Школа анализа данных
Машинное обучение, часть 1
Теоретическое домашнее задание №2

Решите предложенные задачи. Решения необходимо оформить в виде PDF документа. Каждая задача должна быть подробно обоснована, задачи без обоснования не засчитываются. Решения пишутся в свободной форме, однако так, чтобы проверяющие смогли разобраться. Если проверяющие не смогут разобраться в решении какой-нибудь задачи, то она автоматически не засчитывается.

Задача 1 (1 балл) Метрические методы, kNN, устойчивость к шуму.

Известно, что метод ближайших соседей неустойчив к шуму. Рассмотрим модельную задачу бинарной классификации с одним признаком и двумя объектами обучающей выборки: $x_1 = 0.1$, $x_2 = 0.5$. Первый объект относится к первому классу, второй — ко второму. Добавим к объектам новый шумовой признак, распределенный равномерно на отрезке $[0, 1]$. Теперь каждый объект описывается уже двумя признаками. Пусть требуется классифицировать новый объект $u = (0, 0)$ в этом пространстве методом одного ближайшего соседа с евклидовой метрикой. Какова вероятность того, что после добавления шума второй объект окажется ближе к объекту u, чем первый?

Решение: Рассмотрим евклидово расстояние между объектами $a, b : \rho(a, b) = (a_1 - b_1)^2 + (a_2 - b_2)^2$ и учтем, что шумовые значения $\xi \in [0, 1]$, так как они распределены равномерно на отрезке [0,1]. Тогда искомая вероятность вычисляется следующим образом:

$$P(\rho(x_2, u) < \rho(x_1, u)) = P(x_2^2 + {\xi_2}^2 < x_1^2 + {\xi_1}^2) = P({\xi_2}^2 < {\xi_1}^2 - 0.24) = \int\limits_0^{\sqrt{0.76}} \Bigg(\int\limits_{\sqrt{x_2^2+0.24}}^{1} dx_1 \Bigg) dx_2 =$$

$$= \int\limits_0^{\sqrt{0.76}} (1 - \sqrt{x_2^2 + 0.24}) dx_2 = x_2 \Big|_0^{\sqrt{0.76}} - \frac{x_2}{2}\sqrt{0.24 + x^2}\Big|_0^{\sqrt{0.76}} - 0.12 * ln\Bigg(x_2 + \sqrt{0.24 + x_2^2}\Bigg)\Big|_0^{\sqrt{0.76}} \approx 0.275$$

Вычисление интеграла приведено ниже:

$$\int\limits_0^{\sqrt{0.76}} (1 - \sqrt{x^2 + 0.24}) dx = \int\limits_0^{\sqrt{0.76}} dx - \int\limits_0^{\sqrt{0.76}} \sqrt{x^2 + 0.24} dx$$

1. $\int\limits_0^{\sqrt{0.76}} dx = x\Big|_0^{\sqrt{0.76}}$

2. Чтобы посчитать $\int\limits_0^{\sqrt{0.76}} \sqrt{x^2 + 0.24}dx$, сделаем замену переменной: $x = \sqrt{0.24}sh(t)$, $dx = \sqrt{0.24}ch(t)dt$:

$$\int\limits_0^{\sqrt{0.76}} \sqrt{x^2 + 0.24}dx = 0.24 \int\limits_0^{\sqrt{0.76}} ch^2(t)dt = \frac{0.24}{4} \int\limits_0^{\sqrt{0.76}} (e^t + e^{-t})^2 dt = \frac{0.24}{4} \int\limits_0^{\sqrt{0.76}} (e^{2t} + 2 - e^{-2t}) dt =$$

$$= \frac{0.24}{4}(\frac{e^{2t}}{2} + 2t - \frac{e^{-2t}}{2})\Big|_0^{\sqrt{0.76}} = \frac{0.24}{4}(sh(2t) + 2t)\Big|_0^{\sqrt{0.76}} \ (1)$$

Из замены получаем:

$$\frac{x}{\sqrt{0.24}} = sh(t)$$

$$\frac{x}{\sqrt{0.24}} = \frac{e^t - e^{-t}}{2}$$

$$e^{2t} - \frac{2x}{\sqrt{0.24}}e^t - 1 = 0$$

$$e^t = \frac{x}{\sqrt{0.24}} \pm \sqrt{\left(\frac{x}{\sqrt{0.24}}\right)^2 + 1}$$

Решение $e^t = \frac{x}{\sqrt{0.24}} - \sqrt{\left(\frac{x}{\sqrt{0.24}}\right)^2 + 1}$ не подходит, т.к не имеет решений (правая часть меньше 0). Получаем, что:

$$t = ln\left(\frac{x}{\sqrt{0.24}} + \sqrt{\left(\frac{x}{\sqrt{0.24}}\right)^2 + 1}\right) \ (2)$$

$$sh(2t) = 2sh(t)ch(t) = 2sh(t)\sqrt{1 + sh^2(t)} = \frac{2x}{\sqrt{0.24}}\sqrt{1 + \left(\frac{x}{\sqrt{0.24}}\right)^2} \ (3)$$

Подставим (2) и (3) в (1):

$$\frac{0.24}{4}\left(\frac{2x}{\sqrt{0.24}}\sqrt{1 + \left(\frac{x}{\sqrt{0.24}}\right)^2} + 2ln\left(\frac{x}{\sqrt{0.24}} + \sqrt{\left(\frac{x}{\sqrt{0.24}}\right)^2 + 1}\right)\right)\Bigg|_0^{\sqrt{0.76}} = \frac{x}{2}\sqrt{0.24 + x^2}\Big|_0^{\sqrt{0.76}} +$$

$$+0.12 * ln\left(x + \sqrt{0.24 + x^2}\right)\Big|_0^{\sqrt{0.76}}$$

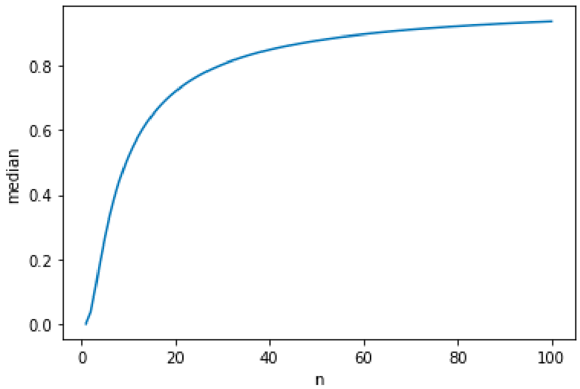
Ответ: ≈ 0.275

Задача 2 (1 балл) Метрические методы, kNN, проклятие размерности.

