Линейная регрессия / Machine learning

Сегодня

Эльбрус буткемп

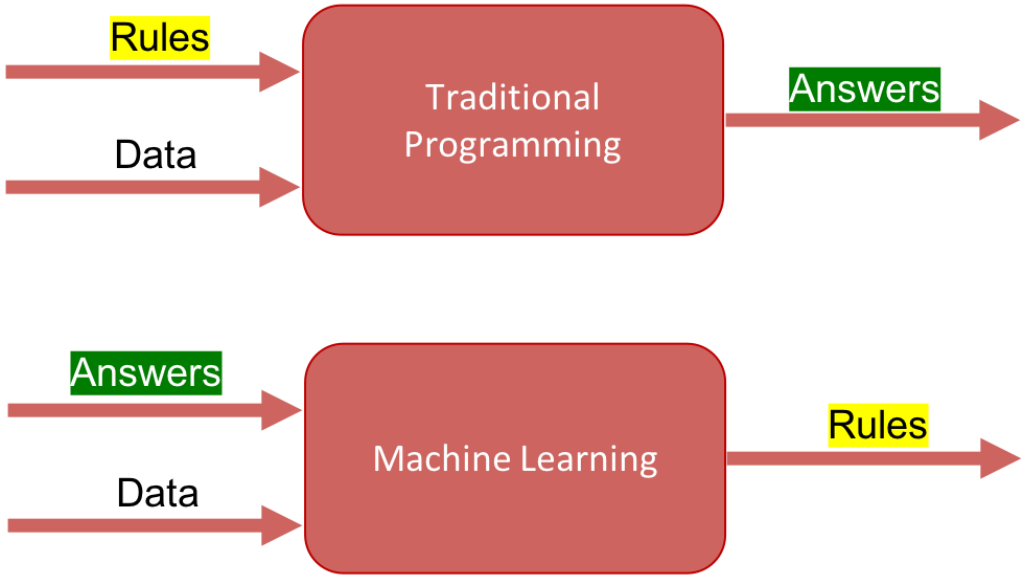
- Краткий экскурс в историю
- Виды задач в машинном обучении
- Примеры результатов

Машинное обучение

Эльбрус буткемп

Программирование vs. Машинное обучение

Эльбрус буткемп



📝Machine Learning vs. Traditional Programming Paradigm

Линейная регрессия

Эльбрус буткемп

Задача: оценить стоимость квартиры по одному признаку (например, площади).

Дана таблица с обучающей выборкой:

Площадь, м2	Цена, млн
50	12
33	8
...	...
76	120

Линейная регрессия

Эльбрус буткемп

Решаем уравнение:

$$y = x_1 w_1 + w_0$$

где x_1 - значение площади, w - "вес" признака, w_0 - свободный параметр.

Если помимо площади есть еще признаки, то уравнение просто увеличивается:

$$y = x_1 w_1 + x_2 w_2 + \cdots + x_n w_n + w_0$$

или короче:

$$y = \sum_{i=1}^{n} x_i w_i + w_0$$

Линейная регрессия: решение

Эльбрус буткемп

- методом градиентного спуска (так обычно и происходит)
  - будем минимизировать *среднеквадратическую ошибку*:

