Программа коллоквиума по курсу «Прикладная статистика в машинном обучении»

А. Артемов, Д. Деркач, М. Шараев,
К. Кондратьева, В. Белавин,
Д. Полонская, А. Хачиянц

30 октября 2018 г.

Глава 1

Введение. Правила проведения коллоквиума

1. Коллоквиум сдается письменно и длится 60 минут (которые считаются от выдачи последнего билета). Коллоквиум состоит из трех частей с увеличивающейся сложностью: теоретического минимума, задач и теоретического максимума (в который включены задачи повышенной сложности). Полная стоимость коллоквиума, как всегда, составляет 10 баллов.

2. Теоретический минимум состоит из трех простых вопросов по теории, выбранных наугад из тестов, написанных на занятиях. Вопросы обязательно из трех различных тем. Каждый вопрос, на которых дан правильный ответ, приносит студенту 1 (один) балл. Таким образом, максимальное количество баллов, которое можно набрать в ходе решения теоретического минимума – **3 (три) балла.** При этом теоретический минимум дает право приступить к решению остального коллоквиума только в том случае, если в нем допущено не более 1 (одной) ошибки. В случае, если студент при решении теорминимума ошибается дважды или трижды, он получает оценку 1 (один) за коллоквиум. При этом решенные задачи и доказанные теоремы не играют роли.

3. Задачи: две задачи, выбранные наугад из задач, выданных на дом (кроме бонусных), и из семинарских задач. Каждая из задач оценивается в 2 (два) балла. Таким образом, максимальное количество баллов, которое можно набрать в ходе решения задач – **4 (четыре) балла.** Задачи не требуют написания кода. Решение задач не влияет на получение баллов по теоретическому максимуму.

4. Теоретический максимум состоит из задач повышенной сложности (бонусных задач из домашних заданий) и теорем, доказанных на лекциях. Он содержит всего один пункт (необходимо решить задачу или сформулировать доказательство теоремы) и оценивается в **3 (три) балла**. Теоремы можно доказывать нестрого: нам важно, чтобы были сформулированы основные пункты доказательства, его техника играет второстепенную роль. Решение теормаксимума не влияет на получение баллов за решенные задачи.

Глава 2

Программа теоретического минимума

2.1 Тест 1

1. Распределение конечной суммы случайных величин $X_1, \ldots, X_n$ с произвольным дискретным распределением $g(\cdot)$ и конечными дисперсией и матожиданием будет:

   (a) нормальным;

   (b) приближенно нормальным;

   (c) асимптотически нормальным;

   (d) биномиальным.

2. Дисперсия выборочного среднего $\overline{X}_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ с ростом n

   (a) растет;

   (b) убывает как $\frac{1}{\sqrt{n}}$;

   (c) убывает как $\frac{1}{n}$.

3. Геометрическое распределение описывает

   (a) число успехов в серии из n испытаний;

   (b) число испытаний до появления первого успеха;

   (c) среднее число испытаний до появления n успехов;

   (d) количество событий, происшедших до момента t.

4. Каждый второй момент стандартного нормального распределения

   (a) нулевой;

   (b) подсчитывается рекуррентно из предыдущего момента;

   (c) конечный;

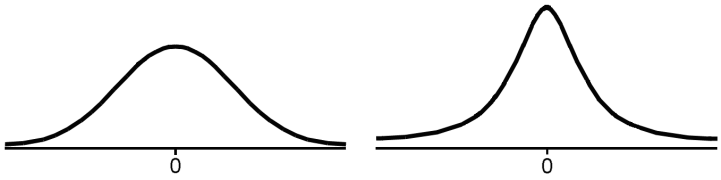
   (d) является функцией своего номера.

5. На рисунке ниже представлены графики плотностей двух распределений. Что можно сказать про их коэффициенты эксцесса?

   (a) коэффициенты эксцесса равны;

   (b) коэффициент эксцесса для плотности слева больше;

   (c) коэффициент эксцесса для плотности справа больше;



(d) коэффициент эксцесса для плотности слева не определен.

6. Пуассоновское приближение для биномиального распределения действует, когда

   (a) $n \gg 1, p \to 0$;

   (b) $n \gg 1, p \to 1$;

   (c) $n \gg 1, p \sim 1, np \to \lambda$;

   (d) $n \gg 1, p \to 0, np \to \lambda$.

7. Функция вида $f(x) = \frac{1}{\sigma_1 \sigma_2 \sqrt{2\pi}} \exp\left\{-\frac{x_1^2}{2\sigma_1^2} - \frac{x_2^2}{2\sigma_2^2}\right\}$:

   (a) задает плотность совместного распределения некоррелированных гауссовских случайных величин;

   (b) задает плотность совместного распределения коррелированных гауссовских случайных величин;

   (c) задает плотность совместного распределения некоррелированных лапласовских случайных величин;

   (d) не является плотностью.