Рассмотрим l точек, распределенных равномерно по объему n-мерного единичного шара с центром в нуле. Предположим, что мы хотим применить метод ближайшего соседа для точки начала координат. Зададимся вопросом, на каком расстоянии будет расположен ближайший объект. Для ответа на этот вопрос выведите выражение для **медианы** расстояния от начала координат до ближайшего объекта. Чтобы проинтерпретировать полученный результат, подставьте в формулу конкретные значения: $l = 500$ и $n = 10$. Покажите, как будет меняться значение медианы при дальнейшем увеличении размерности пространства при фиксированном количестве точек и постройте график этой зависимости. Поясните, в чем состоит проклятие размерности и почему полученная для медианы формула наглядно его демонстрирует. Для размерности n посчитайте, сколько точек $l = f(n)$ необходимо взять, чтобы побороть проклятье размерности.

Решение: Объем n-мерного единичного шара равен $c * 1^n$, где c-некоторая константа. Пусть r-радиус внутреннего шара с центром в начале координат. Тогда вероятность того, что все точки выборки лежат вне внутреннего шара равна $p = \left(\frac{c*1^n - c*r^n}{c*1^n}\right)^l = (1 - r^n)^l \Rightarrow r = (1 - p^{\frac{1}{l}})^{\frac{1}{n}}$. Тогда медиана расстояния от начала координат до ближайшего объекта равна $r = (1 - (\frac{1}{2})^{\frac{1}{l}})^{\frac{1}{n}}$.

При $l = 500$ и $n = 10$: $r = 0.51779$. Таким образом, можно сказать, что ближайший сосед будет находится примерно в середине шара.



На графике видно, что при увеличении размерности медиана расстояния от начала координат до ближайшего соседа будет приближаться к 1, это означает, что выборка концентрируется возле границы шара.

Проклятие размерности связано с экспоненциальным возрастанием количества данных из-за увеличения размерности пространства. В метрических классификаторах вычисляется расстояние между объектами. По графику видно, что точки концентрируются возле границы, то есть расстояние до них будет одинаковым, что в свою очередь ведет к неинформативности расстояния. Но об этом и говорит проклятие размерности: чтобы сохранить информативность, нужно увеличивать размер выборки (то есть с увеличением размерности пространства возрастает количество данных).

$l = f(n) = -\frac{1}{log_2(1-r^n)} = -\frac{1}{log_2(1-(\frac{1}{2})^n)}$ элементов нужно, чтобы побороть проклятие размерности.

Задача 3 (1.5 балла). Метод максимального правдоподобия, равномерное распределение.

Найдите оценку максимального правдоподобия для параметра θ равномерного непрерывного распределения, нарисуйте пример выборки и получившуюся оценку в каждом из случаев:

1. (0.5 балла) $U[a, b], \theta = (a, b)$

   Решение:

   $L = \prod\limits_{i=1}^{n} \frac{1}{b-a} I[a \leq X_i \leq b] = \frac{1}{(b-a)^n} I[a \leq X_{(1)} \& X_{(n)} \leq b] \to max$, поэтому $\hat{a} = X_{(1)}$ и $\hat{b} = X_{(n)}$

   Ответ: $\hat{a} = X_{(1)}$ и $\hat{b} = X_{(n)}$

2. (0.5 балла) $U[-\theta, \theta]$

   Решение:

   $L = \prod\limits_{i=1}^{n} \frac{1}{\theta+\theta} I[-\theta \leq X_i \leq \theta] = \frac{1}{(2\theta)^n} I[-\theta \leq X_{(1)} \& X_{(n)} \leq \theta] \to max$, поэтому $\hat{\theta} = max\{|X_{(1)}|, |X_{(n)}|\}$

   Ответ: $\hat{\theta} = max\{|X_{(1)}|, |X_{(n)}|\}$

3. (0.5 балла) $U[\theta, \theta + 1]$

   Решение:

   $L = \prod\limits_{i=1}^{n} \frac{1}{\theta+1-\theta} I[\theta \leq X_i \leq \theta + 1] = I[\theta \leq X_{(1)} \& X_{(n)} \leq \theta + 1] = I[X_{(n)} - 1 \leq \theta \leq X_{(1)}] \to max$, поэтому $\hat{\theta}$ - любое из $[X_{(n)} - 1, X_{(1)}]$

   Ответ: $\hat{\theta}$ - любое из $[X_{(n)} - 1, X_{(1)}]$

In [1]:

```
import random

import numpy as np
import matplotlib.pyplot as plt
```

Пункт 1

Сгененрируем значения из $U[a, b]$, где a=2 и b=7, и найдем $\hat{a} = X_{(1)}$ и $\hat{b} = X_{(n)}$ - ОМП

In [2]:

```
arr = np.random.uniform(2, 7, 5)
print(arr)
print("a = {}, b = {}".format(np.sort(arr)[0], np.sort(arr)[-1]))
```

```
[3.04399075 3.57910662 4.73883461 4.97576313 4.026861  ]
a = 3.043990750771759, b = 4.975763130113691
```

А теперь попробуем увеличивать размер выборки и посмотрим, что происходит с оценками.

In [3]:

```
arr_10 = np.random.uniform(2, 7, 10)
print("Размер выборки:", arr_10.shape[0])
print("a = {}, b = {}".format(np.sort(arr_10)[0], np.sort(arr_10)[-1]))
```

```
Размер выборки: 10
a = 2.1292314864614696, b = 6.868478506884339
```

In [4]:

```
arr_100 = np.random.uniform(2, 7, 100)
print("Размер выборки:", arr_100.shape[0])
print("a = {}, b = {}".format(np.sort(arr_100)[0], np.sort(arr_100)[-1]))
```

```
Размер выборки: 100
a = 2.00041116976005, b = 6.950006858958406
```

