ДГПС ФКН ВШЭ, 3 курс, 3 модуль

Задание 2. Регрессии.

Дополнительные главы прикладной статистики, весна 2021

Время выдачи задания: 26 марта (пятница).

Срок сдачи: **11 апреля (воскресенье), 18:00.**

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

Правила сдачи

Инструкция по отправке:

1. Решения теоретических и практических задач следует присылать единой IPython-тетрадкой в форматах `ipynb` **и** `html` (присылайте оба формата, т.к. AnyTask из-за высокой загрузки иногда не рендерит часть тетрадки в формате `ipynb` – а если мы не увидим какую-то задачу, мы не сможем её проверить).

Оценивание и штрафы:

1. Максимально допустимая оценка за работу над основными задачами – 20 баллов, которые делением на 2 переводятся в 10-балльную систему.

2. Дедлайн жесткий. Сдавать задание после указанного срока сдачи нельзя.

3. 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) Вы задались целью статистически достоверно сравнить качество двух стохастических алгоритмов машинного обучения (например, алгоритмов из семейства reinforcement learning). Предположим, что качество алгоритма 1 задается (случайной) величиной X_1, а качество алгоритма 2 – величиной X_2 (распределения X_1 и X_2 неизвестны). Алгоритм 1 назовем неразличимым по качеству с алгоритмом 2, если их средние уровни качества равны: $\Delta\mu = \mu_1 - \mu_2 = \mathrm{E}X_1 - \mathrm{E}X_2 = 0$; в противном случае алгоритм 1 лучше (хуже) по качеству, чем алгоритм 2.

   Для того, чтобы сравнивать алгоритмы по качеству, воспользуемся аппаратом проверки статистических гипотез. Таким образом, для сравнения алгоритмов по качеству необходимо проверить гипотезу $\mathcal{H}_0 : \Delta\mu = 0$ против альтернативы $\mathcal{H}_1 : |\Delta\mu| > 0$ по выборкам $(x_1^1, \ldots, x_n^1)$ и $(x_1^2, \ldots, x_n^2)$, показывающим значения их метрик, полученных в эксперименте.

   Проведите статистическое моделирование для сравнения эффективности нескольких распространенных статистических критериев в задаче различения алгоритмов по качеству.

   - В качестве критерия рассмотрите два теста: тест Манна-Уитни-Уилкоксона и t-критерий Уэлча. Критерий Манна-Уитни-Уилкоксона реализуется функцией `scipy.stats.mannwhitneyu(x1, x2, alternative='two-sided')`, а t-критерий Уэлча функцией `scipy.stats.ttest_ind` с `equal_var=False`.
   - В качестве множества постановок задачи рассмотрите ситуации, когда X_1 и X_2 имеют:

  - одинаковый тип распределения и равные стандартные отклонения;
  - одинаковый тип распределения, но неравные стандартные отклонения;
  - различные типы распределения и равные стандартные отклонения;
  - различные типы распределения и неравные стандартные отклонения.
- В качестве типов распределения рассмотрите следующие: стандартное нормальное распределение, логнормальное распределение, распределение Коши (с «тяжелыми хвостами») на отрезке $[-3, 3]$. Все распределения отмасштабируйте так, чтобы их среднее $\mu = 0$, стандартное отклонение $\sigma = 1$. При рассмотрении различных стандартных отклонений положите $\sigma_2 = 2\sigma_1$.
- Проведите следующие эксперименты. Все эксперименты необходимо провести с числом повторений $N_r = 10^3$, для каждого критерия, каждой постановки и размеров выборок $N_s \in \{1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 100\}$.
  - Измерение вероятности ложной тревоги: зафиксируйте $\alpha = 0.05$ и при верной $\mathcal{H}_0$ подсчитайте долю случаев, в которых была отклонена гипотеза $\mathcal{H}_0$.
  - Измерение мощности теста: зафиксируйте $\alpha = 0.05$ и при верной $\mathcal{H}_1$ подсчитайте долю случаев, в которых была отклонена гипотеза $\mathcal{H}_0$. При этом размер сдвига $\Delta\mu$ варьируйте в диапазоне от 0 до 3 с шагом 0.1.

Требования к оформлению результатов в этой задаче:

- Должны быть представлены графики зависимостей вероятно-

сти ошибки I рода от размера выборки N_s для каждой постановки (при этом на одном и том же графике должны быть представлены кривые для каждого критерия). Сгруппируйте графики по типу рассматриваемой постановки (например, в разделе «одинаковый тип распределения и равные стандартные отклонения» должно быть 3 графика, на каждом по 2 кривые, и т.д.).