8. «Отбеливанием» случайного вектора $\mathbf{x} \in \mathbb{R}^n$ со средним $\boldsymbol{\mu}$ и ковариационной матрицей $\boldsymbol{\Sigma}$ называется следующее преобразование

   (a) $\mathbf{z} \leftarrow \mathbf{x} - \boldsymbol{\mu}$;

   (b) $\mathbf{z} \leftarrow (\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$;

   (c) $\mathbf{z} \leftarrow \boldsymbol{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu})$;

   (d) $\mathbf{z} \leftarrow \frac{\mathbf{x} - \boldsymbol{\mu}}{\mathrm{diag}(\boldsymbol{\Sigma})}$ поэлементно.

9. Следующий код на языке `python` вычисляет некоторую величину:

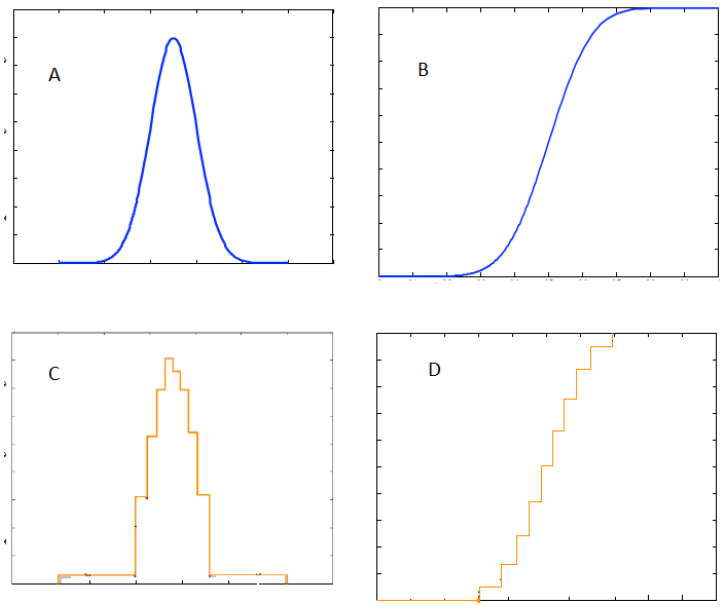
```
from numpy import random, sqrt, log, sin, cos, pi

# transformation function
def transform(u1,u2):
  z1 = sqrt(-2 * log(u1)) * cos(2 * pi * u2)
  z2 = sqrt(-2 * log(u1)) * sin(2 * pi * u2)
  return z1, z2

u1 = random.rand(1000)
u2 = random.rand(1000)

z1, z2 = transform(u1, u2)
```

   Напишите на пустом месте справа кода, какой смысл имеют переменные `z1` и `z2`.

10. Запишите на пустом месте ниже отрицательный логарифм правдоподобия (NLL) для независимой выборки $X_1, \ldots, X_n$ из распределения $\mathcal{N}(\mu_0, \sigma_0^2)$.

2.2 Тест 2

1. Алгоритм непараметрического бутстрепа включает следующие шаги:
   (a) сгенерировать одну выборку случайным образом и пересчитать новое значение статистики на ней
   (b) сгенерировать B выборок случайным образом и пересчитать новое значение статистик на каждой из них
   (c) найти параметры распределения случайной величины
2. Функция распределения дискретной случайной величины показана на картинке
   (a) A
   (b) B
   (c) C
   (d) D



3. Отличия параметрического бутстрепа от непараметрического
   (a) всегда получается одно и то же значение оценок;
   (b) улучшает оценку разброса в маленькой выборке;
   (c) зависит от выбора модели распределения;
   (d) не подходит для оценки медианы распределения.
4. Нормальный интервал для статистики T_n находится по формуле
   (a) $(T_n - z_{\alpha/2}\widehat{se}_{boot}, T_n + z_{\alpha/2}\widehat{se}_{boot})$;

(b) $(T_n - z_{\alpha/2}\sqrt{\widehat{se}_{boot}}, T_n + z_{\alpha/2}\sqrt{\widehat{se}_{boot}})$;

(c) $(2T_n - T^*_{n\,1-\alpha/2}, 2T_n - T^*_{n\,\alpha/2})$;

(d) $(T_n - z_{1-\alpha/2}\widehat{se}_{boot}, T_n + z_{1-\alpha/2}\widehat{se}_{boot})$.

5. Метод складного ножа отличается от бутстрепа тем, что

   (a) всегда получается одно и то же значение оценок;

   (b) улучшает оценку разброса в маленькой выборке;

   (c) применяется при сложных схемах семплирования;

   (d) плохо работает для порядковых статистик.

6. Предположим что для проверки гипотезы выбрано значение $p = 0.01$, какова вероятность получить не более двух ошибок первого рода при тестировании 100 таких независимых гипотез одновременно? Напишите решение ниже.

7. Одновременно тестируется m гипотез, из которых предположительно m_0 истинных. Необходимо обуспечить FWER на уровне α. Скорректированный по Бонферрони уровень α выбирается как:

   (a) $\alpha * m$;

   (b) α/m;

   (c) $\alpha * m_0$;

   (d) α/m_0.