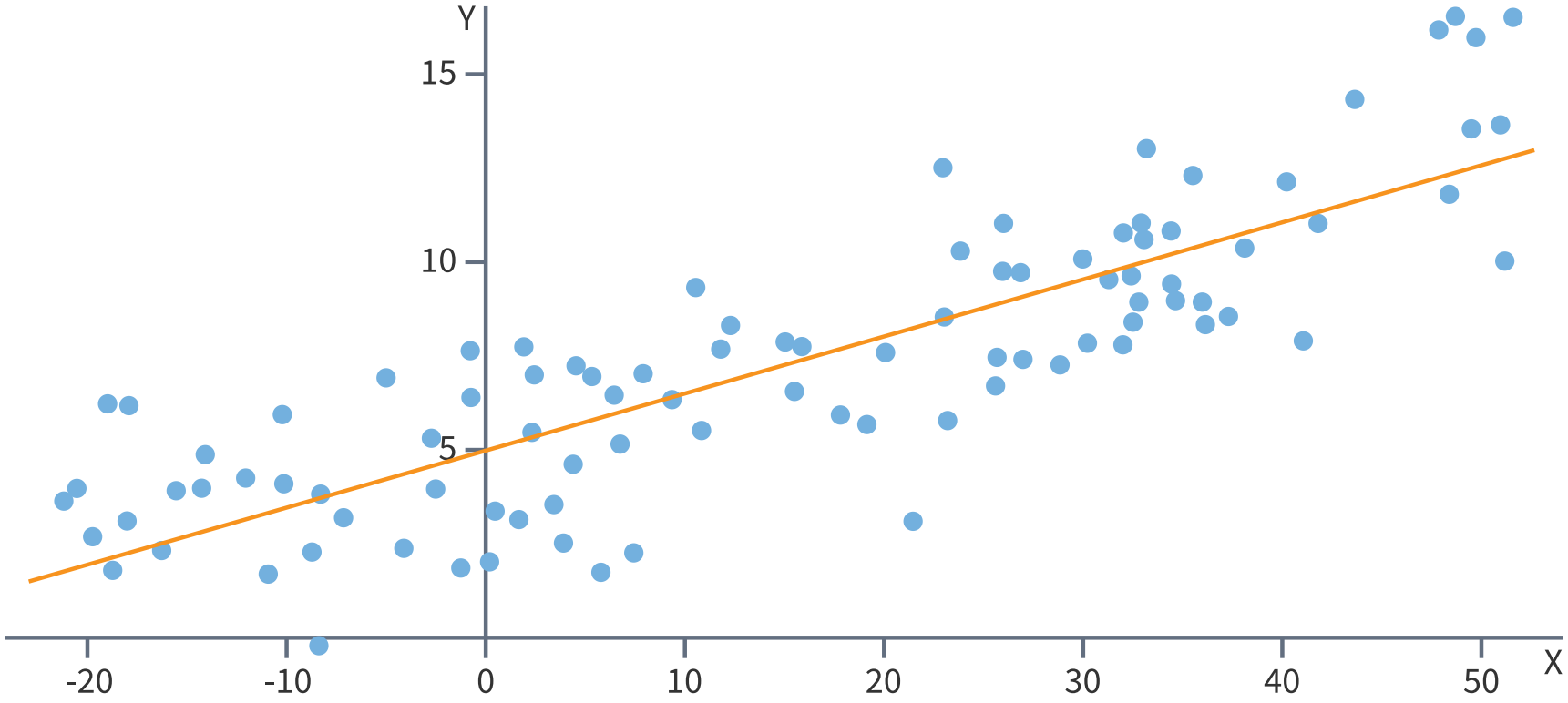
$$L = \frac{1}{K}\sum_{i=1}^{K}(y_i - \hat{y}_i)^2 \to \min$$

$$L = \frac{1}{K}\sum_{i=1}^{K}(y_i - (x_1 \cdot w_1 + x_2 \cdot w_2 + \ldots + x_n \cdot w_n))^2 \to \min$$

y - настоящее значение, $\hat{y}$ - предсказанное моделью значение, K - число объектов в обучающей выборке, $w_1, w_2, \ldots, w_n$ - веса признаков, это и есть наши дифференцируемые парамаметры, частные производные по ним будут составлять наш градиент.

Линейная регрессия: решение

Эльбрус буткемп



Регуляризация линейных моделей

Эльбрус буткемп

- Часто модель может переобучиться на какие-либо выбросы в данных
- Чтобы этого избежать, можно использовать регуляризацию:
  чаще всего это добавление в функцию ошибки каких-либо значений.
- Три самых распространенных метода регуляризации моделей:
  i. $L1$-регуляризация / LASSO: добавляем в функцию ошибки сумму модулей коэффициентов:

$$L(y, \hat{y}) = \frac{1}{K}\sum_{i=1}^{K}(y - \hat{y})^2 + \alpha||w||_1$$

α - коэффициент регуляризации

Регуляризация линейных моделей

Эльбрус буткемп

- Часто модель может переобучиться на какие-либо выбросы в данных
- Чтобы этого избежать, можно использовать регуляризацию:
  чаще всего это добавление в функцию ошибки каких-либо значений.
- Три самых распространенных метода регуляризации моделей:
  2. $L2$-регуляризация / Ridge: добавляем в функцию ошибки сумму квадратов коэффициентов:

