

Собеседование на специализацию «Интеллектуальный анализ данных»

Киселев Никита Б05-002

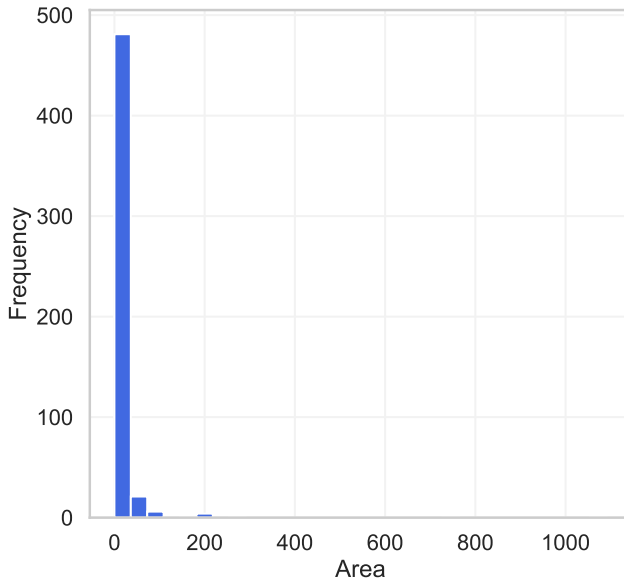
14 апреля 2022 г.

Московский физико-технический институт
(национальный исследовательский университет)

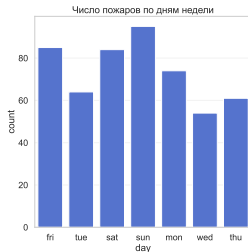
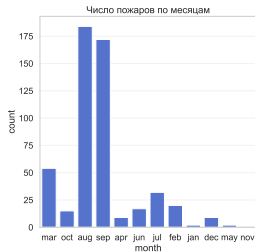
Задача 21

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



- Множество объектов $\mathbb{X} = \mathbb{R}^n$
- Объекту $x \in \mathbb{X}$ соответствует признаковое описание $x = (f_1(x), \dots, f_n(x))$, где $f_j : \mathbb{X} \rightarrow D_j$
- Множество ответов $\mathbb{Y} = \mathbb{R}$
- Выборка $\mathbb{D} = \{(x_i, y_i) \mid x_i \in \mathbb{X}, y_i \in \mathbb{Y}, i = 1, \dots, m\}$
- Матрица объекты-признаки $X = (x_1, \dots, x_m)^T$, вектор ответов $y \in \mathbb{R}^m$
- Вектор параметров модели $w = (w_1, \dots, w_n)^T$
- Ставится задача минимизации ошибки алгоритма $Q(w, X) = \|Xw - y\|_2^2 \rightarrow \min_w$

Метод наименьших квадратов

$$Q(w, X) = \|Xw - y\|_2^2 = (Xw - y)^T (Xw - y) \rightarrow \min_w$$

Приравняем к нулю производную по вектору w :

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

$$X^T Xw = X^T y$$

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

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

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

При этом значении вектора w достигается минимум функционала ошибки

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

Изменение параметра α

График зависимости MSE(α)

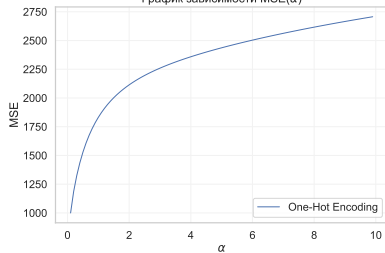


График зависимости весов признаков от α

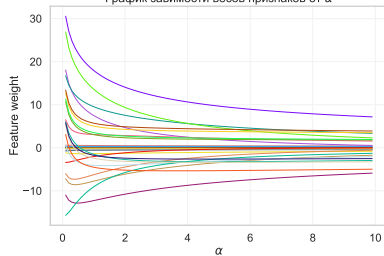


График зависимости MSE(α)

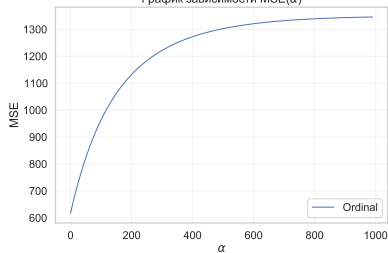
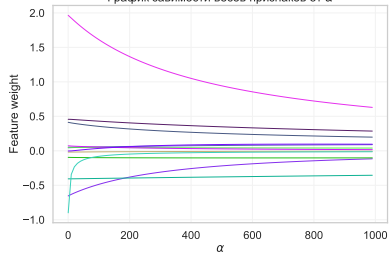