8. Необходимые условия для применения пермутационного теста

   (a) независимость проверяемых гипотез;

   (b) стратифицированность выборки;

   (c) возможность замены тестируемой гипотезы на гипотезу об эквивалентности выборок;

   (d) большой размер выборки.

9. Бэггинг с числом суррогатных выборок B позволяет:

   (a) снизить ошибку модели в B раз;

   (b) снизить разброс модели в B раз;

   (c) снизить смещение модели в B раз;

   (d) убрать корреляцию между ошибками модели.

10. Опишите ниже, каким образом в пермутационном тесте учитывается зависимость в данных?

2.3 Тест 3

1. При решении задачи параметрического оценивания необходимо:

   (a) оценить параметры нормального распределения;

   (b) параметризовать оценку максимального правдоподобия;

   (c) построить оценку некоторой функции параметрического семейства.

2. Пусть задана выборка $X_1, \ldots, X_n \sim F$, при этом у распределения имеется плотность $f(x;\theta)$. Функция правдоподобия задается формулой:

   (a) $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i;\theta)$

   (b) $\mathcal{L}_n(\theta) = \sum_{i=1}^n f(X_i;\theta)$

   (c) $\mathcal{L}_n(\theta) = \log \prod_{i=1}^n f(X_i;\theta)$

   (d) $\mathcal{L}_n(\theta) = \prod_{i=1}^n \log f(X_i;\theta)$

3. Перечислите свойства оценки максимального правдоподобия:

   (a) равна 0 в правильном значении параметра;

   (b) не зависит от параметризации;

   (c) асимпотически нормальна;

   (d) всегда дискретна.

4. Информация Фишера, $I(\theta)$, для плотности $f(x;\theta)$ равна:

   (a) $-\mathbb{E}\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right)$

   (b) $-\mathbb{E}\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)$;

   (c) $\mathbb{E}\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right)$;

   (d) $\mathbb{E}\left(\frac{\partial^2 \log f(X;\theta)}{\partial X^2}\right)$

5. Пусть задано семейство экспоненциальных распределений в виде $f(x;\theta) = h(x)e^{\eta(\theta)T(x)-B(\theta)}$ и его log-partition функция $A(\eta)$. Чему равна вариация величины $T(X)$:

   (a) $A''(\eta)$;

   (b) $A(\eta)|_{\eta=\theta}$;

   (c) $A'(\eta)$;

   (d) $h'(x)$.

6. Покажите, что нормальное распределение относится к экспоненциальному семейству функций.

7. X - случайная величина, с ненулевым матожиданием μ. Оцените матожидание для функции $g(\theta) = \theta^2$.

   (a) μ^2;

(b) 2μ;

(c) $1/\mu$;

(d) μ^3.

8. Для распределения Пуассона с параметром λ найдите значение дисперсии для оценки $\log\left(\widehat{\lambda}\right)$

(a) $1/\lambda$

(b) λ

(c) λ^3

(d) $\log\lambda$

9. Пусть Y_n – последовательность случайных величин для которых $\sqrt{n}[Y_n - \theta] \to \mathcal{N}(0, \sigma^2)$ по распределению. К чему для функции $g(x) = \cos(x)$ будет стремиться $\sqrt{n}[g(Y_n) - g(\theta)]$?

(a) $\mathcal{N}(0, \sigma^2[g'(\theta)]^2)$

(b) $\sigma^2\frac{g''(\theta)}{2}\chi_1^2$

(c) 0

(d) $\sigma^2\frac{g''(\theta)}{2}\chi_2^2$

10. Пусть есть две случайные величины, X и Y, с ненулевыми средними. Оцените матожидание для функции $g(\theta_X, \theta_y) = \theta_x^2\theta_y^2$. ‘

2.4 Тест 4

1. В определении f-дивергенции, $D_f(P \parallel Q) \equiv \int_\Omega f\left(\frac{dP}{dQ}\right) dQ$, используется функци f, выберите все свойства функции f из списка:

(a) f выпуклая;

(b) $f(1) = 0$;

(c) $f(0) = \infty$;

(d) $f(1) = 1$

2. Для рассстояния Хеллингера f задаётся выражением:

(a) $f(t) = t\log t$

(b) $f(t) = t^2 - 1$

(c) $f(t) = (\sqrt{t} - 1)^2$

(d) $f(t) = 1/t$

3. Определите расстояние полной вариации согласно теореме Шеффе:

(a) $D(p_\xi, p_\eta) = \frac{1}{2}\int_{\mathbb{R}^n} |p_\xi(x) - p_\eta(x)|dx$;

(b) $D(p_\xi, p_\eta) = \sup_{\mathrm{A}} \left|\int_{\mathrm{A}} \mathrm{f_X(x)dx} - \int_{\mathrm{A}} \mathrm{g_Y(y)dy}\right|$

(c) $D(p_\xi, p_\eta) = \inf_A \frac{1}{2}\int_{\mathbb{R}^n}(p_\xi(x) - p_\eta(x))^2 dx$;

(d) $D(p_\xi, p_\eta) = \int_{\mathbb{R}^n} \frac{1}{\int_{\mathbb{R}^n}|p_\xi(x) - p_\eta(x)|}dx$.