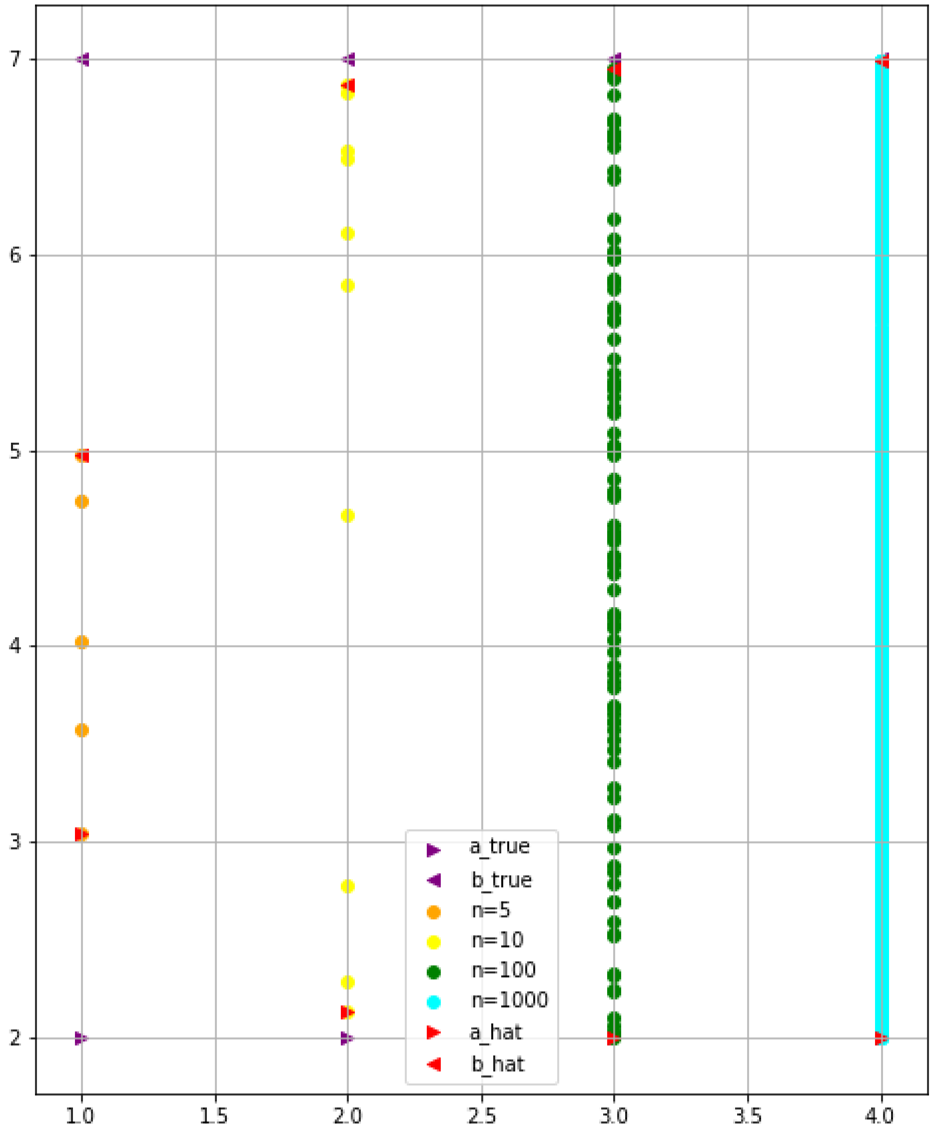
In [5]:

```
arr_1000 = np.random.uniform(2, 7, 1000)
print("Размер выборки:", arr_1000.shape[0])
print("a = {}, b = {}".format(np.sort(arr_1000)[0], np.sort(arr_1000)[-1]))
```

```
Размер выборки: 1000
a = 2.000184415678444, b = 6.9900633335742395
```

In [6]:

```
def plot_X(a, b, calc_a, calc_b):
    fig, ax = plt.subplots(1, figsize=(8, 10))

    ax.scatter(np.arange(1, 5), [a] * 4, c = "purple", marker=">", label="a_true")
    ax.scatter(np.arange(1, 5), [b] * 4, c="purple", marker="<", label="b_true")

    ax.scatter(np.ones(arr.shape), arr, label="n=" + str(arr.shape[0]), c="orange")
    ax.scatter(1, calc_a(arr), c="r", marker=">")
    ax.scatter(1, calc_b(arr), c="r", marker="<")

    ax.scatter(2 * np.ones(arr_10.shape), arr_10, label="n=" + str(arr_10.shape[0]), c="yellow")
    ax.scatter(2, calc_a(arr_10), c="r", marker=">")
    ax.scatter(2, calc_b(arr_10), c="r", marker="<")

    ax.scatter(3 * np.ones(arr_100.shape), arr_100, label="n=" + str(arr_100.shape[0]), c="green")
    ax.scatter(3, calc_a(arr_100), c="r", marker=">")
    ax.scatter(3, calc_b(arr_100), c="r", marker="<")

    ax.scatter(4 * np.ones(arr_1000.shape), arr_1000, label="n=" + str(arr_1000.shape[0]), c="aqua")
    ax.scatter(4, calc_a(arr_1000), c="r", marker=">", label="a_hat")
    ax.scatter(4, calc_b(arr_1000), c="r", marker="<", label="b_hat")

    ax.grid()
    ax.legend()

plot_X(2, 7, lambda arr: np.sort(arr)[0], lambda arr: np.sort(arr)[-1])
```



На самом деле видим, что оценки максимального правдободобия приближаются к реальным значениям с увеличением размера выборки.

Пункт 2

Сгененрируем значения из $U[-\theta, \theta]$, где $\theta = 4$ и найдем $\hat{\theta} = max(|X_{(1)}|, |X_{(n)}|)$ - ОМП

In [7]:

```
arr = np.random.uniform(-4, 4, 5)
arr
theta = max(abs(np.sort(arr)[0]), abs(np.sort(arr)[-1]))
print("theta =", theta)
```

```
theta = 3.448892660143657
```

In [8]:

```
arr_10 = np.random.uniform(-4, 4, 10)
theta_10 = max(abs(np.sort(arr_10)[0]), abs(np.sort(arr_10)[-1]))
print("Размер выборки:", arr_10.shape[0])
print("theta =", theta_10)
```

```
Размер выборки: 10
theta = 3.7553334421387508
```

In [9]:

```
arr_100 = np.random.uniform(-4, 4, 100)
theta_100 = max(abs(np.sort(arr_100)[0]), abs(np.sort(arr_100)[-1]))
print("Размер выборки:", arr_100.shape[0])
print("theta =", theta_100)
```

```
Размер выборки: 100
theta = 3.9934787202104634
```

In [10]:

```
arr_1000 = np.random.uniform(-4, 4, 1000)
theta_1000 = max(abs(np.sort(arr_1000)[0]), abs(np.sort(arr_1000)[-1]))
print("Размер выборки:", arr_1000.shape[0])
print("theta =", theta_1000)
```

```
Размер выборки: 1000
theta = 3.9983754561479383
```