$$L(y, \hat{y}) = \frac{1}{K}\sum_{i=1}^{K}(y - \hat{y})^2 + \alpha||w||_2$$

α - коэффициент регуляризации

Регуляризация линейных моделей

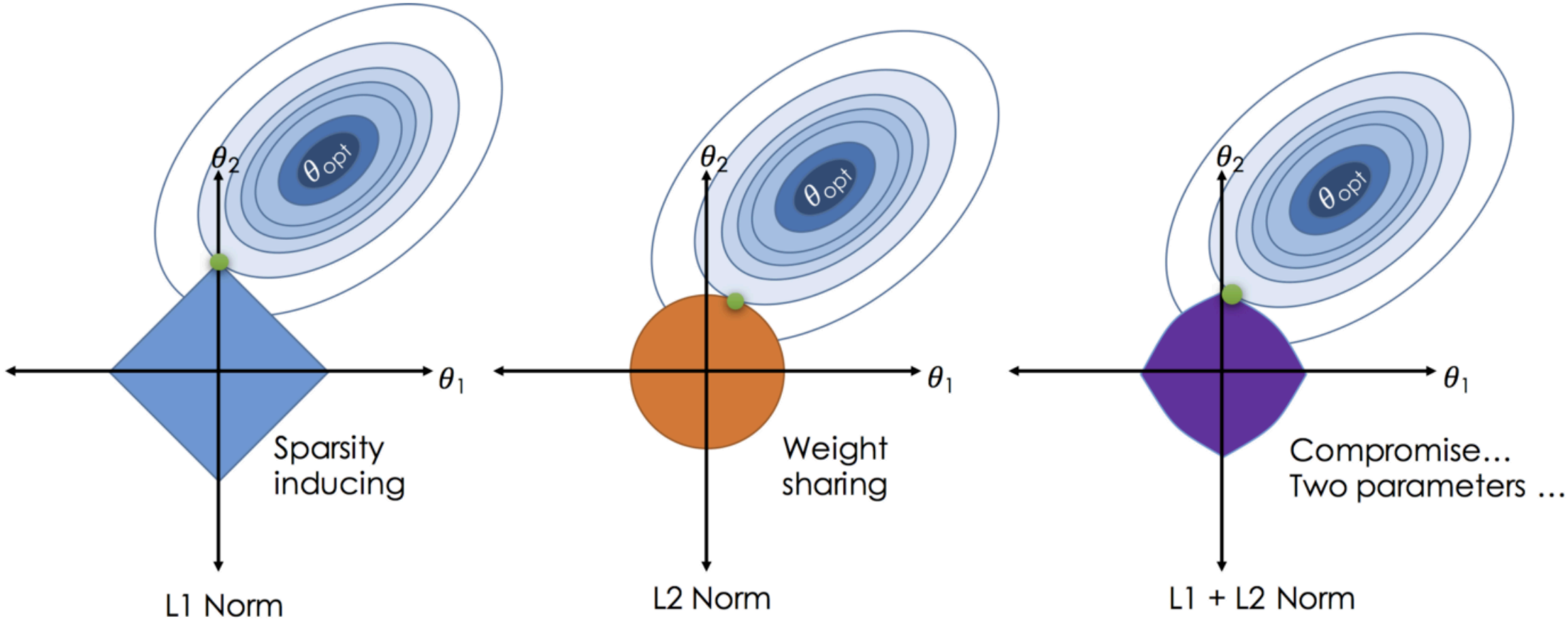
Эльбрус буткемп

- Часто модель может переобучиться на какие-либо выбросы в данных
- Чтобы этого избежать, можно использовать регуляризацию:
  чаще всего это добавление в функцию ошибки каких-либо значений.
- Три самых распространенных метода регуляризации моделей:
  2. ElasticNET: добавляем в функцию ошибки сумму квадратов и сумму модулей коэффициентов:

$$L(y, \hat{y}) = \frac{1}{K}\sum_{i=1}^{K}(y - \hat{y})^2 + \alpha||w||_2 + \beta||w||_1$$