4. Расставьте правильно знаки неравенства между известными расстояниями: полной вариации, D, Кульбака-Лейблера, KL, хи-квадрат, χ^2.

5. Отметьте несимметричные расстояния:

   (a) Кульбака-Лейблера;

   (b) χ^2;

   (c) полной вариации;

   (d) Йенсена-Шеннона.

6. Минимизация расстояния Кульбака-Лейблера между эмпирическим и модельным распределениями эквивалентна:

   (a) уточнению эмпирического распределения;

   (b) переводу модельного распределения в экспоненциальную форму;

   (c) минимизации необходимого количества параметров в параметрическом модельном распределении;

   (d) максимизации оценки максимального правдоподобия для модельного распределения.

7. Какая связь между дивергенциями Йенсена-Шеннона (ЙШ) и Кульбака-Лейблера (КЛ)?

   (a) ЙШ – максимум КЛ на всех возможнымх множествах;

   (b) ЙШ – минимум КЛ на всех возможных множествах;

   (c) ЙШ – симметризованная КЛ;

   (d) КЛ равна 0, если ЙШ больше 2.

8. Что возвратит вызов `A(pk,qk)`:

```
import numpy as np

def entropy_multi(p, q):
  return np.sum(p * np.log(p / q), axis=0)

def entropy_single(p):
    return np.sum(p * np.log(p), axis=0)

def A(pk, qk):
    # arraynise
    pk = np.asarray(pk)
    # normalise
    pk = 1.0*pk / np.sum(pk, axis=0)
    # check to decide if we apply single or multi entorpy
    if qk is None:
        return np.sum(entropy_single(pk), axis=0)
    else:
        # arraynise
        qk = np.asarray(qk)
        if len(qk) != len(pk):
        raise ValueError("qk and pk must have same length.")
        qk = 1.0*qk / np.sum(qk, axis=0)
    return np.sum(entropy_multi(pk, qk), axis=0)
```

(a) дивергенцию Кульбака-Лейблера между pk и qk;

(b) обратную дивергенцию Кульбака-Лейблера между pk и qk;

(c) энтропию двух распределений;

(d) оценку максимальное правдоподобия для параметра qk.

9. В отличие от расстояния полной вариации, расстояние Васерштейна:

(a) учитывает расстояние, на котором находятся отличия в распределениях;

(b) не равна 0 в случае одинаковых распределений;

(c) не является метрикой на пространстве плотностей вероятности;

(d) учитывает только ненулевые носители распредлений.

10. Даны два распределения на вещественной прямой, заданных следующим образом: $f(x) = \delta(x)$, $g(x;\theta) = \delta(x-\theta)$. Качественно нарисуйте, как меняются метрики Йенсена-Шеннона и Васерштейна при изменении θ от -1 до 1.

2.5 Тест 5

1. Размер критерия для простой гипотезы — это то же самое, что и

(a) средняя задержка в обнаружении разладки;

(b) вероятность ошибки первого рода;

(c) вероятность ошибки второго рода;

(d) уровень значимости.

2. Функция мощности критерия описывает

(a) мощность критического множества Ω;

(b) вероятность отклонить нулевую гипотезу в случае, когда она неверна;

(c) вероятность отклонить нулевую гипотезу в случае, когда она верна;

(d) объем выборки, необходимый для отклонения нулевой гипотезы, когда она неверна.

3. Какие из приведенных ниже критериев, используемых для различения двух гипотез, опираются на предположения нормальности наблюдений?

(a) критерий Неймана-Пирсона;

(b) критерий Вальда;

(c) критерий отношения правдоподобия;

(d) t-критерий Стьюдента.

4. При тестировании гипотезы $\mathbb{H}_0 : \theta = \theta_0$ о значении среднего для нормальной модели $\mathcal{N}(\theta, 1)$ против альтернативы $\mathbb{H}_1 : \theta = \theta_1 > \theta_0$ достаточно вычислить следующую статистику выборки:

(a) арифметическое среднее;

(b) медиану;

(c) выборочную дисперсию;

(d) квантиль.

5. Следующий код на языке `python` вычисляет некоторую величину:

```
import numpy as np
from scipy.stats import norm
def quant(n, mu=0, sigma=1., alpha=.05):
    x_alpha = norm.ppf(1 - alpha)
    return mu + sigma * x_alpha / np.sqrt(n)

X = np.loadtxt('data.txt')
t_alpha = quant(n=len(X))
is_exceeds = np.mean(X) >= t_alpha
```

Объясните смысл величины `is_exceeds`, которая здесь вычисляется.

