по адресу <https://vincentarelbundock.github.io/Rdatasets/csv/boot/cd4.csv>

Датасет CD4 содержит информацию о 20 ВИЧ-инфицированных пациентах до и после года лечения на экспериментальном анти-вирусном лекарстве.

- (а) Для коэффициента корреляции Пирсона между данными до и после лечения посчитайте 95% доверительный интервал следующими методами: нормальный, перцентильный, центральный и t-бутстреп. Для подсчёта дисперсии  $\hat{\sigma}$  для t-bootstrap используйте следующую формулу:  $\hat{\sigma}(r) = \sqrt{\frac{1-r^2}{n-2}}$ .

Очень часто для работы бывает удобно применить нормализующее преобразование:  $\frac{1}{2} \log \frac{1+r}{1-r}$ . Сделайте это и посчитайте заново все интервалы. Что изменилось? Стали ли они более согласованными? Для подсчёта дисперсии для t-bootstrap в нормализованном виде используйте следующую оценку дисперсии:  $\hat{\sigma}(r) = \frac{1}{\sqrt{n-3}}$

- (b) В предыдущем пункте для t-bootstrap вы использовали некоторое аналитическое приближение для вычисления дисперсии корреляции. Теперь вам предстоит реализовать двойной бутстрап: над бутстрапными выборками по которым считается  $r$  делайте ещё бутстрап для оценки  $\sigma(r)$ . Сравните дисперсии посчитанные аналитически и бутстрапом. Что вы заметили? Как изменились доверительные интервалы?
- (c) После предыдущего пункта вам наверняка захотелось проверить все наши оценки на смещённость. Оцените смещение (bias) для коэффициента корреляции с помощью jackknife.
6. (4 балла) Фирма, занимающаяся маркетинговыми исследованиями, была нанята производителем автомобилей для определения вероятности того, что семья купит новую машину в течение следующего года. Была получена случайная выборка из 10 семей, у которых узнавали данные о годовом доходе. Опрос, проведённый 12 месяцев спустя, проверял купила ли семья автомобиль. Скачайте данные `car_reduced.table`.
- Постройте логистическую регрессию для предсказания покупки в зависимости от дохода. Укажите проблему. Для понимания природы проблемы постройте логарифм профильной функции правдоподобия для коэффициентов регрессии.
  - Для решения проблемы применяют регуляризованную функцию правдоподобия Фирта (Firth):

$$\log \mathcal{L}^*(\beta) = \log \mathcal{L}(\beta) + \frac{1}{2} \log \det I(\beta)$$

где  $\mathcal{L}$  - стандартная функция правдоподобия,  $I(\beta)$  - информационная матрица Фишера. В случае логистической регрес-

сии с одним фактором, можем записать:

$$I(\beta) = X^T W X,$$

где  $X$  - матрица дизайна эксперимента (признаковое описание объектов), а  $W$  определяется по формуле:

$$W = \text{diag}(\hat{y}_i(1 - \hat{y}_i))$$

Проверьте, что такой способ решает проблему из первого пункта.

- Стандартным решением проблемы полной разделимости данных является получение дополнительного набора данных. Вам удалось получить 23 новых примера, кроме того удалось добавить ещё одну переменную – возраст текущего автомобиля. Скачайте `car.table`, проверьте, что обычная логистическая регрессия работает в случае зависимости только от дохода. Сравните коэффициенты для обычной регрессии и регуляризованной. Обратите внимание, что этот этап можно выполнить двумя способами: напрямую оптимизируя регуляризованное правдоподобие или подсчитав значения правдоподобия на узлах решётки.
- Постройте двухфакторную модель.
- Для проверки качества постройте QQ-график остатков модели против нормального распределения. О чём говорит график? Можно ли его объяснить?