# hw_theor04

March 16, 2021

## 0.1 Машинное обучение

### 0.1.1 Факультет математики НИУ ВШЭ, 2020-21 учебный год

*Илья Щуров, Соня Дымченко, Руслан Хайдуров, Максим Бекетов, Павел Егоров*

Страница курса

## 0.2 Домашнее задание 4. Вокруг линейных моделей

Задание выполнил(а): *(впишите свои фамилию и имя)*

**Внимание!** Домашнее задание выполняется самостоятельно. При попытке сдать хотя бы частично списанный текст, или текст, полученный в результате совместного решения задач, вся работа будет оценена на 0 баллов. Мы также уведомим администрацию факультета и попросим применить дисциплинарное взыскание (предупреждение, выговор, отчисление) ко всем вовлеченным студентам.

### 0.2.1 Задача 1 (20 баллов)

Рассмотрим следующую модель. Значения $x_1, \ldots, x_n \in \mathbb{R}^d$ фиксированы. Вектор $w \in \mathbb{R}^d$ фиксирован. Также фиксирован вектор $\sigma = (\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^n$. Значения $y_i$ определяются следующим образом:

$$y_i = \langle w, x_i \rangle + \varepsilon_i,$$

где $\varepsilon_i$ — независимые случайные величины, распределённые по нормальному закону, $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ (то есть у каждого $\varepsilon_i$ своя дисперсия, равная $\sigma_i^2$, все $\sigma_i$ фиксированы и известны).

1. Найти функцию правдоподобия $p((y_1, \ldots, y_n) \mid w, x, \sigma)$, равную плотности вероятности получения данных $y_1, \ldots, y_n$ при заданных фиксированных $w$, $x$ и $\sigma$.
2. Найти логарифм правдоподобия.
3. Записать задачу максимизации логарфима правдоподобия по $w$. Выкинуть лишние слагаемые и записать аналог $RSS$ для этой задачи.
4. Записать задачу максимизации правдоподобия в матричном виде. Для этого ввести матрицы $X$ (матрица объект-признак, по строкам записаны $x_1, \ldots, x_n$) и $\Sigma$ — диагональная матрица, у которой на диагонали стоят $\sigma_1, \ldots, \sigma_n$.
5. Решить эту задачу в матричном виде. (Найти градиент аналога $RSS$ в матричном виде, приравнять нулю, решить получившееся уравнение. Найти гессиан, показать, что он отрицательно определён в точке максимума.)
6. Является ли полученная оценка для $w$ несмещённой?

7. Найти ковариационную матрицу для оценки $w$.

**Подсказка.** Для самопроверки можеет подставить в качестве вектора $\sigma$ постоянный вектор (все компоненты равны одному и тому же числу). Должны получиться формулы, которые доказывались на лекциях.

*(впишите решение сюда)*

### 0.2.2 Задача 2 (10 баллов)

Рассмотрим такую модель. Значения $x_1, \ldots, x_n \in \mathbb{R}$ — фиксированные числа, $\beta \in \mathbb{R}$ — фиксированное число, $\varepsilon_i \sim \mathcal{N}(0, 1)$ — независимые случайные ошибки,

$$y_i = \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Пусть $\widehat{\beta}$ — МНК-оценка $\beta$ для данной модели. Для предсказания значения $y$ в точке $x_{new}$ используется следующий алгоритм:

$$\widehat{y}_{new} = \gamma \cdot \widehat{\beta} x_{new},$$

где $\gamma \in \mathbb{R}$ — некоторая константа (не зависящая от $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$).

1. Для заданного $x_{new}$ найти квадрат смещения предсказания и разброс (дисперсию) предсказания (как функции от $x_{new}$, $\beta$ и $\gamma$ и $(x_1, \ldots, x_n)$).
2. Найти такое значение $\gamma$, при котором ожидаемая квадратичная ошибка предсказания минимальна. (Эта величина будет зависеть от $\beta$, $x_{new}$ и $(x_1, \ldots, x_n)$)

(Эта задача демонстрирует ещё одно проявление bias-variance tradeoff: мы можем пожертвовать несмещённостью, чтобы уменьшить ошибку предсказания.)

*(впишите решение сюда)*

### 0.2.3 Задача 3 (10 баллов)

Маша, Неля и Катя решают задачу линейной регрессии. Данные у них одинаковые, в них $n$ наблюдений и два признака $x^{(1)}$ и $x^{(2)}$, а также вектор ответов $y$. Признаки имеют нулевое выборочное среднее и нулевую выборочную ковариацию. Маша находит вектор весов $(w_1, w_2)$ как МНК-оценку для задачи $y_i = w_1 x_i^{(1)} + w_2 x_i^{(2)} + \varepsilon_i$. Неля решила выбросить второй признак и находит вес $w_1$ как МНК-оценку для задачи $y_i = w_1 x_i^{(1)} + \varepsilon_i$. Катя выбросила первый признак и находит вес $w_2$ как МНК-оценку для задачи $y_i = w_2 x_i^{(2)} + \varepsilon_i$. Докажите, что $w_1 = w_1$ и $w_2 = w_2$. Будет ли это верно в случае, если признаки будут по-прежнему иметь нулевое среднее, но окажутся скоррелированными (то есть не будут иметь нулевую ковариацию)?