6. В критерии Вальда при увеличении объема выборки при фиксированном уровне ошибки 1-го рода вероятность ошибки 2-го рода

(a) остается неизменной;

(b) увеличивается;

(c) уменьшается.

7. Почему мы всегда говорим о том, чтобы «отклонить нулевую гипотезу», и никогда о том, чтобы ее «принять»?

(a) Согласно принципу фальсифицируемости Поппера, мы никогда не можем доказать истинность теорий (гипотез), но можем опровергнуть их.

(b) Нулевую гипотезу невозможно принять, т.к. мы можем лишь проверять согласие или противоречие гипотезы и экспериментальных данных. Даже если данные не противоречат нулевой гипотезе, мы можем лишь ее «не отклонить».

(c) Потому что критерий Поппера устроен таким образом, что тестируется всегда только одна гипотеза – нулевая.

(d) Пирсон и Фишер, разработчики теории проверки гипотез, использовали такую терминологию, и она сохранилась до наших дней.

8. Ниже приведен фрагмент из работы Huth, Alexander G., et al. "Natural speech reveals the semantic maps that tile human cerebral cortex."Nature 532.7600 (2016): 453.

“Overall, right LPC responds more than left LPC to mental, professional, temporal and locational concepts, but less than left LPC to violent and visual concepts (q(FDR) < 0.05, t-test).”

Определите, какое из утверждений, приведенных ниже, верно.

(a) В тесте Вальда с уровнем значимости 0.05 произошло отклонение нулевой гипотезы.

(b) Критерий Стьюдента с уровнем значимости 0.05 и коррекцией на множественные сравнения привел к отклонению нулевой гипотезы.

(c) z-тест с уровнем значимости 0.05 и коррекцией Бонферрони привел к неотклонению нулевой гипотезы.

(d) Критерий отношения правдоподобия с уровнем значимости 0.05 и коррекцией на множественные сравнения привел к принятию нулевой гипотезы.

9. В каких условиях применим t-критерий (критерий Стьюдента)?

   (a) Проверка гипотезы о равенстве средних двух выборок, данные нормальны, средние и дисперсии неизвестны.

   (b) Проверка гипотезы о значении параметра среднего нормального распределения, среднее и дисперсия неизвестны.

   (c) Проверка гипотезы сдвига для двух произвольных распределений, дисперсия которых известна.

   (d) Проверка нормальности выборки независимых наблюдений, средние и дисперсии неизвестны.

10. Какие из следующих гипотез о параметрах нормального $\mathcal{N}(\mu, \sigma^2)$ распределения являются *простыми?*

    (a) $\mathbb{H} : \mu = \mu_0$

    (b) $\mathbb{H} : \mu = \mu_0, \sigma > \sigma_0$

    (c) $\mathbb{H} : \mu \neq \mu_0, \sigma = \sigma_0$

    (d) $\mathbb{H} : \mu = \mu_0, \sigma = \sigma_0$

Глава 3

Задачи

3.1 Задачи из домашнего задания 1

1. (2 балла) Подсчитать вероятность, что случайно выбранный студент не знает ни Java, ни C++, если известны вероятности того, что он знает Java, C++ и и то, и другое. 28% студентов из университета программируют на Java, 7% на C++ и 5% программируют на обоих языках.

2. (2 балла) Подсчитать вероятность того, что из первых r бит переданной по сети строки, состоящей из m нулей и n единиц, k бит будут единицами. Принять расположение 0 и 1 равновероятным.

3. (2 балла) Подсчитать вероятность того, что в выборке размера k из n чипов, 1 из которых неисправный, будет этот неисправный чип.

4. (2 балла) Дано множество из n элементов. Написать программу, печатающую на экране случайное подмножество этого множества, состоящее из k элементов, причем все C_n^k подмножеств равновероятны.

5. (2 балла) Подсчитать аналитически математическое ожидание числа сравнений при сортировке n различных чисел алгоритмом QuickSort, если в исходном массиве числа находятся в случайном порядке. В ответе должна быть указана функция от n.

6. (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 θ – медиана.

   Покажите что для оценки $\widehat{\theta}$ по бутстрепной выборке верно следующее:

$$\mathsf{P}(\widehat{\theta} > x_{(i)}) = \sum_{j=0}^{5} \mathrm{Bin}\left(j, n, \frac{i}{n}\right),$$

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

7. (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 θ – медиана.

   Покажите, что оценка $\widehat{\theta}$ по бутстрепной выборке равна $x_{(i)}$ с вероятностью:

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

3.2 Задачи из домашнего задания 2

1. (2 балл) Наблюдаемый самолёт характеризуется расстоянием до наблюдателя, r, и углом наблюдения, θ. Пусть есть m измерений R и Θ, найдите вариацию высоты самолёта, вычисляемую по формуле $Y = R\sin\Theta$. Если R фиксирована, когда достигается максимальная вариация Y?