In [11]:

```
arr_10000 = np.random.uniform(-4, 4, 10000)
theta_10000 = max(abs(np.sort(arr_10000)[0]), abs(np.sort(arr_10000)[-1]))
print("Размер выборки:", arr_10000.shape[0])
print("theta =", theta_10000)
```

```
Размер выборки: 10000
theta = 3.9996741815722174
```

In [12]:

```
plot_X(-4, 4,
       lambda arr: -max(abs(np.sort(arr)[0]), abs(np.sort(arr)[-1])),
       lambda arr: max(abs(np.sort(arr)[0]), abs(np.sort(arr)[-1])))
```

Пункт 3

Сгененрируем значения из $U[\theta, \theta + 1]$, где $\theta = 3$. Построим оценки $\hat{\theta}$ только для некоторых значений из отрезка $[X_{(n)} - 1, X_{(1)}]$ - ОМП

In [13]:

```
theta = 3
```

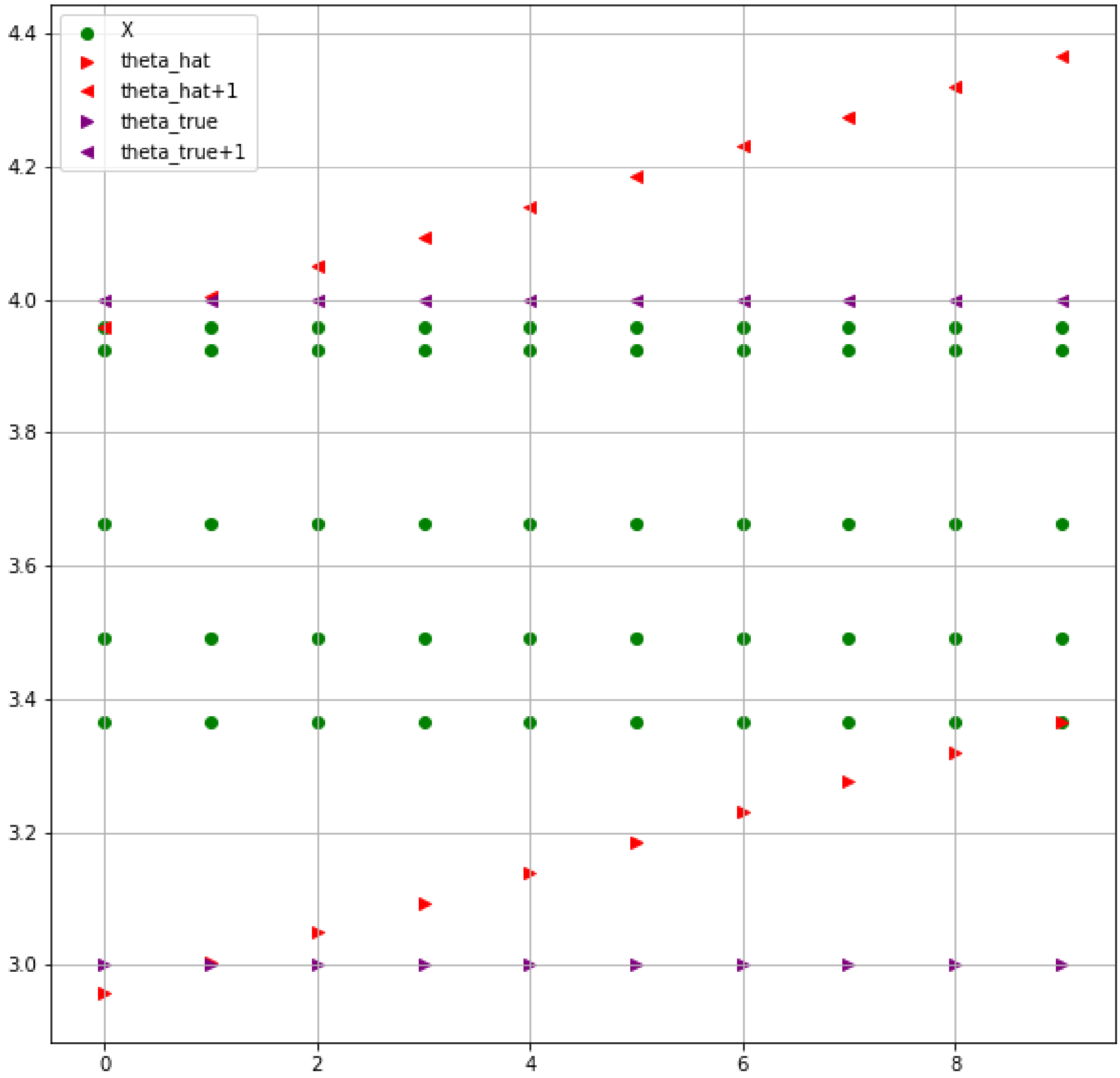
In [14]:

```
def plot_theta(arr, theta_hats):
    fig, ax = plt.subplots(1, figsize=(10, 10))

    x = np.arange(theta_hats.shape[0])
    ax.scatter(np.array(np.arange(x.shape[0]).tolist() * arr.shape[0]).reshape(a
rr.shape[0], x.shape[0]).T,
               np.array(arr.tolist() * x.shape[0]).reshape(x.shape[0], arr.shape
[0]), c = "green", label="X")
    ax.scatter(x, theta_hats, c="r", marker=">", label="theta_hat")
    ax.scatter(x, theta_hats + 1, c="r", marker="<", label="theta_hat+1")
    ax.scatter(x, [theta] * x.shape[0], c="purple", marker=">", label="theta_tru
e")
    ax.scatter(x, [theta + 1] * x.shape[0], c="purple", marker="<", label="theta
_true+1")
    ax.legend()
    ax.grid()
```

Разобьем вышеуказанный отрезок с одинаковым шагом (получили несколько оценок для θ) и для каждой $\hat{\theta}$ на графике покажем выборку, оценки $\hat{\theta}$ и $\hat{\theta} + 1$ и истинную θ