- Должны быть представлены графики зависимостей мощности критерия от размера выборки N_s для каждой постановки (при этом на одном и том же графике должны быть представлены кривые для каждого критерия). Сгруппируйте графики по типу рассматриваемой постановки.
- К отчету должен быть приложен исходный код, реализующий сравнение.

2. (3 балла) Загрузите данные из файла https://raw.githubusercontent.com/SchattenGenie/pic-storage/master/figure_skating.csv, это результаты женского фигурного катания Олимпиады-2014 года в Сочи. Сравните перцентильное и квантильное гауссовы преобразования, а также преобразование Бокса-Кокса. Для этого проведите на полученных данных тест Шапиро-Вилкса и постройте для каждого преобразования график изначальных оценок против преобразованных (QQ график). Сделайте вывод: какое преобразование вы бы предпочли для дальнейшего анализа?

3. (4 балла) Существуют две основные альтернативы классическому параметрическому дисперсионному анализу Фишера:

   - анализ Уэлча;
   - анализ Брауна-Форсайта (Brown-Forsythe).

Для построения теста Уэлча при выборочной дисперсии s_i^2 в группе i с n_i событиями, при общем количестве групп r, вводятся веса $w_i = \frac{n_i}{s_i^2}$. После чего производится подсчёт следующей характеристики:

$$SSTR = \sum_{i=1}^{r} w_i(\bar{Y}_{i\cdot} - \bar{Y}_w)^2,$$

где $\bar{Y}_w$ - взвешенное с w_i среднее средних в группе i; $\bar{Y}_{i\cdot}$ – среднее в группе i. Для удобства расчётов посчитаем коэффициент Λ:

$$\Lambda = \frac{3\sum\limits_{i=1}^{r} \frac{\left(1-\frac{w_i}{\sum\limits_{i=1}^{r} w_i}\right)^2}{n_i-1}}{r^2-1}.$$

В этом случае, F-статистика имеет следующий вид:

$$F_w = \frac{SSTR/(r-1)}{1+2\Lambda(r-2)/3} \sim F(r-1; 1/\Lambda).$$

Сравните:

- Классический дисперсионный анализ (ANOVA).
- Дисперсионный анализ с использованием метода Уэлча.
- Дисперсионный анализ с использованием теста Краскела-Уоллиса.

Рассмотрите случай трёх выборок. Сделайте выводы о регионах применимости F-теста, теста Уэлча, теста Краскела-Уоллиса для однофакторного анализа.

Подсказка:
Возможные эксперименты для рассмотрения:

(a) Сбалансированные выборки, набранные из $\mathcal{N}(0; 1)$, для $n = 5, 10, 20, 100$.

(b) Сбалансированные выборки, набранных из $\mathcal{N}(0;\sigma)$, для $n = 5, 10, 20, 100$. При этом σ для выборок – случайная тройка из элементов $1, 2, 3, 4$.

(c) Несбалансированные выборки, набранных из $\mathcal{N}(0;1)$, для n из набора $5, 10, 20, 100$.

(d) Сбалансированные выборки, со смещённым средним нормального распределения (средние могут принимать значения $0, 1, 2$)

(e) Эксперименты 1-4 для логнормального распределения.

Возможные переменные для вывода (критическое значение $\alpha = 0.05$):

- Эмпирическая ошибка 1-го рода.
- Мощность теста.

4. (5 балла) Рассмотрим данные, описывающие скорости 82 галактик из созвездия Северной Короны (https://vincentarelbundock.github.io/Rdatasets/csv/MASS/galaxies.csv). Мы хотим узнать, есть ли пустоты или суперкластеры в данной части вселенной. Одним из свидетельств наличия суперкластеров является мультимодальность распределения скоростей галактик. Другими словами, нам необходимо проверить гипотезу об 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 α:

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

(b) Используя бутстреп оцените следующую величину:

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

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.

N.B.: 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? (Если нужна подсказка – можете обратиться в личку :)

(c) С каким уровнем значимости отвергается нулевая гипотеза?

5. (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 https://raw.githubusercontent.com/SchattenGenie/pic-storage/master/car_reduced.table.

- Постройте логистическую регрессию (без регуляризации) для предсказания покупки в зависимости от дохода. Укажите проблему. Для понимания природы проблемы постройте логарифм профильной функции правдопободия для коэффициентов регрессии.
- Для решения проблемы применяют регуляризированную фукнцию правдоподобия Фирта (Firth):

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

где $\mathcal{L}$ - станадартная функция правдоподобия, $I(\beta)$ - информационная матрица Фишера. В случае логистической регрессии 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.
- Стандартной решением проблемы полной разделимости данных является получаение дополнительного набора данных. Вам удалось получить 23 новых примера, кроме того удалось добавить ещё одну переменную – возраст текущего автомобиля. Скачайте https://raw.githubusercontent.com/SchattenGenie/pic-storage/master/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?