

# Практическая работа №5: «Модели линейной и логистической регрессии»

## Оглавление

Цель работы .....	1
Задачи работы .....	1
Перечень обеспечивающих средств.....	2
Общие теоретические сведения .....	2
<b>Линейная регрессия</b> .....	2
<b>Градиентный спуск</b> .....	2
<b>Допущения линейной регрессии</b> .....	3
<b>Ограничения линейной регрессии</b> .....	3
<b>Регуляризация</b> .....	3
<b>L1-регуляризация (lasso)</b> .....	3
<b>L2-регуляризация (ridge)</b> .....	4
<b>Elastic Net регуляризация</b> .....	4
Задание .....	5
Требования к отчету .....	5
Литература .....	5

## ***Цель работы***

Получить практические навыки использования линейной регрессии.

## ***Задачи работы***

1. Научиться аналитически решать задачу линейной регрессии.
2. Научиться решать задачу линейной регрессии с помощью библиотеки sklearn.

## **Перечень обеспечивающих средств**

1. ПК.
2. Учебно-методическая литература.
3. Задания для самостоятельного выполнения.

## **Общие теоретические сведения**

### **Линейная регрессия**

Данные: пары значений  $(x_i, y_i)$ , где  $i = 1, \dots, N$ .

$x$  называется предиктором или регрессором,

$y$  называется зависимой переменной.

Задача: Найти такие значения  $a$  и  $b$ , чтобы функция  $f(x) = ax + b$  как можно точнее аппроксимировала  $y$ , т.е. чтобы  $f(x_i) \approx y_i$  для всех  $i$ .

Метрика производительности – среднеквадратичная ошибка:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2$$

где  $y_i$  - значение из данных,  $f(x_i)$  - результат работы модели.

### **Градиентный спуск**

1. Случайным образом выбираем точку  $(a_0, b_0)$ .
2. Вычисляем значения частных производных ошибки.
3. Изменяем координаты так, чтобы двигаться в сторону уменьшения производной:

$$a_i = a_{i-1} - \alpha \frac{\partial \text{MSE}}{\partial a}(a_{i-1}), b_i = b_{i-1} - \beta \frac{\partial \text{MSE}}{\partial b}(b_{i-1}).$$

4. Если  $\text{MSE}(a_i, b_i) - \text{MSE}(a_{i-1}, b_{i-1})$  достаточно мало, то завершаем. Иначе – возвращаемся к шагу 2.

## Допущения линейной регрессии

Остатки: величины  $y_i - f(x_i)$ .

Допущения линейной регрессии

- Между  $x$  и  $y$  есть линейная зависимость.
- Остатки распределены нормальным образом.
- Среднее значение остатков равно нулю.
- Дисперсия остатков постоянна.

## Ограничения линейной регрессии

- Низкая точность при аппроксимации нелинейных функций.
- Нельзя использовать для вычислений вне известного интервала.
- Считаем, что предикторы не содержат ошибок измерений.
- Нет ограничений области значений.

## Регуляризация

Если данных мало, а модель сложная, то высока вероятность переобучения.

Регуляризация – добавление дополнительных слагаемых к метрике производительности для того, чтобы штрафовать модель за излишне сложные решения и, таким образом, препятствовать переобучению.

Смещение увеличивается, разброс уменьшается.

## L1-регуляризация (lasso)

$$\frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda_1 \sum_i |a_i|$$

$$\sum_i -y_i \log f(z_i) - (1 - y_i) \log(1 - f(z_i)) + \lambda_1 \sum_i |a_i|$$

L1-регуляризация обнуляет параметры  $a_i$ , которые вносят в основном шум.

### **L2-регуляризация (ridge)**

$$\frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda_2 \sum_i a_i^2$$

$$\sum_i -y_i \log f(z_i) - (1 - y_i) \log(1 - f(z_i)) + \lambda_2 \sum_i a_i^2$$

L2-регуляризация не даёт значениям параметров  $a_i$  бесконтрольно увеличиваться.

### **Elastic Net регуляризация**

$$\frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda_1 \sum_i |a_i| + \lambda_2 \sum_i a_i^2$$

$$\sum_i -y_i \log f(z_i) - (1 - y_i) \log(1 - f(z_i)) + \lambda_1 \sum_i |a_i| + \lambda_2 \sum_i a_i^2$$

## **Задание**

### **Пояснение**

Для сохранения результатов данной работы вам понадобится два файла: doc/docx – для текста и ipynb – для кода. Назовите их одинаково: «Фамилия – задание 5».

### **Часть 1**

- Решите аналитически задачу линейной регрессии для следующего набора данных  $(x_i, y_i)$ , аналогично тому, как это было сделано в лекции: (0,0), (1,1), (2,3).
- Сохраните результат в своём docx/doc-файле.

### **Часть 2**

- Обновите свой репозиторий, созданный в практической работе №1, из оригинального репозитория:  
[https://github.com/mosalov/Notebook\\_For\\_AI\\_Main](https://github.com/mosalov/Notebook_For_AI_Main).

### **Часть 3**

- Откройте свой репозиторий в Binder (<https://mybinder.org/>).
- Откройте файл «task5.ipynb».
- Изучите, при необходимости – выполните повторно, приведённый в файле код.
- По аналогии с изученным выполните задание, приведённое в последней ячейке.
- Сохраните код в ipynb-файле. Необходимые пояснения опишите в своём docx/doc-файле.

## **Требования к отчету**

Оба файла (doc/docs и ipynb) загрузите в свой репозиторий, созданный в практическом задании №1 по пути: «Notebook\_For\_AI\_Main/2020 Осенний семестр/Практическое задание 5/» и сделайте пул-реквест.

## **Литература**

- <https://habr.com/ru/post/514818/>
- <https://habr.com/ru/post/474602/>
- <http://statistica.ru/theory/osnovy-lineynoy-regressii/>
- <http://statistica.ru/theory/logisticheskaya-regressiya/>

- <https://habr.com/ru/post/485872/>
- <https://habr.com/ru/company/ods/blog/323890/#metod-maksimalnogo-pravdopodobiya>
- <https://dyakonov.org/2018/03/12/%D0%BB%D0%BE%D0%B3%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F-%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F-%D0%BE%D1%88%D0%B8%D0%B1%D0%BA%D0%B8/>