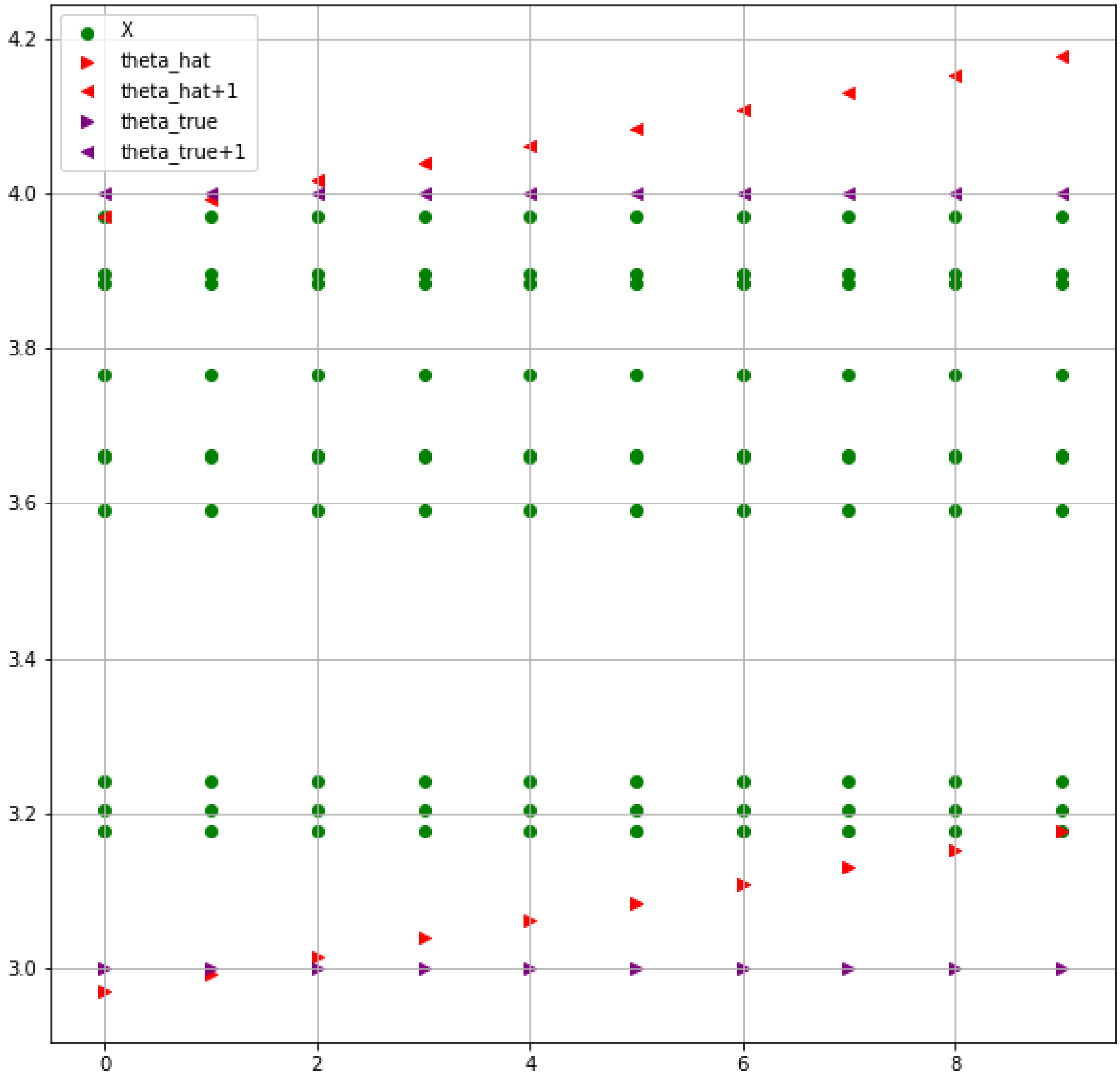
In [15]:

```
arr = np.random.uniform(theta, theta + 1, 5)
x_min = np.sort(arr)[0]
x_max = np.sort(arr)[-1]
plot_theta(arr, np.linspace(x_max - 1, x_min, 10))
```



In [16]:

```
arr = np.random.uniform(theta, theta + 1, 10)
x_min = np.sort(arr)[0]
x_max = np.sort(arr)[-1]
plot_theta(arr, np.linspace(x_max - 1, x_min, 10))
```



Задача 4 (2 балла). Метод максимального правдоподобия, случайные векторы. Пусть вектор x задан следующим образом:

$$x = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}, \quad \text{где } \theta \in [0, \pi/2] \text{ –известный параметр}$$

Зададим случайные векторы $\mathbf{Y_1}, \mathbf{Y_2} \ldots, \mathbf{Y_n}$ для каждого i как

$$\mathbf{Y_i} = \begin{bmatrix} \cos(\theta + \phi) \\ \sin(\theta + \phi) \end{bmatrix} + \mathbf{Z_i}$$

где $\mathbf{Z_i} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ – гауссовский шум с известным параметром σ, причем величины $\mathbf{Z_i}$ независимы (и одинаково распределены). Наша цель – оценить неизвестный параметр $\phi \in [-\pi, \pi]$ по значениям Y_i. Физический смысл этой задачи – выделение фазового сигнала в условиях шума. Похожая задача в более сложном виде решается, например, в Wi-Fi адаптерах в режиме реального времени.

1. (1 балл) Найдите логарифм максимального правдоподобия $l_n(\mathbf{Y_1}, \ldots, \mathbf{Y_n}; \phi)$.

   Решение:

   Правдоподобие:

   $$L = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^2 det(\sigma^2 I)}} e^{-\frac{1}{2}(z_i - 0)^T \Sigma^{-1}(z_i - 0)} = \frac{1}{(2\pi\sigma)^n} e^{-\frac{1}{2}\sum\limits_{i=1}^{n} z_i^T \Sigma^{-1} z_i} = \frac{1}{(2\pi\sigma)^n} e^{-\frac{1}{2} z^T \Sigma^{-1} z},$$

   где $z = \sum\limits_{i=1}^{n} z_i$.

   $$z^T \Sigma^{-1} z = z^T \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}^{-1} z = z^T \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix} z = \begin{pmatrix} z_1 & z_2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \frac{1}{\sigma^2}(z_1^2 + z_2^2)$$

   Тогда логарифм максимального правдоподобия:

   $$\log L = -n \log(2\pi\sigma) - \frac{1}{2\sigma^2}(z_1^2 + z_2^2) = -n\log(2\pi\sigma) - \frac{1}{2\sigma^2}\left((\sum_{i=1}^{n} z_{i1})^2 + (\sum_{i=1}^{n} z_{i2})^2 \right) =$$

   $$= -n\log(2\pi\sigma) - \frac{1}{2\sigma^2}\left((\sum_{i=1}^{n} y_{i1} - n\cos(\theta + \phi))^2 + (\sum_{i=1}^{n} y_{i2} - n\sin(\theta + \phi))^2 \right),$$

   где y_{i1} и y_{i2} - 1 и 2 элементы вектора y_i.

