Собеседование на специализацию «Интеллектуальный анализ данных»

Киселев Никита Б05-002

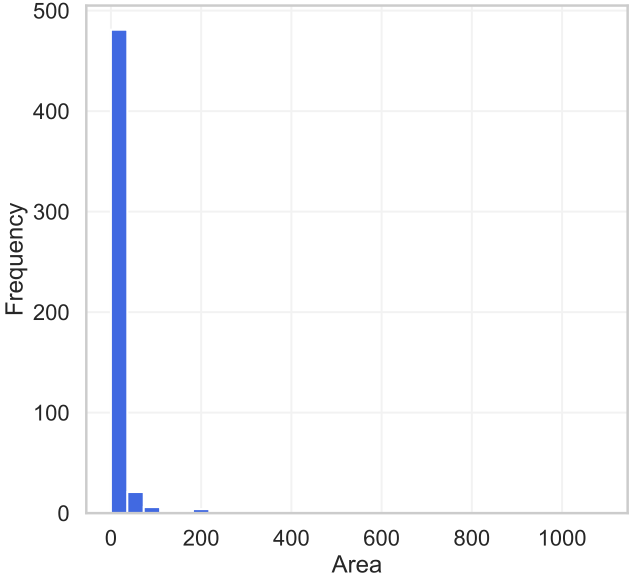
14 апреля 2022 г.

Московский физико-технический институт
(национальный исследовательский университет)

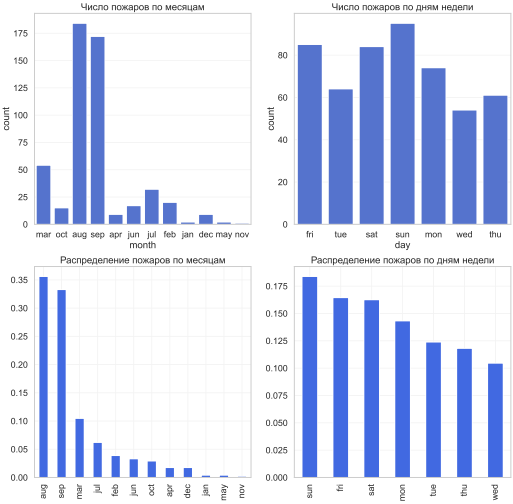
Задача 21

Предсказание площади лесных пожаров. На основе погодных измерений необходимо предсказать объем выгоревших лесных массивов на севере Португалии. Выборка состоит из 13 признаков и 517 объектов. Для решения задачи предлагается использовать метод наименьших квадратов с регуляризацией. Нарисовать график весов признаков и общей ошибки на кросс-валидации при изменении параметра регуляризации. Какие признаки наиболее важны для нашей задачи? Что изменится, если предварительно все признаки стандартизовать?

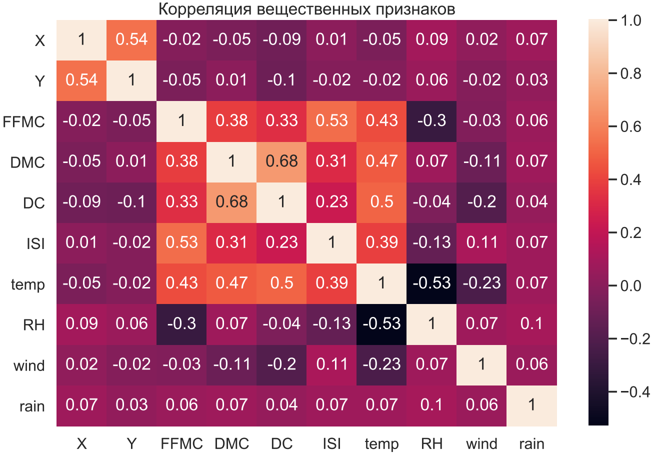
Распределение ответов



Распределение номинальных признаков



Корреляция количественных признаков



Линейная регрессия

- Множество объектов $\mathbb{X} = \mathbb{R}^n$
- Объекту $\mathrm{x} \in \mathbb{X}$ соответствует признаковое описание $\mathrm{x} = (f_1(x), \dots, f_n(x))$, где $f_j : \mathrm{X} \to D_j$
- Множество ответов $\mathbb{Y} = \mathbb{R}$
- Выборка $\mathbb{D} = \{(\mathrm{x}_i, y_i) \mid \mathrm{x}_i \in \mathbb{X}, y_i \in \mathbb{Y}, i = 1, \dots m\}$
- Матрица объекты-признаки $X = (\mathrm{x}_1, \dots, \mathrm{x}_m)^T$, вектор ответов $\mathrm{y} \in \mathbb{R}^m$
- Вектор параметров модели $\mathrm{w} = (w_0, w_1, \dots, w_n)^T$
- Ставится задача минимизации ошибки алгоритма $Q(w, X) = \|X\mathrm{w} - y\|_2^2 \to \min\limits_{\mathrm{w}}$

Метод наименьших квадратов

$$Q(\mathrm{w}, X) = \|X\mathrm{w} - y\|_2^2 = (X\mathrm{w} - y)^T(X\mathrm{w} - y) \to \min_{\mathrm{w}}$$

Приравняем к нулю производную по вектору w:

$$\begin{aligned}\nabla_{\mathrm{w}} Q(\mathrm{w}, X) = \nabla_{\mathrm{w}}(-y^T X\mathrm{w} + \mathrm{w}^T X^T X\mathrm{w} + y^T y - \mathrm{w}^T X^T y) = \\ = -X^T y + (X^T X + X^T X)\mathrm{w} + 0 - X^T y = 0\end{aligned}$$

$$X^T X\mathrm{w} = X^T y$$

$$\boxed{\mathrm{w}^* = (X^T X)^{-1} X^T y}$$

Могут возникнуть проблемы мультиколлинеарности в случае, если матрица $X^T X$ плохо обусловлена. Один из способов решения — добавление к этой матрице диагональной:

$$\mathrm{w}^* = (X^T X + \alpha E_n)^{-1} X^T y$$

При этом значении вектора w достигается минимум функционала ошибки

$$Q(\mathrm{w}, X, \alpha) = \|X\mathrm{w} - y\|_2^2 + \alpha\|\mathrm{w}\|_2^2$$