График зависимости весов признаков от α



При стандартизации происходит преобразование признаков:

$$\hat{f}_j(x_i) = \frac{f_j(x_i) - \bar{f}_j}{S_j},$$

где

$$\bar{f}_j = \frac{1}{m} \sum_{i=1}^m f_j(x_i) \text{ — выборочное среднее,}$$

$$S_j = \sqrt{\frac{1}{m} \sum_{i=1}^m (f_j(x_i) - \bar{f}_j)^2} \text{ — среднеквадратичное отклонение.}$$

Изменение параметра α при стандартизации

График зависимости MSE(α)

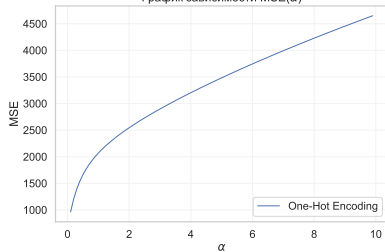


График зависимости весов признаков от α

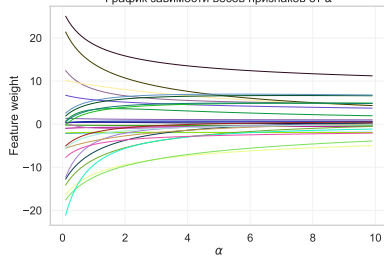


График зависимости MSE(α)

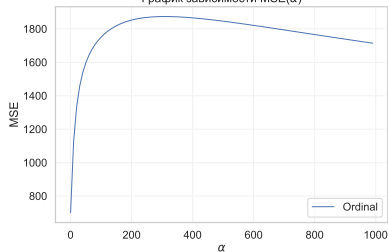
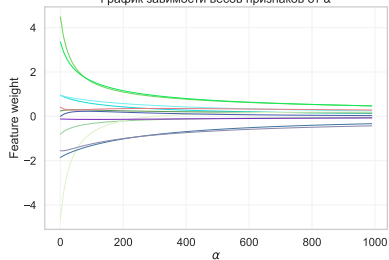
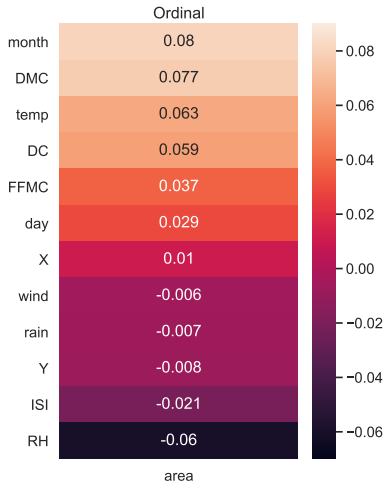
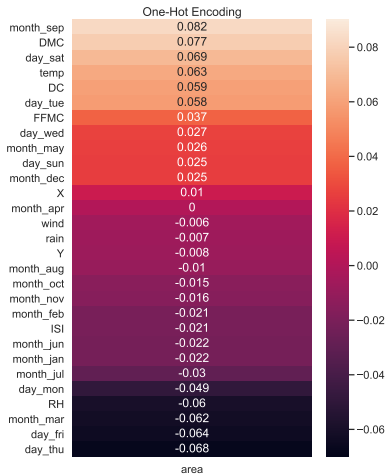


График зависимости весов признаков от α

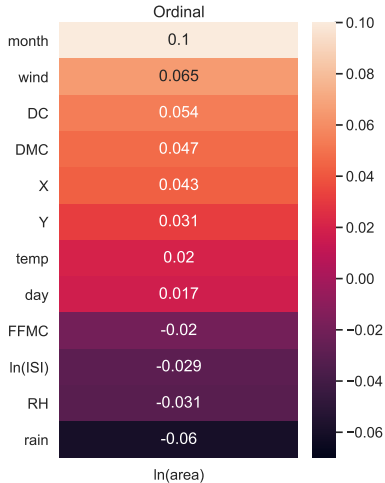
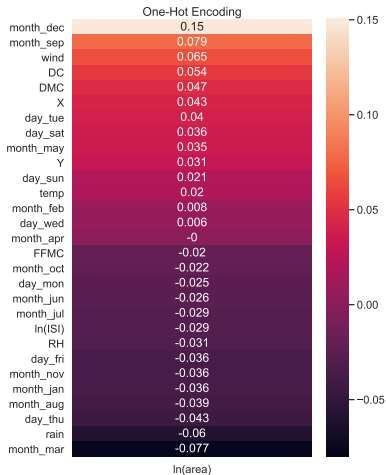