   На самом деле, если σ известна, то максимизация логарифма правдоподобия эквивалентна максимизации:

   $$-\left(\sum_{i=1}^{n} y_{i1} - n\cos(\theta + \phi) \right)^2 - \left(\sum_{i=1}^{n} y_{i2} - n\sin(\theta + \phi) \right)^2$$

   Ответ: $-n\log(2\pi\sigma) - \frac{1}{2\sigma^2}\left((\sum\limits_{i=1}^{n} y_{i1} - n\cos(\theta + \phi))^2 + (\sum\limits_{i=1}^{n} y_{i2} - n\sin(\theta + \phi))^2 \right)$, где y_{i1} и y_{i2} - 1 и 2 элементы вектора y_i.

2. (1 балл) Используя результаты предыдущего пункта, вычислите оценку максимального правдоподобия $\hat{\phi_n}$ для параметра ϕ. Подсказка: рассмотрите сперва случай $\sigma = 0$ (шума нет). Как изменится оценка, если ковариационную матрицу шума $\mathbf{Z}$ растянуть в k раз (умножить на $k\mathbf{I}$)?

Решение:

1) Если шума нет, то вектора Y_i принимают одно значение и максимального правдоподобия нет (т.е. можно сказать, что Y_i измерено точно).

2) Теперь рассмотрим случай, когда дисперсия ненулевая.

$$\frac{\partial \log L}{\partial \phi} = -2\sum_{i=1}^{n} y_{i1} n \sin(\theta+\phi) + 2\sum_{i=1}^{n} y_{i2} n \cos(\theta+\phi) = 0$$

$$\sum_{i=1}^{n} y_{i1} \sin(\theta+\phi) - \sum_{i=1}^{n} y_{i2} \cos(\theta+\phi) = 0$$

$$\operatorname{tg}(\theta+\phi) = \frac{\sum\limits_{i=1}^{n} y_{i2}}{\sum\limits_{i=1}^{n} y_{i1}}$$

$$\theta+\phi = \operatorname{arctg}\frac{\sum\limits_{i=1}^{n} y_{i2}}{\sum\limits_{i=1}^{n} y_{i1}}, \text{ потому что } \theta+\phi \in [0, \pi/2]$$

$$\hat{\phi} = \operatorname{arctg}\frac{\sum\limits_{i=1}^{n} y_{i2}}{\sum\limits_{i=1}^{n} y_{i1}} - \phi$$

На $\hat{\phi}$ растяжение ковариационной матрицы шума Z в k раз не влияет, однако значение правдоподобия изменится:

$$\frac{L_{prev}}{L} = \frac{\frac{1}{(2\pi\sigma)^n} e^{-\frac{1}{2}z^T\Sigma^{-1}z}}{\frac{1}{(2\pi\sigma\sqrt{k})^n} e^{-\frac{1}{2}z^T(\Sigma kI)^{-1}z}} = e^{-\frac{1}{2}z^T\Sigma^{-1}z(1-\frac{1}{k})} k^{\frac{n}{2}}$$

Получаем, что при растяжении ковариационной матрицы в k раз оценка максимального правдоподобия изменяется так: $L = L_{prev}\frac{e^{\frac{1}{2}z^T\Sigma^{-1}z(1-\frac{1}{k})}}{k^{\frac{n}{2}}}$.

Ответ:

1) Если шума нет, оценки максимального правдоподобия нет

2) $\hat{\phi} = \operatorname{arctg}\frac{\sum_{i=1}^{n} y_{i2}}{\sum_{i=1}^{n} y_{i1}} - \phi$

Задача 5 (2 балла) Метод максимального правдоподобия, байесовский метод.

Антон очень любит городской каршеринг. А еще больше он любит считать машины каршеринга, припаркованные у офиса. В понедельник и среду он насчитал 4 машины, во вторник – 5, а в четверг всего одну. Одна машина – это слишком мало, подумал Антон – кто-то забронирует ее раньше меня, и придется ехать на такси. Чтобы не расстраиваться раньше времени, Антон попросил вас посчитать вероятность обнаружить хотя бы две машины рядом с офисом в очередной день.

1. (0.5 балла) Обозначим $n = 4$, $[x_1, x_2, x_3, x_4] = [4, 5, 4, 1]$ - выборка, наблюдаемая Антоном. Положив число машин за случайную величину, распределенную как

$$X \sim Pois(\lambda): \; P(\mathbf{X} = k; \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

найдите λ^{MLE} с помощью метода максимального правдоподобия. Посчитайте условную вероятность $P(X_\lambda \geq 2|\lambda = \lambda^{MLE})$.

Решение:

Найдем оценку максимального правдопободия λ_{MLE}.

$L = \log \prod\limits_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = -\lambda n + \sum\limits_{i=1}^{n} x_i \log \lambda - \sum\limits_{i=1}^{n} \log x_i!$

$\frac{\partial L}{\partial \lambda} = -n + \frac{\sum\limits_{i=1}^{n} x_i}{\lambda} = 0 => \lambda_{MLE} = \frac{\sum\limits_{i=1}^{n} x_i}{n}$

Подставляя заданные значения, получаем: $\lambda_{MLE} = 3.5$

$P(X_\lambda \geq 2|\lambda = \lambda^{MLE}) = 1 - P(X_\lambda = 0|\lambda = \lambda^{MLE}) - P(X_\lambda = 1|\lambda = \lambda^{MLE}) =$

$= 1 - \frac{e^{-\lambda_{MLE}}\lambda_{MLE}^0}{0!} - \frac{e^{-\lambda_{MLE}}\lambda_{MLE}^1}{1!} = 1 - e^{-\lambda_{MLE}}(1 + \lambda_{MLE}) \approx 0.8641$

Ответ: $P(X_\lambda \geq 2|\lambda = \lambda^{MLE}) \approx 0.8641$

2. (1 балл) Для использования байесовского метода нам понадобится зафиксировать априорное распределение. Предлагается взять гамма распределение

$$Y \sim G(\alpha, \beta): \; p(x; \alpha, \beta) = \frac{x^{\alpha-1}e^{-\beta x}\beta^\alpha}{\Gamma(\alpha)}$$

Покажите, что апостериорным распределением при заданных выше условиях будет отрицательное биномиальное распределение со следующими коэффициентами

$$Z \sim NB(\alpha + \sum_i x_i, \frac{1}{1 + \beta + n}) = NB(r, q): \; p(\mathbf{X} = k; r, q) = C_{k+r-1}^k q^k (1-q)^r$$

Решение:

$P(x = k|X) = \int_\lambda p(x = k|\lambda)p(\lambda|X)d\lambda = \int_\lambda p(x = k|\lambda)\frac{P(X|\lambda)p(\lambda)}{P(X)}d\lambda \quad (*)$

Найдем $P(X)$ и подставим его значение в $(*)$.