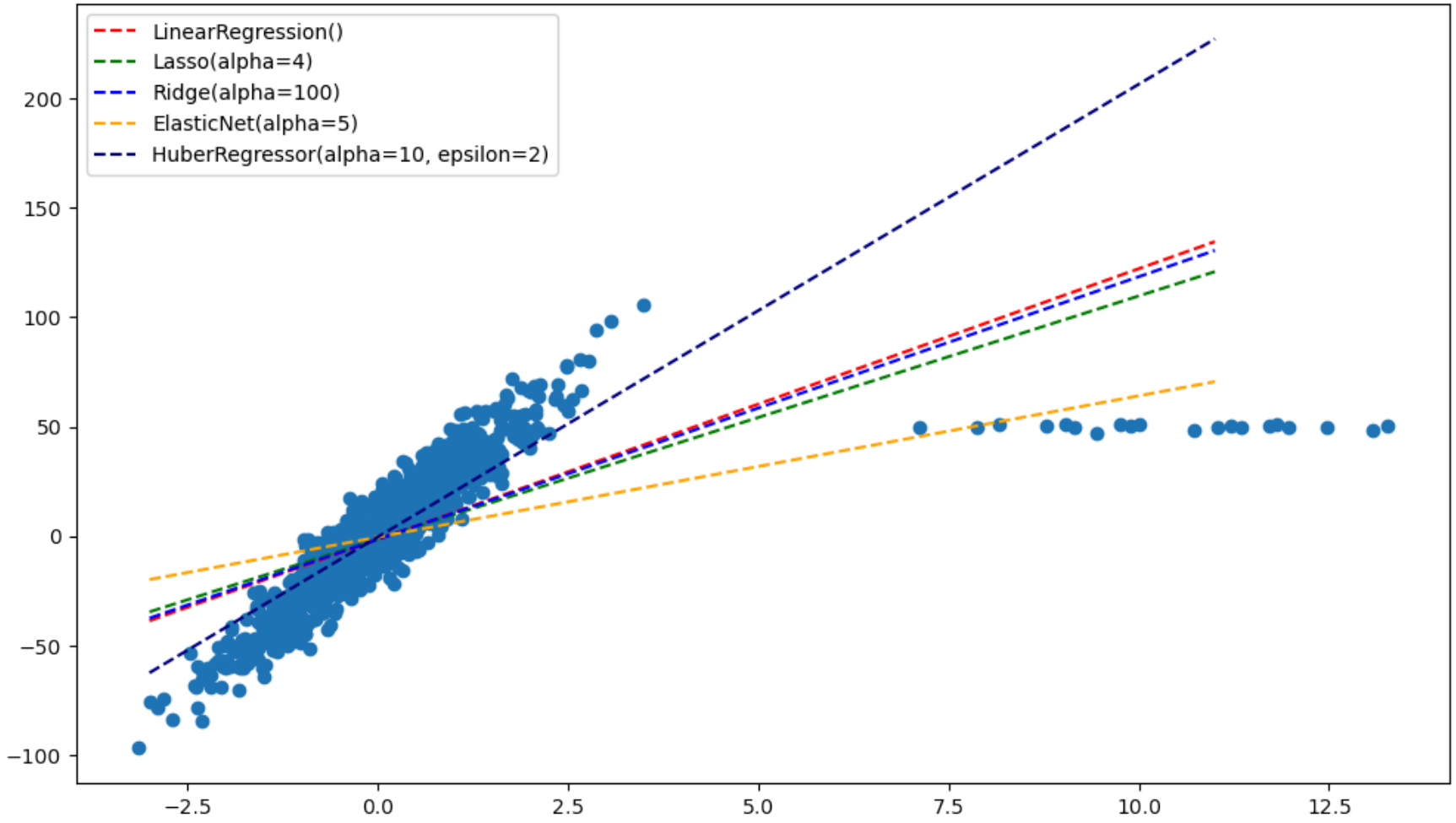
Регуляризация линейных моделей

Эльбрус буткемп



Регуляризация линейных моделей

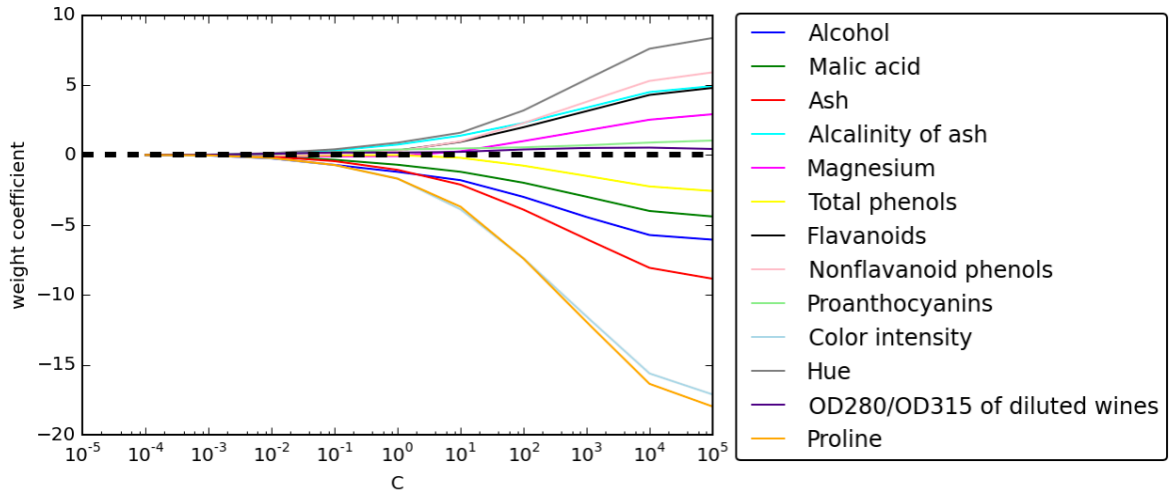
Эльбрус буткемп



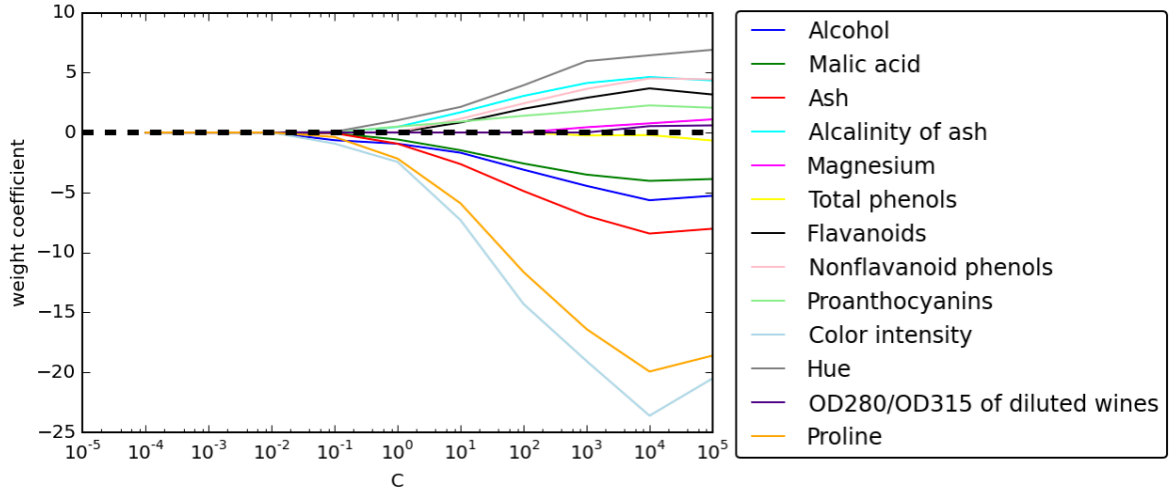
LASSO как способ отбора признаков

Эльбрус буткемп

With penalty l2



With penalty l1



Логистическая регрессия

Эльбрус буткемп

- Решает задачу классификации, не смотря на название!


center

Логистическая регрессия

Эльбрус буткемп

Модель остается линейной:

$$z = \sum_{i=1}^{n} x_i w_i + w_0$$

но полученный z подставляем в *логистическую* функцию:

$$y = \frac{1}{1 + e^{-z}}$$

выход y будет лежать в диапазоне от 0 до 1: $y \in [0, 1]$

это значение мы будем интерпретировать как *вероятность* того, что объект относится к классу **1**