*(впишите решение сюда)*

### 0.2.4 Задача 4 (10 баллов)

Маша и Катя решают задачу линейной регрессии. Изначально у них одинаковый набор данных, состоящий из $n$ наблюдений $x_i$, $i = 1, \ldots, x_n$ по $d$ признаков и вектора ответов $y = (y_1, \ldots, y_n)$. Маша записала линейную модель

$$y_i = x_i^{(1)} w_1 + \ldots + x_i^{(d)} w_d + \varepsilon_i.$$

и стала искать МНК-оценку для $(w_1, \ldots, w_d)$. А Катя считает, что реальная зависимость между $y$ и признаками является нелинейной, поэтому она добавила новые признаки в модель (но не стала убирать старые). В качестве новых признаков она использовала различные линейные и нелинейные функции от старых признаков, которые ей приходили в голову. Таким образом, Катина модель выглядит так:

$$y_i = x_i^{(1)} w_1 + \ldots + x_i^{(d)} w_d + x_i^{(d+1)} w_{d+1} + \ldots + x_i^{(d+k)} w_{d+k} + \varepsilon_i,$$

где $x^{(d+1)}, \ldots, x^{(d+k)}$ — новые признаки, добавленные Катей. Она также ищет вектор весов с помощью метода наименьших квадратов.

После нахождения вектора весов каждая девушка вычислила RSS для своей модели (по обучающей выборке). 1. Докажите, что RSS Кати оказался не больше RSS Маши. 2. Могут ли RSS оказаться одинаковыми, но при этом ненулевыми? Если нет, докажите. Если да, приведите пример.

*(впишите решение сюда)*

### 0.2.5 Задача 5 (10 баллов)

У рассеянного Александра есть вектор ответов $(y_1, \ldots, y_n)$ для задачи классификации, каждый $y_i \in \{0, 1\}$, $n$ нечётное число, всего среди $y_i$ есть $m$ единиц и $(n-m)$ нулей. Александр потерял матрицу признаков, поэтому вынужден использовать алгоритм, обучающийся только по ответам, и везде предсказывающий одно и то же значение. Он хочет, чтобы алгоритм предсказывал вероятность $p$ получения единицы, и думает, какую функцию потерь ему выбрать из двух возможных:

1. Log-loss:
$$L_{LL}(y, p) = \begin{cases} -\log p, & \text{if } y = 1; \\ -\log(1-p), & \text{if } y = 0. \end{cases}$$

2. Абсолютное отклонение:
$$L_{AD}(y, p) = |y - p|.$$

Для нахождения оптимального $p$ Александр решает задачу минимизиации эмпирического риска:

$$\sum_{i=1}^{n} L(y_i, p) \to \min_p,$$

где $L$ — это либо $L_{LL}$, либо $L_{AD}$.

Какое $p$ получится у Александра для каждой из данных функций потерь? Какую из функций потерь следует использовать, если Александр хочет, чтобы $p$ была состоятельной оценкой для вероятности получения единицы?

*(впишите решение сюда)*

### 0.2.6 Задача 6 (5 баллов)

Кларисса решает задачу двухклассовой классификации (классы обозначаются $+1$ и $-1$) с помощью алгоритма машинного обучения, который для $i$-го объекта выдаёт степень уверенности $s_i$ алгоритма в том, что этот объект принадлежит к классу $+1$ (например, это может быть оценка вероятности, данная логистической регрессией). Кларисса выбирает пороговое значение $t$, после чего все объекты, для которых $s_i > t$, относит к положительному классу, а остальные — к отрицательному. Иными словами, окончательное предсказание классификатора имеет вид:

$$\hat{y}_i = [s_i > t] - [s_i \leq t].$$

В таблице для каждого элемента обучающей выборки даны значения $s_i$ и их истинные классы. Построить ROC-кривую для данного алгоритма. (Вы можете сделать это вручную — это хорошее упражнение (посмотрите пример здесь) — либо самостоятельно написать код, который строит ROC-кривую. Использовать готовые решения из библиотек типа scikit-learn нельзя.)

| $s_i$ | $y_i$ |
|---|---|
| 0.6 | $+1$ |
| 0.5 | $-1$ |
| 0.1 | $-1$ |
| 0.2 | $-1$ |
| 0.4 | $+1$ |
| 0.7 | $+1$ |

*(впишите решение сюда)*