$P(X) = \int_\lambda P(X|\lambda)p(\lambda)d\lambda = \int_\lambda \frac{\prod\limits_{i=1}^{n}\lambda^{x_i}e^{-\lambda}}{\prod\limits_{i=1}^{n} x_i!} * \frac{\lambda^{\alpha-1}e^{-\beta\lambda}\beta^\alpha}{\Gamma(\alpha)}d\lambda = \frac{\beta^\alpha}{\prod\limits_{i=1}^{n} x_i!\Gamma(\alpha)} \int_\lambda \frac{\lambda^{\sum\limits_{i=1}^{n} x_i + a - 1} e^{-(n+\beta)\lambda}(n+b)^{\sum\limits_{i=1}^{n} x_i + \alpha}}{\Gamma(\sum\limits_{i=1}^{n} x_i + \alpha)}d\lambda *$

$* \frac{\Gamma(\sum\limits_{i=1}^{n} x_i + \alpha)}{(n+b)^{\sum\limits_{i=1}^{n} x_i + \alpha}} = \frac{\beta^\alpha \Gamma(\sum\limits_{i=1}^{n} x_i + \alpha)}{(n+b)^{\sum\limits_{i=1}^{n} x_i + \alpha}\Gamma(\alpha)\prod\limits_{i=1}^{n} x_i!}$

$(*) = \int_\lambda \frac{\lambda^k e^{-\lambda}}{k!} \frac{\prod\limits_{i=1}^{n}\lambda^{x_i}e^{-\lambda}}{\prod\limits_{i=1}^{n} x_i!} \frac{\lambda^{\alpha-1}e^{-\beta\lambda}\beta^\alpha}{\Gamma(\alpha)} \frac{(n+b)^{\sum\limits_{i=1}^{n} x_i + \alpha}\Gamma(\alpha)\prod\limits_{i=1}^{n} x_i!}{\beta^\alpha\Gamma(\sum\limits_{i=1}^{n} x_i + \alpha)}d\lambda = \int_\lambda \frac{\lambda^k e^{-\lambda}}{k!} \frac{\lambda^{\sum\limits_{i=1}^{n} x_i + \alpha - 1} e^{-(n+\beta)\lambda}(n+b)^{\sum\limits_{i=1}^{n} x_i + \alpha}}{\Gamma(\sum\limits_{i=1}^{n} x_i + \alpha)}d\lambda =$

$= \frac{(n+\beta)^{\sum\limits_{i=1}^{n} x_i + \alpha}\Gamma(k + \sum\limits_{i=1}^{n} x_i + \alpha)}{k!(n+\beta+1)^{k + \sum\limits_{i=1}^{n} x_i + \alpha}\Gamma(\sum\limits_{i=1}^{n} x_i + \alpha)} = C^k_{k+\alpha+\sum\limits_{i=1}^{n} x_i - 1}\left(\frac{1}{n+\beta+1}\right)^k \left(1 - \frac{1}{n+\beta+1}\right)^{\sum\limits_{i=1}^{n} x_i + \alpha}$

При вычислении $(*)$ нужно воспользоваться тем же трюком, что и при вычислении $P(X)$: выделить интеграл от плотности гамма-распределения, который равен 1, так как берем интеграл по всем λ

ч.т.д.

3. (0.5 балла) Для параметров $\alpha = 2$ и $\beta = 2$, вычислите, используя предыдущие пункты, апостериорную оценку вероятности
$$P(X_\lambda \geq 2|X)$$
Сравните ее с оценкой максимального правдоподобия из пункта 1. О чем говорит результат сравнения?

Решение: $P(X_\lambda \geq 2|X) = 1 - P(X_\lambda = 0|X) - P(X_\lambda = 1|X) =$
$$= 1 - \frac{(0+2+14-1)!}{0!(2+14-1)} \left(\frac{1}{4+2+1}\right)^0 \left(1 - \frac{1}{4+2+1}\right)^{14+2} - \frac{(1+2+14-1)!}{1!(2+14-1)} \left(\frac{1}{4+2+1}\right)^1 \left(1 - \frac{1}{4+2+1}\right)^{14+2} \approx 0.72$$

Ответ: 0.72

Задача 6 (1 балл) Байесовский классификатор. Докажите, что наивный байесовский классификатор в случае n бинарных признаков $x_j \in \{0, 1\}$, $j = 1, \ldots, n$ является линейным разделителем: $a(x) = [w_0 + w_1x_1 + \ldots + w_nx_n > 0]$. Выпишите формулы для вычисления коэффициентов w_j, $j = 0, \ldots, n$ по обучающей выборке.

Решение:

Рассмотри бинарный классификатор и объекты x с бинарными признаками.

$P(y = 1|x) = P(y = 1)P(x|y = 1) = P(y = 1) \prod\limits_{j=1}^{n} P(x_i|y = 1) = P(y = 1) \prod\limits_{j=1}^{n} p_j^{x_j}(1 - p_j)^{1-x_j}$, где p_j - вероятность появления j-го признака.

Логарифмирование не меняет максимум, поэтому

$\log P(y = 1|x) = \log P(y = 1) + \sum\limits_{j=1}^{n} x_j \log p_j + (1 - x_j) \log(1 - p_j) =$

$= \log P(y = 1) + \sum\limits_{j=1}^{n} \log(1 - p_j) + \sum\limits_{j=1}^{n} x_j(\log p_j - \log(1 - p_j))$. Получили линейный классификатор (конечно, лучше сказать линейный регрессор, потому что мы предсказываем логарифм вероятности принадлежности объекта x к классу 1).

Стоит отметить, что из полученной формулы можно легко получить формулу, которая дана в условии. Для этого нужно определить порог t, когда мы переходим из класса 0 в 1:
$$a(x) = [\log P(y = 1|x) > t] = [\log P(y = 1|x) - t > 0]$$

Значения, при которых достигается максимум правдоподобия:

1)$p_j = \frac{\sum_{i=1}^{|D|} x_{ij}}{|D|}$, где x_{ij} - j-ый признак i-го объекта, $|D|$ - длина выборки.