Логистическая регрессия

Эльбрус буткемп

- Все положительные значения будут иметь вероятность больше 0.5
- Все отрицательные значения будут иметь вероятность меньше 0.5


center

📝 [Understanding Logistic Regression using R

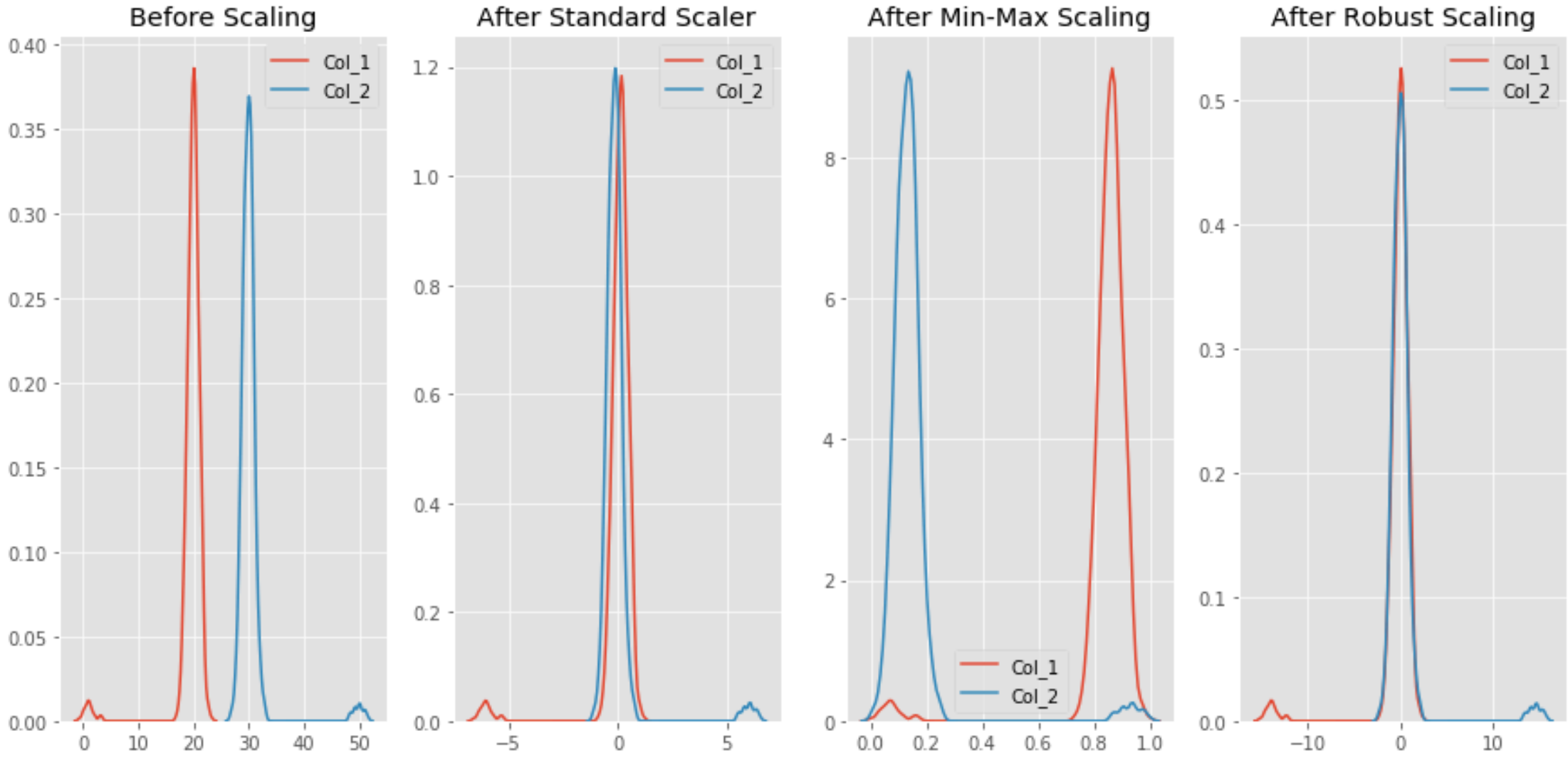
Нормализация признаков

Эльбрус буткемп

- **Standart Scaler** используем:
  - когда данные нормально или почти нормально распределены
- **MinMax Scaler** используем:
  - когда данные распределены не нормально
  - когда стандартное отклонение слишком маленькое
- **Robust Scaler** используем
  - когда в данных есть выбросы

Нормализация признаков

Эльбрус буткемп



📝©Feature Scaling with Scikit-Learn

Нормализация признаков

Эльбрус буткемп

- В случае линейной регресии позволяет интерпретировать коэффициенты линейной регресии
- В случае **любых** алгоритмов, вычисляющих расстояние, нормализует вклад каждого измерения в итоговое расстояние