Отбор признаков

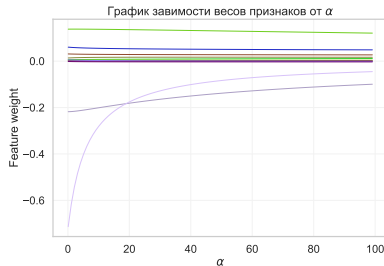
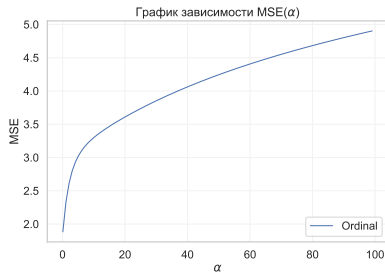
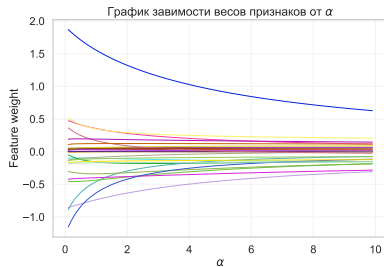
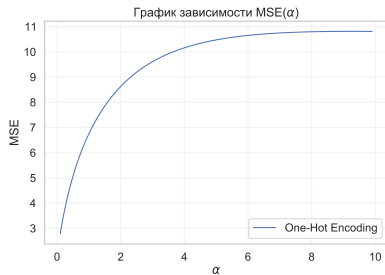


1. rain — номинальный
2. $FFMC \geq 75$
3. $ISI \rightarrow \ln(ISI)$
4. $area \rightarrow \ln(1 + area)$

Взаимосвязь новых признаков и ответов



Изменение параметра α



Изменение параметра α при стандартизации

График зависимости MSE(α)

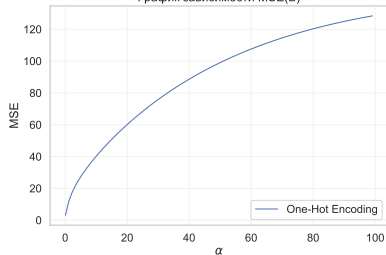


График зависимости весов признаков от α

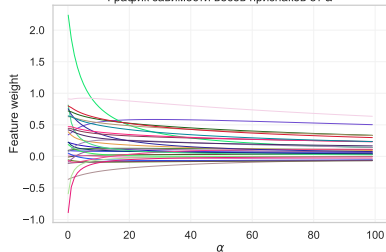


График зависимости MSE(α)

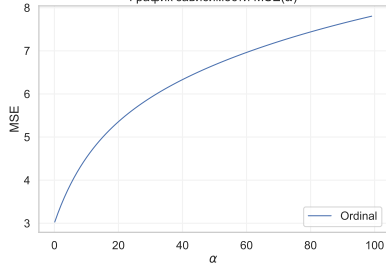
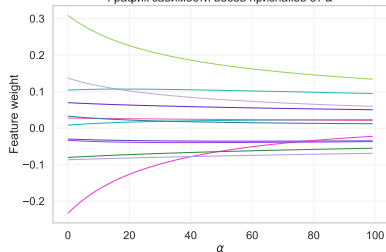


График зависимости весов признаков от α



Сравнение результатов

| Стандартизация \ Преобразование | До | После |
|---------------------------------|--------|-------|
| | До | После |
| - | 616,50 | 1,82 |
| + | 624,78 | 1,89 |

Таблица 1: Лучшее значение MSE на кросс-валидации

Стоит отметить, что после преобразования ответами являются $\ln(1 + area)$.

Значимость признаков при решении задачи лучше всего оценивается на данных после преобразования. Таковыми являются:

- month — месяц года
- wind — скорость ветра
- rain — количество осадков
- DC и DMC — индексы засухи и влажности почвы