2)$P(y = 1) = \frac{\sum_{i=1}^{|D|} [y_i=1]}{|D|}$

Тогда искомые коэффициенты вырыжаются следующим образом:

$w_j = \log p_j - \log(1 - p_j) = \log \frac{\sum_{i=1}^{|D|} x_{ij}}{|D|} - \log \left(1 - \frac{\sum_{i=1}^{|D|} x_{ij}}{|D|}\right), j = 1, ..., n$

$w_0 = \log P(y = 1) + \sum\limits_{j=1}^{n} \log(1 - p_j) - t = \frac{\sum_{i=1}^{|D|} [y_i=1]}{|D|} + \sum\limits_{j=1}^{n} \log \left(1 - \frac{\sum_{i=1}^{|D|} x_{ij}}{|D|}\right) - t$

Задача 7 (1.5 балла) Байесовский классификатор.

Рассмотрим задачу классификации текстов $D = \{d_1, \ldots, d_{|D|}\}$ на K классов $Y = \{1, \ldots, K\}$. Каждый документ d_i представляет собой некоторое подмножество множества возможных слов $W = \{w_1, \ldots, w_{|W|}\}$ (т.е. нас не интересует порядок слов и количество вхождений каждого слова). В качестве признаков для каждого документа выберем индикаторы вхождения слов в него. Матрица «объекты-признаки» задается как

$$x_{ij} = I[w_j \in d_i], \quad i = 1, \ldots |D|, \ j = 1, \ldots K, \ I - \text{индикатор}$$

Для решения задачи воспользуемся наивным байесовским классификатором, который основывается на предположении, что признаки независимы:

$$p(x_i \,|\, y_i) = p(x_{i1} \,|\, y_i) \ldots p(x_{i|W|} \,|\, y_i), \ x_i = (x_{i1}, \ldots, x_{i|W|})$$

Будем считать, что при фиксированном классе каждый признак имеет распределение Бернулли. Таким образом, априорные распределения и функции правдоподобия задаются как

$$p(k \,|\, \pi) = \pi_k, \quad k = 1, \ldots, K;$$
$$p(x_{ij} \,|\, k, \theta) = \theta_{jk}^{x_{ij}} (1 - \theta_{jk})^{1 - x_{ij}}, \quad i = 1, \ldots, |D|, \ j = 1, \ldots, |W|, \ k = 1, \ldots, K.$$

Здесь обучаемые параметры π_k – вероятность k-го класса, $\theta_j k$ – вероятность встретить j-е слово в документе k-го класса. Распределение одного документа записывается следующим образом:

$$p(d_i, y_i \,|\, \pi, \theta) = p(y_i \,|\, \pi) \prod_{j=1}^{|W|} p(x_{ij} \,|\, y_i, \theta) = \prod_{k=1}^{K} \pi_k^{[y_i = k]} \prod_{j=1}^{|W|} \prod_{k=1}^{K} p(x_{ij} \,|\, k, \theta_{jk})^{[y_i = k]}.$$

Докажите, что оценки максимального правдоподобия на параметры π и θ имеют вид

$$\hat{\pi}_k = \frac{\sum_i [y_i = k]}{|D|},$$
$$\hat{\theta}_{jk} = \frac{\sum_i [y_i = k][x_{ij} = 1]}{\sum_i [y_i = k]},$$

где все суммирования ведутся по документам от 1 до $|D|$.

Решение:

1) Получим оценку для π_k, для этого обозначим $\prod_{j=1}^{|W|} \prod_{k=1}^{K} p(x_{ij} \,|\, k, \theta_{jk})^{[y_i = k]}$ за константу C (она не зависит от значений π_i) для упрощения вычислений.

Логарифм максимального правдоподобия:

$$\log L = \sum_{i=1}^{|D|} \log \prod_{i=1}^{K} \pi_k^{[y_i = k]} C = \sum_{i=1}^{|D|} \sum_{k=1}^{K} ([y_i = k] \log \pi_k + C)$$

Упростим произведение. Каждый документ принадлежит только одному классу, поэтому рассмотрим вероятности, что документ принадлежит классу k или не принадлежит. Тогда производная получится следующая:

$$\frac{\partial \log L}{\partial \pi_k} = \frac{\partial}{\partial \pi_k} \left(\sum_{i=1}^{|D|} [y_i = k] \log \pi_k + [y_i \neq k] \log(1 - \pi_k) + C \right) = \sum_{i=1}^{|D|} \left(\frac{[y_i = k]}{\pi_k} - \frac{[y_i \neq k]}{1 - \pi_k} \right) =$$

$$= \sum_{i=1}^{|D|} \frac{[y_i = k](1 - \pi_k) - \pi_k [y_i \neq k]}{\pi_k (1 - \pi_k)} = \sum_{i=1}^{|D|} \frac{[y_i = k](1 - \pi_k) - \pi_k (1 - [y_i = k])}{\pi_k (1 - \pi_k)} = 0 => \hat{\pi}_k = \frac{\sum\limits_{i=1}^{|D|} [y_i = k]}{|D|}$$

ч.т.д.
2)

$$\log L = \sum_{i=1}^{|D|}\sum_{k=1}^{K}[y_i = k]\log \pi_k + \sum_{i=1}^{|D|}\sum_{j=1}^{|W|}\sum_{k=1}^{K}[y_i = k]\Bigg(x_{ij}\log \theta_{jk} + (1 - x_{ij})(1 - \theta_{jk})\Bigg)$$

$$\frac{\partial \log L}{\theta_{jk}} = \sum_{i=1}^{|D|}[y_i = k]\left(\frac{x_{ij}}{\theta_{jk}} - \frac{1 - x_{ij}}{1 - \theta_{jk}}\right) = \sum_{i=1}^{|D|}[y_i = k]\frac{x_{ij}(1 - \theta_{jk}) - \theta_{jk}(1 - x_{ij})}{\theta_{jk}(1 - \theta_{jk})} =$$

x_{ij} принимает значения 1 или 0 (по условию), тогда:

$$= \sum_{i=1}^{|D|}\frac{[y_i = k]\Bigg([x_{ij} = 1](1 - \theta_{jk}) - \theta_{jk}(1 - [x_{ij} = 1])\Bigg)}{\theta_{jk}(1 - \theta_{jk})} = \sum_{i=1}^{|D|}\frac{[y_i = k]([x_{ij} = 1] - \theta_{jk})}{\theta_{jk}(1 - \theta_{jk})} = 0 =>$$

$$\sum_{i=1}^{|D|}[y_i = k]([x_{ij} = 1] - \theta_{jk}) = \sum_{i=1}^{|D|}[y_i = k][x_{ij} = 1] - \sum_{i=1}^{|D|}[y_i = k]\theta_{jk} = 0 => \hat{\theta}_{jk} = \frac{\sum\limits_{i=1}^{|D|}[y_i = k][x_{ij} = 1]}{\sum\limits_{i=1}^{|D|}[y_i = k]}$$

ч.т.д.