2. (2 балла) Для случайной величины $X \sim \mathcal{N}(\mu, \sigma^2)$ и функции $g(x) = \exp(x)/(1+\exp(x))$ получите выражение для среднего и дисперсии $g(x)$ с помощью дельта-метода.

3. (2 балла) Для случайной величины $X \sim \mathcal{N}(\mu, \sigma^2)$ и функции $g(x) = \exp(x)/(1+\exp(x))$ получите выражение для среднего и дисперсии $g(x)$ с помощью бутстрапа (опишите схему их получения).

4. (2 балла) У Вас есть две монеты, P и Q, одна из которых имеет дефект. Из-за этого дефекта вероятность выпадения орла в монете P выше на 2ϵ, чем вероятность выпадения решки. Для второй монеты Q эти вероятности равны 0.5. Вы можете подбрасывать монету и смотреть на результат. Существует алгоритм $A(x_1, .., x_m) \to \{0; 1\}$, который говорит, является ли монета деффектной ($A = 0$) или настоящей ($A = 1$) на основании m независимых подбрасываний. С помощью неравенства Пинскера и свойств дивергенции Кульбака-Лейблера найдите минимальное m, для которого A сможет получить ответ с вероятностью больше 90% $P_{x\in Q}(A(x) = 1) > 0.9$, $P_{x\in P}(A(x) = 1) > 0.9$. Докажите, что для любых распределений $\tilde{P}$ и $\tilde{Q}$ над U и функции $f(x) : U \to [0; B]$ выполнено:

$$|\mathbb{E}_{\tilde{P}}[f(x)] - \mathbb{E}_{\tilde{Q}}[f(x)]| \leq B||\tilde{P} - \tilde{Q}||,$$

   где $||\cdot||$ – расстояние полной вариации. Для доказательства можно использовать дискретные распределения.

5. (2 балла) У Вас есть две монеты, P и Q, одна из которых имеет дефект. Из-за этого дефекта вероятность выпадения орла в монете P выше на 2ϵ, чем вероятность выпадения решки. Для второй монеты Q эти вероятности равны 0.5. Вы можете подбрасывать монету и смотреть на результат. Существует алгоритм $A(x_1, .., x_m) \to \{0; 1\}$, который говорит, является ли монета деффектной ($A = 0$) или настоящей ($A = 1$) на основании m независимых подбрасываний. С помощью неравенства Пинскера и свойств дивергенции Кульбака-Лейблера найдите минимальное m, для которого A сможет получить ответ с вероятностью больше 90% $P_{x\in Q}(A(x) = 1) > 0.9$, $P_{x\in P}(A(x) = 1) > 0.9$. Найдите верхнюю границу $KL(P||Q)$ с точностью до ϵ^2, считая $\epsilon < 0.25$.

6. (2 балла) У Вас есть две монеты, P и Q, одна из которых имеет дефект. Из-за этого дефекта вероятность выпадения орла в монете P выше на 2ϵ, чем вероятность выпадения решки. Для второй монеты Q эти вероятности равны 0.5. Вы можете подбрасывать монету и смотреть на результат. Существует алгоритм $A(x_1, .., x_m) \to \{0; 1\}$, который говорит, является ли монета деффектной ($A = 0$) или настоящей ($A = 1$) на основании m независимых подбрасываний. С помощью неравенства Пинскера и свойств дивергенции Кульбака-Лейблера найдите минимальное m, для которого A сможет получить ответ с вероятностью больше 90% $P_{x\in Q}(A(x) = 1) > 0.9$, $P_{x\in P}(A(x) = 1) > 0.9$. Используя свойства КЛ-дивергенции для большого количества семплов и неравенство Пинскера, найдите нижнюю оценку на m через ϵ с точностью до $1/\epsilon^2$

3.3 Задача с семинара 6

1. (2 балла) Построить критерий для проверки гипотезы $H_0 : p = 1/2$ при альтернативной гипотезе $H_a : p \neq 1/2$ по результатам восьми испытаний, подчиняющихся схеме Бернулли. Вероятность ошибки первого рода α положить равной 0.05.

Глава 4

Программа теоретического максимума

4.1 Теоремы

Теорема 1. *Пусть $\theta = T(F)$ и $\hat{\theta}_n = T(\hat{F}_n)$. Далее, пусть $\hat{\theta}^*_{n,1}, \ldots, \hat{\theta}^*_{n,B}$ — получены итерированием шагов 1 и 2 алгоритма бутстрепа. Пусть θ^*_β — обозначает β - квантиль для $(\theta^*_{n,1}, \ldots, \theta^*_{n,B})$. Тогда при некоторых несильных условия на $T(F)$,*

$$\mathsf{P}_F(T(F) \in C_n) \to 1 - \alpha, n \to \infty,$$

$$C_n = (2\hat{\theta}_n - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta}_n - \hat{\theta}^*_{\alpha/2})$$

Теорема 2 (Дельта-метод)**.** *Пусть Y_n – последовательность случайных величин для которых $\sqrt{n}[Y_n - \theta] \to \mathcal{N}(0, \sigma^2)$ по распределению. Тогда для дифференциируемой в θ функции $g(.)$ с ненулевой производной, $\sqrt{n}[g(Y_n) - g(\theta)] \to \mathcal{N}(0, \sigma^2[g'(\theta)]^2)$ по распределению.*

Теорема 3 (Многопараметрический дельта-метод)**.** *Пусть $\mathbf{X_1}, .., \mathbf{X_n}$ выборка случайных векторов, причём $EX_{ij} = \mu_i$ и $\mathrm{Cov}(X_{ik}, X_{jk}) = \sigma_{ij}$. Для функции g с непрывными первыми производными и значения μ, для которого $\tau^2 = \sum\sum \sigma_{ij} \frac{\partial g(\mu)}{\partial \mu_i} \frac{\partial g(\mu)}{\partial \mu_j} > 0$ выполняется:*

$$\sqrt{n}[g(\bar{\mathbf{X}}_\mathbf{1}, .., \bar{\mathbf{X}}_\mathbf{s}) - \mathbf{g}(\mu)] \to \mathcal{N}(\mathbf{0}, \boldsymbol{\tau}^\mathbf{2})$$

Теорема 4 (Шеффе)**.** *Если существуют плотности распределения $p_\xi(x)$, $p_\eta(x)$, $x \in \mathbb{R}^n$ случайных величин ξ и η, то:*

$$D(p_\xi, p_\eta) = \frac{1}{2} \int\limits_{\mathbb{R}^n} |p_\xi(x) - p_\eta(x)| dx,$$

Теорема 5 (Неравенство Пинскера)**.** *Если случайные величины X и Y имеют плотности $p_\xi(x)$ и $p_\eta(x)$, где $x \in \mathbb{R}^n$, то*

$$KL(p_\xi, p_\eta) \geq 2 \left[D(p_\xi, p_\eta)\right]^2.$$

Теорема 6 (КЛ-дивергенция для нескольких случайных величин)**.** *Пусть $X^n = X_1, ..X_n$ и $Y^n = Y_1, ..Y_n$ – случайные векторы, $X_1, ..X_n \sim p_X$, $Y_1, ..., Y_n \sim p_Y$. Тогда*

$$KL(p_{X^n}, p_{Y^n}) = nKL(p_X, p_Y)$$

Теорема 7. *Пусть θ_* — реальное значение параметра θ. Обозначим через*

$$M_n(\theta) = \frac{1}{n} \sum_i \log \frac{f(X_i; \theta)}{f(X_i; \theta_*)}$$

и $M(\theta) = -KL(\theta_*, \theta)$.

Допустим, что $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathsf{P}} 0$ и для каждого $\epsilon > 0$ $\sup_{\theta : |\theta - \theta_*| \geq \epsilon} M(\theta) < M(\theta_*)$.

Пусть $\widehat{\theta}_n$ обозначает ОМП, тогда $\widehat{\theta}_n \xrightarrow{\mathsf{P}} \theta_*$.

Теорема 8 (Лемма Неймана-Пирсона)**.** *Наиболее мощный критерий уровня α задается критическим множеством*

$$G^* = G_{c_\alpha} = \left\{ \boldsymbol{x} \in \mathbb{R}^\ell : \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} \geqslant c_\alpha \right\}$$

Теорема 9 (Критерий Стьюдента (t-test))**.** *Пусть $X_1, ..., X_n \sim N(\mu, \sigma^2)$, где параметры (μ, σ^2) неизвестны.*

$$\mathbb{H}_0 : \mu = \mu_0 \quad vs. \quad \mathbb{H}_1 : \mu \neq \mu_0$$

Обозначим через S_n^2 выборочную дисперсию. Тогда статистика критерия:

$$T = \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{S_n}$$

Основная гипотеза отвергается, если $|T| > t_{n-1,\alpha/2}$, где $t_{n-1,\alpha/2}-$ квантиль распределения Стьюдента с $n-1$ степенями свободы.

Теорема 10 (bias-variance разложение MSE и MISE)**.**

$$MSE(\hat{p}_n, p, x_0) = bias^2(x_0) + \mathrm{Var_p}\hat{\mathrm{p}}_\mathrm{n}(\mathrm{x}_0) =$$
$$= [\mathbb{E}_p \hat{p}_n(x_0) - p(x_0)]^2 + \mathbb{E}_p[\hat{p}_n(x_0) - \mathbb{E}_p \hat{p}_n(x_0)]^2$$

$$MISE(\hat{p}_n, p) = \int\limits_{\mathbb{R}} bias^2(x)dx + \int\limits_{\mathbb{R}} \mathrm{Var_p}\hat{\mathrm{p}}_\mathrm{n}(\mathrm{x})\mathrm{dx}$$

Теорема 11 (Приближенная несмещённость оценки риска с помощью кросс-валидации)**.** *Пусть*

$$\mathcal{J}(h) = \int\limits_{\mathbb{R}} \hat{p}_n(x)^2 dx - 2 \int\limits_{\mathbb{R}} \hat{p}_n(x) p(x) dx.$$

Оценка риска с помощью кросс-валидации:

$$\hat{\mathcal{J}}(h) = \int\limits_{\mathbb{R}} [\hat{p}_n(x)]^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{p}_{(-i)}(X_i),$$

где $\hat{p}_{(-i)}$ - оценка гистограммы по выборке без $i-$ого наблюдения. Тогда $\mathbb{E}\hat{\mathcal{J}}(h) \approx \mathbb{E}\mathcal{J}(h)$

Теорема 12. *Пусть $M = M(n)$ - число ячеек в гистограмме $\hat{p}_n$, причем $M(n) \to \infty$ и $\frac{M(n)\log(n)}{n} \to \infty$ при $n \to \infty$.*

Определим

$$p_-(x) = (\max\{\sqrt{\hat{p}_n(x)} - C, 0\})^2, \quad p_+(x) = (\sqrt{(x)} + C)^2, \text{ где } C = \frac{1}{2} z_{\frac{\alpha}{2M}} \sqrt{\frac{M}{n(b-a)}}$$

Тогда $(p_-(x), p_+(x))$ является $1-\alpha$ доверительным интервалом.

Теорема 13.

$$MISE(\hat{p}_n, p) \approx \frac{1}{4}\sigma_K^4 h^4 \int\limits_{\mathbb{R}} (p''(x))^2 dx + \frac{1}{nh} \int\limits_{\mathbb{R}} (K(x))^2 dx$$

Минимум достигается при $h = h^*$:

$$h^* = \left(\frac{1}{n} \frac{\int\limits_{\mathbb{R}} (K(x))^2 dx}{\left(\int\limits_{\mathbb{R}} x^2 K(x) dx \right)^2 \left(\int\limits_{\mathbb{R}} p''(x))^2 dx \right)} \right)^{\frac{1}{5}}$$

При этом $MISE(\hat{p}_n, p) = O\left(n^{-\frac{4}{5}}\right)$

4.2 Бонусные задачи из домашнего задания 1

1. (3 балла) Пусть $T_n = \overline{X}_n^2$, $\mu = (X_1)$, $\alpha_k = \int |x-\mu|^k dF(x)$ и $\widehat{\alpha}_k = \sum\limits_{i=1}^{n} \left|X_i - \overline{X}_n\right|^k$. Докажите, что матожидание оценки дисперсии функционала T_n с помощью бутстрапа(т.е. матожидание по эмпирической функции распределения) равна:

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

2. (3 балла) Доказать эффективность бэггинга можно на следующем игрушечном примере. Рассмотрим задачу классификации и предиктор («решающий пень», т.е. решающее дерево глубины 1) вида:

$$\widehat{\theta}_n(x) = \mathbf{1}_{[\widehat{d}_n \leqslant x]}, \quad x \in \mathbb{R}.$$

Здесь $\widehat{d}_n$ – действительное число, оцененное по выборке $\mathbf{X}^\ell = \{Y_i, X_i\}_{i=1}^\ell$. Пусть оценка $\widehat{d}_n$ асимптотически нормальна, причем скорость сходимости к нормальному распределению у нее b_n^{-1}, т.е.

$$b_n(\widehat{d}_n - d_0) \to_D \mathcal{N}(0, \sigma_\infty^2),$$

где σ_∞^2 – ее асимптотическая дисперсия.

Рассмотрим некоторый x в b_n^{-1}-окрестности параметра d_0, т.е. $x = x_n(c) = d_0 + c\sigma_\infty b_n^{-1}$.

Подсчитайте:

- асимптотические математическое ожидание и дисперсию классификатора $\widehat{\theta}_n(x)$ для таких x;
- асимптотические математическое ожидание и дисперсию бэггинг-классификатора $\widehat{\theta}_{n;B}(x) = \frac{1}{J}\sum\limits_{j=1}^{J}\widehat{\theta}_{n;(j)}(x)$ для таких x.

Асимптотики рассматривать при $n \to \infty$.

4.3 Бонусные задачи из домашнего задания 2

1. (3 балла) Найдите выражение для дивергенции Кульбака-Лейблера между двумя нормальными распределениями $\mathcal{N}(\mu_1; \sigma_1)$ и $\mathcal{N}(\mu_2; \sigma_2)$.

2. (3 балла) (По стопам семинара про натуральный градиент) В семинаре матрица Фишера считалась эмпирически по выборке. Этот подход имеет преимущество в случаях, когда плотность распределения неизвестна или сложно вычислима, но посчитать градиент достаточно просто (к примеру, когда распределение порождается нейронной сетью). Альтернативно, если плотность распределения известна, то матрица Фишера вычислима аналитически.

   Вывести аналитическое выражение матрицы Фишера для вектора параметров: $\theta = (\mu_1, \mu_2, \sigma_{11}, \rho, \sigma_{22})$.